American Political Science Review (2022) 1–19
doi:10.1017/S0003055422000880

Elements of External Validity: Framework, Design, and Analysis

NAOKI EGAMI *Columbia University, United States*
ERIN HARTMAN *University of California, Berkeley, United States*

The external validity of causal findings is a focus of long-standing debates in the social sciences. Although the issue has been extensively studied at the conceptual level, in practice few empirical studies include an explicit analysis that is directed toward 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.*

INTRODUCTION

Over the last few decades, social scientists have developed and applied a host of statistical methods to make valid causal inferences, known as the credibility revolution. This trend has focused primarily on *internal* validity—researchers seek 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.

Although 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 focused primarily 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) in the section Formal Framework for External Validity. 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 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 attempts to assess whether the direction of causal effects is generalizable. The former goal is important when researchers want to generalize the substantive or policy effect 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

Naoki Egami, Assistant Professor, Department of Political Science, Columbia University, United States, naoki.egami@columbia.edu.
Erin Hartman, Assistant Professor, Department of Political Science and of Statistics, University of California, Berkeley, United States, ekhartman@berkeley.edu.

Received: December 21, 2020; revised: August 06, 2021; accepted: July 20, 2022.

[1] See Appendix J 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.

required by each (in the section Effect-Generalization). 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 the section Sign-Generalization, 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—for example, 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 (in the section Motivating Empirical Applications). Using them, we illustrate how to implement our proposed methods and provide practical recommendations in the section Empirical Applications and Appendix C. All of our methods can be implemented via the companion R package `evalid`. Finally, in the section Discussion, we discuss several important extensions. First, although 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 focused primarily on changes in populations—that is, X-validity (Cole and Stuart 2010; Egami and Hartman 2021; Imai, King, and Stuart 2008). Although 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 type of validity 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 (Dahabreh et al. 2019; Hartman et al. 2015; Kern et al. 2016; Tipton 2013) and provide practical guidance. To account for changes in populations and contexts together—that is, X- and C-validity, we use identification results from the causal diagram approach (Bareinboim and Pearl 2016) and develop new estimators in the section Effect-Generalization. The third and main methodological contribution is to provide a formal approach to sign-generalization. Although 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.

MOTIVATING EMPIRICAL APPLICATIONS

Field Experiment: Reducing Transphobia

Prejudice can negatively affect social, political, and health outcomes of out-groups 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 preexperiment baseline survey. They randomly assigned canvassers to either encourage voters to actively take the perspective of transgender people (*perspective taking*) or to have a placebo conversation with respondents. To measure attitudes toward 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 10-minute conversation substantially reduced transphobia, and the effects persisted for three months.

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.

Lab Experiment: Effect of Emotions on Dissent in Autocracy

Many authoritarian countries employ various frightening acts of repression to deter dissent. To disentangle the psychological foundations 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?" (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.

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 pretreatment 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 the subsection External Validity of Observational Studies.

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 $C_i = (\text{Miami}, 2015)$.

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).

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

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 interested not only 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 US, or in any other populations of theoretical interest. The same question applies to other dimension—that is, 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:

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

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

TABLE 1. Summary of Typology

	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)

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 the 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.

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.

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 2008). When relying on convenience samples or nonprobability 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 pretreatment 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 (Hartman 2020). The first approach attempts to account for how subjects are sampled into the experiment, including the common practice of using sampling weights (Miratrix et al. 2018; Mutz 2011). 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. Relatedly, for some questions in survey experiments, recent studies find that causal estimates from convenience samples are similar to those estimated from nationally representative samples due to little treatment heterogeneity, despite the significant difference in their sample characteristics (Coppock, Leeper, and Mullinix 2018; Mullinix et al. 2015). 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.

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 the 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" (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.

Although the target treatments differ depending on the types of experiments and corresponding research goals, the 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)

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

It states that the assigned treatment T and the target treatment T^* induce the same average treatment effects. For example, the causal effect of the perspective-taking intervention is the same regardless of whether canvassers are recruited by established LGBT organizations.

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 the section Sign-Generalization—that is applicable to concerns about T-validity regardless of whether they arise in field, survey, or lab experiments.

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—for example, support for hypothetical immigrants, policies, and politicians. In such cases, Y-validity analyses 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)

$$\begin{aligned}&\mathbb{E}_{\mathcal{P}}\left[Y_i^*(T=1,\ c)-Y_i^*(T=0,\ c)\right]\\&\quad=\mathbb{E}_{\mathcal{P}}[Y_i(T=1,\ c)-Y_i(T=0,\ c)].\end{aligned}\tag{5}$$

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—that is, $Y = Y^*$. For example, many Get-Out-the-Vote experiments in the US 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 the section Sign-Generalization, 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 one year or two 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.

