Elements of External Validity: Framework, Design, and Analysis*

Naoki Egami† Erin Hartman‡

First Version: June 30, 2020
This Version: March 8, 2022

Abstract

External validity of causal findings is a focus of long-standing debates in the social sciences. While the issue has been extensively studied at the conceptual level, in practice, few empirical studies have explicit analysis aimed towards externally valid inferences. In this article, we make three contributions to improve empirical approaches for external validity. First, we propose a formal framework that encompasses four dimensions of external validity; X-, T-, Y-, and C-validity (populations, treatments, outcomes, and contexts). The proposed framework synthesizes diverse external validity concerns. We then distinguish two goals of generalization. To conduct *effect-generalization* — generalizing the magnitude of causal effects — we introduce three estimators of the target population causal effects. For *sign-generalization* — generalizing the direction of causal effects — we propose a novel multiple-testing procedure under weaker assumptions. We illustrate our methods through field, survey, and lab experiments as well as observational studies.

Keywords: Causal inference, External validity, Generalization

Word Count: 11930

---

*The proposed methodology is implemented via our forthcoming open-source software R package, `evalid`. We would like to thank Martin Bisgaard, Graeme Blair, David Broockman, Ryan Brutger, Juan Correa, Michael Findley, Nikhar Gaikwad, Don Green, Jens Hainmueller, Dan Hopkins, Joshua Kalla, Kevin Munger, Rocío Titiunik, Abby Wood, and Lauren Young, for their thoughtful comments. We would also like to thank participants at Polmeth 2020, APSA 2020, and seminars at Princeton, Stanford, University of California, Berkeley, and University of Texas, Austin.

†Assistant Professor, Department of Political Science, Columbia University, New York, NY 10027. Email: naoki.egami@columbia.edu, URL: `https://naokiegami.com`

‡Assistant Professor, Department of Political Science and of Statistics, University of California, Berkeley, Berkeley, CA 94720. Email: ekhartman@berkeley.edu, URL: `www.erinhartman.com`

1 Introduction

Over the last few decades, social scientists have developed and applied a host of statistical methods to make valid causal inference, known as a credibility revolution. This trend has primarily focused on *internal* validity — researchers aim to unbiasedly estimate causal effects *within* a study, without making strong assumptions. One of the most important long-standing methodological debates is about *external validity* — how scientists can generalize causal findings beyond a specific study.

While concepts of external validity are widely discussed in the social sciences, there are few empirical applications where researchers explicitly incorporate external validity into the design or analysis. Only 11% of all experimental studies and 13% of all observational causal studies published in the American Political Science Review from 2015 to 2019 contain a formal analysis of external validity in the main text, and none discuss conditions under which generalization is credible.[1] The lack of empirical approaches for external validity has remained, potentially because social science studies have diverse goals and concerns surrounding external validity, and yet, most existing methodologies have primarily focused on the subset of threats that are statistically more tractable. In many applications, important concerns about external validity receive no empirical evaluation.

In this article, we develop a framework and methodologies to improve empirical approaches for external validity. Building on the classical experimental design literature (Campbell and Stanley, 1963; Shadish, Cook and Campbell, 2002), we begin by proposing a unified causal framework that decomposes external validity into four components; X-, T-, Y-, and C-validity (populations, treatments, outcomes, and contexts/settings) (Section 3). With the proposed framework, we formally synthesize a variety of external validity concerns researchers face in practice and relate them to causal assumptions; to name a few examples, convenience samples (X-validity), differences in treatment implementations (T-validity), survey versus behavioral

[1]See Appendix K for more details on our literature review. A review paper by Findley, Kikuta and Denly (2020) also finds that only an exceptional few papers contained a dedicated external validity discussion.

outcomes (Y-validity), and differences in causal mechanisms across time, geography, and/or institutions (C-validity). We clarify conditions under which analysts can and cannot account for each type of validity.

After researchers identify the most relevant dimensions of external validity using our proposed framework, they can determine the goal of the external validity analysis: effect- or sign-generalization. *Effect-generalization* considers how to generalize the magnitude of causal effects, and *sign-generalization* aims to assess whether the direction of causal effects is generalizable. The former goal is important when researchers want to generalize the substantive or policy impact of treatments. The latter is relevant when analysts wish to test substantive theories that have observable implications only on the direction of treatment effects but not on the exact magnitude. Sign-generalization is also sometimes a practical compromise when effect-generalization, which requires stronger assumptions, is not feasible.

To enable effect-generalization, we introduce three classes of estimators and clarify the assumptions required by each (Section 5). Weighting-based estimators adjust for selection into experiments, outcome-based estimators control for treatment effect heterogeneity, and doubly robust estimators combine both to mitigate the risk of model misspecification.

In Section 6, we propose a new approach to sign-generalization. It is increasingly common to include variations in relevant dimensions of external validity at the design stage, e.g., measuring multiple outcomes, treatments, contexts and diverse populations within each study. We formalize this common practice as the design of purposive variations and discuss why and when it is effective for testing the generalizability of the sign of causal effects. By extending a partial conjunction test (Benjamini and Heller, 2008; Karmakar and Small, 2020), we then propose a novel sign-generalization test that combines purposive variations to quantify the extent of external validity. Because the design of purposive variations is already common in practice, application of the sign-generalization test can provide formal measures of external validity, while requiring little additional practical cost.

To focus on issues of external validity, we use three randomized experiments, covering field, survey, and lab experiments, as our motivating applications (Section 2). Using them, we illustrate how to implement our proposed methods and provide practical recommendations in

Section 7 and Appendix C. All of our methods can be implemented via the companion R package `evalid`. Finally, in Section 8, we discuss severeral key extensions. First, while the primary concern in observational studies is about internal validity, external validity is equally important for experimental and observational studies (Westreich et al., 2019). We discuss how to analyze the same four dimensions of external validity in observational studies. Second, we discuss how our proposed methods are related to and helpful for meta-analysis and recent efforts toward scientific replication of experiments, such as the EGAP Metaketa initiative.

Our contributions are threefold. First, we formalize all four dimensions of external validity within the potential outcomes framework (Neyman, 1923; Rubin, 1974). Existing causal methods using potential outcomes have primarily focused on changes in populations, i.e. X-validity (Imai, King and Stuart, 2008; Cole and Stuart, 2010). While a typology of external validity and different research goals of generalization are not new and have been discussed in the classical experimental design literature (Campbell and Stanley, 1963; Shadish, Cook and Campbell, 2002), this literature has focused on providing conceptual clarity and did not use a formal causal framework. We relate each validity type to explicit causal assumptions, which enables us to develop statistical methods that researchers can use in practice for generalization. Second, for effect-generalization of X-validity, we build on a large existing literature (Tipton, 2013; Hartman et al., 2015; Kern et al., 2016; Dahabreh et al., 2019) and provide practical guidance. To account for changes in populations and contexts together, i.e. X- and C-validity, we use identification results from the causal diagram approach (Bareinboim and Pearl, 2016) and develop new estimators in Section 5. The third and main methodological contribution is to provide a formal approach to sign-generalization. While this important goal has been informally and commonly discussed in practice, to our knowledge, no method has been available. Finally, our work is distinct from and complementary to a recent review paper by Findley, Kikuta and Denly (2020). The main goal of their work is to review how to *evaluate* external validity and how to report such evaluation in papers. In contrast, our paper focuses on how to *improve* external validity by proposing concrete methods (e.g., estimators and tests) that researchers can use in practice to implement effect- or sign-generalization.

2 Motivating Empirical Applications

2.1 Field Experiment: Reducing Transphobia

Prejudice can negatively impact social, political, and health outcomes of outgroups experiencing discrimination. Yet, the prevailing literature has found intergroup prejudices highly resistant to change. In a recent study, Broockman and Kalla (2016) use a field experiment to study whether and how much a door-to-door canvassing intervention can reduce prejudice against transgender people. It was conducted in Miami-Dade County, Florida, in 2015 among voters who answered a pre-experiment baseline survey. They randomly assigned canvassers to either encourage voters to actively take the perspective of transgender people ("perspective-taking") or have a placebo-conversation with respondents. To measure attitudes towards transgender people as outcome variables, they recruited respondents to four waves of follow-up surveys. The original authors find that the intervention involving a single approximately ten-minute conversation substantially reduced transphobia, and the effects persisted for three months.

2.2 Survey Experiment: Partisan-Motivated Reasoning

Scholars have been interested in how citizens perceive reality in ways that reflect well on their party, called partisan-motivated reasoning. Extending this literature, Bisgaard (2019) theorizes that partisans can acknowledge the same economic facts, and yet they rationalize reality using partisan-motivated reasoning. Those who support an incumbent party engage in blame-avoidant (credit-seeking) reasoning in the face of negative (positive) economic information, and opposition supporters behave conversely. To test this theory, the original author ran a total of four survey experiments across two countries, the United States and Denmark, to investigate whether substantive findings are consistent across different contexts where credit attribution of economic performance behaves differently. In each experiment, he recruited representative samples of the voting-age population, and then randomly assigned subjects to receive either positive or negative news about changes in GDP. He measured how respondents update their economic beliefs and how they attribute responsibility for the economic changes to a ruling party. Across four experiments, he finds support for his hypotheses.

2.3 Lab Experiment: Effect of Emotions on Dissent in Autocracy

Many authoritarian countries employ various frightening acts of repression to deter dissent. To unpack the psychological underpinnings of this authoritarian repression strategy, Young (2019) asks, "Does the emotion of fear play an important role in shaping citizens' willingness to dissent in autocracy, and if so, how?" (p. 140). She theorizes that fear makes citizens more pessimistic about the risk of repression and, consequently, less likely to engage in dissent. To test this theory, the original author conducted a lab experiment in Zimbabwe in 2015. She recruited a hard-to-reach population of 671 opposition supporters using a form of snowball sampling. The experimental treatment induced fear using an experimental psychology technique called the autobiographical emotional memory task (AEMT); at its core, an enumerator asks a respondent to describe a situation that makes her relaxed (control condition), or afraid (treatment condition). As outcome variables, she measured propensity to dissent with a host of hypothetical survey outcomes and real-world, low-stakes behavioral outcomes. She finds that fear negatively affects dissent decisions, particularly through pessimism about the probability that other opposition supporters will also engage in dissent.

3 Formal Framework for External Validity

In external validity analysis, we ask whether causal findings are generalizable to other (1) populations, (2) treatments, (3) outcomes, and (4) contexts (settings) of theoretical interest. We incorporate all four dimensions into the potential outcomes framework (Neyman, 1923; Rubin, 1974) by extending the classical experimental design literature (Shadish, Cook and Campbell, 2002). We will refer to each aspect as X-, T-, Y- and C-validity, where X represents pre-treatment covariates of populations, T treatments, Y outcomes, and C contexts. We will use an experimental study as an example because it helps us focus on issues of external validity. We discuss observational studies in Section 8.3.

3.1 Setup

Consider a randomized experiment with a total of n units, each indexed by $i \in \{1, \ldots, n\}$. We use $\mathcal{P}$ to denote this experimental sample, within which a treatment variable T_i is randomly assigned to each respondent. For notational clarity, we focus on a binary treatment $T_i \in \{0, 1\}$, but the same framework is applicable to categorical and continuous treatments with appropriate notational changes. Researchers measure outcome variable Y_i. We use C_i to denote a context to which unit i belongs. For example, the field experiment by Broockman and Kalla (2016) was conducted in Miami-Dade County in Florida in 2015, and $C_i = (\mathsf{Miami, 2015})$.

We then define $Y_i(T = t, c)$ to be the potential outcome variable of unit i if the unit were to receive the treatment $T_i = t$ within context $C_i = c$ where $t \in \{0, 1\}$. In contrast to the standard potential outcomes, our framework explicitly shows that potential outcomes also depend on context C. This allows for the possibility that causal mechanisms of how the treatment affects the outcome can vary across contexts.

Under the random assignment of the treatment variable T within the experiment, we can use simple estimators, such as difference-in-means, to estimate the *sample average treatment effect* (SATE).

$$\mathsf{SATE} \equiv \mathbb{E}_{\mathcal{P}}\{Y_i(T = 1, c) - Y_i(T = 0, c)\}. \tag{1}$$

This represents the causal effect of treatment T on outcome Y for the experimental population $\mathcal{P}$ in context $C = c$. The main issue of external validity is that researchers are not only interested in this within-experiment estimand but also whether causal conclusions are generalizable to other populations, treatments, outcomes, and contexts.

We define the *target* population, treatment, outcome, and context to be the targets against which external validity of a given experiment is evaluated. These targets are defined by the goal of the researcher or policy-maker. For example, Broockman and Kalla (2016) conducted an experiment with voluntary participants in Miami-Dade County in Florida. For X-validity, the target population could be adults in Miami, in Florida, in the U.S., or any other populations of theoretical interest. The same question applies to other dimensions, i.e., T-, Y-, and C-validity. Specifying targets is equivalent to clarifying studies' scope conditions, and thus, this

choice should be guided by substantive research questions and underlying theories of interest (Wilke and Humphreys, 2020).

Formally, we define the *target population average treatment effect* (T-PATE) as follows.

$$\mathsf{T\text{-}PATE} \equiv \mathbb{E}_{\mathcal{P}^*}\{Y_i^*(T^* = 1, c^*) - Y_i^*(T^* = 0, c^*)\}, \tag{2}$$

where $*$ denotes the target of each dimension. Note that the methodological literature often defines the population average treatment effect by focusing only on the difference in populations $\mathcal{P}$ and $\mathcal{P}^*$, but our definition of the T-PATE explicitly considers all four dimensions.

Therefore, we formalize a question of external validity as follows. "Would we obtain the same causal conclusion (e.g., the magnitude or sign of causal effects) if we use the target population $\mathcal{P}^*$, target treatment T^*, target outcome Y^*, and target context c^*?" Most importantly, external validity is defined with respect to specific targets researchers specify. This is essential because no experiment is universally externally valid; a completely different experiment should, of course, return a different result. Therefore, to empirically evaluate external validity of experiments in a fair way, both analysts and evaluators should clarify the targets against which they evaluate experiments. If the primary goal of the experiment is theory testing, these targets can be abstract theoretical concepts (e.g., incentives). On the other hand, if the goal is to generate policy recommendations for a real-world intervention, these targets are often more concrete.

3.2 Typology of External Validity

Building on a typology that has been influential conceptually (Campbell and Stanley, 1963), we provide a formal way to analyze practical concerns about external validity with the potential outcomes framework introduced in the previous section. We decompose external validity into four components, X-, T-, Y-, and C-validity, and we show how practical concerns in each dimension are related to fundamental causal assumptions. Table 1 previews a summary of the four dimensions.

3.2.1 X-validity

The difference in the composition of units in experimental samples and the target population is arguably the most well-known problem in the external validity literature (Imai, King and Stuart,

	Practical Concerns (examples)	Causal Assumptions (formalization)
X-validity	Convenience samples, Survey non-response, Attrition	Ignorability of Sampling and Treatment Effect Heterogeneity (Assumption 1)
T-validity	Realistic treatments, Bundled treatments, Difference in implementations	Ignorable Treatment-Variations (Assumption 2)
Y-validity	Behavioral or hypothetical survey outcomes, Short- or long-term outcomes	Ignorable Outcome-Variations (Assumption 3)
C-validity	Mechanisms differ across time, geography, political institutions, and so on	Contextual Exclusion Restriction (Assumption 4)

Table 1: Summary of Typology.

2008). When relying on convenience samples or non-probability samples, such as undergraduate samples and online samples (e.g., Mechanical Turk and Lucid), many researchers worry that estimated causal effects for such samples may not generalize to other target populations.

Bias due to the difference between experimental sample $\mathcal{P}$ and the target population $\mathcal{P}^*$ can be addressed when selection into the experiment and treatment effect heterogeneity are unrelated to each other after controlling for pre-treatment covariates $\mathbf{X}$ (Cole and Stuart, 2010).

Assumption 1 (Ignorability of Sampling and Treatment Effect Heterogeneity)

$$Y_i(T=1,c) - Y_i(T=0,c) \perp\!\!\!\perp S_i \mid \mathbf{X}_i \tag{3}$$

where $S_i \in \{0,1\}$ indicates whether units are sampled into the experiment or not.

