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

Naoki Egami† Erin Hartman‡

First Version: June, 30, 2020
This Version: October 20, 2020

Abstract

External validity of randomized experiments 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 (units, treatments, outcomes, and contexts). The proposed framework synthesizes diverse external validity concerns that arise in practice. 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* — assessing whether the direction of causal effects is generalizable, we propose a novel multiple-testing procedure under weaker assumptions. We illustrate our methods through three applications covering field, survey, and lab experiments.

---

*The proposed methodology is implemented via our forthcoming open-source software R package, `evalid`. We would like to thank Martin Bisgaard, David Broockman, Graeme Blair, Brandon de la Cuesta, Don Green, Jens Hainmueller, Dan Hopkins, Joshua Kalla, Ian Lundberg, Kevin Munger, Santiago Olivella, Abby Wood, Yang-Yang Zhou, and Stephanie Zonszein for their thoughtful comments. We would also like to thank panel participants and attendees at Polmeth 2020 and APSA 2020.

†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 Statistics and of Political Science, University of California, Los Angeles, Los Angeles, CA 90095. Email: ekhartman@ucla.edu, URL: `www.erinhartman.com`

1 Introduction

Over the last few decades, social scientists have extensively used randomized experiments to learn about causal relationships. The central advantage of experimental studies is their *internal* validity — due to the randomization of treatment, researchers can unbiasedly estimate causal effects *within* an experiment, without making strong modeling assumptions. One of the most important long-standing methodological debates is about *external validity* of randomized experiments — whether and how scientists can generalize causal findings beyond a specific experiment. The problem is so fundamental that the vast majority of textbooks on experimental methods have at least one chapter on external validity (see e.g., Morton and Williams, 2010; Mutz, 2011; Gerber and Green, 2012).

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 12% of all experimental studies published in American Political Science Review from 2015 to 2019 contain any type of formal analyses for external validity,[1] and none discuss conditions under which generalization is credible. The lack of empirical approaches for external validity has remained because social science experiments have diverse goals and concerns surrounding external validity, yet most existing methodology has primarily focused on a subset of threats that are statistically more tractable. As a consequence, many 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 (see Figure 1). 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). A fundamental question about external validity is "if we could do an experiment with other relevant populations, treatments, outcomes, and contexts of theoretical interest, would we obtain the same causal conclusion?" With the proposed framework, we formally synthesize a variety of external validity concerns we face in practice and relate them to causal assumptions; to name a few examples, con-

[1]Papers that included formal analyses for external validity were all survey experiments and used survey weights to adjust for sample representativeness.

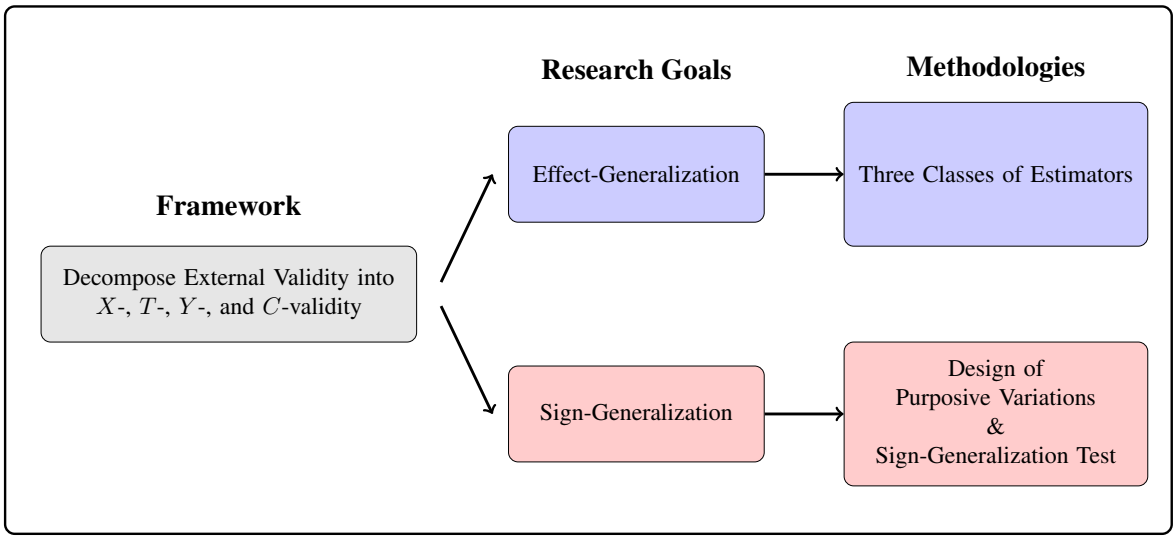

Figure 1: Proposed Approach Toward External Validity.

venience samples (X-validity), differences in treatment implementations (T-validity), survey versus behavioral outcomes (Y-validity), and differences in causal mechanisms across time, geography, and institutions (C-validity). We clarify conditions under which analysts can and cannot account for each type of validity.

Using the proposed framework, we then develop tailored methods by distinguishing two central research goals. *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 is not feasible, such as when identification assumptions are untenable or when required data are not available. We believe both goals are important in any given study, but we distinguish them to clarify different required assumptions and methods.

For effect-generalization, we introduce three classes of estimators; weighting-based, outcome-based, and doubly robust estimators (Section 4). While weighting-based estimators adjust for selection into experiments, outcome-based estimators model treatment effect heterogeneity. Doubly robust estimators combine both to further mitigate concerns about model misspecifica-

tion. For the well-researched problem of X-validity, we provide practical guidance in choosing estimators by balancing estimators' standard errors and biases due to model misspecification. Given that generalization is often involved with the differences both in populations (X) and contexts (C), we also provide new estimators that account for X- and C-validity together.

Finally, we propose a new approach to sign-generalization (Section 5). It is increasingly common to measure multiple outcomes, treatments, contexts and diverse units within experiments. 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 a little additional practical cost.

To illustrate the general applicability of our proposed approach, we discuss the external validity of three randomized experiments, covering field, survey, and lab experiments. We briefly introduce them in Section 2 and provide a full reanalysis in Section 6 and Appendix B. We discuss how the proposed approach can be used in the context of meta-analysis and observational studies in Section 7.

Our contributions are threefold. First, we formalize all four dimensions of external validity within the potential outcomes framework (Neyman, 1923; Rubin, 1974). While existing causal methods use the potential outcomes, they have focused on X-validity (Imai *et al.*, 2008; Cole and Stuart, 2010; Tipton, 2013). The classical experimental design literature (Campbell and Stanley, 1963; Shadish *et al.*, 2002) has proposed a typology and different research goals for generalization. While this literature primarily 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 the large literature (Hartman *et al.*, 2015; Kern *et al.*, 2016; Dahabreh *et al.*, 2019) and provide practical guidance. To account for X- and C-validity together, we use identification results from the causal diagram approach (Bareinboim and Pearl, 2016) and develop new estimators in Sec-

tion 4. 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. The main advantage is that the same proposed approach is applicable to all four dimensions, and it requires much weaker assumptions than those necessary for effect-generalization.

2 Motivating Empirical Applications

Before presenting the proposed approach, we describe three motivating studies representing field, survey, and lab experiments. We connect each study to questions of external validity throughout this article.

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 (Paluck and Green, 2009). 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, often called partisan-motivated reasoning (Bartels, 2002; Bullock *et al.*, 2015). 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 of the economic changes to a ruling party. Across four experiments, he finds support for his hypothesis.