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's (2016) study from Miami in 2015 to another context, such as New York City in 2020. Establishing C-validity is challenging because a randomized experiment is done in one context c and researchers need to generalize or transport experimental results to another context c^*, where they did not run the experiment. Formally, C-validity is a question about covariates that have no variation within an 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 the subsection Setup.

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,

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

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 New York City 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 the context moderators.[2] This name reflects its similarity to the exclusion restriction well known in the instrumental variable literature.

Assumption 4 (Contextual Exclusion Restriction)

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

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,

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

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 posttreatment 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. The main question is whether the variable has any variation within an experiment—if the variable has some variation, it is an X-validity problem, and it is a C-validity problem otherwise. For example, suppose we conduct a Get-Out-The-Vote 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 statewide experiment in Florida where some

[2] This formalization builds on the st-adjustment (Correa, Tian, and Bareinboim 2019). Although 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 literature on instrumental variables.

districts are electorally competitive and others are safe. Then, if we want to generalize this result to another state—for example, the state of New York—where the proportion of electorally competitive districts differs, the electoral competitiveness of districts can be addressed as an X-validity problem.[3] This is because our experimental data has both electorally competitive and safe districts and what differs across the two states is their distribution. 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 exist.

THE PROPOSED APPROACH TO EXTERNAL VALIDITY: OUTLINE

In the section Formal Framework for External Validity, we developed a formal framework and discussed concerns for external validity. In this section, we outline our proposed approach to external validity, reserving details of our methods to the sections Effect-Generalization and Sign-Generalization.

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 focus primarily 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), whereas we discuss all four dimensions in Appendix C. We also provide additional examples of how to identify relevant dimensions in the section Empirical Applications 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 the section Formal Framework for External Validity on the specifics of how to conceptualize each dimension.

Once relevant dimensions are identified, analysts should decide the *goal* of an external validity analysis, whether effect- or sign-generalization. Effect-Generalization—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 because 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. The assumptions required for effect-generalization include Assumptions 1–4 detailed in the section Formal Framework for External Validity, whereas we describe assumptions necessary for sign-generalization in the section Sign-Generalization. In some settings, researchers can design experiments such that the 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 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 the sections Effect-Generalization and Sign-Generalization, 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 to external validity in practice.

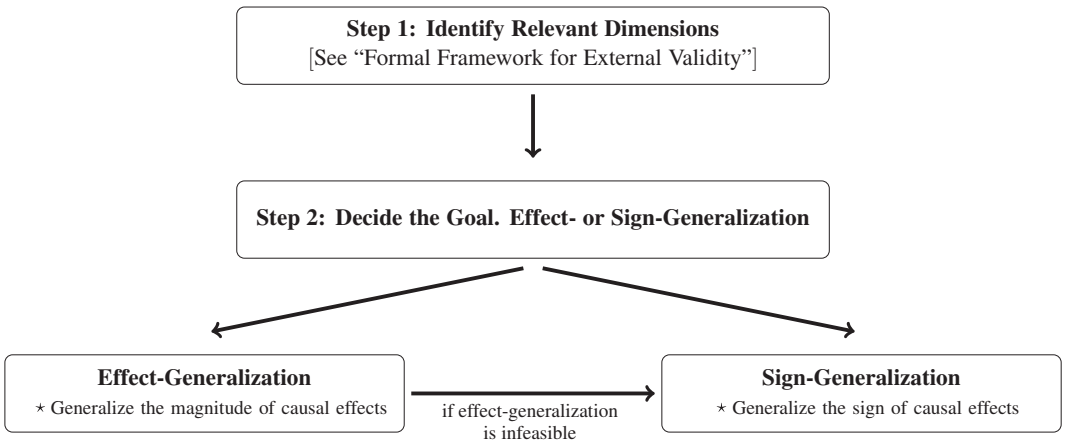
EFFECT-GENERALIZATION

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 seek 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 the subsection Addressing Multiple Dimensions Together).

For X- and C-validity, we start by asking *whether* effect-generalization is feasible by evaluating the required assumptions (Assumption 1 for X-validity, and Assumption 4 for C-validity). If the required assumptions hold, researchers can employ three classes of estimators—weighting-based, outcome-based, and doubly robust estimators. We provide practical

[3] 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.

FIGURE 1. The Proposed Approach to External Validity



guidance on how to choose an estimator in the subsection How to Choose a T-PATE Estimator. Importantly, because the required assumptions are often strong, credible effect-generalization might be impossible. In such cases, sign-generalization might still be feasible because it requires weaker assumptions (see the section Sign-Generalization).

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 in the section Sign-Generalization that sign-generalization 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.

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.

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—that is, 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 the sampling weights are correctly specified.

Outcome-Based Estimator

Although 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.

