ANNUAL REVIEWS

Annual Review of Political Science

External Validity

Michael G. Findley,[1] Kyosuke Kikuta,[2] and Michael Denly[1]

[1]Department of Government, University of Texas, Austin, Texas 78712, USA; email: mikefindley@utexas.edu

[2]Osaka School of International Public Policy, Osaka University, Osaka 560-0043, Japan

Annu. Rev. Political Sci. 2021. 24:365–93

The *Annual Review of Political Science* is online at polisci.annualreviews.org

https://doi.org/10.1146/annurev-polisci-041719-102556

Keywords

external validity, generalizability, generalization, transportability, population, sample

Abstract

External validity captures the extent to which inferences drawn from a given study's sample apply to a broader population or other target populations. Social scientists frequently invoke external validity as an ideal, but they rarely attempt to make rigorous, credible external validity inferences. In recent years, methodologically oriented scholars have advanced a flurry of work on various components of external validity, and this article reviews and systematizes many of those insights. We first clarify the core conceptual dimensions of external validity and introduce a simple formalization that demonstrates why external validity matters so critically. We then organize disparate arguments about how to address external validity by advancing three evaluative criteria: model utility, scope plausibility, and specification credibility. We conclude with a practical aspiration that scholars supplement existing reporting standards to include routine discussion of external validity. It is our hope that these evaluation and reporting standards help rebalance scientific inquiry, such that the current obsession with causal inference is complemented with an equal interest in generalized knowledge.

Crude classifications and false generalizations are the curse of the organized life.

H.G. Wells (1905, p. 191)

External validity: captures the extent to which inferences drawn from a given study's sample apply to a broader population or other target populations

Generalizability: refers to inferences based on a sample drawn from a defined population

Transportability: refers to inferences based on a sample but targeted at a different population

1. INTRODUCTION

Lewis Carroll's *Alice's Adventures in Wonderland* tells the story of Alice falling down a rabbit hole into a psychedelically strange location that became all-encompassing to the point that Alice confused reality with fantasy. Echoing Carroll (1865), in a 2015 issue of *The New Yorker*, Kathryn Schulz suggests that "these days. . .when we say that we fell down the rabbit hole, we seldom mean that we wound up somewhere psychedelically strange. We mean that we got interested in something to the point of distraction—usually by accident, and usually to a degree that the subject in question might not seem to merit" (Schulz 2015).

We fear that the social sciences have fallen down an internal validity rabbit hole to an unmerited point of distraction. Over the past 30 years, the credibility revolution—focused on design-based, internal validity—has contributed to major scientific breakthroughs across various fields (Angrist & Pischke 2010, Pearl & Mackenzie 2018). Indeed, tens of thousands of causally well-identified studies are now complete or in progress, which is unequivocally a positive development (see Druckman et al. 2006, Angrist & Pischke 2010, Samii 2016). However, it is often unclear how the results of many internally valid studies apply beyond their immediate objects of investigation. This is a grave problem. Although social scientists study particular features of the social world, ultimately they care about making inferences beyond the data at hand, which is "the ultimate goal of all good social science" (King et al. 1994, pp. 8, 34). Indeed, inference is what sets social science apart from history and other idiographic disciplines. Without the ability to apply findings from specific studies to the wider world, an inference is of little interest. It is time that political scientists (and other social scientists) take external validity more seriously.

The concept of external validity has existed for decades (e.g., Campbell 1957). In its most basic form, external validity captures the extent to which inferences drawn from a given study's sample apply to a broader population or other target populations. When an inference concerns the broader population of a predefined sample, the literature refers to it as one of generalizability (Lesko et al. 2017). By contrast, when an inference applies to other target populations, it corresponds to transportability (Pearl & Bareinboim 2014). All credible external validity inferences—whether they refer to generalizability or transportability—need to account for multiple dimensions. To date, social scientists have grouped these dimensions under the UTOS framework from Cronbach & Shapiro (1982); the acronym captures units, treatments, outcomes, and settings. Given that the more recent literature stresses the importance of mechanisms and time for external validity, we add two letters to UTOS—mechanisms and time—and regroup these dimensions under the broader framework of M–STOUT.