2.3 Lab Experiment: The Effect of Emotions on Dissent in Autocracy

Many authoritarian countries employ various frightening acts of repression to deter dissent. To unpack 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-in-the-field 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 experimental results 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 (Campbell and Stanley, 1963; Shadish *et al.*, 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.

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 notional 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 $C_i = (\mathsf{Miami},\ \mathsf{2015})$. While this context variable C_i might have some variations within the experiment (so-called sub-contexts), we focus on a more common setting where every unit within the experiment has the same value $C_i = c$.

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 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 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 how much 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 (Hartman, 2020; Wilke and Humphreys, 2020). Formally, we define the *target population average causal effect* (T-PATE) as follows.

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

where $*$ denotes the target of each dimension. Importantly, 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. A natural question is about the difference between the SATE and the T-PATE, which we turn to next.

3.2 Typology of External Validity

With this formal setup, we can now analyze sources of biases in the SATE. 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.

We begin by decomposing external validity into four components, X-, T-, Y-, and C-validity. Formally, we can decompose total external validity bias as follows.[2]

$$\underbrace{\textsf{T-PATE} - \textsf{SATE}}_{\text{Total External Validity Bias}} = \underbrace{\Delta_X}_{X\text{-Bias}} + \underbrace{\Delta_T}_{T\text{-Bias}} + \underbrace{\Delta_Y}_{Y\text{-Bias}} + \underbrace{\Delta_C}_{C\text{-Bias}} \tag{3}$$

where each bias component corresponds to one of the four dimensions of external validity.

$$\Delta_X \equiv \mathbb{E}_{\mathcal{P}^*}\{Y_i(T = 1, c) - Y_i(T = 0, c)\} - \mathbb{E}_{\mathcal{P}}\{Y_i(T = 1, c) - Y_i(T = 0, c))\}, \qquad \text{(from } \mathcal{P} \text{ to } \mathcal{P}^*\text{)}$$

[2]The decomposition is not unique. However, the main point — each bias term refers to only one dimension, keeping other dimensions fixed —- is the same regardless of how we express the decomposition.

$$\Delta_T \equiv \mathbb{E}_{\mathcal{P}^*}\{Y_i(T^*=1,c) - Y_i(T^*=0,c)\} - \mathbb{E}_{\mathcal{P}^*}\{Y_i(T=1,c) - Y_i(T=0,c)\}, \qquad \text{(from } T \text{ to } T^*\text{)}$$

$$\Delta_Y \equiv \mathbb{E}_{\mathcal{P}^*}\{Y_i^*(T^*=1,c) - Y_i^*(T^*=0,c)\} - \mathbb{E}_{\mathcal{P}^*}\{Y_i(T^*=1,c) - Y_i(T^*=0,c)\}, \qquad \text{(from } Y \text{ to } Y^*\text{)}$$

$$\Delta_C \equiv \mathbb{E}_{\mathcal{P}^*}\{Y_i^*(T^*=1,c^*) - Y_i^*(T^*=0,c^*)\} - \mathbb{E}_{\mathcal{P}^*}\{Y_i^*(T^*=1,c) - Y_i^*(T^*=0,c)\}. \qquad \text{(from } c \text{ to } c^*\text{)}$$

In the following subsections, we show how practical concerns in each dimension are related to fundamental causal assumptions. As we see below, those assumptions are strong, and they directly reflect the inherent difficulty and importance of addressing external validity concerns. Table 1 previews a summary of the four dimensions.

	Practical Concerns (examples)	Causal Assumptions (formalization)
X-validity	Convenience samples, Selection bias, Survey non-response, Attrition	Ignorability of Sampling and Treatment Effect Heterogeneity
T-validity	Realistic treatments, Bundled treatments, Difference in implementations	Ignorable Treatment-Variations
Y-validity	Proxies, Short- or long-term outcomes, Cross-national comparability	Ignorable Outcome-Variations
C-validity	Mechanisms differ across time, geography, political institutions, ...	Contextual Exclusion Restriction

Table 1: Summary of Typology.

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 (Heckman, 1979; Imai *et al.*, 2008). Randomized experiments are often criticized for using convenience samples, and many researchers are worried that estimated causal effects for such samples might not be generalizable to other target populations.

X-bias, Δ_X, is the difference in the average causal effects induced by the difference between experimental samples $\mathcal{P}$ and the target population $\mathcal{P}^*$. X-bias is zero when selection into the experiment and treatment effect heterogeneity are unrelated to each other after adjusting for pre-treatment covariates $\mathbf{X}$ (Cole and Stuart, 2010; Pearl and Bareinboim, 2014).

Assumption 1 (Ignorability of Sampling and Treatment Effect Heterogeneity)

$$Y_i(T=1,c) - Y_i(T=0,c) \perp\!\!\!\perp S_i \mid \mathbf{X}_i \tag{4}$$

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

Importantly, Assumption 1 clarifies that both the sampling process and the sources of treatment effect heterogeneity are important when analyzing X-validity.

This formal expression synthesizes two common approaches for addressing X-validity (Egami and Hartman, 2018). The first approach aims to account for how subjects are sampled into the experiment, including the common practice of using sampling weights (Mutz, 2011; Hartman *et al.*, 2015; Miratrix *et al.*, 2018). Random sampling is a well-known special case. The second common approach is based on treatment effect heterogeneity (Kern *et al.*, 2016; Nguyen *et al.*, 2017). If analysts can adjust for all the variables explaining treatment effect heterogeneity, Assumption 1 holds. 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.

Finally, we clarify the important distinction between X- and C-validity, which often arise together in practice. For example, when analysts want to generalize experimental results from one geography to another, typically the populations and underlying mechanisms are different across two locations. X-validity, defined above, relates to differences in the populations of units, which is distinct from C-validity — which concerns whether treatment effects on the *same units* changes across contexts — which we discuss further in Section 3.2.4. In Section 4, we discuss how to adjust for X- and C-validity together to estimate the T-PATE.

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 as the 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 implementations have important effects on the treatment effectiveness.

The intensity of the treatment is also a question about T-validity. Broockman and Kalla (2016)'s treatment involved a single ten-minute conversation. For a cost-benefit analysis, researchers might be interested in whether the treatment can achieve a similar effect size with five-minute conversations. Given that analysts do not know a priori whether this target treatment can induce the same treatment effect as the one implemented in the experiment, this is a concern about T-validity.

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, the target treatment.

In lab-experiments, this concern is often about bundled treatments. To test theoretical mechanisms in experiments, it is important to experimentally manipulate targeted variables of interest and 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 discussions, 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, T-bias is zero when the treatment-variation is irrelevant to treatment effects.

Assumption 2 (Ignorable Treatment-Variations)

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

It states that the assigned treatment T and the target treatment T^* induce the same 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 the potential violation of this same assumption. Thus, we can develop a general method — a new sign-generalization test in Section 5 — 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, 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. Y-validity analysis asks 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 and how much outcome measures used in an experimental study are informative about the target outcomes of interest. Similar to T-bias, Y-bias is zero when the outcome-variation is irrelevant to treatment effects.

Assumption 3 (Ignorable Outcome-Variations)

$$\mathbb{E}_{\mathcal{P}}[Y_i^*(T=1,c) - Y_i^*(T=0,c)] \;=\; \mathbb{E}_{\mathcal{P}}[Y_i(T=1,c) - Y_i(T=0,c)]. \tag{6}$$

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., Green and Gerber, 2008).