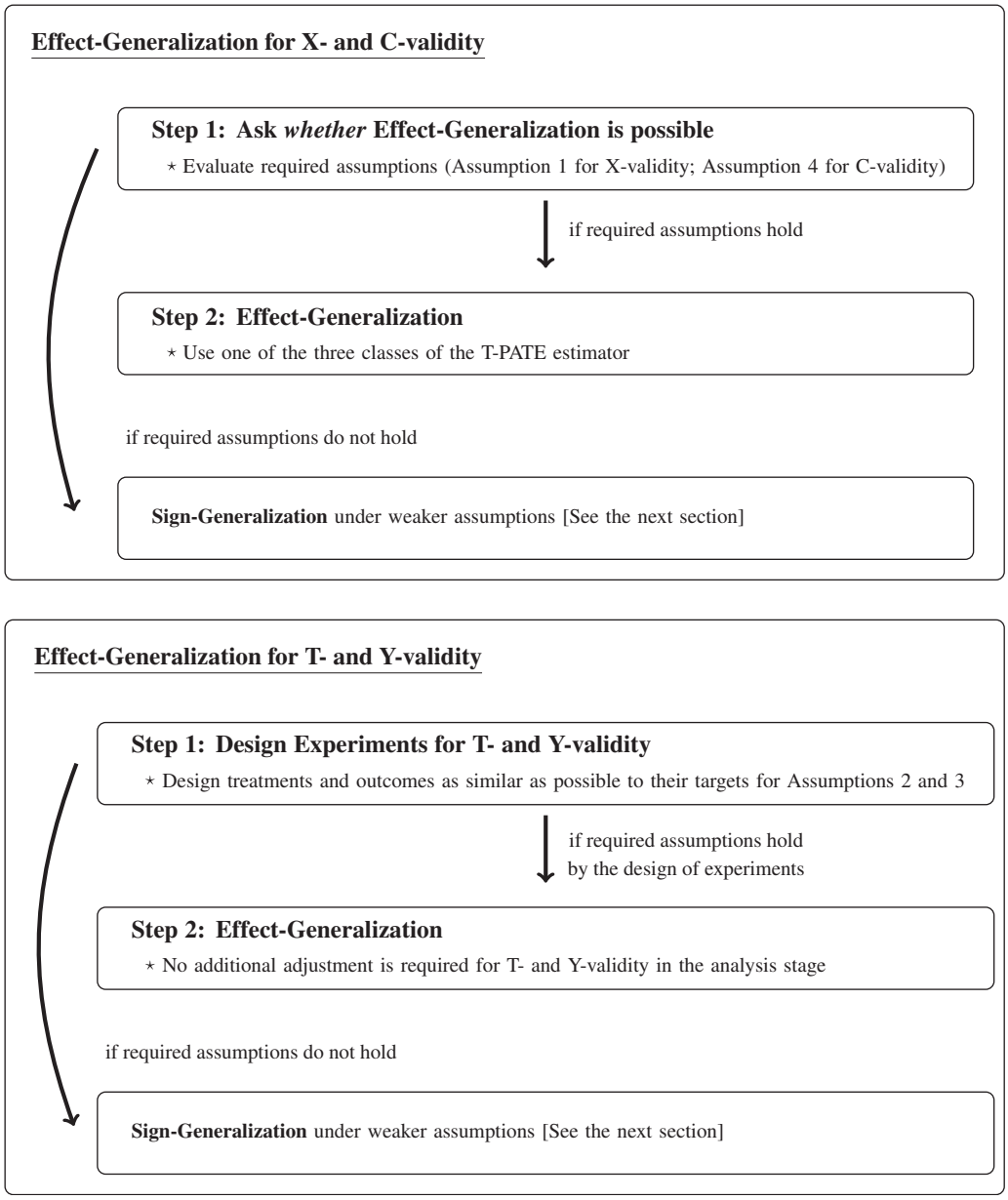
Formally, in the first step, we estimate the outcome model $\widehat{g}_t(\mathbf{X}_i) \equiv \widehat{\mathbb{E}}(Y_i | T_i = t, \mathbf{X}_i, S_i = 1)$ for $t \in \{0, 1\}$, where $S_i = 1$ indicates an experimental unit. This outcome model can be as simple as ordinary least squares or rely on more 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, $\widehat{T\text{–}PATE_{OUT}} = \frac{1}{N}\sum_{j \in \mathcal{P}^*}(\widehat{Y}_j(1) - \widehat{Y}_j(0))$, where the sum is over the target population data $\mathcal{P}^*$, and N is the size of the target population data.