As with internal validity, which begins from a fundamental problem of causal inference,[1] external validity faces a similar challenge. Although some have suggested a fundamental solution to the problem of external validity (Bareinboim & Pearl 2013, 2016; Marcellesi 2015), it is generally impossible to fully approximate or accurately account for external validity in its entirety. Notably,

[1]Causal inference involves counterfactual reasoning, but the fundamental problem of causal inference is that the counterfactual for any internal validity inference can never be observed. Scholars attempt to overcome the fundamental problem of causal inference through random assignment of the independent variable of interest to treatment and control groups, which helps approximate the unknown counterfactual as the sample size increases to infinity. For more information, see Holland (1986) and Imbens & Rubin (2015).

inferences about the future can never be verified at the time of a study, and causal interaction between the mechanism M and other STOUT dimensions makes external validity more difficult to obtain in practice than conventional random sampling presumes (Muller 2015).[2]

This article on external validity advances three core goals. First, on a broad scale, the article serves as a call to make external validity an equal partner with internal validity/causal inference in scientific progress. A fast-emerging methodological literature on external validity covers generalizability, transportability, and the M–STOUT dimensions. However, this literature is not very coherent, which is perhaps one reason why social scientists in all disciplines rarely apply the insights from methodological work on external validity. To better understand the disconnect between the methodological and applied work, we coded more than 1,000 randomly sampled articles from 12 social science journals to gauge how they address external validity.[3] Roughly 65% of the articles in this sample contain direct or indirect mention of external validity. Nevertheless, only an exceptional few contained a dedicated external validity discussion, and even those articles made limited and sometimes inaccurate external validity inferences.

Second, we review the literature on the conceptual dimensions of external validity and then clarify the meaning of external validity. In particular, the article synthesizes the insights from the methodological literature on external validity to clarify the various dimensions of M–STOUT, the causal interactions across them, and the distinction between generalizability and transportability. Little applied work makes the distinctions that the methodological studies articulate, which has led to many inaccurate inferences about external validity.

Third, precisely because achieving external validity is not straightforward, Section 4 organizes theoretical and methodological advice from the literature by advancing three evaluative criteria. Model utility refers to the utility of a model that organizes the inference(s) from a sample or research synthesis. Scope plausibility refers to the extent to which a study's sample dimensions and population counterparts are plausibly selected and developed. Specification credibility refers to the extent to which theoretical and empirical methods provide defensible inferences that inform a theoretical population(s) of interest. Studies that make credible inferences about generalizability need to approach the ideal of estimating the population average treatment effect (PATE). By contrast, studies that make credible inferences about transportability need to approach the ideal of estimating the target (population) average treatment effect (TATE).

A robust set of evaluative criteria not only provides guidance to help avoid the above pitfalls but also can ensure that all studies are held to the same standards. Studies on the United States or France, for example, need to be subject to the same scrutiny as work on Malawi, Myanmar, or historical Palau. Giving studies a free pass on external validity because they are studying Western, educated, industrialized, rich, democratic, or populous countries is unscientific and yet unquestionably the norm (see Henrich et al. 2010, Wilson & Knutsen 2020). Nevertheless, we wish to emphasize that the results derived from samples need not apply universally to justify their importance. A study of Malawi may not apply far and wide, just as a study of the United States may not. If a study in Malawi applies to only two other countries, and a study of the United States applies to only two other countries, that is acceptable. An elevated role for external validity in the social sciences means that scholars should rigorously understand the level of external validity, report on it accurately, and apply those same standards to all studies.