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 5, we discuss how to address this common concern about Assumption 3 by using multiple outcomes.

As discussed in Section 3.1, we reemphasize the importance of specifying the target outcome against which we evaluate Y-validity, which is the same as specifying the scope condition. If researchers are interested in outcomes that capture a different concept (e.g., support for ethnic minorities rather than the one for transgender people studied in Broockman and Kalla (2016)), they should conduct a different experiment. Y-validity is about the robustness to outcome-variations that capture the same concept and that are implied by the same substantive theories. This is critical because there are often many outcome measures that have equally high construct validity and share the same observable implications.

We note that there are many issues about measurement that are important, independent of external validity concerns, that is, even when only considering the quality of outcome measures within experiments. Examples include measurement error (e.g., Ansolabehere *et al.*, 2008), social desirability bias (Blair and Imai, 2012), and construct validity (Shadish *et al.*, 2002). Notwithstanding the importance of these issues, we think it is important to distinguish concerns about outcome measures themselves and the issue of Y-validity, which concerns the relationship between two outcome measures Y and Y^*.

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). Although many different concerns are often labeled as related to "context," we define C-validity as a question about mechanisms; how do treatment effects on the *same* units change across contexts? This is also known as transportability (Bareinboim and Pearl, 2016).

Social scientists often discuss geography and time as important contexts (e.g., Deaton and Cartwright, 2018; Munger, 2019; Wilke and Humphreys, 2020). 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.

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 Pearl and Barein-

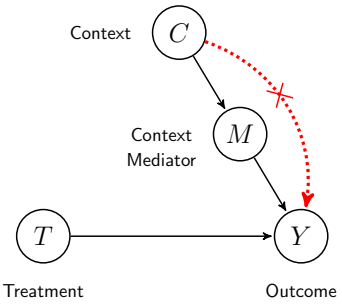

Figure 2: Causal DAG for Contextual Exclusion Restriction.

boim (2011) and further developed in Bareinboim and Pearl (2016) using a causal graphical approach. Building on this emerging literature, we formalize the C-validity within the potential outcomes framework introduced in Section 3.1.[3]

Intuitively, in order to generalize across contexts, we need to adjust for variables related to mechanisms through which contexts affect outcomes. We refer to such intermediate variables between contexts and outcomes as *context-mediators*, which are different from the standard mediator that is between treatments and outcomes (see Figure 2). More formally, we need to assume that contexts affect outcomes only through measured context-mediators. We call this the *contextual exclusion restriction* (Assumption 4), whose name reflects its similarity to the exclusion restriction well known in the instrumental variable literature. This assumption formally captures long-standing discussions about the importance of mechanisms in external validity analysis (e.g., Wilke and Humphreys, 2020).

Formally, we introduce the contextual exclusion restriction, which states that the context variable C_i has no direct causal effect on the outcome once fixing context-mediators.[4]

Assumption 4 (Contextual Exclusion Restriction)

$$Y_i(T = t, \mathbf{M} = \mathbf{m}, c) = Y_i(T = t, \mathbf{M} = \mathbf{m}, c^*), \tag{7}$$

[3]One benefit of the proposed framework is that we can discuss the four dimensions of external validity using the same formalization, and derive a decomposition (equation (3)). The graphical approach is more suitable when researchers are willing to assume the underlying causal directed acyclic graphs.

[4]This formalization builds on the st-adjustment (Correa *et al.*, 2019). While their representation uses conditional independence within a selection diagram framework, we use nested counterfactuals in the potential outcomes framework. This helps us connect it to the broader literature on instrumental variables.

where the potential outcome $Y_i(T = t, c)$ is expanded with the potential context-mediators $\mathbf{M}_i(c)$ 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-mediators, and thus, researchers can incorporate any number of variables to satisfy the contextual exclusion restriction (we use one variable in Figure 2 for simplicity).

For example, in Broockman and Kalla (2016), C_i may denote a combination (county, year), with c = (Miami, 2015) and the target context c^* = (NYC, 2020). Suppose the context-mediator M_i represents the number of transgender individuals living in unit i's neighborhood. If the context (Miami, 2015) has no direct causal effect on outcomes other than through this context mediator, Assumption 4 holds. In contrast, if there are other channels through which contexts affect outcomes, such as the issue salience of transgender rights in a city, and they are not adjusted for, the assumption is violated (denoted by a red dotted line in Figure 2).

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 *et al.*, 1996).

Second, in order to avoid post-treatment bias (Rosenbaum, 1984), *context*-mediators $\mathbf{M}_i$ cannot be affected by treatment T_i (in Figure 2, there is no causal arrow from T to M). 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-mediator. However, if we also have to condition on another context-mediator, such as the issue salience of transgender rights in a city, to block causal paths from C to Y, this context-mediator might be affected by the intervention.

See Appendix A.1 for the proof of the identification of the T-PATE under this contextual exclusion restriction and other standard identification assumptions.

4 Effect-Generalization

In Section 3, we developed a formal framework and discussed concerns for external validity. In this section, we propose methods for one of the central research goals in external validity analysis; effect generalization — how to identify and estimate the T-PATE. This analysis is a central concern for randomized experiments that have policy implications ("Experiments for

Policy Interventions"; see also Roth (1995)). 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. Building on assumptions in the previous section, we discuss estimators for the T-PATE and their practical implementations. We start with addressing the well-researched problem of X-validity in Section 4.1, and then consider estimation strategies to simultaneously account for X- and C-validity in Section 4.2. We discuss T- and Y-validity in Section 4.3.

4.1 Three Classes of Estimators for X-validity

To focus our discussion on X-validity in this subsection, we only consider moving from the experimental population $\mathcal{P}$ to the target population $\mathcal{P}^*$. Our main quantity of interest is the T-PATE defined as $\mathbb{E}_{\mathcal{P}^*}\{Y_i(T=1,c) - Y_i(T=0,c)\}$.

Researchers need to adjust for differences between experimental samples and the target population, using covariates $\mathbf{X}$ measured in both data, to address X-validity (Assumption 1). In Section 3.2.1, we discussed two approaches; accounting for sampling processes (how experimental units are sampled), or treatment effect heterogeneity (how treatment effects vary across units). Here, extending this general thinking, we provide three classes of estimators for the T-PATE (see Figure 3 for a summary).

4.1.1 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. This approach is similar to those methods used to adjust for non-representativeness in traditional survey studies.

Two widely-used estimators in this class are (1) an inverse probability weighted (IPW) estimator, which can be computed as the weighted difference-in-means (Cole and Stuart, 2010), and (2) an ordinary least squares estimator with sampling weights (Särndal *et al.*, 2003). 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 (Cole and Stuart, 2010). Both estimators also require a modeling assumption that sampling weights are correctly specified.