The formal expression synthesizes two common approaches for addressing X-validity (Egami and Hartman, 2021). The first approach aims to account for how subjects are sampled into the experiment, including the common practice of using sampling weights (Miratrix et al., 2018). Random sampling is a well-known special case where no explicit sampling weights are required. The second common approach is based on treatment effect heterogeneity (e.g., Kern et al., 2016). If analysts can adjust for all variables explaining treatment effect heterogeneity, Assumption 1 holds. A special case is when treatment effects are homogeneous: when true, the difference between the experimental sample and the target population does not matter, and no

adjustment is required. Combining the two ideas, a general approach for X-validity is to adjust for variables that affect selection into an experiment and moderate treatment effects. The required assumption is violated when unobserved variables affect both sampling and treatment effect heterogeneity.

3.2.2 T-validity

In social science experiments, due to various practical and ethical constraints, the treatment implemented within an experiment is not necessarily the same as the target treatment that researchers are interested in for generalization.

In field experiments, this concern often arises due to difference in implementations. For example, when scaling up the perspective-taking treatment developed in Broockman and Kalla (2016), researchers might not be able to partner with equally established LGBT organizations and to recruit canvassers of similar quality. Many field experiments have found that details of implementation have important effects on treatment effectiveness.

In survey experiments, analysts are often concerned with whether randomly assigned information is realistic and whether respondents process it as they would do in the real world. For instance, Bisgaard (2019) designs treatments by mimicking contents of newspaper articles that citizens would likely read in everyday life, which are the target treatments.

In lab experiments, this concern is often about bundled treatments. To test theoretical mechanisms, it is important to experimentally activate a specific mechanism. However, in practice, randomized treatments often act as a bundle, activating several mechanisms together. For instance, Young (2019) acknowledges that "[a]lthough the AEMT [the treatment in her experiment] is one of the best existing ways to induce a specific targeted emotion, in practice it tends to induce a bundle of positive or negative emotions" (p. 144). In this line of discussion, researchers view treatments that activate specific causal mechanisms as the target and consider an assigned treatment as a combination of multiple target treatments. The concern is that individual effects cannot be isolated because each target treatment is not separately randomized.

While the target treatments differ depending on the types of experiments and corresponding research goals, practical challenges discussed above can be formalized as concerns over the same

causal assumption. Formally, bias due to concerns of T-validity is zero when the treatment-variation is irrelevant to treatment effects.

Assumption 2 (Ignorable Treatment-Variations)

$$\mathbb{E}_{\mathcal{P}}[Y_i(T=1,c) - Y_i(T=0,c)] \;=\; \mathbb{E}_{\mathcal{P}}[Y_i(T^*=1,c) - Y_i(T^*=0,c)]. \tag{4}$$

It states that the assigned treatment T and the target treatment T^* induce the same average treatment effects. For example, the causal impact of the perspective-taking intervention is the same regardless of whether canvassers are recruited by established LGBT organizations or not.

Most importantly, a variety of practical concerns outlined above are about potential violations of this same assumption. Thus, we develop a general method — a new sign-generalization test in Section 6 — that is applicable to concerns about T-validity, regardless of whether they arise in field, survey, or lab experiments.

3.2.3 Y-validity

Concerns of Y-validity arise when researchers cannot measure the target outcome in experiments. For example, in her lab experiment, Young (2019) could not measure actual dissent behaviors, such as attending opposition meetings, for ethical and practical reasons. Instead, she relies on a low-risk behavioral measure of dissent (wearing a wristband with a pro-democracy slogan) and a host of hypothetical survey measures that span a range of risk levels.

Similarly, in many experiments, even when researchers are inherently interested in behavioral outcomes, they often need to use hypothetical survey-based outcome measures, e.g., support for hypothetical immigrants, policies, and politicians. In such cases, Y-validity analysis might ask whether causal effects learned with these hypothetical survey outcomes are informative about causal effects on the support for immigrants, policies, and politicians in the real world.

The difference between short-term and long-term outcomes is also related to Y-validity. In many social science experiments, researchers can only measure short-term outcomes and not the long-term outcomes of main interest.

Formally, a central question is whether outcome measures used in an experimental study are informative about the target outcomes of interest. Bias due to the difference in an outcome

measured in the experiment Y and the target outcome Y^* is zero when the outcome-variation is irrelevant to treatment effects.

Assumption 3 (Ignorable Outcome-Variations)

$$\mathbb{E}_{\mathcal{P}}[Y_i^*(T=1,c) - Y_i^*(T=0,c)] \;=\; \mathbb{E}_{\mathcal{P}}[Y_i(T=1,c) - Y_i(T=0,c)]. \tag{5}$$

This assumption substantively means that the average causal effects are the same for outcomes measured in the experiment Y and for the target outcomes Y^*. The assumption naturally holds if researchers measure the target outcome in the experiment, i.e., $Y = Y^*$. For example, many Get-Out-of-the-Vote experiments in the U.S. satisfy this assumption by directly measuring voter turnout with administrative records (e.g., Gerber and Green, 2012).

Thus, when analyzing Y-validity, researchers should consider how causal effects on the target outcome relate to those estimated with outcome measures in experiments. In Section 6, we discuss how to address this common concern about Assumption 3 by using multiple outcome measures.

We note that there are many issues about measurement that are related to but different from Y-validity, such as measurement error, social desirability bias, and most importantly, construct validity. Following Morton and Williams (2010), we argue that high construct validity helps Y-validity, but it is not sufficient. This is because the target outcome is often chosen based on theory, and thus, experiments with high construct validity are more likely to be externally valid in terms of outcomes. However, construct validity does not imply Y-validity. For example, as repeatedly found in the literature, practical differences in outcome measures (e.g., outcomes measured 1 year or 2 years after administration of a treatment) are often indistinguishable from a theoretical perspective, and yet, they can induce large variation in treatment effects. We also provide further discussion on the relationship between external validity and other related concepts in Appendix G.

3.2.4 C-validity

Do experimental results generalize from one context to another context? This issue of C-validity is often at the heart of debates in external validity analysis (e.g., Deaton and Cartwright,

2018). Social scientists often discuss geography and time as important contexts. For example, researchers might be interested in understanding whether and how we can generalize Broockman and Kalla (2016)'s study from Miami in 2015 to another context, such as New York City in 2020. C-validity is challenging because a randomized experiment is done in one context c, and researchers often want to generalize or transport experimental results to another context c^*, where they did not run the experiment.

Even though this concern about contexts has a long history (Campbell and Stanley, 1963), to our knowledge, the first general formal analysis of C-validity is given by Bareinboim and Pearl (2016) using a causal graphical approach. Building on this emerging literature, we formalize C-validity within the potential outcomes framework introduced in Section 3.1.

We define C-validity as a question about mechanisms; how do treatment effects on the *same* units change across contexts? For example, in Broockman and Kalla (2016), even the same person might be affected differently by the perspective-taking intervention depending on whether she lives in New York City in 2020, or in Miami in 2015. Formally,

$$\underbrace{Y_i(T=1,c) - Y_i(T=0,c)}_{\text{Causal effect for unit } i \text{ in context } c} \quad \neq \quad \underbrace{Y_i(T=1,c^*) - Y_i(T=0,c^*)}_{\text{Causal effect for unit } i \text{ in context } c^*}$$

In order to generalize experimental results to another unseen context, we need to account for variables related to mechanisms through which contexts affect outcomes and moderate treatment effects. We refer to such variables as *context-moderators*. Specifically, researchers need to assume that contexts affect outcomes only through measured context-moderators. This implies that the causal effect for a given unit will be the same regardless of contexts, as long as the values of the context-moderators are the same. For example, in Broockman and Kalla (2016), the context-moderator could be the number of transgender individuals living in each unit's neighborhood. Then, analysts might assume that the causal effect for a given unit will be the same regardless of whether she lives in NYC in 2020 or in Miami in 2015, as long as we adjust for the number of transgender individuals living in her neighborhood.

We formalize this assumption as the *contextual exclusion restriction* (Assumption 4), which states that the context variable C_i has no direct causal effect on the outcome once fixing

context-moderators. This name reflects its similarity to the exclusion restriction well known in the instrumental variable literature.

Assumption 4 (Contextual Exclusion Restriction)

$$Y_i(T = t, \mathbf{M} = \mathbf{m}, c) = Y_i(T = t, \mathbf{M} = \mathbf{m}, c^*), \tag{6}$$

where the potential outcome $Y_i(T = t, c)$ is expanded with the potential context-moderators $\mathbf{M}_i(c)$ as $Y_i(T = t, c) = Y_i(T = t, \mathbf{M}_i(c), c)$, and then, $\mathbf{M}_i(c)$ is fixed to $\mathbf{m}$. We define $\mathbf{M}_i$ to be a vector of context-moderators, and thus, researchers can incorporate any number of variables to satisfy the contextual exclusion restriction. See Appendix H.2 for the proof of the identification of the T-PATE under this contextual exclusion restriction and other standard identification assumptions.

Most importantly, this assumption implies that the causal effect for a given unit will be the same regardless of contexts, as long as the values of the context-moderators are the same. Formally,

$$\underbrace{Y_i(T = 1, \mathbf{M} = \mathbf{m}, c) - Y_i(T = 0, \mathbf{M} = \mathbf{m}, c)}_{\text{Causal effect for unit } i \text{ with } \mathbf{M} = \mathbf{m} \text{ in context } c} = \underbrace{Y_i(T = 1, \mathbf{M} = \mathbf{m}, c^*) - Y_i(T = 0, \mathbf{M} = \mathbf{m}, c^*)}_{\text{Causal effect for unit } i \text{ with } \mathbf{M} = \mathbf{m} \text{ in context } c^*}$$

This assumption is plausible when the measured context-moderators capture all the reasons why causal effects vary across contexts. In other words, after conditioning on measured context-moderators, there is no remaining context-level treatment effect heterogeneity. In contrast, if there are other channels through which contexts affect outcomes and moderate treatment effects, the assumption is violated.

Several points about Assumption 4 are worth clarifying. First, there is no general randomization design that makes Assumption 4 true. This is similar to the case of instrumental variables in that the exclusion restriction needs justification based on domain knowledge even when instruments are randomized (Angrist, Imbens and Rubin, 1996). Second, in order to avoid post-treatment bias, context-moderators $\mathbf{M}_i$ cannot be affected by treatment T_i. In Broockman and Kalla (2016), it is plausible that the door-to-door canvassing interventions do not affect the number of transgender people in one's neighborhood, a context-moderator.

Finally, we clarify the subtle yet important difference between X- and C-validity. Most importantly, the same variables may be considered as issues of X- or C-validity depending on the nature of the problem and data at hand. For example, suppose we conduct a GOTV experiment in an electorally safe district in Florida. If we want to generalize this experimental result to another district in Florida that is electorally competitive, the competitiveness in the district is a question about C-validity. This is because our experimental data does not contain any data from an electorally competitive district, which defines the target context. However, suppose we conduct a state-wide experiment in Florida where some districts are electorally competitive and others are safe. Then, if we want to generalize this result to another state, e.g., the state of New York, where the proportion of electorally competitive districts differ, the electoral competitiveness of districts can be addressed as the X-validity problem.[2] This is because our experimental data has both electorally competitive and safe districts, and what differs across the two states is their distribution. This example shows that whether a given variable should be considered as an X- or C-validity question depends on the application.

In general, X-validity is a question about the representativeness of the experimental data. Thus, X-validity is of primary concern when we ask whether the *distribution* of certain variables in the experiment is similar to the target population distribution of the same variables. In contrast, C-validity is a question about transportation (Bareinboim and Pearl, 2016) to a new context. Thus, C-validity is the main concern when we ask whether the experimental result is generalizable to a context where no experimental data exists.

4 The Proposed Approach toward External Validity: Outline

In Section 3, we developed a formal framework and discussed concerns for external validity. In this section, we outline our proposed approach toward external validity, reserving details of our methods to Section 5 (effect-generalization) and Section 6 (sign-generalization).

[2]To generalize experimental results from the state of Florida to the state of New York, we have to consider other context moderators based on Assumption 4 as well. Here, we focus only on electoral competitiveness of districts as an example.

The first step of external validity analysis is to ask *which* dimensions of external validity are most relevant in one's application. For example, in the field experiment by Broockman and Kalla (2016), we primarily focus on X-validity (their experimental sample was restricted to Miami-Dade registered voters who responded to a baseline survey) and Y-validity (the original authors are interested in effects on both short- and long-term outcomes), while we also discuss all four dimensions in Appendix C. We also provide additional examples of how to identify relevant dimensions in Section 7 and Appendix C. Regardless of the type of experiment, researchers should consider all four dimensions of external validity and identify relevant ones. We refer readers to Section 3 on the specifics of how to conceptualize each dimension.

Once relevant dimensions are identified, analysts should decide the *goal* of external validity analysis, whether effect- or sign-generalization. Effect-generalization — how to estimate the T-PATE, i.e., generalizing the magnitude of the causal effect — is a central concern for randomized experiments that have policy implications. For example, in the field experiment by Broockman and Kalla (2016), effect-generalization is essential as cost-benefit considerations will be affected by the actual effect size. Sign-generalization — evaluating whether the sign of causal effects is generalizable — is relevant when researchers are testing theoretical mechanisms, and substantive theories have observable implications on the direction or the order of treatment effects but not on the effect magnitude. For example, our motivating examples of Bisgaard (2019) and Young (2019) explicitly write main hypotheses in terms of the sign of causal effects.

Given the goal, the next step is to ask *whether* the specified goal is achievable by evaluating the assumptions required for each goal in relevant external validity dimensions. Assumptions required for effect-generalization include Assumptions 1–4 detailed in Section 3, while Section 6 describes assumptions necessary for sign-generalization. In some settings, researchers can design experiments such that required assumptions are plausible, which is often the preferred approach. Importantly, even if effect-generalization is infeasible, sign-generalization might be possible in a wide range of applications as it requires much weaker assumptions. Thus, sign-generalization is also sometimes a practical compromise when effect-generalization is not feasible.

We emphasize that, even if external validity concerns are acute, credible effect- or sign-generalization might be impossible given the design of the experiment, available data, and the

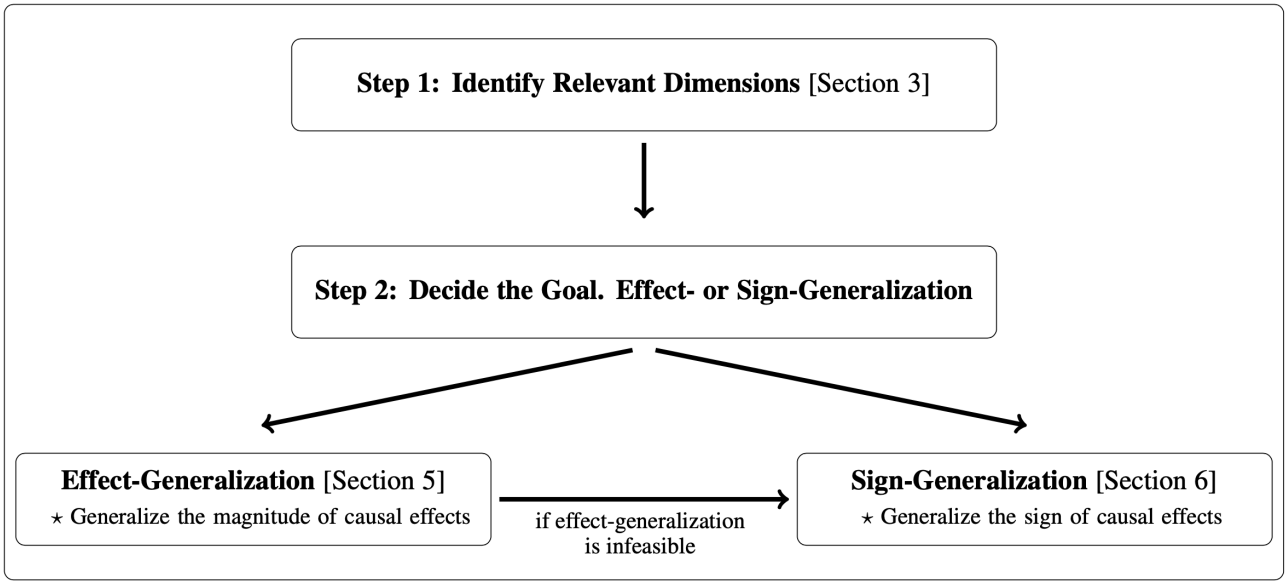


Figure 1: The Proposed Approach toward External Validity.

nature of the problem. In such cases, we recommend that researchers clarify which dimensions of external validity are most concerning and why effect- and sign-generalization are not possible (e.g., required assumptions are untenable, or required data on target populations, treatments, outcomes, or contexts are not available).