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

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 (Dahabreh et al. 2019; Robins, Rotnitzky, and Zhao 1994). Specifically, to use weighting-based estimators, we have to assume the sampling model is correctly specified (the pink area in Figure 3a). Similarly, outcome-based estimators assume the correct outcome model (the orange area). In contrast, doubly robust estimators are consistent for

FIGURE 2. Summary of Effect-Generalization



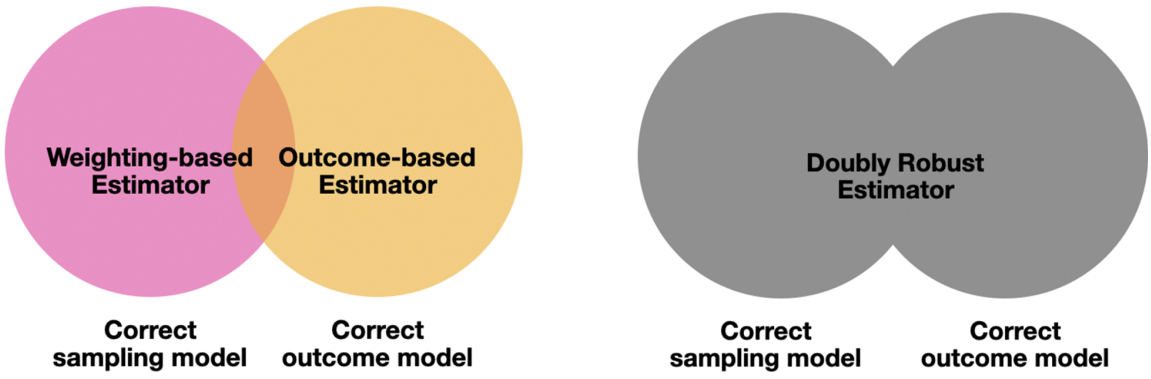
the T-PATE as long as either the outcome model or the sampling model is correctly specified; furthermore, analysts need not know which one is, in fact, correct. Figure 3b shows that the doubly robust estimator is consistent in much wider applications (the gray area in Figure 3b). Therefore, this estimator significantly relaxes the modeling assumptions of the previous two methods. Although they weaken the modeling assumptions, we restate that doubly robust estimators also require Assumption 1 for the identification of the T-PATE.

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

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

where π_i is the sampling weight of unit i, and $\widehat{g}_t(\cdot)$ is an outcome model estimated in the experimental data. The first two terms represent the IPW estimator based

FIGURE 3. Properties of Doubly Robust Estimator



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

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

Note: The doubly robust estimator is consistent as long as the sampling or outcome model is correctly specified (gray area in panel b).

on residuals $Y_i - \widehat{g}_t(\mathbf{X}_i)$, and the last term is equal to the outcome-based estimator.

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. However, there are scenarios when the alternative classes of estimators may be more appropriate. In particular, the weighted OLS can incorporate pretreatment covariates that are only measured in the experimental sample, which can greatly increase the precision in the estimation of the T-PATE (see the section Empirical Applications), 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 and the other two estimators tend to have large standard errors.

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 $g(\cdot)$.

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

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

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

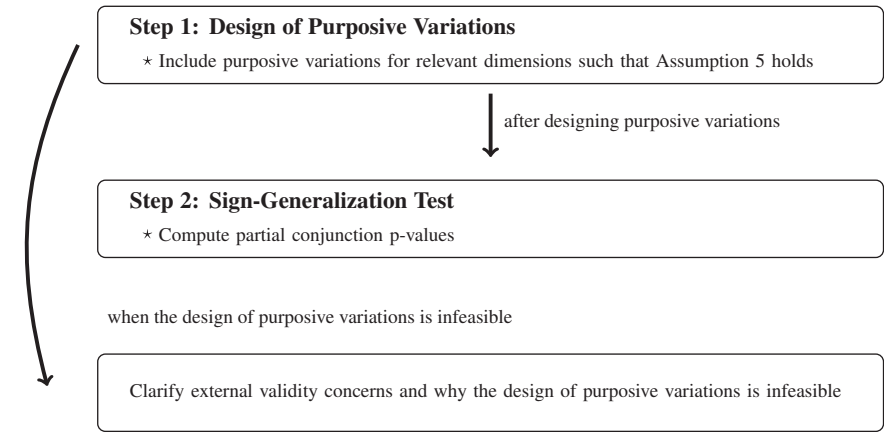
$\widehat{\pi}_i$ is the same as sampling weights used for *X*-validity, but it should be multiplied by $\widehat{\theta}_i$, which captures the difference in the distribution of $(\mathbf{X}_i, \mathbf{M}_i)$ in the experimental context c and the target context c^*. The outcome model $\widehat{g}_t(\cdot)$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

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 the section Formal Framework for External Validity (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,

FIGURE 4. Summary of Sign-Generalization



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 conducting 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 in the next section.

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 (in the subsection Design of Purposive Variations). The required range assumption (Assumption 5) is much weaker than are the assumptions required for effect-generalization.