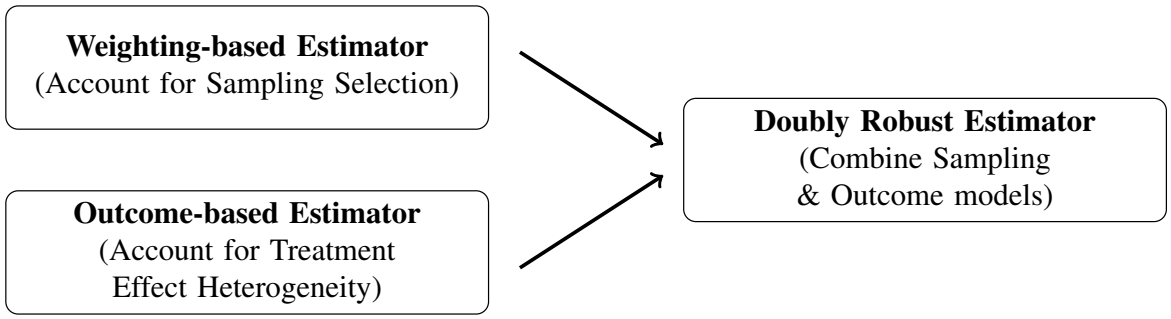

Figure 3: Three Classes of T-PATE Estimators.

4.1.2 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 (Kern *et al.*, 2016; Nguyen *et al.*, 2017). We first estimate an outcome model in the experiment, and then use it to predict treatment effects in the target population data.

A general two-step estimator is as follows. First, we estimate outcome models for the treatment and control groups, separately, in the experimental data; $\widehat{g}_1(\mathbf{X}_i) \equiv \widehat{\mathbb{E}}(Y_i \mid T_i = 1, \mathbf{X}_i, S_i = 1)$ and $\widehat{g}_0(\mathbf{X}_i) \equiv \widehat{\mathbb{E}}(Y_i \mid T_i = 0, \mathbf{X}_i, S_i = 1)$, where $S_i = 1$ indicates an experimental unit. This outcome model can be as simple as ordinary least squares, or more flexible estimators, such as BART (Chipman *et al.*, 2010). In the second step, we use the estimated models to predict potential outcomes for the target population data. For unit j in the target population data $\mathcal{P}^*$, $\widehat{Y}_j(1) = \widehat{g}_1(\mathbf{X}_j)$ and $\widehat{Y}_j(0) = \widehat{g}_0(\mathbf{X}_j)$, and therefore, $\widehat{\mathsf{T\text{-}PATE}}_{\mathtt{out}} = \frac{1}{N} \sum_{j \in \mathcal{P}^*} \left(\widehat{Y}_j(1) - \widehat{Y}_j(0) \right)$, 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 also assumes Assumption 1 for identification of the T-PATE. However, this estimator relies on a different modeling assumption from the one used in the weighting-based estimator; namely, the outcome models $g_1(\mathbf{X}_i)$ and $g_0(\mathbf{X}_i)$ are correctly specified. Conventionally, researchers considered modeling outcomes as more difficult than modeling sampling processes, and thus, many have emphasized weighting-based estimators. However, some recent studies suggest that there is little treatment effect heterogeneity for some domains in survey experiments (Mullinix *et al.*, 2015; Coppock *et al.*, 2018). For such

areas, it may be more reliable to model outcomes than sampling processes.

4.1.3 Doubly Robust Estimator

Finally, we discuss a class of doubly robust estimators, which relaxes the modeling assumptions of the first two approaches (Robins *et al.*, 1994; Dahabreh *et al.*, 2019). In particular, doubly robust estimators are consistent for the T-PATE as long as either the outcome model or the sampling weights are correctly specified; furthermore, analysts need not know which one is in fact correct. This is important in practice. When either the sampling weight or outcome model is incorrect, then weighting-based estimators and outcome-based estimators will provide different results and the researcher cannot determine which is more credible. In contrast, a doubly robust estimator will still provide a consistent estimate for the T-PATE even in this setting. It is important to note that, naturally, doubly robust estimators are also inconsistent when both outcome models and sampling weights are misspecified, and thus, it is always important to assess potential model misspecification in practice. Finally, while they weaken modeling assumptions, we restate that doubly robust estimators also require Assumption 1 for identification of the T-PATE.

We now introduce two estimators in this class (Robins *et al.*, 1994; Dahabreh *et al.*, 2019), which synthesize weighting-based and outcome-based estimators we discussed so far.

- Augmented IPW estimator:

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

  where π_i is the sampling weight of unit i. $\widehat{g}_1(\cdot)$ and $\widehat{g}_0(\cdot)$ are outcome models for the treatment and control groups, respectively, and they are estimated in the experimental data as discussed in Section 4.1.2.

- Weighted least squares projection:

  (1) Estimate weighted least squares for the treatment and control groups, separately, within the experimental data. For units in the experiment ($\mathcal{P}$),

$$\begin{aligned} Y_i &\sim \mathbf{X}_i^\top \gamma_1 \quad \text{with weights} \quad \pi_i \quad \text{for} \quad T_i = 1 \\ Y_i &\sim \mathbf{X}_i^\top \gamma_0 \quad \text{with weights} \quad \pi_i \quad \text{for} \quad T_i = 0 \end{aligned}$$

Estimators	Appropriate when:
Weighting-based	• Sampling weights can be reliably estimated • Many pre-treatment covariates are only available in the experimental sample
Outcome-based	• Treatment effect heterogeneity is limited • Sampling weights are extreme
Doubly Robust	• Sampling or outcome models may be misspecified

Table 2: Practical considerations for choosing a T-PATE estimator.

(2) We then project the estimated models to the target population data ($\mathcal{P}^*$).

$$\widehat{\text{T-PATE}}_{\texttt{wLS-proj}} = \frac{1}{N} \sum_{j \in \mathcal{P}^*} \mathbf{X}_j^\top \{\widehat{\gamma}_1 - \widehat{\gamma}_0\}.$$

4.1.4 Practical Considerations

In practice, researchers often do not know the true model for the sampling process or treatment effect heterogeneity. For this reason, we suggest researchers implement doubly robust estimators to mitigate model misspecification, whenever possible. There are, however, practical considerations for when the alternative classes of estimators may be more appropriate, which we outline in Table 2. In particular, the weighting-based estimators can incorporate pre-treatment covariates that are only measured in the experimental sample, which can greatly increase the precision in estimation of the T-PATE (see Section 6.1). The tradeoff lies in whether the researcher can correctly specify the sampling model. As long as treatment effect heterogeneity is limited, the outcome-based estimator is also appropriate, especially when sampling weights are extreme (i.e., variance of sampling weights is high), which is exactly the settings where estimators using sampling weights (weighting-based and doubly robust estimators) tend to have large standard errors.

4.1.5 Inference

To compute standard errors of each estimator, 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.

4.2 Accounting for 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 differ.

To account for X- and C-validity together, we provide a new estimator by extending the three classes of estimators with generalized sampling weights $\pi_i \times \theta_i$ and outcome models $g(\cdot)$ defined as follows.

$$\widehat{\pi}_i = \frac{1}{\widehat{\Pr}(S_i = 1 \mid C_i = c, \mathbf{M}_i, \mathbf{X}_i)}, \quad \text{and} \quad \widehat{\theta}_i = \frac{\widehat{\Pr}(C_i = c^* \mid \mathbf{M}_i, \mathbf{X}_i)}{\widehat{\Pr}(C_i = c \mid \mathbf{M}_i, \mathbf{X}_i)}$$

$$\widehat{g}_t(\mathbf{X}_i, \mathbf{M}_i) \equiv \widehat{\mathbb{E}}(Y_i \mid T_i = t, \mathbf{X}_i, \mathbf{M}_i, S_i = 1, C_i = c), \quad \text{for} \quad t \in \{0, 1\}.$$

We provide the proof in Appendix A.1.