In Sections 5 and 6, we discuss *how* to conduct effect- and sign-generalization, respectively, when researchers can credibly justify the required assumptions. Our proposed workflow is summarized in Figure 1, and we refer readers there for a holistic view of our approach toward external validity in practice.

5 Effect-Generalization

In this section, we discuss *how* to conduct effect-generalization — including how to identify and estimate the T-PATE. This goal is most relevant for randomized experiments that aim to make policy recommendations. To keep the exposition clear, we first consider each dimension separately to highlight the difference in required assumptions and available solutions (we discuss how to address multiple dimensions together in Section 8.1).

For X- and C-validity, we start by asking *whether* effect-generalization is feasible by evaluating required assumptions (Assumption 1 for X-validity, and Assumption 4 for C-validity). If

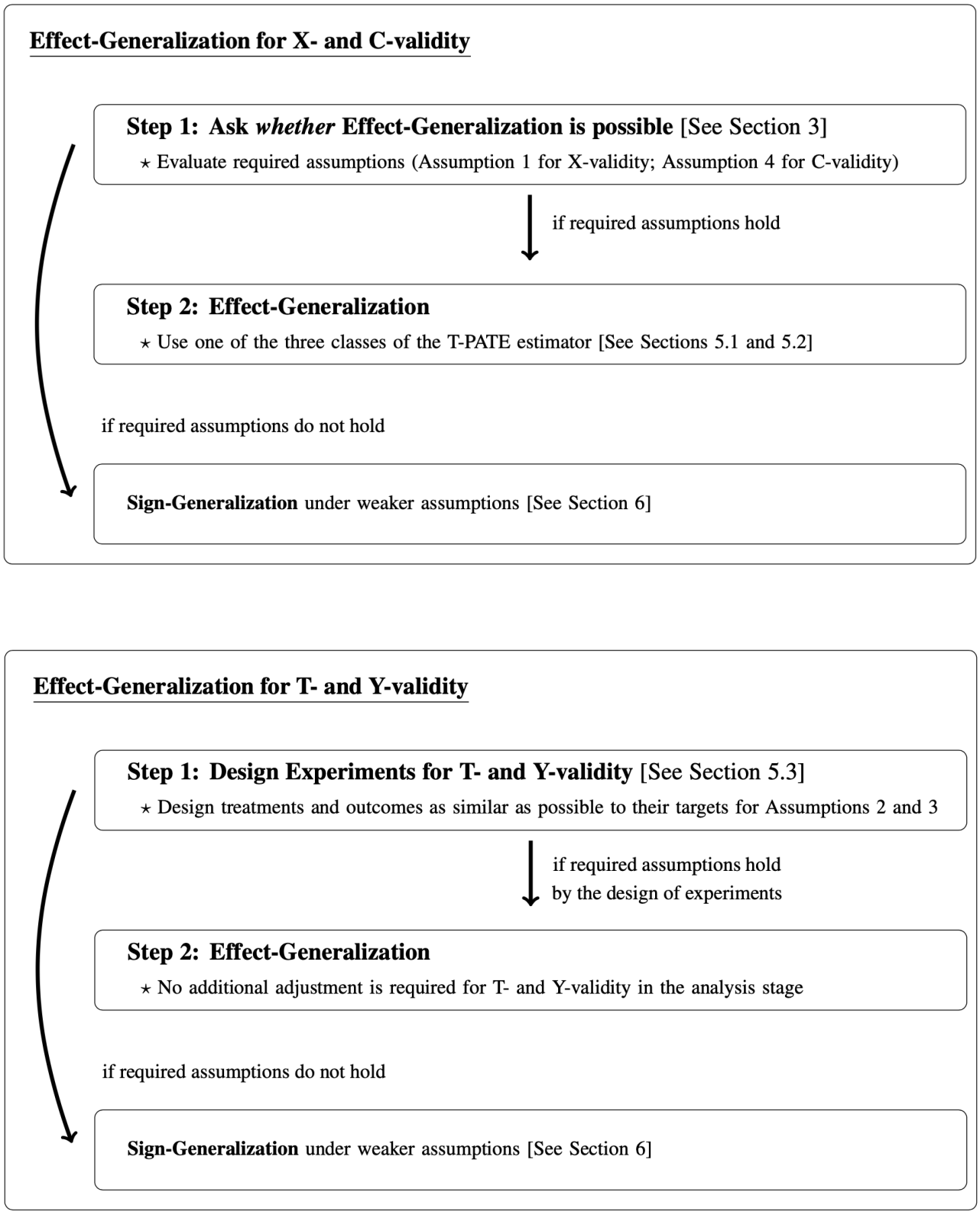


Figure 2: Summary of Effect-Generalization.

the required assumptions hold, researchers can employ three classes of estimators — weighting-based, outcome-based, and doubly robust estimators. We provide practical guidance on how to choose an estimator in Section 5.1.4. Importantly, because the required assumptions are often strong, credible effect-generalization might be impossible. In such cases, sign-generalization might still be feasible as it requires weaker assumptions (see Section 6).

For T- and Y-validity, we argue the required assumptions are much more difficult to justify *after* experiments are completed. Therefore, we emphasize the importance of *designing* experiments such that their required assumptions (Assumptions 2 and 3) are plausible by designing treatments and measuring outcomes as similar as possible to their targets. We also highlight that sign-generalization in Section 6 is more appropriate for addressing T- and Y-validity when researchers cannot modify their experiment to satisfy the required assumptions.

Our proposed approach is summarized in Figure 2, separately for X- and C-validity and T- and Y-validity.

5.1 X-validity: Three Classes of Estimators

Researchers need to adjust for differences between experimental samples and the target population to address X-validity (Assumption 1). We provide formal definitions of estimators and technical details in Appendix H.2.

5.1.1 Weighting-based Estimator

The first is a weighting-based estimator. The basic idea is to estimate the probability that units are sampled into the experiment, which is then used to weight experimental samples to approximate the target population. A common example is the use of survey weights in survey experiments.

Two widely-used estimators in this class are (1) an inverse probability weighted (IPW) estimator (Cole and Stuart, 2010), and (2) an ordinary least squares estimator with sampling weights (weighted OLS). Without weights, these estimators are commonly used for estimating the SATE, i.e., causal effects within the experiment. When incorporating sampling weights, these estimators are consistent for the T-PATE under Assumption 1. Both estimators also require a modeling assumption that sampling weights are correctly specified.

5.1.2 Outcome-based Estimator

While the weighting-based estimator focuses on the sampling process, we can also adjust for treatment effect heterogeneity to estimate the T-PATE (e.g., Kern et al., 2016). A general two-step estimator is as follows. First, we estimate outcome models for the treatment and control groups, separately, in the experimental data. In the second step, we use the estimated models to predict potential outcomes for the target population data.

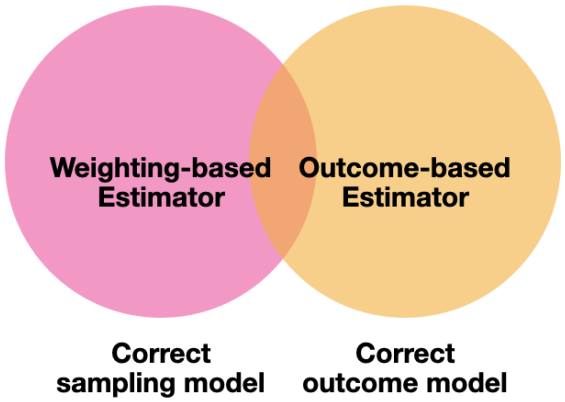
Formally, in the first step, we estimate the outcome model $\widehat{g}_t(\mathbf{X}_i) \equiv \widehat{\mathbb{E}}(Y_i \mid T_i = t, \mathbf{X}_i, S_i = 1)$ for $t \in \{0, 1\}$ where $S_i = 1$ indicates an experimental unit. This outcome model can be as simple as ordinary least squares, or rely on more flexible estimators. In the second step, for unit j in the target population data $\mathcal{P}^*$, we predict its potential outcome $\widehat{Y}_j(t) = \widehat{g}_t(\mathbf{X}_j)$, and thus, $\widehat{\mathsf{T\text{-}PATE}}_{\mathtt{out}} = \frac{1}{N} \sum_{j \in \mathcal{P}^*} \left(\widehat{Y}_j(1) - \widehat{Y}_j(0) \right)$, where the sum is over the target population data $\mathcal{P}^*$, and N is the size of the target population data.

It is worth re-emphasizing that this estimator requires Assumption 1 for identification of the T-PATE, and it also assumes the outcome models are correctly specified.

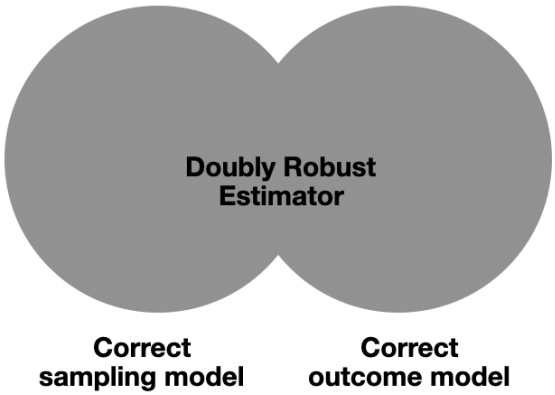
5.1.3 Doubly Robust Estimator

Finally, we discuss a class of doubly robust estimators, which reduces the risk of model misspecification common in the first two approaches (Robins, Rotnitzky and Zhao, 1994; Dahabreh et al., 2019). Specifically, to use weighting-based estimators, we have to assume the sampling model is correctly specified (the pink area in Figure 3 (a)). Similarly, outcome-based estimators assume the correct outcome model (the orange area). In contrast, doubly robust estimators are consistent for the T-PATE as long as either the outcome model or the sampling model is correctly specified; furthermore, analysts need not know which one is, in fact, correct. Figure 3 (b) shows that the doubly robust estimator is consistent in much wider applications (the gray area in Figure 3 (b)). Therefore, this estimator significantly relaxes modeling assumptions of the previous two methods. While they weaken modeling assumptions, we restate that doubly robust estimators also require Assumption 1 for identification of the T-PATE.

We now introduce the augmented IPW estimator (AIPW) in this class (Robins, Rotnitzky and Zhao, 1994; Dahabreh et al., 2019), which synthesizes weighting-based and outcome-based



(a) Weighting- and Outcome-based Estimators:
Consistent only in each circle



(b) Doubly Robust Estimator:
Consistent in the union of circles

Figure 3: Properties of Doubly Robust Estimator. *Note*: The doubly robust estimator is consistent as long as the sampling or outcome model is correctly specified (gray area in (b)).

estimators we discussed so far.

$$\widehat{\text{T-PATE}}_{\texttt{AIPW}} = \underbrace{\frac{\sum_{i\in\mathcal{P}} \pi_i T_i\{Y_i - \widehat{g}_1(\mathbf{X}_i)\}}{\sum_{i\in\mathcal{P}} \pi_i T_i} - \frac{\sum_{i\in\mathcal{P}} \pi_i (1-T_i)\{Y_i - \widehat{g}_0(\mathbf{X}_i)\}}{\sum_{i\in\mathcal{P}} \pi_i (1-T_i)}}_{\text{Weighting-based Estimator using Residuals}} + \underbrace{\frac{1}{N}\sum_{j\in\mathcal{P}^*}\{\widehat{g}_1(\mathbf{X}_j) - \widehat{g}_0(\mathbf{X}_j)\}}_{\text{Outcome-based Estimator}},$$

where π_i is the sampling weight of unit i. $\widehat{g}_t(\cdot)$ is an outcome model estimated in the experimental data. The first two terms are the IPW estimator based on residuals $Y_i - \widehat{g}_t(\mathbf{X}_i)$, and the last term is equal to the outcome-based estimator.

5.1.4 How to choose a T-PATE estimator

In practice, researchers often do not know the true model for the sampling process (e.g., when using online panels or work platforms) or treatment effect heterogeneity. For this reason, we recommend doubly robust estimators to mitigate the risk of model misspecification, whenever possible. There are, however, scenarios when the alternative classes of estimators may be more appropriate. In particular, the weighted OLS can incorporate pre-treatment covariates that are only measured in the experimental sample, which can greatly increase the precision in the estimation of the T-PATE (see Section 7.1), while this estimator requires correctly specified sampling weights. As long as treatment effect heterogeneity is limited, the outcome-based estimator is also appropriate, especially when variance of sampling weights is large, which is

exactly the settings where the other two estimators tend to have large standard errors.

5.2 X- and C-validity Together

In external validity analysis, concerns over X- and C-validity often arise together. This is because when we consider a target context different from the experimental context, both underlying mechanisms and populations often differ. To account for X- and C-validity together, we propose new estimators by generalizing sampling weights $\pi_i \times \theta_i$ and outcome models $g(\cdot)$.

$$\widehat{\pi}_i \equiv \underbrace{\frac{1}{\widehat{\Pr}(S_i = 1 \mid C_i = c, \mathbf{M}_i, \mathbf{X}_i)}}_{\text{Conditional sampling weights}}, \quad \text{and} \quad \widehat{\theta}_i \equiv \underbrace{\frac{\widehat{\Pr}(C_i = c^* \mid \mathbf{M}_i, \mathbf{X}_i)}{\widehat{\Pr}(C_i = c \mid \mathbf{M}_i, \mathbf{X}_i)}}_{\substack{\text{Difference in the distributions} \\ \text{across contexts}}}$$

$$\widehat{g}_t(\mathbf{X}_i, \mathbf{M}_i) \equiv \underbrace{\widehat{\mathbb{E}}(Y_i \mid T_i = t, \mathbf{X}_i, \mathbf{M}_i, S_i = 1, C_i = c)}_{\text{Outcome model using both } \mathbf{X}_i \text{ and } \mathbf{M}_i}, \quad \text{for} \quad t \in \{0, 1\},$$

where $\mathbf{X}_i$ are covariates necessary for Assumption 1 and $\mathbf{M}_i$ are context-moderators necessary for Assumption 4.

$\widehat{\pi}_i$ is the same as sampling weights used for X-validity, but it should be multiplied by $\widehat{\theta}_i$, which captures the difference in the distribution of $(\mathbf{X}_i, \mathbf{M}_i)$ in the experimental context c and the target context c^*. Outcome model $\widehat{g}_t(\cdot)$ use both $\mathbf{X}_i$ and $\mathbf{M}_i$ to explain outcomes. Note that estimators for X-validity alone (discussed in Section 5.1) or for C-validity alone are special cases of this proposed estimator. We provide technical details and proofs in Appendix H.

5.3 T- and Y-validity

Issues of T- and Y-validity are even more difficult in practice, which is naturally reflected in the strong assumptions discussed in Section 3.2 (Assumptions 2 and 3). This inherent difficulty is expected because defining a treatment and an outcome are the most fundamental pieces of any substantive theory; they formally set up potential outcomes, and they are directly defined based on research questions.

Therefore, we emphasize the importance of *designing* experiments such that the required assumptions are plausible by designing treatments and measuring outcomes as similar as possible to their targets. For example, to improve T-validity, Broockman and Kalla (2016) studied

door-to-door canvassing conversations that typical LGBT organizations can implement in a real-world setting. To safely measure outcomes as similar as possible to the actual dissent decisions in autocracy, Young (2019) carefully measured real-world, low-stakes behavioral outcomes in addition to asking hypothetical survey outcomes. This design-based approach is essential because, if the required assumptions hold by the design of the experiment, no additional adjustment is required for T- and Y-validity in the analysis stage. If such design-based solutions are not available, there is no general approach to conduct effect-generalization for T- and Y-validity without making stringent assumptions.

Importantly, even when effect-generalization is infeasible, researchers can assess external validity by examining the question of sign-generalization under weaker assumptions, which we discuss next in Section 6.

6 Sign-Generalization

We now consider the second research goal in external validity analysis; sign-generalization — evaluating whether the sign of causal effects is generalizable. This goal is most relevant when researchers are testing theoretical mechanisms, and substantive theories have observable implications on the direction or the order of treatment effects but not on the effect magnitude. Sign-generalization is also sometimes a practical compromise when effect-generalization is not feasible.

The first step of sign-generalization is to include variations in relevant external validity dimensions at the design stage of experiments. To address X-, T-, Y-, and C-validity, researchers can include diverse populations, multiple treatments, outcomes, and contexts into experiments, respectively. Incorporating such explicit variations has a long history and is already standard in practice. We formalize this common practice as the *design of purposive variations* and show what assumption is necessary for using such purposive variations for sign-generalization (Section 6.1). The required overlap assumption (Assumption 5) is much weaker than assumptions required for effect-generalization.