Model utility: the utility of a model that organizes the inference(s) from a sample or research synthesis

Scope plausibility: the extent to which a study's sample dimensions and population counterparts are plausibly selected and developed

Specification credibility: the extent to which theoretical and empirical methods provide defensible inferences that inform a theoretical population(s) of interest

[2] Leamer (2010, p. 34) refers to causal interaction issues as ones related to "interactive confounding variables."
[3] *American Political Science Review*, *American Journal of Political Science*, *Journal of Politics*, *American Economic Review*, *Quarterly Journal of Economics*, *Journal of Political Economy*, *Journal of Personality and Social Psychology*, *Psychological Review*, *Annual Review of Psychology*, *American Sociological Review*, *American Journal of Sociology*, and *Social Science Research*. See Findley et al. (2022) for a fuller report.

A key theme in the literature is the difference in how experimentalists and observationalists approach external validity. In reviewing this literature, we conclude that both groups need to improve their external validity inferences. Experimentalists (including natural experimentalists) appear largely not to care much about external validity, and their work has huge limitations with respect to sample selection bias (i.e., the S, U, and T in M–STOUT). Observational researchers, for their part, appear to have a false sense of security about external validity. For example, contrary to popular belief, large-*N* time series cross-sectional (TSCS) data do not build in external validity by design. Even if we assume that observationalists solve sample selection issues using pooled or random samples that preserve the integrity of the representativeness (a heroic assumption), observational work often relies on poor indicators for treatments and outcomes, suggesting large problems with variable selection bias (i.e., the T and O in M–STOUT). Eliding the seriousness of the time dimension and failing to engage external validity in a transparent, falsifiable way are drawbacks for both experimentalists and observationalists.

As the internal validity boom crowds out external validity, much hangs in the balance. Typically, skeptics ask: Why worry about external validity if you cannot identify causality? Our response is: Why worry about causality if you cannot discern external validity? Indeed, one of our main arguments is that bias due to external validity can be just as severe as bias due to internal validity. Identifying causality is critically important but, by itself, falls short in producing generalized knowledge. Researchers, reviewers, and editors need to embrace external validity as a core objective of scientific inquiry.

2. WHAT IS EXTERNAL VALIDITY?

Validity refers to the approximate truth or usefulness of an inference (Trochim & Donnelly 2006). Validity is not a property of a theory or design but rather of a study's inferences. That is, inferences could be valid in some studies but not in others, even if the studies have exactly the same theory or design (Shadish et al. 2002).

External validity captures the extent to which inferences drawn from a given study's sample apply to a broader population or other target populations. We make the distinction between a broader population and other target populations because external validity takes on two different forms. Generalizability refers to inferences based on a sample drawn from a defined population (Lesko et al. 2017), and transportability refers to inferences based on a sample but targeted at a different population (Pearl & Bareinboim 2014). The credibility of both generalizability and transportability inferences depends on the extent to which each can account for multiple dimensions, including mechanisms, settings, treatments, outcomes, units, and time (M–STOUT).

The evolving terminology of external validity has created significant confusion over the past 60 years. Influenced heavily by Brunswik (1947) on representative design, Campbell (1957) coined the distinction between external and internal validity. That dichotomy endured until Cook & Campbell (1979) separated out (and added) the distinct term of statistical conclusion validity[4] from internal validity, and construct validity[5] from external validity. Since 1979, those four terms have endured, though social scientists continue to privilege internal validity over the other three. More recently, some scholars simply use a separate term, generalizability, to refer to external validity writ

[4]Statistical conclusion validity refers to the appropriate use of statistical methods to assess whether a causal relationship exists and generalizes.

[5]Construct validity refers to a variable that is operationalized such that it corresponds to the larger theoretical concept of interest (see Trochim & Donnelly 2006).

large. Unfortunately, social scientists mostly conflate generalizability with the related concept of transportability. Although many social scientists do not yet use the term transportability, it has become standard in the statistics literature and provides a much-needed clarification.

2.1. Scope, Populations, and Samples

A first step in establishing external validity is to clearly identify a study's scope, which refers to the applicability and limitations of a theory or argument (Walker & Cohen 1985, Goertz & Mahoney 2012). Though poorly understood, the identification of scope conditions is perhaps the most common approach for making external validity inferences. The correct identification of scope is crucial because it sets the parameters for the definition of population(s) and sample(s). Defining scope, including the identification of populations and samples, is among the most fundamental research design decisions, and one taught in basic research design courses (see, for example, Trochim & Donnelly 2006).