4.3 T- and Y-validity

Although adjusting for X- and C-validity already requires strong assumptions, 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 define research questions, and formally setup potential outcomes. Given the essential nature of outcomes and treatments, there is no general approach to remove T- and Y-biases without making stringent identification or modeling assumptions.

Although effect-generalization may be infeasible, researchers can assess external validity by examining the question of sign-generalization under much weaker assumptions, which we discuss next in Section 5.

5 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 relevant when

researchers are testing theoretical mechanisms, and substantive theories have observable implications of 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. Sign-generalization is also sometimes a practical compromise when effect-generalization is not feasible, such as when identification assumptions (discussed in Section 3.2) are untenable or when required data on target populations, treatments, outcomes, or contexts are not available.

In this section, we show how to test the sign of the T-PATE by measuring multiple outcomes, and incorporating diverse units, treatments, and contexts into experiments. We formalize this common practice as the design of *purposive variations* and examine assumptions behind it in Section 5.1. This formalization helps us develop a new sign-generalization test that can statistically assess the direction of the T-PATE (Section 5.2).

5.1 Design of Purposive Variations

Without loss of generality, suppose a substantive theory predicts that the T-PATE is positive. Then, the proposed sign-generalization test aims to test the following hypothesis.

$$\begin{aligned} & H_0^* : \mathbb{E}_{\mathcal{P}^*}\{Y_i^*(T^*=1, c^*) - Y_i^*(T^*=0, c^*)\} \leq 0 \\ \text{versus} \qquad & H_1^* : \mathbb{E}_{\mathcal{P}^*}\{Y_i^*(T^*=1, c^*) - Y_i^*(T^*=0, c^*)\} > 0. \end{aligned} \tag{8}$$

If we can provide statistical evidence against the null hypothesis H_0^*, we support the substantive theory predicting the positive effect (the alternative hypothesis H_1^*).

However, when we cannot observe target populations, treatments, outcomes, or contexts, we cannot directly test the sign of the T-PATE. Even in such scenarios, we can indirectly test this hypothesis by using multiple outcomes, and incorporating diverse units, treatments, and contexts into experiments. A central idea is that if we consistently find positive (negative) causal effects across variations in all four dimensions (X-, Y-, T-, C-validity), they together bolster evidence for the positive (negative) T-PATE (Shadish *et al.*, 2002). We call this approach the *design of purposive variations*.[5] Incorporating variations has a long history

[5]This is known as a purposive sampling in Shadish *et al.* (2002). We explicitly use "variations" instead of "sampling" to clarify that the external validity concern is not only about populations (X-validity) but also for all four dimensions.

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 had purposive variations on at least one dimension. For example, Young (2019) considers a low-risk behavioral measure of dissent as well as twelve survey measures that span a range of risk levels, and tests whether causal estimates have the same sign across all outcomes. At its core, multi-site experiments can be viewed as an example of designs inducing variations in contexts (e.g., Blair and McClendon, 2020).

Purposive variations are directly useful for showing robustness of findings across a range of observed variations in four dimensions. However, without additional assumptions, the purposive variations are inherently *local* in that the variations are measured only within experiments, but by definition, external validity concerns are about variations we *do not* observe in the experiment. Therefore, we need to understand the conditions under which purposive variations measured *within* the experiment help us infer about the sign of the T-PATE, which is *external* to the experiment.

A practical question is; how should we incorporate 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 provide a discussion for similar designs and tests regarding other dimensions in Section 5.3.

While there are many valid ways to choose the K 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 $\left\{\mathbb{E}_{\mathcal{P}}\{Y_i^k(T=1,c) - Y_i^k(T=0,c)\}\right\}_{k=1}^K$.[6]

Although this assumption might seem strong at first, its substantive meaning is natural. Intuitively, we choose the K outcomes on which treatment effects serve as a lower bound and an upper bound for the T-PATE, meaning the T-PATE is within a range of the K causal effects we can estimate in the experiment (see Figure 4). This is akin to the overlap assumption

[6]Without loss of generality, we can also apply arbitrary monotone re-scaling functions f_k to each outcome to match the scales of the sample outcomes and the target outcomes (e.g., from binary to continuous outcomes).

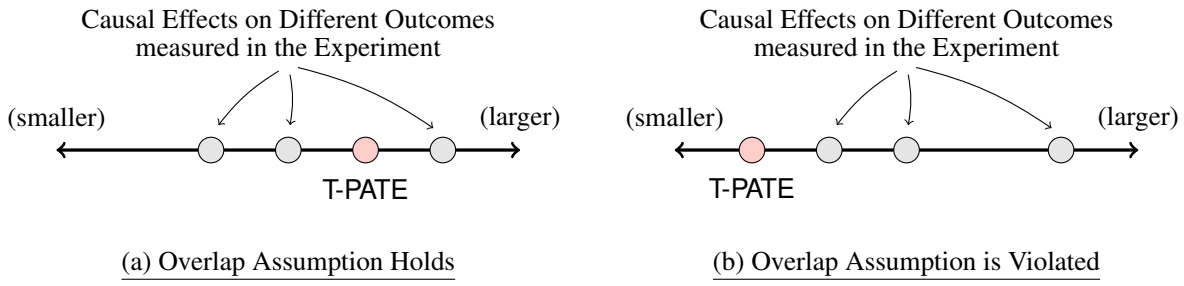

Figure 4: Overlap Assumption.

required in standard observational causal inference; pre-treatment covariates for the treatment and control groups should overlap (Imbens and Rubin, 2015). Similarly, in sign-generalization, we require that the target outcome and the purposive variations overlap. Without this assumption, inferences will heavily depend on extrapolation and modeling assumptions, 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. Importantly, this intuitive strategy is simple to implement, and thus already applied in practice. 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 and general justification for such designs 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), she considers a total of thirteen outcomes. 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, and thus, testing the sign of causal estimates on the K outcomes is not directly informative about the T-PATE. At the same time, in this scenario, no external validity analysis is possible without using extrapolation. Our proposed approach guards against such model-dependent extrapolation by clarifying underlying assumptions.

5.2 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.

Focusing again on Y-validity, our target null hypothesis is written as,

$$H_0^* : \mathbb{E}_{\mathcal{P}}\{Y_i^*(T=1,c) - Y_i^*(T=0,c)\} \leq 0. \tag{9}$$

If 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 $k \in \{1, \ldots, K\}$,

$$H_0^k : \mathbb{E}_{\mathcal{P}}\{Y_i^k(T=1,c) - Y_i^k(T=0,c)\} \leq 0. \tag{10}$$

5.2.1 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 (10)) is a valid test for the target null hypothesis (equation (9)) under the overlap assumption.

This result is important because, as we show below, this means that a common practice of separately checking each of the K p-values against a prespecified significance level α (e.g., $\alpha = 0.05$) is valid as a sign-generalization test, without additional multiple testing corrections.

Theorem 1 (Validity of Union-based Sign-Generalization Test)
Consider the union of the K null hypotheses.

$$H_0 : \bigcup_{k=1}^{K} H_0^k. \tag{11}$$

Under Assumption 5, a valid test of the union null H_0 is also a valid test of the target null hypothesis H_0^*.

We provide the proof in Appendix A.2.1. In practice, Theorem 1 implies that rejecting the union of the K null hypotheses (equation (11)) implies the rejection of the target null hypothesis about the T-PATE (equation (9)).

We can obtain a p-value for this union null hypothesis using the maximum p-value (Berger *et al.*, 1996). In Appendix A.2.1, we review formal discussions for why this simple procedure properly accounts for multiple comparisons and does not require additional multiple testing corrections.