If researchers can include purposive variations to satisfy the required assumption, the final step is to conduct a new sign-generalization test, which computes partial conjunction p-values

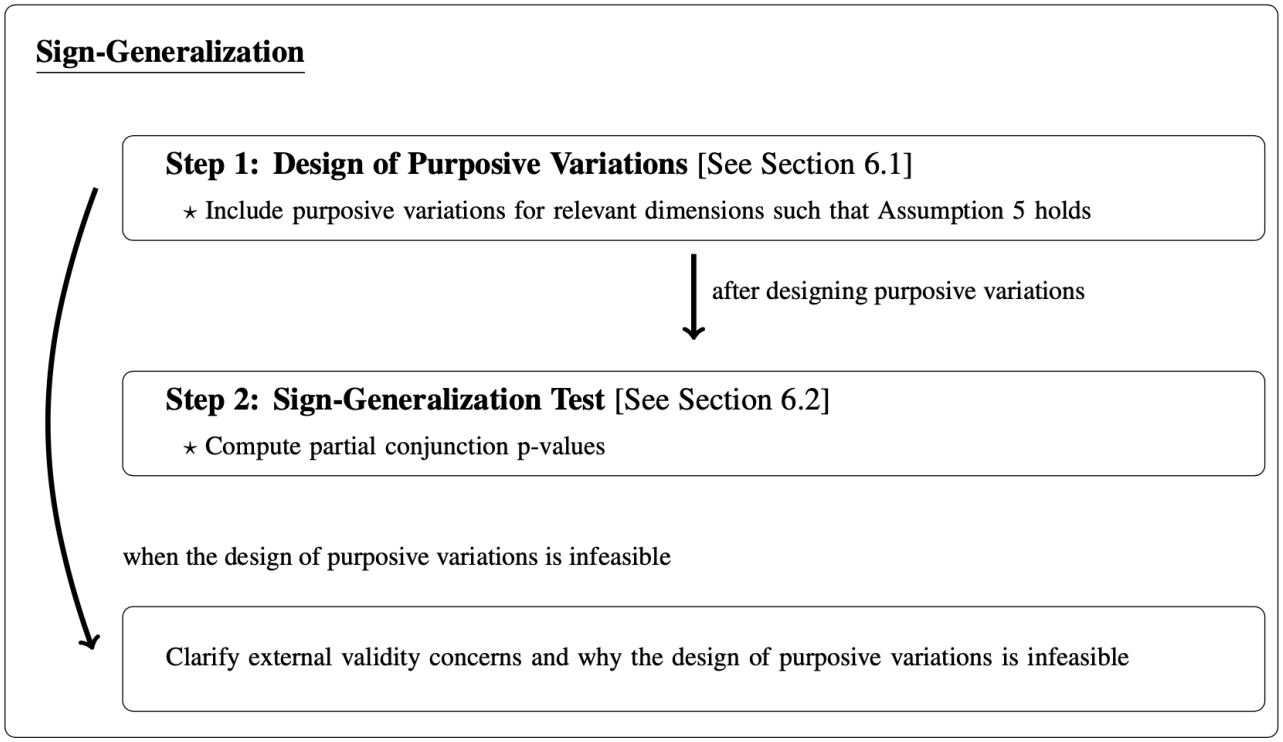


Figure 4: Summary of Sign-Generalization.

(Benjamini and Heller, 2008). Using these adjusted p-values, researchers can assess the direction of the T-PATE while accounting for multiple comparisons correctly. We detail their practical implementation and describe how to interpret them in Section 6.2. The key advantage is that the same proposed approach is applicable to all four dimensions. Our proposed approach is summarized in Figure 4, reserving methodological details for below.

6.1 Design of Purposive Variations

If possible, we woud like to test the sign of the T-PATE directly. However, it is infeasible in many applications as we often cannot observe target populations, treatments, outcomes, or contexts. Even in such scenarios, we can indirectly test the sign of the T-PATE by using multiple outcomes and incorporating diverse units, treatments, and contexts into experiments. The central idea is that if we consistently find positive (negative) causal effects across variations in all four dimensions, they together bolster evidence for a positive (negative) T-PATE (Shadish, Cook and Campbell, 2002). We call this approach the *design of purposive variations*. Incorporating

variations has a long history and is already standard in practice. In our review of all the experiments published in the APSR between 2015 and 2019, we found that at least 80% of articles included variations on at least one dimension.

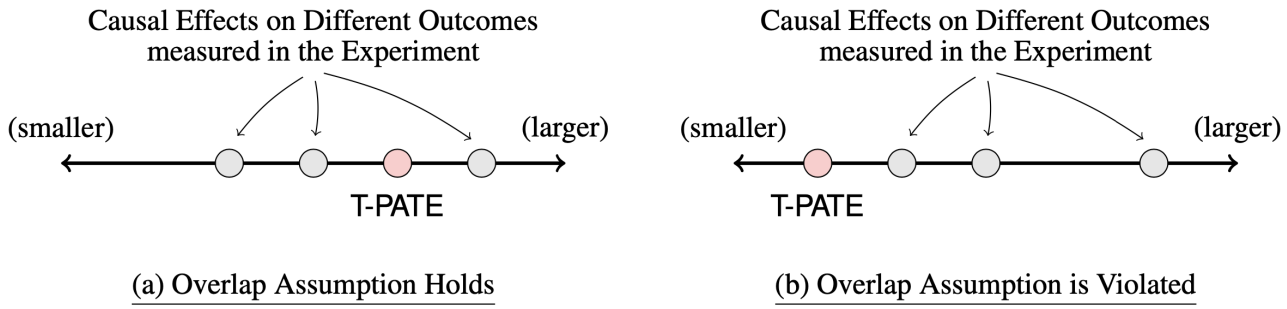
A practical question is: how should we incorporate *purposive* variations into experiments for testing the sign of the T-PATE? To answer this, we now formally introduce the design of purposive variations. For the sake of clear presentation, we focus on Y-validity. We discuss other dimensions in Section 6.3.

While there are many valid ways to choose variations for outcomes, we propose a simple approach based on a convex combination.

Assumption 5 (Overlap Between Target Outcomes and Purposive Variations) Choose K outcomes, $\{Y^1, \ldots, Y^K\}$, such that the T-PATE, $\mathbb{E}_{\mathcal{P}}\{Y_i^*(T=1,c) - Y_i^*(T=0,c)\}$, is within a convex hull of the K causal effects $\left\{\mathbb{E}_{\mathcal{P}}\{Y_i^k(T=1,c) - Y_i^k(T=0,c)\}\right\}_{k=1}^K$.

Although this assumption might seem strong at first, its substantive meaning is natural. Intuitively, we choose the K outcomes such that the T-PATE is within a range of the K causal effects we estimate in the experiment (see Figure 5). This is akin to the overlap assumption required in standard observational causal inference. Similarly, in sign-generalization, we require that the target outcome and the purposive variations overlap. Without this assumption, inferences will heavily depend on extrapolation, which we wish to avoid. In practice, because we do not know the T-PATE, researchers can make this assumption more plausible by choosing a range of outcomes on which treatment effects are expected to be smaller and larger than the T-PATE. For example, Young (2019) writes, "the items were selected to be contextually relevant and to span a range of risk levels" (p. 145). Assumption 5 provides a formal justification for such a design of purposive variations.

This assumption is violated when the T-PATE is outside a range of causal effects covered by the K outcomes. For example, in Young (2019), if the target outcome is a real-world high-risk dissent behavior and the intervention effect on this outcome is much smaller than those studied in the experiment, the overlap assumption is violated. At the same time, in this scenario, no external validity analysis is possible without using extrapolation. Our proposed approach



(a) Overlap Assumption Holds

(b) Overlap Assumption is Violated

Figure 5: Overlap Assumption.

guards against such model-dependent extrapolation by clarifying underlying assumptions.

6.2 Sign-Generalization Test

We now propose a new sign-generalization test. The goal here is to use purposive variations to test whether the sign of causal effects is generalizable.

Without loss of generality, suppose a substantive theory predicts that the T-PATE is positive. We focus again on Y-validity, and thus, our target null hypothesis can be written as,

$$H_0^* : \mathbb{E}_{\mathcal{P}}\{Y_i^*(T=1,c) - Y_i^*(T=0,c)\} \leq 0. \tag{7}$$

If we can provide statistical evidence against the null hypothesis H_0^*, we support the substantive theory predicting a positive effect.

When we cannot measure the target outcome Y^* in the experiment to directly evaluate this target hypothesis, we rely on the K hypotheses, corresponding to the K outcomes in experiments; for $k \in \{1, \ldots, K\}$,

$$H_0^k : \mathbb{E}_{\mathcal{P}}\{Y_i^k(T=1,c) - Y_i^k(T=0,c)\} \leq 0. \tag{8}$$

6.2.1 Connecting Purposive Variations to Sign-Generalization

We first show that when causal effects are positive (negative) for all K outcomes, the causal effect on the target outcome is also positive (negative) under the overlap assumption (Assumption 5). It implies that testing the union of the K null hypotheses (equation (8)) is a valid test for the target null hypothesis (equation (7)) under the overlap assumption. In practice, this means that a common approach of checking whether all K causal estimates are statistically

significant at a prespecified significance level α (e.g., $\alpha = 0.05$) is valid as a sign-generalization test, without additional multiple testing corrections. Details and derivations are presented in Appendix H.

6.2.2 Partial Conjunction Test

While checking whether all p-values are smaller than α is easy to implement, it can be too stringent in practice. For example, even if an estimated causal effect on just one out of many outcomes is not statistically significant, the method above is inconclusive about sign-generalization. However, intuitively, finding positive effects on most outcomes provides strong evidence for Y-validity.

To incorporate such flexibility, we build on a formal framework of partial conjunction tests, which was recently formalized by Benjamini and Heller (2008) and extended to observational causal inference in Karmakar and Small (2020). We extend the partial conjunction test framework to external validity analysis.

In the partial conjunction test, our goal is to provide evidence that the treatment has a positive effect on at least r out of K outcomes. Formally, the partial conjunction null hypothesis is as follows.

$$\widetilde{H}_0^r : \sum_{k=1}^{K} \mathbf{1}\{H_0^k \text{ is false}\} < r \tag{9}$$

where $r \in [1, K]$ is a threshold specified by researchers, and $\sum_{k=1}^{K} \mathbf{1}\{H_0^k \text{ is false}\}$ counts the number of true non-nulls. By rejecting this partial conjunction null, researchers can provide statistical evidence that the treatment has positive causal effects on at least r outcomes. For example, when $r = 0.8K$, researchers can assess whether the treatment has positive effects on at least 80% of outcomes.

How can we obtain a p-value for this partial conjunction test? We only need one-sided p-values computed separately for each of K outcomes $\{p_1, \ldots, p_K\}$. We first sort them such that $p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(K)}$. Then, we define the partial conjunction p-value as follows.

$$\begin{aligned} \widetilde{p}_{(1)} &\equiv K p_{(1)} \\ \widetilde{p}_{(r)} &\equiv \max\{(K - r + 1)p_{(r)}, \widetilde{p}_{(r-1)}\} \quad \text{for } r \geq 2. \end{aligned} \tag{10}$$

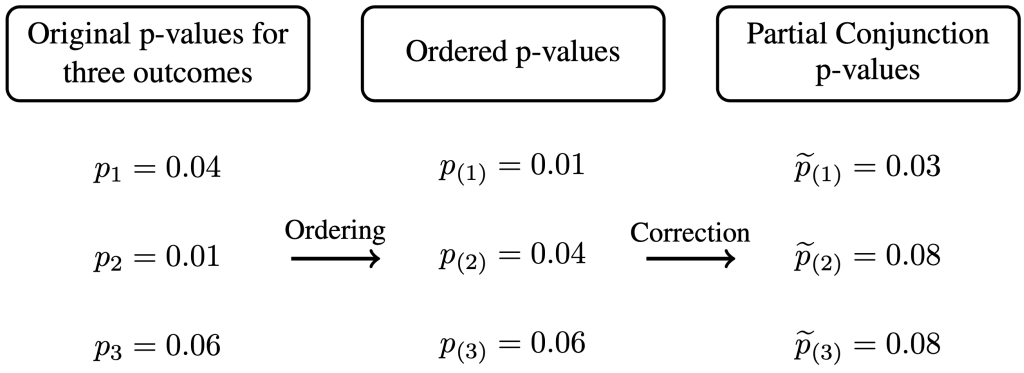


Figure 6: Example of Partial Conjunction Test with Three Outcomes. *Note:* The second step of "Correction" is based on equation (10).

The p-value for $\widetilde{H}_0^r$ is $\widetilde{p}_{(r)}$ (see Figure 6 for an example). This procedure is valid under any dependence across p-values (see Appendix H.3). In Appendix H.3, we also discuss scenarios in which p-values are independent across variations.

Finally, it is important to emphasize that researchers do not need to specify the threshold r. Rather, we recommend reporting partial conjunction p-values $\widetilde{p}_{(r)}$ for every threshold r (see equation (10) and examples in Section 7.2). For instance, in Figure 6, we would report all three partial conjunction p-values $\{0.03, 0.08, 0.08\}$, each testing whether 1, 2, or 3 out of our three outcomes have positive effects. While researchers might be worried about a multiple testing problem, no further adjustment to p-values is required due to the monotonicity properties of the partial conjunction p-value (see Appendix H.3 and Benjamini and Heller (2008)). In addition, using the K partial conjunction p-values, researchers can also directly estimate the number of outcomes for which the treatment has positive effects by counting the number of outcomes whose corresponding partial conjunction p-values are less than α. For example, in Figure 6, the estimated number of outcomes that have positive effects is one because only one out of the three outcomes is significant at $\alpha = 0.05$. We provide the details and proofs in Appendix H.3.

6.3 Other Dimensions

While this section focused on Y-validity for clear presentation, researchers can use the same sign-generalization test for other dimensions as long as purposive variations are included for each dimension of external validity. For purposive X-variations, researchers can explicitly

sample distinct subgroups that they expect to have different treatment effects. For instance, in Broockman and Kalla (2016), researchers could explicitly recruit respondents who have transgender friends and those who do not. For purposive T-variations, researchers can include treatment versions that change only one aspect at a time. For example, Young (2019) induced fear in respondents with two versions of the treatment: "general fear condition" unrelated to politics and "political fear condition" directly related to politics. Finally, purposive C-variation is gaining popularity in political science. It has recently become more feasible to run survey experiments in multiple countries at multiple time points (e.g., Bisgaard, 2019), and an increasing number of researchers conduct multi-site field experiments (e.g., Dunning et al., 2019; Blair and McClendon, 2020). It is important to emphasize that researchers can also assess multiple dimensions together (e.g., Y- and T-validity together) with the same approach. We provide examples of doing so in Section 7.

7 Empirical Applications

We now report a reanalysis of Broockman and Kalla (2016) as an example of effect-generalization, and Bisgaard (2019) as an example of sign-generalization. In Appendix C, we provide results for Young (2019), which focuses on sign-generalization.

7.1 Field Experiment: Reducing Transphobia

Broockman and Kalla (2016) find that a 10-minute perspective-taking conversation can lead to a durable reduction in transphobic beliefs. Typical of modern field experiments, their experimental sample was restricted to Miami-Dade registered voters who responded to a baseline survey, answered a face-to-face canvassing attempt, and responded to the subsequent survey waves, raising common concerns about X-validity. Unlike many other field experiments, their experiment provides a rare opportunity to evaluate Y-validity, in particular, whether the intervention has both short- and long-term effects, by measuring outcomes over time (3 days, 3 weeks, 6 weeks, and 3 months after the intervention). For the main outcome variable, the original authors computed a single index in each wave based on a set of survey questions on attitudes toward transgender people. Given the significant policy implication of the effect magnitude,

we study effect-generalization (Section 5), while addressing concerns of X- and Y-validity together. Given space constraints, we focus on these two dimensions which are most insightful for illustrating the proposed approach, and we discuss T- and C-validity in Appendix C.1.

While there are many potentially important target populations, we specify our target population to be all adults in Florida, defined using the common content data from the 2016 Cooperative Congressional Election Study (CCES).