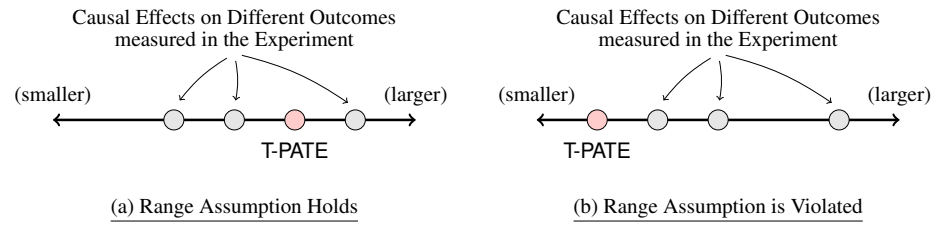
If researchers can include purposive variations to satisfy the required assumption, the final step is to conduct a new sign-generalization test, which computes partial conjunction *p*-values (Benjamini and Heller 2008). Using these adjusted *p*-values, researchers can assess the direction of the T-PATE while accounting for multiple comparisons correctly. We detail their practical implementation and describe how to interpret them in the subsection Sign-Generalization Test. The main 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.

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 because 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 more than 80% of articles included variations on at least one dimension.

Purposive variations are directly useful for showing the robustness of findings across the range of observed variations. However, without additional assumptions, the purposive variations are inherently *local* in that the

FIGURE 5. Range Assumption



(a) Range Assumption Holds (b) Range Assumption is Violated

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 conditions under which purposive variations measured *within* the experiment help us infer the sign of the T-PATE, which is *external* to the experiment.

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 the subsection Other Dimensions.

Although there are many valid ways to choose variations for outcomes, we propose a simple approach based on a range.

Assumption 5 (Target Outcomes within a Range of 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 range of the K causal effects $\{\mathbb{E}_{\mathcal{P}}\{Y_i^k(T=1,\ c) - Y_i^k(T=0,\ c)\}\}_{k=1}^K$.[4]

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).

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" (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 range assumption is violated. 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.

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

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

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

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

Connecting Purposive Variations to Sign-Generalization

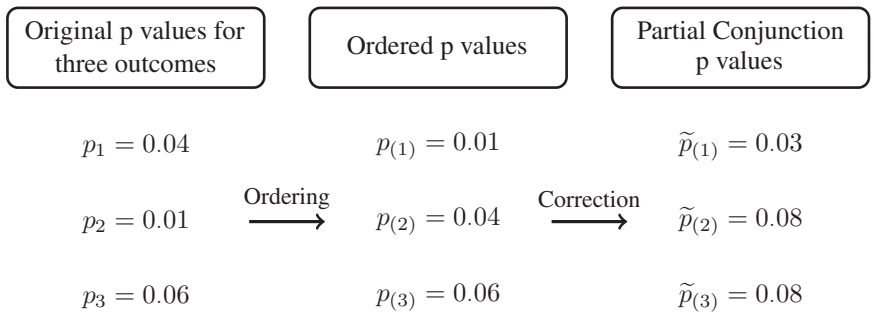
We first show that when causal effects are positive (negative) for all K outcomes, the causal effect on the target outcome is also positive (negative) under the range assumption (Assumption 5). It implies that testing the union of the K null hypotheses (Equation 8) is a valid test for the target null hypothesis (Equation 7) under the range 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 (Berger and Hsu 1996). Details and derivations are presented in Appendix H.

Partial Conjunction Test

Although checking whether all p-values are smaller than α is easy to implement, it can be too stringent in

[4] Without loss of generality, we can also apply arbitrary monotone rescaling functions f_k to match the scales of the K outcomes and the target outcomes (e.g., from binary to continuous outcomes).

FIGURE 6. Example of Partial Conjunction Test with Three Outcomes



Note: The second step of Correction is based on Equation 10.

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:

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

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 nonnulls. 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-values as follows:

$$\tilde{p}_{(1)} \equiv K p_{(1)}$$
$$\tilde{p}_{(r)} \equiv max\left\{(K-r+1)p_{(r)}, \tilde{p}_{(r-1)}\right\} \text{ for } r \geq 2. \tag{10}$$

The p-value for $\tilde{H}_0^r$ is $\tilde{p}_{(r)}$ (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 $\tilde{p}_{(r)}$ for every threshold r (see Equation 10 and examples in the section Empirical Applications). For instance, in Figure 6, we would report all three partial conjunction p-values {0.03, 0.08, 0.08}, each testing whether at least 1, 2, or 3 out of our three outcomes have positive effects. Although 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.

Other Dimensions

Although 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 points (e.g., Bisgaard 2019), and an increasing number of researchers conduct multisite field experiments (e.g., Blair and McClendon 2020; Dunning et al. 2019). 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 the next section.

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.

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 (three days, three weeks, six weeks, and three 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 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.

Although 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. Although 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 that affect both sampling and treatment effect heterogeneity.