5.2.2 Partial Conjunction Test

While the test based on the maximum p-value is easy to implement, the union null hypothesis (equation (11)) 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 at conventional significance levels, we cannot reject the union null hypothesis because we rely on the maximum p-value. 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). To our knowledge, this paper is the first to extend the partial conjunction test framework to external validity analysis.

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

$$\widetilde{H}_0^r : \sum_{k=1}^{K} \mathbf{1}\{H_0^k \text{ is false}\} < r \tag{12}$$

where $r \in [1, K]$ is a threshold specified by researchers and $\sum_{k=1}^{K} \mathbf{1}\{H_0^k \text{ is false}\}$ simply counts the number of true non-nulls. This partial conjunction null states that the treatment of interest has a positive effect on at most $r-1$ outcomes. 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. Importantly, the union-based sign-generalization test in Section 5.2.1 is a special case when $r = K$ — assessing whether the treatment has a positive effect on *all* K outcomes.

How can we obtain a p-value for this partial conjunction test? Given the p-values for each of the K null hypotheses $\{p_1, \ldots, p_K\}$, 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.

$$\begin{aligned} \widetilde{p}_{(1)} &\equiv K p_{(1)} \\ \widetilde{p}_{(r)} &\equiv \max\{(K-r+1)p_{(r)}, \widetilde{p}_{(r-1)}\} \qquad \text{for } r \geq 2. \end{aligned} \tag{13}$$

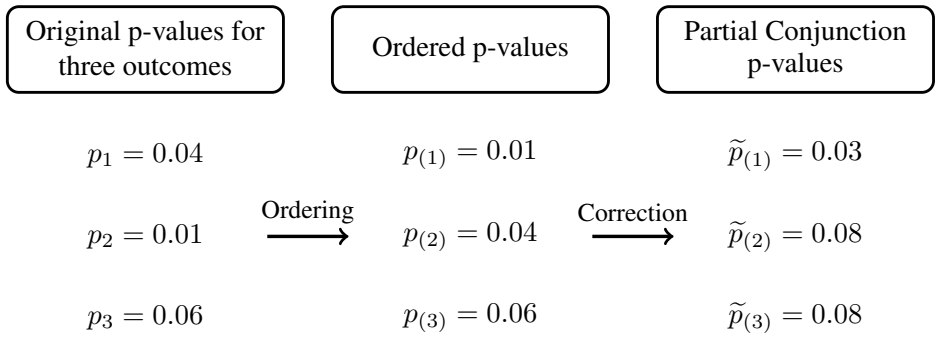

Figure 5: Example of Partial Conjunction Test with Three Outcomes. *Note:* The second step of "Correction" is based on equation (13).

The p-value for $\widetilde{H}_0^r$ is $\widetilde{p}_{(r)}$. This procedure is valid under any dependence across p-values. In Appendix A.2, we also discuss scenarios in which p-values are independent across variations, and describe a statistically more powerful way to compute the partial conjunction p-values in such settings.

One natural question is the choice of the threshold r. Although researchers can pre-register a threshold before data collection, we recommend reporting p-values for all $r \in [1, K]$. There are two key advantages. First, maybe surprisingly, we do not need further adjustment to p-values for multiple comparisons across different thresholds r. This is because the partial conjunction p-value satisfies the monotonicity requirement and the null hypotheses are nested (i.e., when the null hypothesis $\widetilde{H}_0^r$ is accepted, it is always true that $\widetilde{H}_0^{r+1}$ is also accepted) (Benjamini and Heller, 2008). We review a formal proof in Appendix A.2.2.

Second, using the K partial conjunction p-values, researchers can directly estimate the number of outcomes for which the treatment has positive effects. Formally, a simple estimator is defined as,

$$r_{\max} \equiv \max\{r : \widetilde{p}_{(r)} \leq \alpha\}. \tag{14}$$

It simply counts the number of outcomes whose corresponding partial conjunction p-values are less than α. Then, an interval $[r_{\max}, K]$ is the $(1-\alpha)$ confidence interval for the true number of outcomes for which the treatment has positive effects.

In sum, researchers can summarize evidence from the sign-generalization test with two sets of quantities; (1) p-values for all thresholds, $\{\widetilde{p}_{(r)}\}_{r=1}^K$, to present which hypotheses have a sign consistent with the substantive theory of interest and (2) the proportion of outcomes on

which treatment effects are positive $r_{\max}/K$, which provides a one-value summary for the sign stability of causal effects across the purposive variations.

5.3 Summary

While this section focused on Y-validity for clear presentation, similar arguments and methods apply to the other dimensions. To conduct sign-generalization, we first incorporate purposive variations at the experimental design stage. To account for X-, T-, Y-, and C-validity, researchers should incorporate diverse units, multiple treatments, multiple outcomes, and different contexts into the experiment, respectively. With the purposive variations, analysts can then use the partial conjunction test to estimate the proportion of variations that have the same sign of causal effects. We can also assess multiple dimensions together (e.g., Y- and T-validity together) with the same approach, an example of which is provided in Section 6.

5.4 Practical Considerations

One key practical consideration is the number of purposive variations to include. On one hand, the larger number of purposive variations (e.g., many different contexts are studied within one paper) increases the credibility of sign-generalization because the required overlap assumption (Assumption 5) 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 directly means a smaller sample size for each treatment level and each context (e.g., when running experiments in three separate contexts, each experiment receives $n/3$ units instead of $n/2$ in the case of two-context experiments).

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.

6 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 B, we provide results for Young (2019), which focuses on sign-generalization.

6.1 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, Broockman and Kalla (2016) provide 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, Broockman and Kalla (2016) 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 4), while addressing concerns of X- and Y-validity together. Given space constraints, we focus on these two dimensions that are most insightful for illustrating the proposed approach, and we discuss T- and C-validity in Appendix B.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 (Ansolabehere and Schaffner, 2017).

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.