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. While these variables are similar to what applied researchers usually adjust for, we have to carefully assess the necessary identification assumption (Assumption 1). If unobserved variables, such as political interest, affect both sampling and effect heterogeneity, the assumption is untenable. Researchers can make this required assumption more plausible by measuring variables affecting both sampling and treatment effect heterogeneity.

7.1.1 Effect-Generalization

We estimate the T-PATE using the three classes of estimators discussed in Section 5.1. Weighting-based estimators include IPW and weighted OLS that adjusts for control variables pre-specified in the original authors' pre-analysis plan. Sampling weights are estimated via calibration (Hartman et al., 2015). For the outcome-based estimators, we use OLS and a more flexible model, BART. Finally, we implement two doubly robust estimators; the AIPW with OLS and the AIPW with BART. We use block bootstrap to compute standard errors clustered at the household level as in the original study. All estimators are implemented by our companion R package `evalid`.

Figure 7 presents point estimates and their 95% confidence intervals using different estimators. Broockman and Kalla (2016) create an outcome index such that the value of one represents one standard deviation of the index outcome in the control group. Therefore, estimated effects should be interpreted relative to outcomes in the control group. The first column shows estimates of the SATE for four time periods, and the subsequent three columns present estimates of the T-PATE using the three classes of estimators from above.

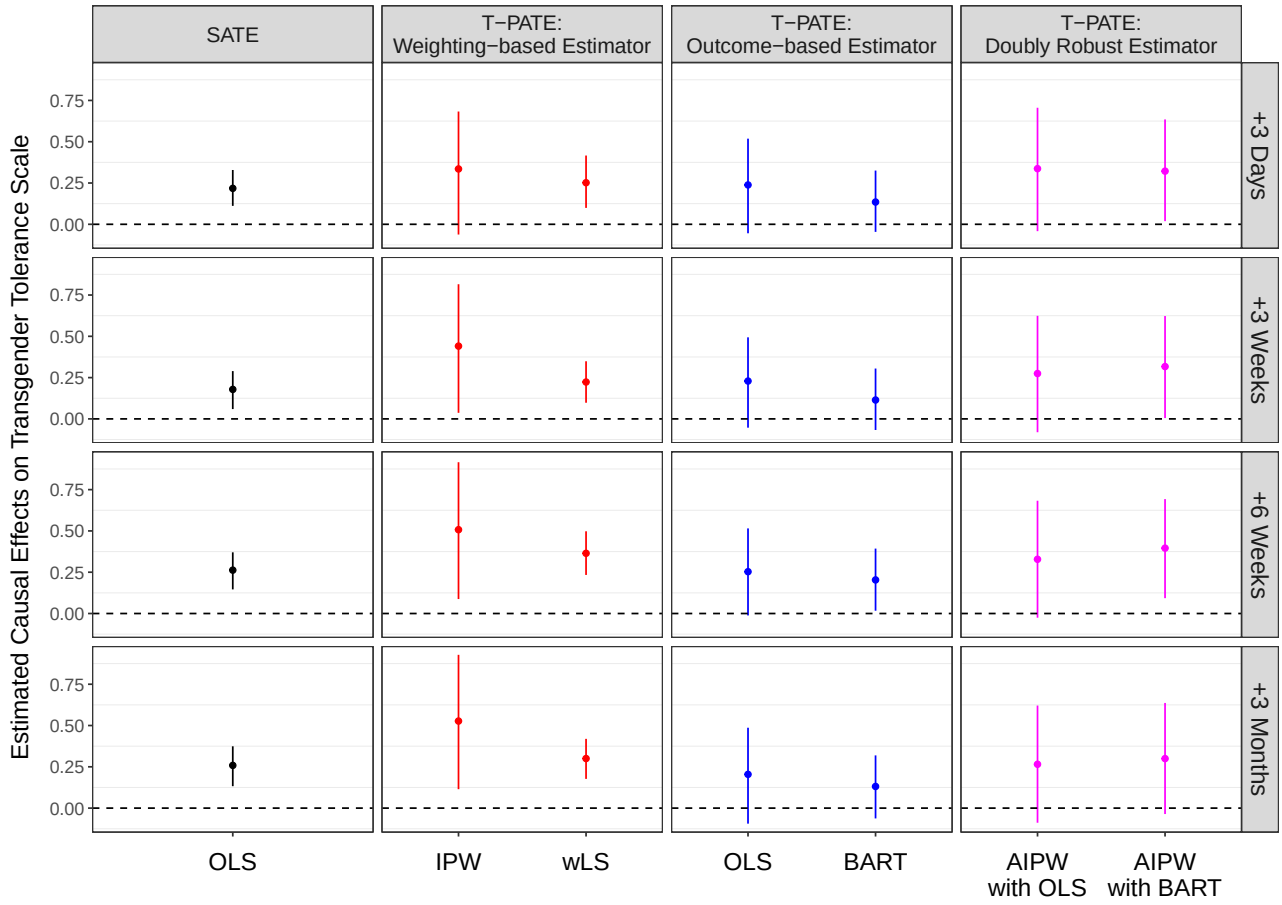


Figure 7: Estimates of the T-PATE for Broockman and Kalla (2016). *Note:* The first column shows estimates of the SATE, and the subsequent three columns present estimates of the T-PATE for three classes of estimators. Rows represent different post-treatment survey waves.

Several points are worth noting. First, the T-PATE estimates are similar to the SATE estimate, and this pattern is stable across all time periods. By accounting for X- and Y-validity, this analysis suggests that Broockman and Kalla (2016)'s intervention has similar effects in the target population across all time periods. We emphasize that, while the SATE estimate and the T-PATE estimates are similar in this application, bias in the SATE estimates can be large in many applications (see Appendix J for illustrations). Thus, we recommend estimating the T-PATE formally and comparing it against the SATE estimate.

Second, in general, estimates of the T-PATE have larger standard errors compared to that of the SATE. This is natural and necessary because the estimation of the T-PATE must also account for differences between the experimental sample and the target population. Importantly,

both the point estimate and the standard error of the T-PATE affect cost-benefit analysis. Thus, even though point estimates are similar, cost-benefit analysis for the target population has more uncertainty due to the larger standard error of the T-PATE.

Finally, we can compare the three classes of estimators. We generally recommend doubly robust estimators because the sampling and outcome models are often unknown in practice. However, in this example, the weighted least squares estimator (wLS in Figure 7) also has a desirable feature; it is the most efficient estimator because it can incorporate many pre-treatment covariates measured only in the experiment, while other estimators cannot. Note that this estimator assumes the correct specification of sampling weights. Outcome-based estimators are also effective here because there is limited treatment effect heterogeneity as found in the original article. Indeed, all estimators provide relatively stable T-PATE estimates, which are close to the SATE in this example. By following similar reasoning, researchers can determine an appropriate estimator in each application (see also Section 5.1.4).

7.2 Survey Experiment: Partisan-Motivated Reasoning

Bisgaard (2019) finds that, even when partisans agree on the facts, partisan-motivated reasoning influences how they internalize those facts and attribute credit (or blame) to incumbents. In terms of external validity analysis, Bisgaard (2019) provides several great opportunities to evaluate sign-generalization in terms of C- and Y-validity. We discuss X- and T-validity in Appendix C.2.

For C-validity, the study incorporates purposive variations by running a total of four survey experiments across two countries, the United States and Denmark (Study 1 in the U.S., and Studies 2–4 in Denmark. See Table 1 of the original study for more details). They differ both in terms of political and economic settings; 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.

While generalization to a new target context was not a clear goal of the original paper, there are potentially many relevant target contexts. For example, Germany shares political and geographic features with Denmark and its global economic power with the United States.

	Variations for C-Validity	Variations for Y-Validity
Study 1	United States	Close-ended (1), Open-ended (1), Argument Rating (6)
Study 2	Denmark	Close-ended (1), Open-ended (1)
Study 3	Denmark	Close-ended (1)
Study 4	Denmark	Open-ended (1)

Table 2: Design of Purposive Variations for Bisgaard (2019). *Note:* The number of the purposive outcome variations is in parentheses.

Thus, if researchers are interested in generalizing results to Germany, it may be reasonable to assume that the purposive contextual variations in Bisgaard (2019) satisfy the required overlap assumption (Assumption 5).

In terms of Y-validity, to measure how citizens attribute responsibilities to incumbents, the original author uses three different sets of outcomes; closed-ended survey responses, open-ended-survey responses, and argument rating tasks. The target outcome is citizens' attribution of responsibility to incumbents when they read economic news in everyday life. The three sets of outcomes provide reasonable variations to capture this target outcome by balancing specificity and reality. We assume that the three sets of outcomes jointly satisfy the required overlap assumption, and we use all the outcomes for the sign-generalization test.

7.2.1 Sign-Generalization Test

The theory of Bisgaard (2019) can be summarized into two hypotheses, one for supporters of the incumbent party and the other for those of the opposition party. In the face of positive economic facts: (H1) Supporters of the incumbent party will be more likely, and (H2) supporters of the opposition party will be less likely, to believe the incumbent party is responsible for the economy. We estimate the treatment effect of showing positive economic news on the attribution of responsibility, relative to showing negative economic news. Thus, for supporters of the incumbent party, the first hypothesis (H1) predicts that the treatment effects are positive, and for supporters of the opposition party, the second hypothesis (H2) predicts that the treatment effects are negative.

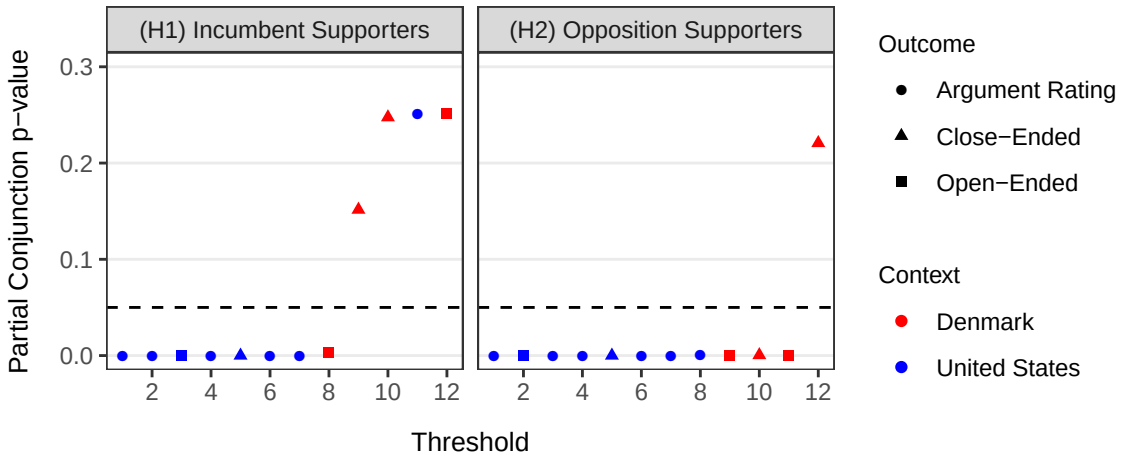


Figure 8: Sign-Generalization Test for Bisgaard (2019). *Note:* We combine causal estimates on multiple outcomes across four survey experiments in two countries. Following Section 6, we report partial conjunction p-values for all thresholds.

For our external validity analysis, we test each hypothesis by considering C- and Y-validity together using the sign-generalization test. The combination of multiple outcomes across four survey experiments in two countries yields twelve causal estimates corresponding to each hypothesis (see Table 2). We then assess the proportion of positive causal effects for the first hypothesis and that of negative causal effects for the second hypothesis using the proposed partial conjunction test.

For each hypothesis, Figure 8 presents results from the partial conjunction test for all thresholds. Each p-value is colored by context, with Denmark in red and the United States in blue. Variations in outcome are represented by symbols. For incumbent supporters, we find eight out of twelve outcomes (66%) have partial conjunction p-values less than the conventional significance level 0.05. It is notable that most of the estimates that do not support the theory are from Denmark, which we might expect since partisan-motivated reasoning would be weaker in Denmark. In contrast, for opposition supporters, the results show eleven out of twelve outcomes (92%) have partial conjunction p-values less than 0.05, and there is stronger evidence across outcomes and contexts.

Therefore, even though there exists some support for both hypotheses, Bisgaard (2019)'s theory is more robust for explaining opposition supporters; opposition supporters engage more

in partisan-motivated reasoning than incumbent supporters.

8 Discussion

8.1 Addressing Multiple Dimensions Together

As illustrated by our empirical applications in Section 7, we often have to consider multiple dimensions of external validity together in practice. In general, we recommend thinking about each dimension separately and sequentially because each dimension requires different types of assumptions as discussed in Section 3.2. Importantly, the proposed methodologies for each dimension can be combined naturally by applying them sequentially. To conduct effect-generalization, it is often easier to address X- and C-validity first before thinking about T- and Y-validity. In Section 7.1, we addressed X-validity using three classes of the T-PATE estimator and then evaluated Y-validity by checking whether estimates are stable across outcomes measured at different points in time.

For sign-generalization, researchers can address multiple dimensions simultaneously as long as they include purposive variations for relevant dimensions. This is one of the key advantages of sign-generalization. In Section 7.2, we examined C- and Y-validity together via the partial conjunction test (see Figure 8). See another example based on Young (2019) in Appendix C.

Finally, we emphasize that it is not always possible to empirically address all relevant dimensions of external validity because the required identification assumptions can be untenable or because required data are not available. In such cases, it is important to clarify which dimension of external validity researchers cannot address empirically and why.

8.2 Relationship to Replication and Meta-Analysis

Meta-analysis is a method for summarizing statistical findings from multiple papers or research literature. While still rare, political scientists have begun using it to aggregate results from randomized experiments (e.g., Dunning et al., 2019; Paluck, Green and Green, 2019). Meta-analysis can be based on the most common, "uncoordinated scientific replication experiments" (different researchers conduct similar experiments over time without explicit coordination across

researchers), or increasingly relevant, "coordinated scientific replication experiments" (e.g., the EGAP Metaketa studies) (Blair and McClendon, 2020).[3] Even though we have so far focused on how to improve external validity of individual experiments, the proposed approach can also be useful for conducting meta-analysis.

First, meta-analysts must also consider the same four dimensions of external validity. Scientific replication of experiments is a powerful tool because researchers can incorporate purposive variations across experiments and design later experiments to overcome external validity concerns of earlier experiments. But, to maximize the utility of scientific replication, researchers have to examine the same four dimensions of external validity and associated assumptions to design experiments that can credibly address external validity concerns. For example, the Metaketa initiative can select sites by explicitly diversifying context-moderators such that the overlap assumption is more plausible.

Second, both effect- and sign-generalization are important for meta-analysis. Some studies, such as Dunning et al. (2019), clearly aim to provide policy recommendations and evaluate the cost-effectiveness of particular interventions. Estimators for the T-PATE (Section 5) are essential when meta-analysts want to predict causal effects in new target sites. Sign-generalization (Section 6) is useful when meta-analysis focuses on synthesizing scientific knowledge (e.g., Paluck, Green and Green (2019) examine whether intergroup contact typically reduces prejudice).

To illustrate how our proposed approach can also be useful for meta-analysis, we consider the Metaketa I (Dunning et al., 2019) as an application. Building on the original analysis, we discuss how researchers might conduct effect-generalization to a new context and how to conduct sign-generalization for coordinated experiments. We report all details in Appendix D.

[3]Replication experiments are still sometimes too costly. For example, researchers might not be able to run multiple studies due to limited resources or because an experiment needs to be done in a rare context. Our proposed approach can be applied to one experiment and does not assume multiple experiments.

8.3 External Validity of Observational Studies

For observational studies, researchers can decompose total bias into internal validity bias and external validity bias (Westreich et al., 2019). Thus, the same four dimensions of external validity are also relevant in observational studies. For example, widely-used causal inference techniques, such as instrumental variables and regression discontinuity, make identification strategies more credible by focusing on a subset of units, which often decreases X-validity. While effect-generalization requires even stronger assumptions in observational studies, sign-generalization is possible in many applications as far as purposive variations exist in observational data.

As a concrete example, we examine two large scale observational studies based on a natural experiment (Dehejia, Pop-Eleches and Samii, 2021) and instrumental variables (Bisbee et al., 2017). Using these two studies, we discuss in Appendix E how to use the proposed sign-generalization test to combine estimates across contexts and evaluate sign-generalization in observational studies. An effect-generalization type analysis is reported in the original studies mentioned above.