Scholars sometimes declare the scope at the outset of a study, perhaps through a deductive exercise that focuses a research question and associated argument. Other times, scholars define the scope inductively, including through robustness checks or other ways of identifying the applicability and limitations of some causal or descriptive inference. Irrespective of the method by which scholars define a study's scope, the more crucial aspect for external validity purposes is that they define it fully. An essential task in this endeavor is to distinguish between the theoretical and accessible population(s), as well as the associated samples. Without clear theoretical populations, it is unclear how inferences travel. Knowing populations that a researcher can access can aid this endeavor and provides clearer understandings of samples, but the researcher must clearly articulate the theoretical population to understand the accessible population and sample.

Another crucial task is to delineate clear scope conditions that specify the bounds of an inference with respect to M–STOUT. Although scholars sometimes use the term scope conditions, similar to the broader concern about external validity, such discussion is frequently incomplete or imprecise. When scholars are precise about scope conditions, most often they pertain to units and, at times, settings, but such scope conditions are overly narrow from the perspective of theory or knowledge accumulation. Indeed, in our survey of the literature, almost no articles raised or discussed scope conditions in any comprehensive way.

Figure 1 illustrates the distinctions among the scope, populations, and samples. The distinction between scope and populations is subtle; scope is broader and thus includes multiple populations and samples, especially for the M–STOUT dimensions. The figure depicts several other core components of external validity, including generalizability and transportability, to which we now turn.

2.2. Generalizability and Transportability

Generalizability refers to the validity of an inference based on a sample that is randomly or nonrandomly drawn from a defined population (Lesko et al. 2017). Because a sample is drawn from a defined population, the sample must be a subset of population: $S \subseteq P$. In **Figure 1**, generalizability is an inference from the sample S_1 to population P_1. Unfortunately, in **Figure 1**, the sample was not randomly selected and thus less representative.

The other variant of external validity, transportability, relates to extending inferences based on a given sample to another population (Pearl & Bareinboim 2014). Because the sample at hand is not drawn from the target population, the sample is not a subset of the population: $S \nsubseteq P$. In **Figure 1**, transportability refers to the validity of an inference from a sample S_1 to population

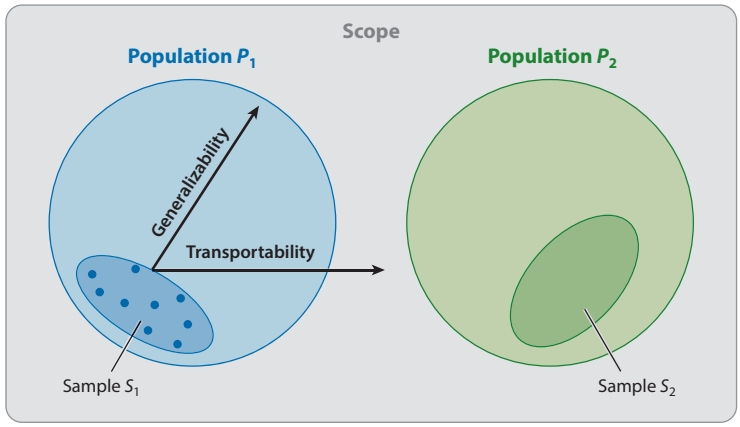

Figure 1

The difference between scope, populations, and samples. The light gray area is the scope of a study (entity of theoretical interests), the blue and green circles are populations (entity of empirical interest), the darker ellipses are sample spaces, and the blue dots are observations in sample S_1. Sample S_1 could be drawn from a defined population P_1, or it may reflect on an unknown (or poorly defined) population P_1. The inference about population P_1 based on the sample S_1 concerns generalizability, because the sample is a subset of the population. By contrast, the inference about population P_2 based on the sample S_1 relates to transportability, because the sample S_1 is not a subset of a predefined population. When one is interested in extending the results of S_1 to another sample S_2, the target population is the same as the sample: $P_2 = S_2$. This is a special case of a transportability inference. Note that the sample S_2 is unnecessary for generalization, and even for the inference of P_2, the sample S_2 does not always exist—though it can certainly help.