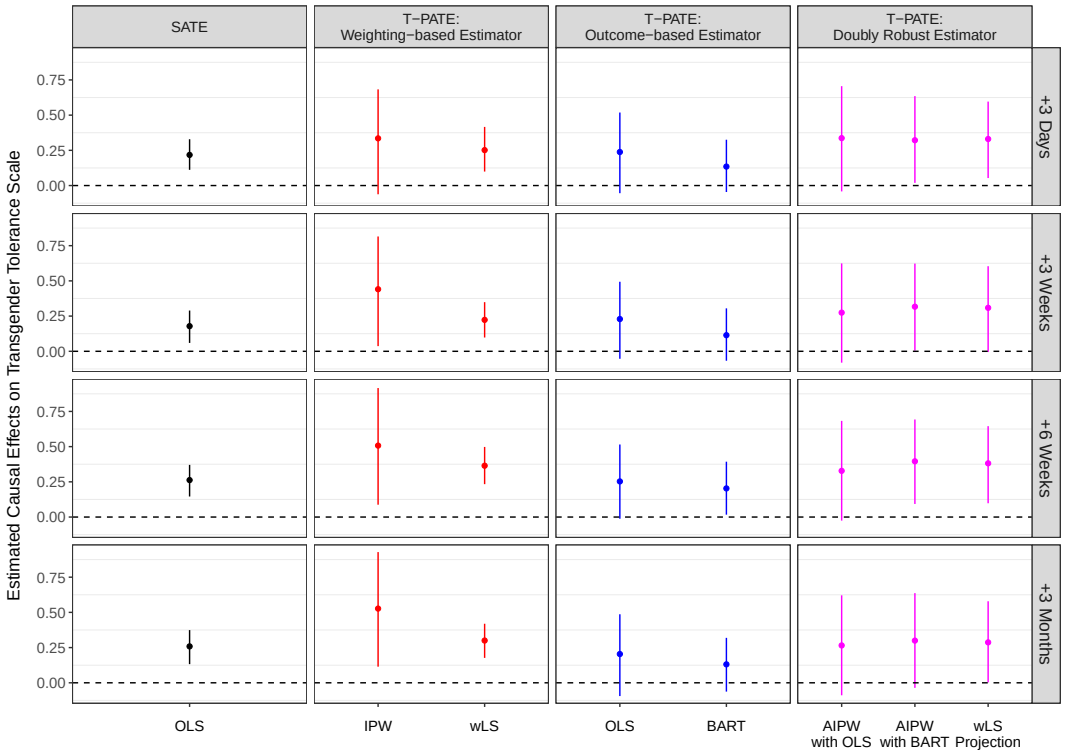

Figure 6: Estimates of the T-PATE for Broockman and Kalla (2016). *Note:* The first column shows estimates of the SATE, and the subsequent three columns present estimates of the T-PATE for three classes of estimators. Rows represent different post-treatment survey waves.

6.1.1 Effect-Generalization

We estimate the T-PATE using the three classes of estimators discussed in Section 4.1. Weighting-based estimators include IPW and weighted least squares that adjusts for control variables pre-specified in the original authors' pre-analysis plan. Sampling weights are estimated via calibration (Hainmueller, 2012; Hartman *et al.*, 2015). For the outcome-based estimators, we use OLS and a more flexible model, BART. Finally, we implement three doubly robust estimators; the AIPW with OLS, the AIPW with BART, and the weighted least squares projection.

Figure 6 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.

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.

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

Finally, we can compare the three classes of estimators. As summarized in Table 2, we generally recommend doubly robust estimators because the sampling and outcome models are often unknown in practice. However, in this example, a weighted least squares estimator (wLS in Figure 6) 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.

6.2 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 B.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.

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

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. 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. While the argument rating tasks force respondents to think specifically about the responsibilities of incumbents, the open-ended questions are more unobtrusive and do not mention political actors. The close-ended questions are in the middle. We assume that the three set of outcomes jointly satisfy the required overlap assumption, and we use all the outcomes for the sign-generalization test.

6.2.1 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, rather than negative, 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

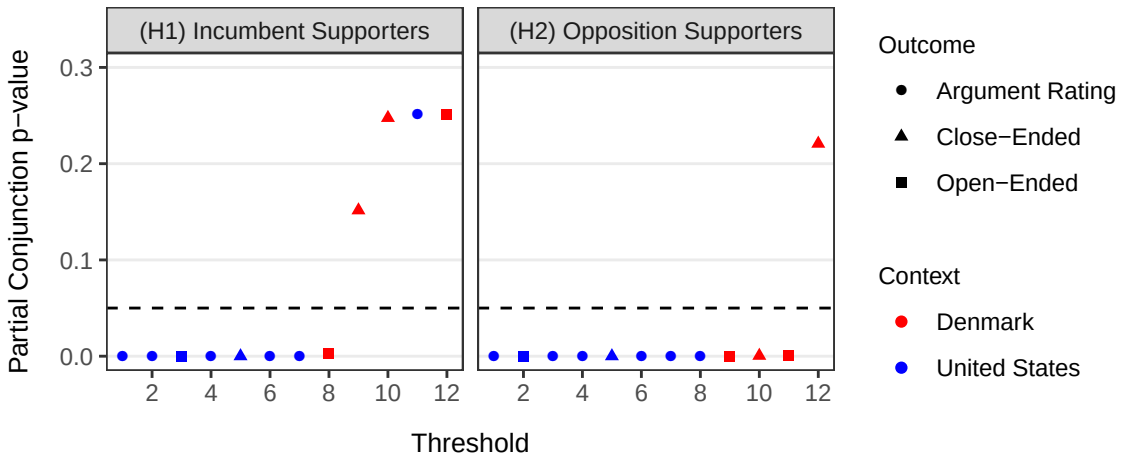

Figure 7: Sign-Generalization Test for Bisgaard (2019). *Note:* We combine causal estimates on multiple outcomes across four survey experiments in two countries. Following Section 5, we report partial conjunction p-value s for all thresholds.

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.

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 3). 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 7 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 as partisan-motivated reasoning would be weaker in Denmark, where the incumbent party's political responsibility is less clear and the level of polarization among citizens is lower. 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.

7 Discussion

7.1 Relationship to Meta-Analysis

Meta-analysis is a method for summarizing statistical findings from multiple papers or research literature (Cooper *et al.*, 2019). While still rare, political scientists have begun using it to aggregate results from randomized experiments (e.g., Blair *et al.*, 2020; Dunning *et al.*, 2019). Even though we have so far focused on how to improve external validity of individual experiments, the proposed approach can also be useful for meta-analysis. First, meta-analysts must also consider the four dimensions of external validity, and therefore the proposed framework (Section 3) can clarify conditions under which meta-analysis can credibly address concerns about external validity. 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 4) are essential when meta-analysts want to predict causal effects in new target sites. Sign-generalization (Section 5) is useful when meta-analysis focuses on synthesizing scientific knowledge (e.g., Paluck *et al.* (2019) examine whether intergroup contact typically reduces prejudice).

The quality of meta-analysis heavily depends on that of each study included. If each experimental study only considers internal validity and has no explicit discussion of assumptions, design, or analysis for external validity, it is impossible for meta-analysis to credibly aggregate experimental results. Therefore, while meta-analysis is a powerful method for tackling external validity, it does not mean that each experimental study can focus on internal validity alone. Rather, the opposite is true. To enable credible accumulation of knowledge through meta-analysis, we have to explicitly incorporate design and analysis for external validity into each experiment. The proposed approach in this paper can help researchers improve both each

experimental study and meta-analysis.

7.2 External Validity of Observational Studies

For observational studies, researchers can decompose the total bias into internal validity bias and external validity bias (Westreich *et al.*, 2019). Thus, the same four dimensions of external validity is also relevant in observational studies. For example, widely-used causal inference techniques, such as instrumental variables, regression discontinuity, and matching estimators, make identification strategies more credible by focusing on a subset of units, which often decreases X-validity. Dehejia *et al.* (2019) study over 100 replications of natural experiments across countries and time, and empirically show that the issue of C-validity is essential for observational studies as well.

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. The same partial conjunction test can be applied to quantify the sign-generalization even in observational studies.

8 Concluding Remarks

External validity of randomized experiments 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. The proposed methods are illustrated with three randomized experiments, covering field, survey, and lab experiments.

Addressing external validity is inherently difficult because it aims to infer whether experimental results are generalizable to other populations, treatments, outcomes, and contexts that we do not observe. In this paper, we formally clarify conditions under which this challenging yet essential inference is possible, and we propose new methods to improve the external validity of randomized experiments.

References

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

Ansolabehere, S., Rodden, J., and Snyder, J. M. (2008). The Strength of Issues: Using Multiple Measures to Gauge Preference Stability, Ideological Constraint, and Issue Voting. *American Political Science Review* **102**, 2, 215–232.