9 Concluding Remarks

External validity has been a focus of long-standing debates in the social sciences. However, in contrast to extensive discussions at the conceptual level, there have been few empirical applications where researchers explicitly incorporate design or analysis for external validity. In this article, we aim to improve empirical approaches for external validity by proposing a framework and developing tailored methods for effect- and sign-generalization. We clarify underlying assumptions required to account for concerns about X-, T-, Y-, and C-validity. We then describe three classes of estimators for effect-generalization and propose a new test for sign-generalization.

Addressing external validity is inherently difficult because it aims to infer whether causal findings are generalizable to other populations, treatments, outcomes, and contexts that we do not observe in our data. In this paper, we formally clarify conditions under which this

challenging yet essential inference is possible, and we propose new methods to improve external validity.

References

Angrist, Joshua D, Guido W Imbens and Donald B Rubin. 1996. "Identification of Causal Effects Using Instrumental Variables." *Journal of the American statistical Association* 91(434):444–455.

Bareinboim, Elias and Judea Pearl. 2016. "Causal Inference and the Data-Fusion Problem." *Proceedings of the National Academy of Sciences* 113(27):7345–7352.

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 and Gwyneth McClendon. 2020. Experiments in Multiple Contexts. In *Handbook of Experimental Political Science*, ed. Donald P. Green and James Druckman. Cambridge University Press.

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

Campbell, Donald T and Julian C Stanley. 1963. *Experimental and quasi-experimental designs for research*. RandMcNally.

Cole, Stephen R and Elizabeth A Stuart. 2010. "Generalizing Evidence From Randomized Clinical Trials to Target PopulationsThe ACTG 320 Trial." *American Journal of Epidemiology* 172(1):107–115.

Dahabreh, Issa J, Sarah E Robertson, Eric J Tchetgen Tchetgen, Elizabeth A Stuart and Miguel A Hernán. 2019. "Generalizing Causal Inferences From Individuals In Randomized Trials to All Trial-Eligible Individuals." *Biometrics* 75(2):685–694.

Deaton, Angus and Nancy Cartwright. 2018. "Understanding and Misunderstanding Randomized Controlled Trials." *Social Science & Medicine* .

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.

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.

Egami, Naoki and Erin Hartman. 2021. "Covariate Selection for Generalizing Experimental Results." *Journal of the Royal Statistical Society, Series A* .

Findley, Michael G, Kyosuke Kikuta and Michael Denly. 2020. "External Validity." *Annual Review of Political Science* .

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

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.

Imai, Kosuke, Gary King and Elizabeth A Stuart. 2008. "Misunderstandings Between Experimentalists and Observationalists About Causal Inference." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 171(2):481–502.

Karmakar, Bikram and Dylan S. Small. 2020. "Assessement of The Extent of Corroboration of An Elaborate Theory of A Causal Hypothesis Using Partial Conjunctions of Evidence Factors." *Annals of Statistics* .

Kern, Holger L, Elizabeth A Stuart, Jennifer Hill and Donald P Green. 2016. "Assessing Methods for Generalizing Experimental Impact Estimates to Target Populations." *Journal of Research on Educational Effectiveness* 9(1):103–127.

Miratrix, Luke W, Jasjeet S Sekhon, Alexander G Theodoridis and Luis F Campos. 2018. "Worth Weighting? How to Think About and Use Weights in Survey Experiments." *Political Analysis* 26(3):275–291.

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

Neyman, Jerzy. 1923. "On the Application of Probability Theory to Agricultural Experiments. Essay on Principles (with discussion). Section 9 (translated)." *Statistical Science* 5(4):465–472.

Paluck, Elizabeth Levy, Seth A Green and Donald P Green. 2019. "The Contact Hypothesis Re-Evaluated." *Behavioural Public Policy* 3(2):129–158.

Robins, James M, Andrea Rotnitzky and Lue Ping Zhao. 1994. "Estimation of Regression Coefficients When Some Regressors Are Not Always Observed." *Journal of the American Statistical Association* 89(427):846–866.

Rubin, Donald B. 1974. "Estimating Causal Effects of Treatments in Randomized and Non-randomized Studies." *Journal of Educational Psychology* 66(5):688.

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

Tipton, Elizabeth. 2013. "Improving Generalizations From Experiments Using Propensity Score Subclassification: Assumptions, Properties, and Contexts." *Journal of Educational and Behavioral Statistics* 38(3):239–266.

Westreich, Daniel, Jessie K Edwards, Catherine R Lesko, Stephen R Cole and Elizabeth A Stuart. 2019. "Target Validity and The Hierarchy of Study Designs." *American Journal of Epidemiology* 188(2):438–443.

Wilke, Anna and Macartan Humphreys. 2020. Field Experiments, Theory, and External Validity. In *The SAGE Handbook of Research Methods in Political Science and International Relations*, ed. Luigi Curini and Robert Franzese. Transaction Publishers.

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

Online Supplementary Appendix

Elements of External Validity:
Framework, Design, and Analysis

Table of Contents

A Effect-Generalization **1**

A.1 Identification of the T-PATE 1

A.2 Three Classes of Estimators 2

A.3 Inference with Bootstrap 3

B Sign-Generalization **3**

B.1 Fisher's Combined p-value 3

B.2 Statistical Power and Purposive Variations 4

C Empirical Applications: Full Analysis **4**

C.1 Field Experiment: Reducing Transphobia 4

C.2 Survey Experiment: Partisan-Motivated Reasoning 8

C.3 Lab Experiment: The Effect of Emotions on Dissent in Autocracy 10

D Metaketa **14**

D.1 Motivating Example 14

D.2 Sign-Generalization Test 14

D.3 Effect-Generalization 16

E External Validity Analysis of Observational Studies **17**

E.1 Motivating Example 17

E.2 Sign-Generalization 18

F Economics-Type Lab Experiment **20**

F.1 General Discussion 20

F.2 Motivating Example 20

G Relationship to Other Concepts **22**

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

H Statistical Details of Proposed Methodologies

I Validation Study Using Multi-Site Experiment

J Simulations

K Literature Review of *American Political Science Review*

L Numeric Results and Model Specification

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,

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

  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 $\mathbf{x} \in \mathcal{X}$, $\mathbf{m} \in \mathcal{M}$,

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

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

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

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

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

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

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

- Consistency: For all units,

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

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.

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

where the 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:

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

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 overlap 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 overlap 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]

[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 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

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). 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. 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 than a conversation with a non-transgender individual. The authors partnered with an LGBT

---

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.

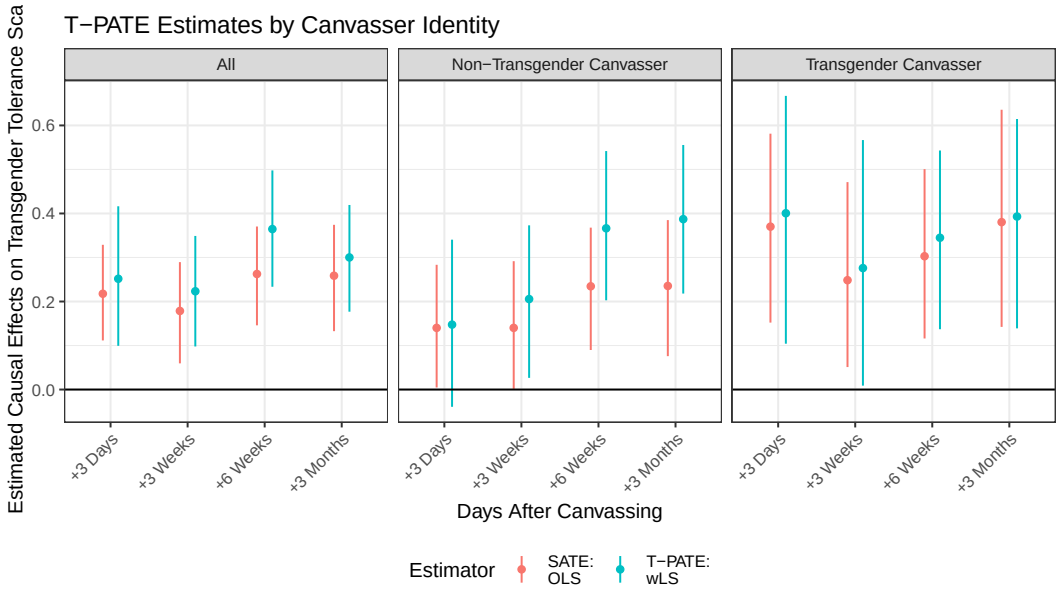


Figure A1: T-PATE Estimates for Broockman and Kalla (2016) By Canvasser Identity *Note:* The x-axis within panels represents survey waves (3 days, 3 weeks, 6 weeks, 3 months). Panels present canvasser identity. Estimates are for the SATE, with pre-specified controls (pink) and the T-PATE with weighted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

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 overlap 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 $b \times \mathsf{T\text{-}PATE} - \frac{1}{N}\sum_{i=1}^{N} c_i$. Therefore, the T-PATE estimate is directly useful for the cost-benefit analysis. In particular, we can estimate the average utility by $b \times \widehat{\mathsf{T\text{-}PATE}} - \frac{1}{N}\sum_{i=1}^{N} c_i$ where $\widehat{\mathsf{T\text{-}PATE}}$ is estimated by one of the three classes of estimators discussed in Section 5. The standard error is $b \times \widehat{\mathsf{se}}(\widehat{\mathsf{T\text{-}PATE}})$ where $\widehat{\mathsf{se}}(\widehat{\mathsf{T\text{-}PATE}})$ 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

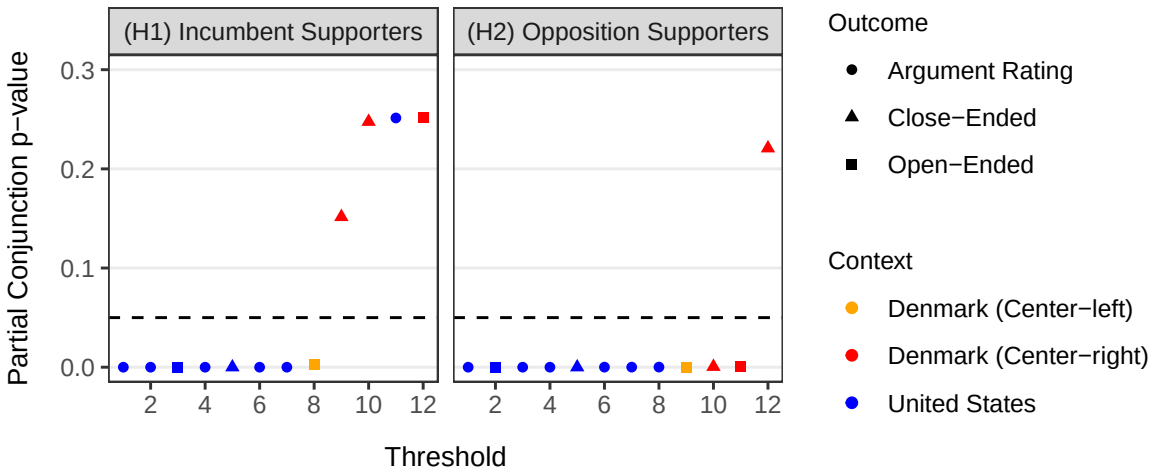


Figure A2: Sign-Generalization Test for Bisgaard (2019). *Note:* We combine causal estimates on multiple outcomes across four survey experiments in three contexts. Following Section 6, we report partial conjunction p-values for all thresholds.

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 overlap 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 overlap 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) $\times$ 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.

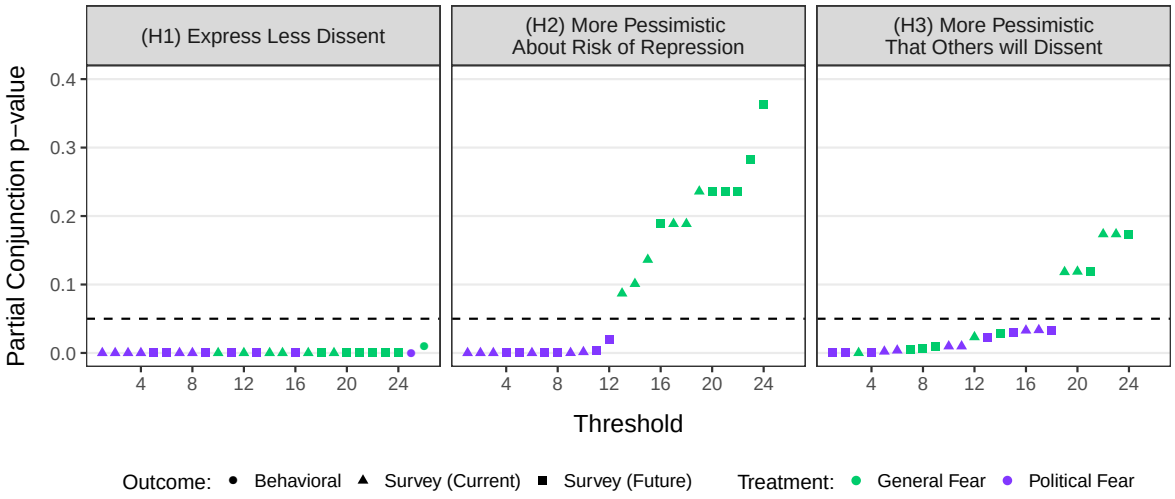


Figure A3: Sign-Generalization Test for Young (2019). *Note:* We combine causal estimates on multiple outcomes across two treatment-variations. In total, we have 26 causal estimates for the first hypothesis and 24 for the second and third hypotheses. Following Section 6, we report partial conjunction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

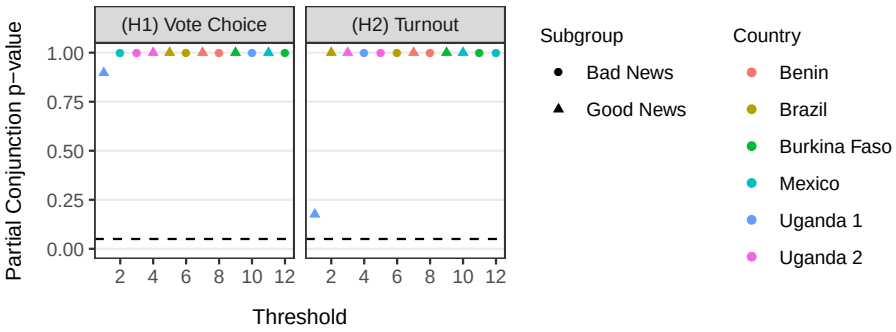


Figure A4: Sign-generalization test for Dunning et al. (2019) for the primary (vote for incumbent) and secondary (voter turnout) outcomes. Country is represented by color and subgroup by symbol.

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 overlap 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.

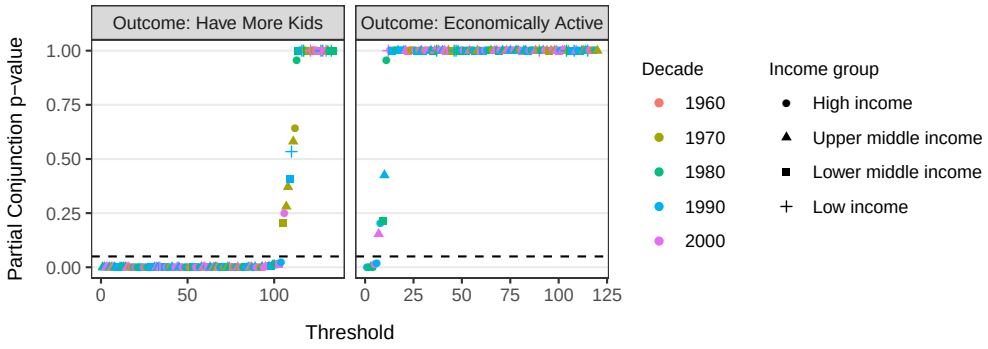


Figure A5: Sign-generalization test for Dehejia, Pop-Eleches and Samii (2021). Outcomes by study-design are represented by columns, country classification from the World Bank is represented by symbol, and color represents the decade of census.