P_2.[6] Because sample S_1 is not a drawn from the population of interest P_2, the external validity may not be high, but it still needs explicit evaluation. In general, transportability poses more empirical challenges than generalizability.[7]

2.3. External Validity Dimensions: M–STOUT

External validity inferences involve analysis of several dimensions. Most existing work uses some variation of the UTOS framework from Cronbach & Shapiro (1982). UTOS gives appropriate attention to units, treatments, outcomes, and settings, and these dimensions need to remain central to external validity inferences. However, the UTOS focus mostly neglects the essential roles of time and mechanisms. We first review each of the UTOS dimensions and then add a discussion of mechanisms and time, underscoring why each is essential for external validity.

2.3.1. Units. Unit-specific external validity inferences are perhaps the most common approach in social science and concern: For which population units do sample inferences hold? A first step along these lines is to identify who (or what) are the units as well as other characteristics of the units that may matter for external validity. One effective means of doing so is to compare a sample's summary statistics against those at the population level through a balance test (e.g., Biglaiser

[6]When one is interested in extending inferences about S_1 to another sample S_2, the population is the same as the targeted sample: $P_2 = S_2$. This is a special case of transportability inference.

[7]Although clearly defining a population of research is recommended, our notation does not assume that populations would always be well defined. What is necessary for distinguishing generalizability and transportability is to articulate whether a sample is a subset of a population.

& Staats 2010). If the sample statistics are generally not that different, then the sample inference may travel to the population level. Alternatively, scholars can make an inference that is primarily related to one country and then rerun their models using data from other countries to assess an inference's transportability. The Metaketa Initiative from Evidence in Governance and Politics (EGAP) follows such a model for field experiments (see Dunning et al. 2019), and Klašnja & Titiunik (2017) provide a relevant example in an observational setting (see also Guardado & Wantchékon 2018). Finally, population-based survey experiments, including those supported by Time-sharing Experiments for the Social Sciences (TESS), provide researchers with the ability to make internally valid inferences at the population level, which helps with external validity as well. Of course, scholars study a wide range of units, such as individuals, households, neighborhoods, municipalities, provinces, countries, and regions. In some cases, scholars include units at multiple levels (e.g., studies with individuals nested in municipalities). Accordingly, many unit-based external validity inferences are more challenging than they may first appear.

2.3.2. Treatments. The external validity of a treatment principally concerns the extent to which inferences hold across different operationalizations of the main explanatory variable of interest to a study.[8] Essentially, an externally valid treatment variable must have construct validity—that is, the treatment must be operationalized such that it corresponds to the larger theoretical concept of interest. Alas, scholars often prioritize convenience over construct validity in variable operationalization, which makes the task of assessing external validity inferences across operationalizations difficult. Such a task is especially difficult when the treatment lacks mundane realism—that is, correspondence to the the actual treatment that people receive in their everyday lives (see Druckman & Kam 2011). Whereas field experiments and observational studies generally have high mundane realism, some—but not all!—laboratory and survey experiments do not and thus provide a poor basis for external validity inferences (see also Falk & Heckman 2009, Mutz 2011, Camerer 2015, Kessler & Lise 2015).

The literature also stresses that treatments may provoke heterogeneous effects among subgroups, making it sometimes necessary to devise a correction or weighting method for external validity inferences (Cole & Stuart 2010; Imai & Ratkovic 2013; Coppock et al. 2018, 2020). In some cases, particularly under imperfect compliance or selection into a treatment, effect heterogeneity may necessitate a change in the estimand of interest away from the average treatment effects. For such instances, Heckman & Vytlacil (2001, 2005) recommend the use of marginal treatment effects and policy-relevant treatment effects. In these same articles and a series of others, these authors show how to integrate these estimands within structural models, which provide a direct correspondence to theory and thus a stronger basis for external validity inferences.