Ansolabehere, S. and Schaffner, B. F. (2017). CCES Common Content, 2016.

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

Bartels, L. M. (2002). Beyond The Running Tally: Partisan Bias In Political Perceptions. *Political Behavior* **24**, 2, 117–150.

Benjamini, Y. and Heller, R. (2008). Screening for Partial Conjunction Hypotheses. *Biometrics* **64**, 4, 1215–1222.

Berger, R. L., Hsu, J. C., *et al.* (1996). Bioequivalence Trials, Intersection-Union Tests and Equivalence Confidence Sets. *Statistical Science* **11**, 4, 283–319.

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

Blair, G., Coppock, A., and Moor, M. (2020). When to Worry About Sensitivity Bias: Evidence From 30 years of List Experiments. *American Political Science Review* .

Blair, G. and Imai, K. (2012). Statistical Analysis of List Experiments. *Political Analysis* **20**, 1, 47–77.

Blair, G. and McClendon, G. (2020). Experiments in Multiple Contexts. In D. P. Green and J. Druckman, eds., *Handbook of Experimental Political Science*. Cambridge University Press.

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

Bullock, J. G., Gerber, A. S., Hill, S. J., and Huber, G. A. (2015). Partisan Bias in Factual Beliefs About Politics. *Quarterly Journal of Political Science* .

Campbell, D. T. and Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research*. RandMcNally.

Chipman, H. A., George, E. I., and McCulloch, R. E. (2010). BART: Bayesian Additive Regression Trees. *The Annals of Applied Statistics* **4**, 1, 266–298.

Cole, S. R. and Stuart, E. A. (2010). Generalizing Evidence From Randomized Clinical Trials to Target PopulationsThe ACTG 320 Trial. *American journal of epidemiology* **172**, 1, 107–115.

Cooper, H., Hedges, L. V., and Valentine, J. C. (2019). *The Handbook of Research Synthesis and Meta-Analysis*. Russell Sage Foundation.

Coppock, A., Leeper, T. J., and Mullinix, K. J. (2018). Generalizability of Heterogeneous Treatment Effect Estimates Across Samples. *Proceedings of the National Academy of Sciences* **115**, 49, 12441–12446.

Correa, J., Tian, J., and Bareinboim, E. (2019). Adjustment Criteria for Generalizing Experimental Findings. In *Proceedings of the 36th International Conference on Machine Learning*, vol. 97, 1361–1369, Long Beach, CA. PMLR.

Dahabreh, I. J., Robertson, S. E., Tchetgen Tchetgen, E. J., Stuart, E. A., and Hernán, M. A. (2019). Generalizing Causal Inferences From Individuals In Randomized Trials to All Trial-Eligible Individuals. *Biometrics* **75**, 2, 685–694.

Deaton, A. and Cartwright, N. (2018). Understanding and Misunderstanding Randomized Controlled Trials. *Social Science & Medicine* .

Dehejia, R., Pop-Eleches, C., and Samii, C. (2019). From Local to Global: External Validity in a Fertility Natural Experiment. *Journal of Business & Economic Statistics* 1–27.

Dunning, T., Grossman, G., Humphreys, M., Hyde, S. D., *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, B. and Tibshirani, R. J. (1994). *An Introduction To The Bootstrap*. CRC press.

Egami, N. and Hartman, E. (2018). Covariate Selection for Generalizing Experimental Results. Working Paper available at `https://arxiv.org/abs/1909.02669`.

Gerber, A. S. and Green, D. P. (2012). *Field Experiments: Design, Analysis, and Interpretation*. WW Norton.

Green, D. P. and Gerber, A. S. (2008). *Get Out The Vote: How To Increase Voter Turnout*. Brookings Institution Press.

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

Hartman, E. (2020). Generalizing Experimental Results. In J. Druckman and D. P. Green, eds., *Advances in Experimental Political Science*. Cambridge University Press.

Hartman, E., Grieve, R., Ramsahai, R., and Sekhon, J. S. (2015). From Sample Average Treatment Effect to Population Average Treatment Effect on the Treated: Combining Experimental with Observational Studies to Estimate Population Treatment Effects. *Journal of the Royal Statistical Society. Series A (Statistics in Society)* **178**, 3, 757–778.

Heckman, J. (1979). Sample Selection Bias as a Specification Error. *Econometrica* .

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

Imbens, G. W. and Rubin, D. B. (2015). *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge University Press.

Karmakar, B. and Small, D. S. (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, H. L., Stuart, E. A., Hill, J., and Green, D. P. (2016). Assessing Methods for Generalizing Experimental Impact Estimates to Target Populations. *Journal of Research on Educational Effectiveness* **9**, 1, 103–127.

Miratrix, L. W., Sekhon, J. S., Theodoridis, A. G., and Campos, L. F. (2018). Worth Weighting? How to Think About and Use Weights in Survey Experiments. *Political Analysis* **26**, 3, 275–291.

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

Mullinix, K. J., Leeper, T. J., Druckman, J. N., and Freese, J. (2015). The Generalizability of Survey Experiments. *Journal of Experimental Political Science* **2**, 2, 109–138.

Munger, K. (2019). Knowledge Decays: Temporal Validity and SocialScience in a Changing World. *Working Paper* .

Mutz, D. C. (2011). *Population-based Survey Experiments*. Princeton University Press.

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

Nguyen, T. Q., Ebnesajjad, C., Cole, S. R., and Stuart, E. A. (2017). Sensitivity analysis for an unobserved moderator in RCT-to-target-population generalization of treatment effects. *The Annals of Applied Statistics* **11**, 1, 225–247.

Paluck, E. L. and Green, D. P. (2009). Prejudice Reduction: What works? A Review and Assessment of Research and Practice. *Annual Review of Psychology* **60**, 339–367.

Paluck, E. L., Green, S. A., and Green, D. P. (2019). The Contact Hypothesis Re-Evaluated. *Behavioural Public Policy* **3**, 2, 129–158.

Pearl, J. and Bareinboim, E. (2011). Transportability of Causal and Statistical Eelations: A Formal Approach. In *2011 IEEE 11th International Conference on Data Mining Workshops*, 540–547. IEEE.

Pearl, J. and Bareinboim, E. (2014). External Validity: From Do-Calculus to Transportability Across Populations. *Statistical Science* **29**, 4, 579–595.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994). Estimation of Regression Coefficients When Some Regressors Are Not Always Observed. *Journal of the American Statistical Association* **89**, 427, 846–866.

Rosenbaum, P. 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.

Roth, A. E. (1995). Introduction To Experimental Economics. *The Handbook of Experimental Economics* **1**, 3–109.

Rubin, D. B. (1974). Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies. *Journal of Educational Psychology* **66**, 5, 688.

Särndal, C.-E., Swensson, B., and Wretman, J. (2003). *Model Assisted Survey Sampling*. Springer Science & Business Media.

Shadish, W. R., Cook, T. D., and Campbell, D. T. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston: Houghton Mifflin.

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

Westreich, D., Edwards, J. K., Lesko, C. R., Cole, S. R., and Stuart, E. A. (2019). Target Validity and The Hierarchy of Study Designs. *American journal of epidemiology* **188**, 2, 438–443.

Wilke, A. and Humphreys, M. (2020). Field Experiments, Theory, and External Validity. In L. Curini and R. Franzese, eds., *The SAGE Handbook of Research Methods in Political Science and International Relations*. Transaction Publishers.

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