Effect-Generalization

We estimate the T-PATE using the three classes of estimators discussed in the subsection X-Validity: Three Classes of Estimators. Weighting-based estimators include IPW and weighted OLS that adjust for control variables prespecified in the original authors' preanalysis plan. Sampling weights are estimated via calibration (Hartman et al. 2015). For the outcome-based estimators, we use OLS and a more flexible model, Baysian additive regression trees (BART). Finally, we implement two doubly robust estimators; the augmented inverse probability weighted estimator (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`.

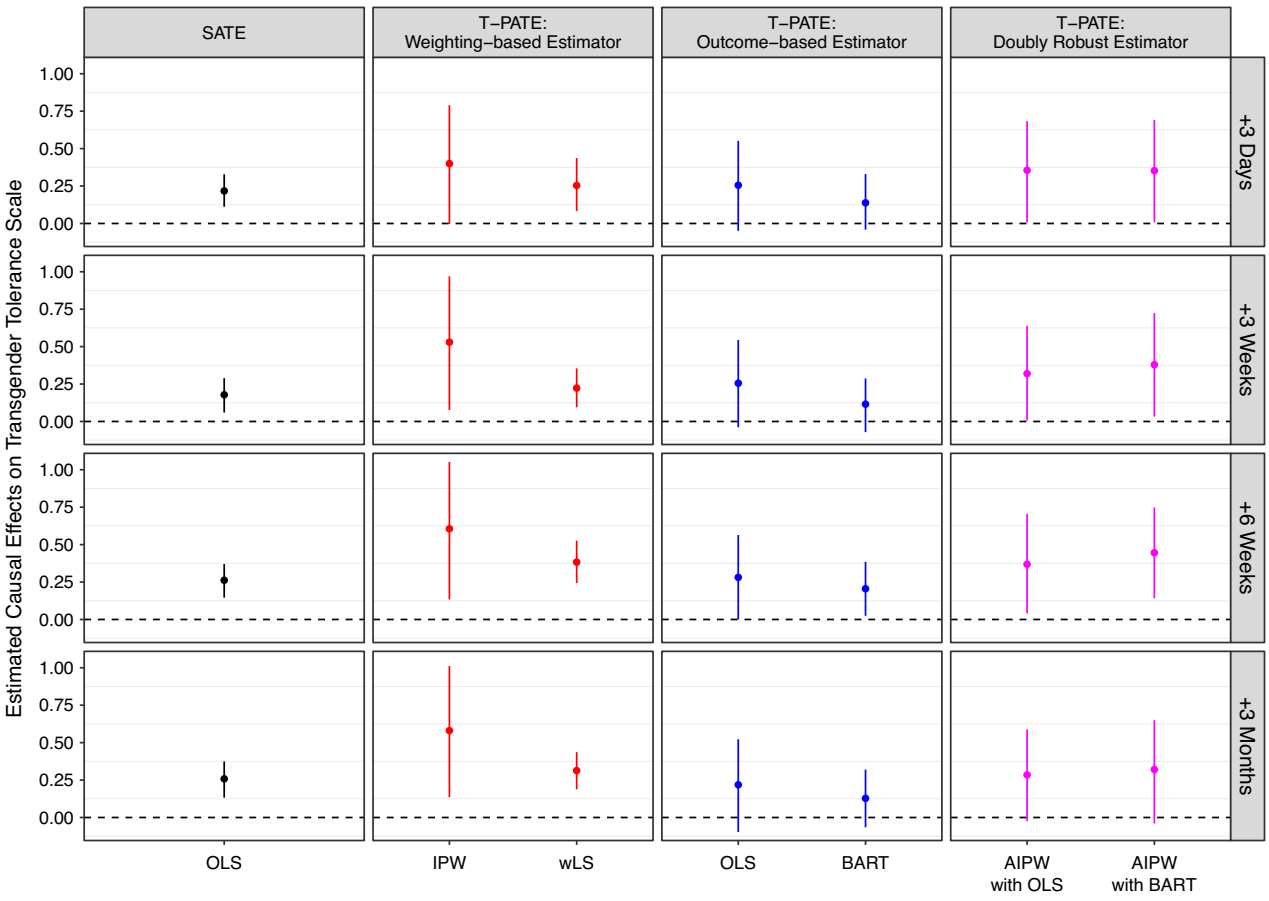
Figure 7 presents point estimates and their 95% confidence intervals using different estimators. Broockman and Kalla (2016) create an outcome index such that the value of one represents one standard deviation of the index outcome in the control group. Therefore, the estimated effects should be interpreted relative to outcomes in the control group. The first column shows estimates of the SATE for four 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 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 periods. We emphasize that, whereas 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 I 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 with 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 the cost–benefit analysis. Thus, even though point estimates are similar, the 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 pretreatment covariates measured only in the experiment, whereas 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

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 posttreatment survey waves.

an appropriate estimator in each application (see also the subsection How to Choose a T-PATE Estimator).