2.3.3. Outcomes. External validity for outcomes is similar to that of treatments in that it mainly pertains to whether inferences hold across different operationalizations of the dependent variable. As with treatments, outcome-related external validity requires some form of mundane realism.[9] Perhaps the most prominent way to assess the external validity of outcomes in the literature is

[8] Of course, it is better if control variables are also well measured, but we specify a focus on the main explanatory variable of interest because, as Pearl (2009) shows using directed acyclic graphs (DAGs), there is a need to distinguish active treatments involving an intervention from passive control variables. Pearl (2009) makes such a distinction with the do-operator. For a gentle introduction to DAGs, see Pearl & Mackenzie (2018).

[9] Incidentally, lack of mundane realism is likely a prominent reason why political scientists have not adopted DAGs, despite their solution to transportability and data fusion problems (see Pearl & Bareinboim 2014, Bareinboim & Pearl 2016).

through replication (McDermott 2011). Outside of replication, scholars have proposed making individual transportability inferences using matching, weighting, and simulation (Hotz et al. 2005, Cole & Stuart 2010, Tipton 2013, Tipton et al. 2014, Allcott 2015, Hartman et al. 2015, Pritchett & Sandefur 2015, Dahabreh et al. 2016, Westreich et al. 2017, Buchanan et al. 2018, Dehejia et al. 2021).

2.3.4. Settings. Settings refer to the environments in which a study's data are generated, which could be places such as a laboratory, country, or village. As numerous scholars underscore, observational, laboratory experimental, survey experimental, and field experimental settings yield different levels of external validity (e.g., Mutz 2011, Coppock & Green 2015, Findley et al. 2017b, Breskin et al. 2019). The bulk of the literature, however, centers on a large and vibrant debate on the merits and drawbacks of observational and experimental settings (e.g., Rodrik 2009, Ravallion 2012, Pritchett & Sandefur 2015, Breskin et al. 2019). Barring attrition, noncompliance, or spillover, experiments tend to be stronger than observational settings with regard to internal validity, but that can come at a cost in terms of external validity. With the exception of some types of survey experiments,[10] scholars have experimented at scale only to a limited extent (Muralidharan & Niehaus 2017). Accordingly, experimental estimates generally concern smaller segments of a study's theoretical population. In turn, Gisselquist (2020) goes so far as to suggest that experiments are precisely estimated case studies, from which it is difficult to make any external validity inference. Banerjee et al. (2017b) suggest that experimentalists can overcome such hurdles by subgroup analysis, weighting, and stratification on theoretically relevant subpopulations, but these strategies usually require stronger assumptions, which experimentalists may or may not deem plausible. For their part, observational studies can overcome many of these challenges, but to do so the relevant sample must approximate a random selection of a study's theoretical population (Breskin et al. 2019). For many observational studies, however, that is just as unrealistic as it is for experiments.

2.3.5. Time. Although the time dimension has not received as much scholarly attention as some of the others, time constitutes an essential dimension of external validity. As K. Munger (unpublished manuscript) convincingly argues, to neglect time is to essentially assume that (*a*) treatment effects, including those for relevant subgroups, do not change; (*b*) the composition of the population of interest remains static; and (*c*) all relevant confounders are already identifiable and measurable. Clearly, these are untenable assumptions. At least since Pierson's (2000) famous article, most political scientists have known that social science is subject to contingencies, path dependency, multiple equilibria, and inertia/feedback effects. Although these issues may seem difficult to disentangle from an external validity perspective, Grzymala-Busse (2011) provides relevant guidance, showing how such time-related factors affect the salience and duration of mechanisms. As we highlight throughout the article, at its core, external validity relates to how mechanisms travel. Perhaps even more fundamentally, though, scholars cannot elide time because the target population of interest for any inference pertains to some future state.