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

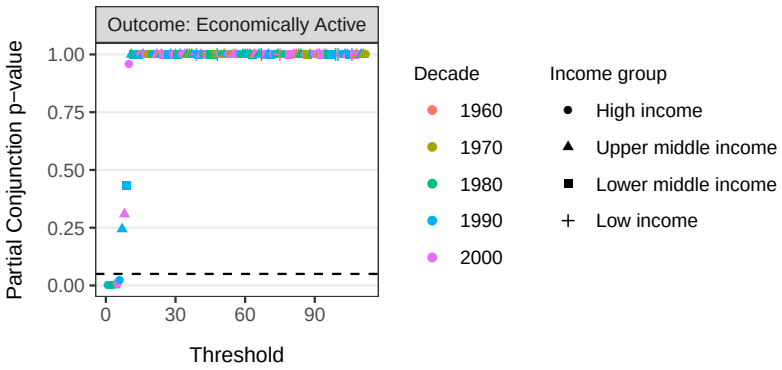


Figure A6: Sign-generalization test for Bisbee et al. (2017) for the "Economically Active" outcome. Country classification from the World Bank is represented by symbol, and color represents the decade of census.

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 undergrad-

uate 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.

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.

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 Validation Study Using Multi-Site Experiment **10**

I.1 Design of Evaluation 10

I.2 Results 12

J Simulations **13**

J.1 Full Simulations 13

J.2 Naturalistic Simulations 18

K Literature Review of *American Political Science Review* **23**

L Numeric Results and Model Specification **29**

L.1 Results for Broockman and Kalla (2016) analysis in Figure 7 29

L.2 Results for Bisgaard (2019) analysis in Figure 8 33

L.3 Results for Broockman and Kalla (2016) analysis in Figure A1 35

L.4 Results for Bisgaard (2019) analysis in Figure A2 36

L.5 Results for Young (2019) analysis in Figure A3 38

L.6 Results for Dehejia, Pop-Eleches and Samii (2021) analysis in Figure A5 . . 41

L.7 Results for Bisbee et al. (2017) analysis in Figure A6 49

L.8 Results for Dunning et al. (2019) analysis in Figure A4 53

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.

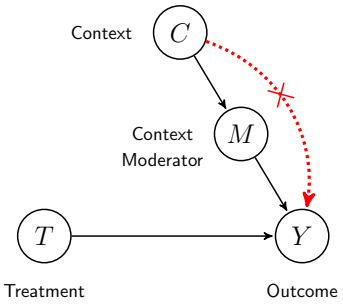


Figure A7: Causal DAG for Contextual Exclusion Restriction.

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

$$\begin{aligned} & \frac{1}{R} \sum_{i=1}^{R} \sum_{i=1}^{R} \theta(\mathbf{M}_i, \mathbf{X}_i; \widehat{\psi}_C)\pi(\mathbf{M}_i, \mathbf{X}_i; \widehat{\psi}_S)\delta_i \mathbf{1}\{C_i = c\} S_i T_i \{Y_i - g_1(\mathbf{M}_i, \mathbf{X}_i; \widehat{\xi}_1)\} \\ \xrightarrow{p} \quad & \mathbb{E}[\theta(\mathbf{M}_i, \mathbf{X}_i; \tilde{\psi}_C)\pi(\mathbf{M}_i, \mathbf{X}_i; \tilde{\psi}_S)\delta_i \mathbf{1}\{C_i = c\} S_i T_i \{Y_i - g_1(\mathbf{M}_i, \mathbf{X}_i; \psi_1^*)\}]. \end{aligned}$$

Now, we can rewrite the expression as follows.

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

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 Validation Study Using Multi-Site Experiment

In this section, we evaluate the performance of the T-PATE estimators using multi-site experiments by Meager (2019) as an experimental benchmark. At its core, we view one site as the target population data and the average treatment effect in this target site as the T-PATE, i.e., our target of inference. We can then evaluate how well the T-PATE estimator can generalize from the remaining sites to this target site.

There are two key advantages of using multi-site experiments to evaluate the performance of the T-PATE estimator. First, in contrast to conventional simulation studies, we use the data from the real empirical application (Meager, 2019), and thus, the data generating process and sampling models are realistic. Most importantly, identification and modeling assumptions necessary for estimating the T-PATE are not guaranteed to be satisfied as in the real empirical application. Thus, it provides a clean evaluation design. Second, unlike the direct application of the T-PATE estimator to a single experiment, we have an experimental benchmark estimate of the T-PATE, as it is simply the SATE in the target population data. Therefore, we can clearly evaluate whether the T-PATE estimators can recover this experimental benchmark.

In summary, this validation study helps us evaluate whether both identification and modeling assumptions are plausible in the real-world application, and the proposed estimators can estimate the T-PATE without bias. We find that the outcome-based and doubly robust estimators could recover the experimental benchmark, and yet weighting-based estimators did not perform well. This result implies that modeling assumptions are as important as the identification assumptions for generalization. As emphasized in Section 5.1.4, we see the benefit of the doubly robust estimator — consistent when either the outcome or sampling model is correctly specified.

Finally, we also note here the scope condition of this validation study; this is an empirical validation study with one multi-site experiment. Thus, like any other causal inference methodologies, whether identification and modeling assumptions are plausible in each study depends on applications, and should be evaluated with domain knowledge.

I.1 Design of Evaluation

Using Meager (2019), we re-analyze microcredit experiments conducted by different authors across four sites.[1] These studies all involve a similar treatment — expanding access to credit — implemented across different countries. We consider causal effects on three economic outcomes: household-level revenues, profits, and expenditures.[2]

Estimation of the T-PATE requires both identification and modeling assumptions. If any of the following assumptions are violated, we will be able to detect such violations because the

[1] We focus on the four studies from her original analysis that have overlapping baseline outcome measures, including Angelucci, Karlan and Zinman (2015), Attanasio et al. (2015), Augsburg et al. (2015), and Crépon et al. (2015).

[2] To improve comparability across sites, we standardize the outcomes by the mean and standard deviation of control group baseline outcome.

T-PATE estimator will differ from the true T-PATE. First, we consider the identification of the T-PATE, which requires strong assumptions regarding each dimension of external validity. For X-validity, analysts need to assume Ignorability of Sampling and Treatment Effect Heterogeneity (Assumption 1). To make this assumption plausible, we adjust for seven covariates that are measured across sites, i.e., gender, age, income, the amount spent on food, and baseline measures of the three economic outcomes. For T-validity, we need to assume ignorable treatment-variations (Assumption 2). While the treatment is similar in each study, each site implemented a slightly different treatment regime, such as whether the loan had individual or group liability; whether it was collateralized; the interest rate; and if the loans were targeted at specific groups, such as women. Therefore, it is possible that this required assumption is violated. Similarly, for Y-validity, we need to assume ignorable outcome-variations (Assumption 3). While we follow Meager (2019)'s standardization protocol across sites, each site varied in endpoint length, and outcome measures were not specifically coordinated across sites. Therefore, it is possible that the required assumption of ignorable outcome-variations is violated. Finally, for C-validity, these four sites are from four different countries — Bosnia and Herzegovina, Mexico, Mongolia, and Morocco — with different demographic, political, and economic characteristics. Therefore, this provides us with challenging generalization tasks. While we cannot directly evaluate these assumptions, by evaluating how well we can recover the T-PATE, we can indirectly see whether these necessary assumptions are plausible.

We will consider three classes of estimators proposed in Section 5: weighting-based, outcome-based, and doubly-robust estimators, each having different modeling assumptions. Weighting-based estimators include IPW and weighted least squares. Sampling weights are estimated via calibration.[3] For the outcome-based estimators, we use OLS and a more flexible model, BART. We implement two doubly robust estimators; the AIPW with OLS and the AIPW with BART. As explained above, in each estimator, we adjust for the following seven variables: gender, age, income, the amount spent on food, and baseline measures of the three outcomes. For each monetary variable (income, the amount spent on food, and the baseline outcome measures), we log the variable and standardize it, within each site, using the control group mean and standard deviation. We also include dummy variables for if the monetary measure is zero, and for if the baseline outcome measure is missing. Finally, following Meager (2019), we also include a dummy variable for whether a household has an existing business, a dimension on which the author finds significant treatment effect heterogeneity.

We define one site as the target population data, and then combine the three remaining sites as the experimental data. We then use the three classes of the T-PATE estimators (proposed in Section 5) on the experimental data in order to estimate the T-PATE in the target site. Finally, we compare this estimate to the average causal effect in the target site,

[3]We had to drop some variables that lead calibration weights to fail to converge. These variables typically involved the indicators for whether a monetary variable was zero or missing. But calibration also required dropping some of the monetary values in certain sites. This instability of the weight estimation results in the poor performance of weighting-based estimators, as we see below.

which can be estimated via the difference-in-means. We repeat this exercise using all four sites as the target population data, and estimate the T-PATE for all three outcomes of interest. We have 12 estimates in total.

Formally, we denote data from site k by $\mathcal{D}_k$ where $k \in \{1, 2, 3, 4\}$, and we denote outcome variable by Y_j where $j \in \{1, 2, 3\}$ as there are three economic outcomes. Suppose we view Site 1 as the target population, and data from Site 2, Site 3, Site 4 as the experimental data. Then, for outcome Y_j, the $\mathsf{T\text{-}PATE}_{j1}$ is the average causal effect on outcome Y_j in Site 1, which can be estimated with the difference-in-means using data $\mathcal{D}_1$. We will use the data from Site 2, Site 3, Site 4 $\{\mathcal{D}_2, \mathcal{D}_3, \mathcal{D}_4\}$, to construct an estimator $\widehat{\mathsf{T\text{-}PATE}}_{j1}$ and its 95% confidence interval $\widehat{\mathrm{CI}}_{j1}$. By repeating this exercise using all four sites and three outcomes of interest, we are going to evaluate the average absolute bias and the coverage of the T-PATE estimator across 12 estimates.

$$\begin{aligned} \text{Average Absolute Bias} &= \frac{1}{12}\sum_{j=1}^{3}\sum_{k=1}^{4} |\mathsf{T\text{-}PATE}_{jk} - \widehat{\mathsf{T\text{-}PATE}}_{jk}| \\ \text{Coverage} &= \frac{1}{12}\sum_{j=1}^{3}\sum_{k=1}^{4} \mathbf{1}\{\mathsf{T\text{-}PATE}_{jk} \in \widehat{\mathrm{CI}}_{jk}\} \end{aligned}$$

I.2 Results

Table A1 presents results. As a reference, we also report the SATE estimate, which is the difference-in-means of the experimental data without any adjustment. With this reference, we can understand how much bias the proposed T-PATE estimator can reduce in practice.

First, we find that all of the estimators, except for the IPW estimator, reduce bias by a large amount. For example, the outcome-based BART estimator reduces by about 95% compared to the SATE. Second, when we consider the coverage of the 95% confidence intervals, we find that outcome-based and doubly robust estimators have close to the nominal coverage. In contrast, the weighting-based estimators exhibit significant under-coverage as low as the SATE estimate.[4]

Several points are worth emphasizing. First, the fact that both outcome-based and doubly robust estimators could recover the experimental benchmark with small bias and desirable coverage rates suggests that identification and modeling assumptions necessary for the T-PATE estimation are plausible in this application even with the limited set of control variables. This is not trivial because there exists a significant amount of the external validity bias to be corrected as we see that the SATE estimator has a large bias and severe under-coverage.

[4]The poor performance of the weighting based estimators indicates that the sampling weights in our example are likely to be misspecified. Within the data, the distributions of the logged and standardized monetary variables, particularly the previous outcomes, are bimodal since a large fraction of families has no previous business revenues, profits, or expenditures. We included binary variables to capture this bimodal nature, but the calibration still led to extreme weights. Even with a maximum weight cap of 10, the distribution of weights for these sites has a long tail, putting relatively extreme weight on a few observations. This demonstrates the difficulty of correctly specifying the sampling model in this example.

Estimator	Average Absolute Bias	Coverage
SATE	0.067	0.50 (6/12)
Weighting-based		
IPW	0.082	0.58 (7/12)
wLS	0.027	0.50 (6/12)
Outcome-based		
OLS	0.038	0.92 (11/12)
BART	0.004	0.92 (11/12)
Doubly Robust		
AIPW with OLS	0.044	0.92 (11/12)
AIPW with BART	0.012	0.92 (11/12)

Table A1: Validation with Microcredit Multisite Experiments.

Second, the finding that the outcome-based and doubly robust estimators performed well and weighting-based estimators did not perform well implies that modeling assumptions are as important as the identification assumptions, and that outcome-models are more likely to be correctly specified and the sampling model (i.e., sampling weights) is likely to be misspecified.

Finally, as emphasized in Section 5.1.4, this empirical validation study shows the benefit of the doubly robust estimator, which is consistent when either the sampling or outcome model is correctly specified. Even though sampling weights are likely to be misspecified, the AIPW estimators reduce, or nearly eliminate, bias and exhibit near nominal coverage of the experimental benchmark because the outcome model is likely to be properly specified. This validation study shows the importance of this double robustness property for applied researchers.

J 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.

J.1 Full Simulations

J.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.

J.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).

J.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 A2. 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.