Survey Experiment: Partisan-Motivated Reasoning

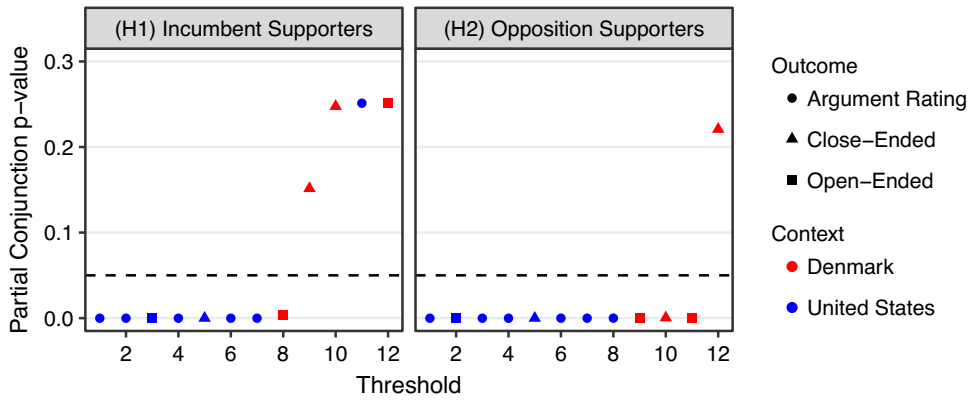
Bisgaard (2019) finds that, even when partisans agree on the facts, partisan-motivated reasoning influences how they internalize those facts and attribute credit (or blame) to incumbents. In terms of external validity analysis, Bisgaard (2019) provides several great opportunities to evaluate sign-generalization 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 US and Studies 2–4 in Denmark. See Table 1 of the original study for more details). They differ in terms of both 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.

Although 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 range 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 range assumption, and we use all the outcomes for the sign-generalization test.

FIGURE 8. Sign-Generalization Test for Bisgaard (2019)



Note: We combine causal estimates on multiple outcomes across four survey experiments in two countries. Following the suggestions in the section Sign-Generalization, we report partial conjunction *p*-values for all thresholds.

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, supporters of the incumbent party will be more likely (H1) and supporters of the opposition party will be less likely (H2) 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 that of 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 12 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.

TABLE 2. Design of Purposive Variations for Bisgaard (2019)

	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)

Note: The number of the purposive outcome variations is in parentheses.

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 8 out of 12 outcomes (66%) have partial conjunction *p*-values less than the conventional significance level of 0.05 . It is notable that most of the estimates that do not support the theory are from Denmark, which we might expect because partisan-motivated reasoning would be weaker in Denmark. In contrast, for opposition supporters, the results show 11 out of 12 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's (2019) theory is more robust for explaining opposition supporters; opposition supporters engage more in partisan-motivated reasoning than do incumbent supporters.

DISCUSSION

Addressing Multiple Dimensions Together

As illustrated by our empirical applications in the previous section, 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 the section Formal Framework for External Validity. 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. For the field experiment in our empirical application, 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.

For sign-generalization, researchers can address multiple dimensions simultaneously as long as they include purposive variations for relevant dimensions. This is one of the main advantages of sign-generalization. For the survey experiment in our empirical application, 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.

Relationship to Replication and Meta-Analysis

Meta-analysis is a method for summarizing statistical findings from multiple papers or research literature. Although 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" (different researchers conduct similar experiments over time without explicit coordination across researchers) or increasingly relevant, "coordinated scientific replication" experiments like the EGAP Metaketa studies (Blair and McClendon 2020).[5] 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-analyses.

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 the 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 range assumption is more plausible.

Second, both effect- and sign-generalization are important for meta-analysis. Some studies, such as Dunning et al. (2019), clearly attempt to provide policy recommendations and evaluate the cost effectiveness of particular interventions. Estimators for the T-PATE are essential when meta-analysts want to predict causal effects in new target sites. Sign-generalization is useful when a meta-analysis focuses on synthesizing scientific knowledge—for example, 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.

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. Although 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.

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 seek to improve empirical approaches for external validity by proposing a framework and developing tailored methods for effect- and sign-generalization. We clarify the 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 seeks 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

[5] 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.

which this challenging yet essential inference is possible, and we propose new methods for improving external validity.

SUPPLEMENTARY MATERIALS

To view supplementary material for this article, please visit http://doi.org/10.1017/S0003055422000880.

DATA AVAILABILITY STATEMENT

Research documentation and/or data that support the findings of this study are openly available at the American Political Science Review Dataverse: https://doi.org/10.7910/DVN/3EKRSI.

ACKNOWLEDGMENTS

The proposed methodology is implemented via the open-source software R package `evalid`, available at https://github.com/naoki-egami/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.

CONFLICT OF INTEREST

The authors declare no ethical issues or conflicts of interest in this research.