2.3.6. Mechanisms. External validity inferences across each UTOS dimension, in addition to time, require attention to the characteristics of the inference, necessitating inclusion of mechanisms. The concept of a mechanism is fraught with complexity (Gerring 2008), but it also holds tremendous promise for producing generalized social scientific knowledge (Cartwright 2020). In

[10]We are notably referring to population-based survey experiments and online survey experiments (Mutz 2011, Berinsky et al. 2012, Huff & Tingley 2015). Scholars using online survey experiments through Amazon's Mechanical Turk show that these experiments represent the US population rather well. However, the extent to which these online experiments represent populations from other countries is very limited.

Table 1 Examples of M–STOUT

Dimension[a]	Study A[b]	Study B[b]	Target inference[c]
Mechanisms	Women's empowerment	NA	Youth empowerment
Settings	Survey experiment in Liberia	TSCS regression of African countries	Field experiment in Guinea
Treatments	CDD projects	Any aid projects	Direct budget support
Outcomes	Self-reported social trust	WVS indicators of social trust	Results of a social trust game
Units	Individuals within villages	African countries	Individuals within counties
Time	Year of 2000	Years of 2000–2020	Year of 2020

[a]Each M–STOUT element is listed in this column.
[b]The corresponding elements in a study about Liberia (Study A) and a TSCS study about African countries, 2000–2020 (Study B).
[c]Examples of possible external validity targets.
Abbreviations: CDD, community-driven development; NA, not applicable; TSCS, time series cross-sectional; WVS, World Value Survey.

the narrowest and most limiting sense, mechanisms are considered to be mediators (Imai et al. 2011), occurring after a treatment and before an outcome. Then, the treatment works through the mediator to affect the outcome. In a broader sense, mechanisms take on many other forms including constraints, equifinality, and interactions (Weller & Barnes 2014, Goertz 2017). Irrespective of how scholars define mechanisms, they most commonly refer to regularly occurring causal relationships, not idiosyncratic chains of events. Accordingly, mechanisms must be capable of traveling across each of the remaining STOUT dimensions for an external validity inference to be credible, and we thus reorganize the label away from UTOS to M–STOUT.

2.3.7. Example studies. For stylized examples, consider two studies: (*a*) a survey experiment about the effects of an economic aid program on social trust, implemented with a village-level community-driven development (CDD) council in Liberia in the year 2000 (Study A), and (*b*) a TSCS regression of social trust indicators in the World Value Survey on aid projects in African countries, 2000–2020 (Study B). **Table 1** presents each of the M–STOUT dimensions in the studies and how they can be generalized or transported to different contexts. Scholars may wonder, for instance, whether findings can hold in different settings (e.g., field experiment), times (e.g., year of 2020), or units of the analysis (e.g., individuals in Guinea). Similarly, scholars may also be interested in whether the results can be transported to different treatment variables (e.g., aid programs using direct budget support) or outcome variables (e.g., social trust measured by a social trust game). Moreover, mechanisms (e.g., the effect via women's empowerment versus youth empowerment) may vary across the STOUT dimensions. For these reasons, the results of the survey experiment are feasibly transportable to exactly the same survey experiment in a different year or to another survey experiment but with a different measure of social trust. However, these caveats do not necessarily mean that the results of the survey experiment in 2000 could be transported to the survey experiment conducted in 2020 and with the different measure of the outcome. In this case, the time and outcome interact; hence, we can change either time or outcome but not both.

3. WHY DOES EXTERNAL VALIDITY MATTER?

The social science literature has long demonstrated mathematically the importance of internal validity, emphasizing in particular the analytic leverage gained within the potential outcomes framework (Imbens & Rubin 2015). In recent years, the literature has generated many advances toward understanding external validity, but they have yet to be formally integrated into a model connected directly to internal validity, enabling a demonstration of whether and how each type of validity matters. We thus formalize a single model that captures the core dimensions of both internal and

external validity. It demonstrates that when external validity is ignored, the traditional focus on internal validity can lead to biased estimates.