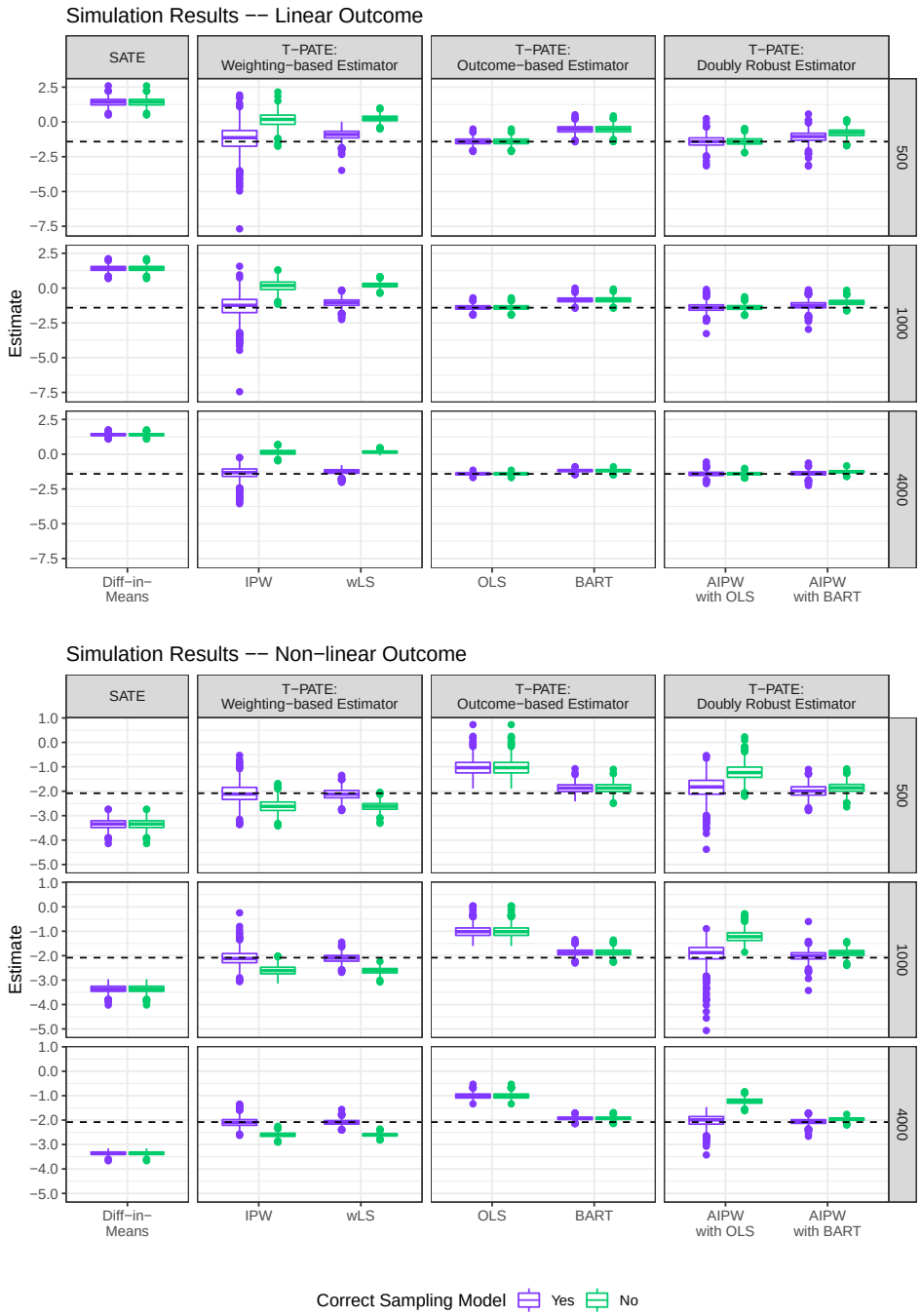


Figure A8: Full Simulation Results. The top panel presents results when the true outcome model is linear; the bottom panel presents results when the true outcome model is non-linear. Results for the correct sampling model are presented in purple, and those for the incorrectly specified sampling model 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 A2: 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	2.835	0.289	8.119	1.134	-1.265	0.206	1.644	0.802
Diff-in-Means	1000	2.831	0.207	8.058	0.801	-1.275	0.152	1.650	0.572
	4000	2.823	0.103	7.980	0.401	-1.275	0.075	1.631	0.287
	500	0.200	0.952	0.945	3.100	-0.014	0.378	0.143	1.229
IPW	1000	0.098	0.780	0.618	2.538	-0.014	0.290	0.084	0.974
	4000	0.055	0.461	0.216	1.543	-0.007	0.180	0.033	0.595
	500	0.495	0.368	0.381	1.343	-0.037	0.220	0.050	0.810
wLS	1000	0.355	0.293	0.212	1.086	-0.024	0.176	0.032	0.632
	4000	0.156	0.193	0.061	0.709	-0.010	0.110	0.012	0.385
	500	0.008	0.246	0.060	0.948	1.060	0.321	1.227	1.129
OLS	1000	0.016	0.179	0.032	0.665	1.076	0.238	1.215	0.831
	4000	0.001	0.088	0.008	0.331	1.072	0.119	1.163	0.437
	500	0.886	0.265	0.855	1.089	0.211	0.197	0.083	0.792
BART	1000	0.568	0.193	0.359	0.780	0.213	0.138	0.064	0.568
	4000	0.239	0.095	0.066	0.399	0.154	0.068	0.028	0.292
	500	0.009	0.391	0.153	1.343	0.212	0.463	0.259	1.498
AIPW with OLS	1000	0.014	0.309	0.096	1.036	0.144	0.425	0.202	1.228
	4000	-0.002	0.174	0.030	0.614	0.052	0.244	0.062	0.813
	500	0.335	0.398	0.270	1.802	0.098	0.258	0.076	1.282
AIPW with BART	1000	0.173	0.306	0.123	1.375	0.068	0.207	0.048	0.982
	4000	0.036	0.173	0.031	0.742	0.022	0.119	0.015	0.535
Incorrect Sampling Model									
	500	2.835	0.289	8.119	1.134	-1.265	0.206	1.644	0.802
Diff-in-Means	1000	2.831	0.207	8.058	0.801	-1.275	0.152	1.650	0.572
	4000	2.823	0.103	7.980	0.401	-1.275	0.075	1.631	0.287
	500	1.576	0.511	2.746	1.971	-0.527	0.254	0.343	0.950
IPW	1000	1.581	0.383	2.646	1.425	-0.525	0.183	0.309	0.686
	4000	1.552	0.192	2.445	0.740	-0.521	0.095	0.281	0.355
	500	1.665	0.238	2.830	0.932	-0.538	0.170	0.319	0.661
wLS	1000	1.626	0.180	2.677	0.691	-0.532	0.129	0.300	0.486
	4000	1.571	0.094	2.478	0.373	-0.522	0.070	0.278	0.258
	500	0.008	0.246	0.060	0.948	1.060	0.321	1.227	1.129
OLS	1000	0.016	0.179	0.032	0.665	1.076	0.238	1.215	0.831
	4000	0.001	0.088	0.008	0.331	1.072	0.119	1.163	0.437
	500	0.886	0.264	0.854	1.091	0.214	0.197	0.085	0.786
BART	1000	0.567	0.193	0.359	0.784	0.215	0.138	0.065	0.569
	4000	0.239	0.096	0.066	0.399	0.154	0.067	0.028	0.294
	500	0.015	0.272	0.074	1.046	0.866	0.317	0.850	1.158
AIPW with OLS	1000	0.017	0.198	0.039	0.745	0.869	0.239	0.812	0.859
	4000	0.003	0.099	0.010	0.379	0.855	0.120	0.746	0.453
	500	0.631	0.282	0.478	1.248	0.223	0.211	0.094	0.904
AIPW with BART	1000	0.387	0.206	0.193	0.879	0.193	0.147	0.059	0.643
	4000	0.153	0.103	0.034	0.441	0.121	0.072	0.020	0.323

J.2 Naturalistic Simulations

While the analyses in Section J.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.

J.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.

J.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.

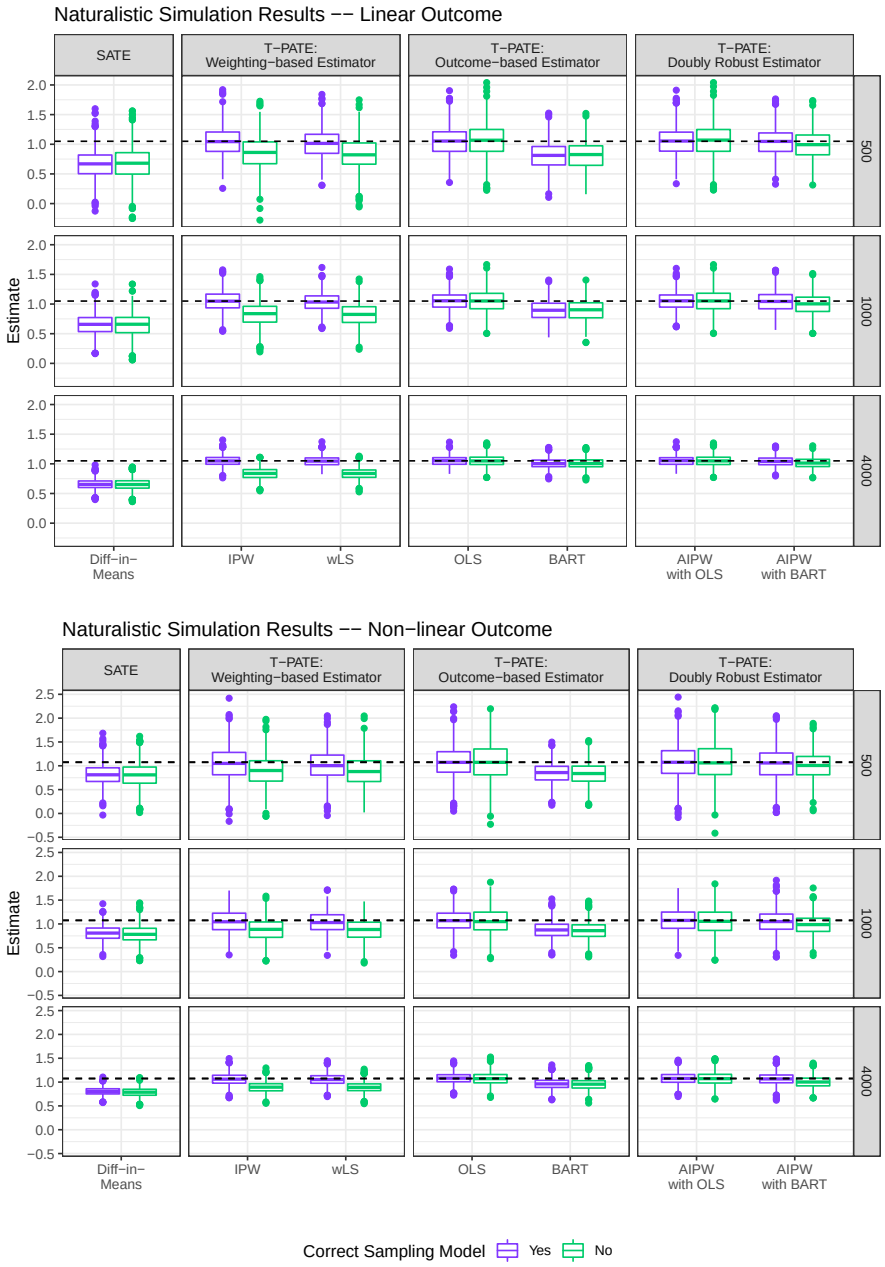


Figure A9: Naturalistic Simulation Results. In the top panel, we report results when the true outcome model is linear; in the bottom panel, we present results when the true outcome model is non-linear. Purple represents results when the sampling weights are correctly specified, and green represents results 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 $n = \{500, 1000, 4000\}$.

J.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 A3. 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.

Table A3: 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	-0.388	0.236	0.206	0.931	-0.266	0.216	0.117	0.862
Diff-in-Means	1000	-0.395	0.173	0.186	0.656	-0.272	0.159	0.099	0.609
	4000	-0.394	0.084	0.163	0.328	-0.273	0.078	0.081	0.305
	500	-0.004	0.247	0.061	1.004	-0.024	0.352	0.125	1.307
IPW	1000	0.000	0.175	0.031	0.702	-0.029	0.242	0.059	0.942
	4000	-0.001	0.088	0.008	0.348	-0.021	0.125	0.016	0.486
	500	-0.036	0.235	0.057	0.953	-0.061	0.313	0.101	1.180
wLS	1000	-0.018	0.162	0.026	0.663	-0.043	0.227	0.053	0.868
	4000	-0.006	0.082	0.007	0.327	-0.026	0.119	0.015	0.463
	500	0.000	0.235	0.055	0.927	0.004	0.331	0.110	1.292
OLS	1000	0.001	0.162	0.026	0.637	-0.006	0.224	0.050	0.885
	4000	-0.001	0.081	0.007	0.312	0.000	0.109	0.012	0.434
	500	-0.243	0.229	0.111	0.922	-0.231	0.218	0.101	0.990
BART	1000	-0.156	0.172	0.054	0.662	-0.206	0.177	0.074	0.746
	4000	-0.043	0.083	0.009	0.371	-0.115	0.114	0.026	0.465
	500	0.000	0.235	0.055	0.929	0.005	0.356	0.127	1.367
AIPW with OLS	1000	0.001	0.162	0.026	0.638	-0.001	0.243	0.059	0.958
	4000	-0.001	0.081	0.007	0.313	-0.004	0.123	0.015	0.482
	500	-0.013	0.231	0.054	0.977	-0.038	0.326	0.108	1.379
AIPW with BART	1000	-0.011	0.172	0.030	0.677	-0.032	0.235	0.056	0.974
	4000	-0.009	0.084	0.007	0.369	-0.011	0.128	0.016	0.518
Incorrect Sampling Model									
	500	-0.379	0.270	0.217	0.379	-0.269	0.251	0.135	0.351
Diff-in-Means	1000	-0.404	0.196	0.202	0.277	-0.285	0.182	0.115	0.257
	4000	-0.396	0.098	0.167	0.138	-0.283	0.090	0.088	0.130
	500	-0.196	0.281	0.118	0.395	-0.179	0.321	0.135	0.431
IPW	1000	-0.222	0.202	0.090	0.287	-0.187	0.227	0.086	0.309
	4000	-0.214	0.100	0.056	0.143	-0.178	0.112	0.044	0.159
	500	-0.213	0.271	0.119	0.383	-0.183	0.313	0.132	0.408
wLS	1000	-0.226	0.193	0.088	0.273	-0.194	0.220	0.086	0.303
	4000	-0.215	0.094	0.055	0.135	-0.179	0.110	0.044	0.153
	500	0.020	0.276	0.076	0.387	0.003	0.389	0.151	0.531
OLS	1000	-0.004	0.194	0.038	0.267	-0.011	0.266	0.071	0.373
	4000	-0.001	0.095	0.009	0.133	0.001	0.130	0.017	0.183
	500	-0.238	0.240	0.114	0.674	-0.236	0.227	0.107	0.714
BART	1000	-0.154	0.179	0.056	0.477	-0.208	0.182	0.076	0.540
	4000	-0.043	0.087	0.009	0.259	-0.116	0.116	0.027	0.332
	500	0.020	0.276	0.077	0.387	0.004	0.392	0.154	0.541
AIPW with OLS	1000	-0.004	0.194	0.038	0.267	-0.011	0.272	0.074	0.377
	4000	-0.001	0.095	0.009	0.133	0.002	0.133	0.018	0.186
	500	-0.063	0.242	0.062	0.755	-0.070	0.286	0.087	0.918
AIPW with BART	1000	-0.056	0.176	0.034	0.476	-0.088	0.211	0.052	0.618
	4000	-0.037	0.086	0.009	0.244	-0.069	0.123	0.020	0.340

K 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 A4. 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 A4.

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 A4: 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 A4: 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 A4: 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 A4: 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 A4: 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 A4: 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?

L 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.

L.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 A5 below.

Table A5: Numeric Values for T-PATE Estimates for Broockman and Kalla (2016).

Estimator	Estimate	SE	95% CI
Measurement at +3 Days			
OLS	0.218	0.055	[0.112, 0.329]
IPW	0.335	0.191	[-0.062, 0.683]
wLS	0.252	0.080	[0.099, 0.416]
OLS	0.238	0.151	[-0.054, 0.519]
BART	0.135	0.095	[-0.046, 0.325]
AIPW with OLS	0.337	0.195	[-0.042, 0.706]
AIPW with BART	0.322	0.160	[0.019, 0.635]
Measurement at +3 Weeks			
OLS	0.179	0.059	[0.059, 0.289]
IPW	0.441	0.200	[0.037, 0.815]
wLS	0.223	0.062	[0.098, 0.349]
OLS	0.229	0.141	[-0.053, 0.494]
BART	0.114	0.095	[-0.067, 0.305]
AIPW with OLS	0.275	0.175	[-0.081, 0.624]
AIPW with BART	0.317	0.158	[0.005, 0.623]
Measurement at +6 Weeks			
OLS	0.263	0.056	[0.146, 0.37]
IPW	0.507	0.210	[0.088, 0.916]
wLS	0.365	0.068	[0.234, 0.498]
OLS	0.253	0.142	[-0.012, 0.515]
BART	0.203	0.095	[0.017, 0.393]
AIPW with OLS	0.328	0.182	[-0.025, 0.683]
AIPW with BART	0.396	0.156	[0.093, 0.693]
Measurement at +3 Months			
OLS	0.259	0.061	[0.133, 0.374]
IPW	0.527	0.202	[0.114, 0.928]
wLS	0.300	0.063	[0.177, 0.419]
OLS	0.204	0.145	[-0.094, 0.487]
BART	0.131	0.097	[-0.062, 0.319]
AIPW with OLS	0.265	0.182	[-0.089, 0.621]
AIPW with BART	0.300	0.172	[-0.036, 0.637]

L.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 A6 are used in the partial conjunction test.

Table A6: 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

L.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 L.1, conducted separately by whether the randomly assigned canvasser self identifies as transgender. The resulting point estimates and intervals are presented in Table A7.

Table A7: 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	0.252	0.080	[0.099, 0.416]
+3 Weeks	SATE	0.179	0.059	[0.059, 0.289]
+3 Weeks	T-PATE: Weighting-based Estimator	0.223	0.062	[0.098, 0.349]
+6 Weeks	SATE	0.263	0.056	[0.146, 0.37]
+6 Weeks	T-PATE: Weighting-based Estimator	0.365	0.068	[0.234, 0.498]
+3 Months	SATE	0.259	0.061	[0.133, 0.374]
+3 Months	T-PATE: Weighting-based Estimator	0.300	0.063	[0.177, 0.419]
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]

L.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 L.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 A8: 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

L.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 A9, 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 A9: 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 A9: 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 A9: 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

L.6 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

L.7 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

L.8 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 A12.

Table A12: 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

References

Angelucci, Manuela, Dean Karlan and Jonathan Zinman. 2015. "Microcredit impacts: Evidence from a randomized microcredit program placement experiment by Compartamos Banco." *American Economic Journal: Applied Economics* 7(1):151–82.

Attanasio, Orazio, Britta Augsburg, Ralph De Haas, Emla Fitzsimons and Heike Harmgart. 2015. "The impacts of microfinance: Evidence from joint-liability lending in Mongolia." *American Economic Journal: Applied Economics* 7(1):90–122.

Augsburg, Britta, Ralph De Haas, Heike Harmgart and Costas Meghir. 2015. "The impacts of microcredit: Evidence from Bosnia and Herzegovina." *American Economic Journal: Applied Economics* 7(1):183–203.

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.

Crépon, Bruno, Florencia Devoto, Esther Duflo and William Parienté. 2015. "Estimating the impact of microcredit on those who take it up: Evidence from a randomized experiment in Morocco." *American Economic Journal: Applied Economics* 7(1):123–50.

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.

Meager, Rachael. 2019. "Understanding the Average Impact of Microcredit Expansions: A Bayesian Hierarchical Analysis of Seven Randomized Experiments." *American Economic Journal: Applied Economics* 11(1):57–91.

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.