ETHICAL STANDARDS

The authors affirm this research did not involve human subjects.

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): 444–55.

Bareinboim, Elias, and Judea Pearl. 2016. "Causal Inference and the Data-Fusion Problem." *Proceedings of the National Academy of Sciences* 113 (27): 7345–52.

Benjamini, Yoav, and Ruth Heller. 2008. "Screening for Partial Conjunction Hypotheses." *Biometrics* 64 (4): 1215–22.

Berger, Roger L., and Jason C. Hsu. 1996. "Bioequivalence Trials, Intersection-Union Tests and Equivalence Confidence Sets." *Statistical Science* 11 (4): 283–319.

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): 824–39.

Blair, Graeme, and Gwyneth McClendon. 2020. "Conducting Experiments in Multiple Contexts." In *Handbook of Experimental Political Science*, eds. Donald P. Green and James Druckman, 411–28. Cambridge: Cambridge University Press.

Broockman, David, and Joshua Kalla. 2016. "Durably Reducing Transphobia: A Field Experiment on Door-to-Door Canvassing." *Science* 352 (6282): 220–24.

Campbell, Donald T., and Julian C. Stanley. 1963. *Experimental and Quasi-Experimental Designs for Research*. Chicago, IL: Rand McNally.

Cole, Stephen R., and Elizabeth A. Stuart. 2010. "Generalizing Evidence from Randomized Clinical Trials to Target Populations: The ACTG 320 Trial." *American Journal of Epidemiology* 172 (1): 107–15.

Coppock, Alexander, Thomas J. Leeper, and Kevin J. Mullinix. 2018. "Generalizability of Heterogeneous Treatment Effect Estimates across Samples." *Proceedings of the National Academy of Sciences* 115 (49): 12441–446.

Correa, J., J. Tian, and E. Bareinboim. 2019. "Adjustment Criteria for Generalizing Experimental Findings." In *Proceedings of Machine Learning Research, Vol. 97: International Conference on Machine Learning, 9–15 June 2019, Long Beach, California, USA*, eds. Kamalika Chaudhuri and Ruslan Salakhutdinov, 1361–69. Long Beach, CA: PMLR.

Dahabreh, Issa J., Sarah E. Robertson, Eric J. 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): 685–94.

Deaton, Angus, and Nancy Cartwright. 2018. "Understanding and Misunderstanding Randomized Controlled Trials." *Social Science & Medicine* 210 (August): 2–21.

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): 217–43.

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

Egami, Naoki, and Erin Hartman. 2022. "Replication Data for: Elements of External Validity: Framework, Design, and Analysis." Harvard Dataverse. Dataset. https://doi.org/10.7910/DVN/3EKRSI.

Egami, Naoki, and Erin Hartman. 2021. "Covariate Selection for Generalizing Experimental Results: Application to a Large-Scale Development Program in Uganda." *Journal of the Royal Statistical Society: Series A* 184 (4): 1524–48.

Findley, Michael G., Kyosuke Kikuta, and Michael Denly. 2020. "External Validity." *Annual Review of Political Science* 24: 365–930.

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

Hartman, Erin. 2020. "Generalizing Experimental Results." Chap. 21 in *Advances in Experimental Political Science*, eds. James Druckman and Donald P. Green. Cambridge: Cambridge University Press.

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* 178 (3): 757–78.

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* 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* 48 (6): 3283–311.

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): 103–27.

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): 275–91.
Morton, Rebecca B., and Kenneth C. Williams. 2010. *Experimental Political Science and the Study of Causality: From Nature to the Lab*. Cambridge: Cambridge University Press.
Mullinix, Kevin J., Thomas J. Leeper, James Druckman, and Jeremy Freese. 2015. "The Generalizability of Survey Experiments." *Journal of Experimental Political Science* 2 (2): 109–38.
Mutz, Diana C. 2011. *Population-Based Survey Experiments*. Princeton, NJ: Princeton 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): 465–72.
Paluck, Elizabeth Levy, Seth A. Green, and Donald P. Green. 2019. "The Contact Hypothesis Re-Evaluated." *Behavioural Public Policy* 3 (2): 129–58.
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): 846–66.
Rubin, Donald B. 1974. "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies." *Journal of Educational Psychology* 66 (5): 688–701.
Shadish, William R., Thomas D. Cook, and Donald T. Campbell. 2002. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston, MA: 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): 239–66.
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): 438–43.
Wilke, Anna, and Macartan Humphreys. 2020. "Field Experiments, Theory, and External Validity." Chap. 53 in *The SAGE Handbook of Research Methods in Political Science and International Relations*, eds. Luigi Curini and Robert Franzese. Piscataway, NJ: Transaction Publishers.
Young, Lauren E. 2019. "The Psychology of State Repression: Fear and Dissent Decisions in Zimbabwe." *American Political Science Review* 113 (1): 140–55.