There is no a priori reason to believe that internal validity is more important than external validity. In fact, ignoring external validity can potentially be as harmful as ignoring internal validity. To make this concrete, consider the simple difference-in-means estimator. Suppose, for example, we are interested in the effect of a treatment (e.g., an aid program) on an outcome (e.g., social trust). For simplicity, we assume that the treatment is dichotomous and the sample is divided into treated and control groups—e.g., villages that did or did not receive the aid program. The difference-in-means estimator is then the difference in the averages of the outcome variable between the treated and control groups.

The challenge is that the difference-in-means estimator almost never yields the desired quantity of interest for making broad, general inferences. To do so, it is necessary to distinguish between the different biases that affect the difference-in-means estimator (Imai et al. 2008, Cole & Stuart 2010, Hartman et al. 2015, Westreich et al. 2017, Andrews & Oster 2019). Using the potential outcomes framework (see Imbens & Rubin 2015),[11] we particularly distinguish four biases, two of which relate to internal validity and the rest of which pertain to external validity:[12]

$$\hat{\delta}_S = \delta_P + b_{S1} + b_{S2} + b_P + b_V. \qquad 1.$$

In Equation 1, $\hat{\delta}_S$ refers to the difference-in-means estimator for the sample. The δ_P term is the causal effect in the population of interest—that is, the PATE for generalizability inferences and the TATE for transportability inferences.[13] The difference-in-means estimator, however, does not always unbiasedly estimate the PATE or TATE. From an internal validity perspective, biases can arise from selection into the assignment of the treatment (b_{S1}) and treatment effect heterogeneity within a sample (b_{S2}). If an estimate is internally valid ($b_{S1} = b_{S2} = 0$), it is possible to unbiasedly estimate the sample average treatment effect (SATE). The existence of an unbiased SATE, however, does not mean that the PATE and TATE can also be unbiasedly estimated. In fact, the lack of a representative sample (b_P) and the difference between the variables at hand and those of substantive interest (b_V) can cause biases related to external validity, making the SATE different from the PATE or TATE. As detailed in the next subsection, without eliminating both internal and external biases, we cannot unbiasedly estimate either the PATE or the TATE.

3.1. Biases from Internal and External Validity

A notable goal of social science research is to make inferences about the effect of a treatment on an outcome in a population P. Formally, when a sample S is a subset of P ($S \subseteq P$), the inference relates to generalizability, and the quantity of interest is the PATE. By contrast, when a sample S

[11] Other than the addition of the variable selection bias b_V, our formalization follows Imai et al. (2008) and draws from Hotz et al. (2005) and Allcott (2015). For formalization using the DAG approach, readers are referred to Pearl & Bareinboim (2014, 2019) and Bareinboim & Pearl (2016).

[12] See Section 7 for the proof. $\delta_P = E_P[Y_i(1) - Y_i(0)]$ for $i \in P$, where P is a population, $Y(1)$ and $Y(0)$ are the potential outcomes with and without a treatment, and i is a unit in a population. $b_{S1} = E_S[Y_j(0) \mid D_j = 1] - E_S[Y_j(0) \mid D_j = 0]$ for $j \in S$, where S is a sample, D is a dichotomous treatment variable, and j is a unit in a sample. $b_{S2} = \Pr(D_j = 0)\{E_S[Y_j(1) - Y_j(0) \mid D_j = 1] - E_S[Y_j(1) - Y_j(0) \mid D_j = 0]\}$ for $j \in S$. $b_P = \Pr(W_i = 0)\{E_P[Y_i(1) - Y_i(0) \mid W_i = 1] - E_P[Y_i(1) - Y_i(0) \mid W_i = 0]\}$ for $i \in \mathrm{P}$, where W is an indicator that takes a value of 1 if a unit is included in a sample or 0 otherwise. $b_V = E_P[\tilde{Y}_i(\tilde{D}_i = 1) - \tilde{Y}_i(\tilde{D}_i = 0)] - E_P[Y_i(D_i = 1) - Y_i(D_i = 0)]$ for $i \in P$, where $\tilde{Y}_i$ and $\tilde{D}_i$ are the variables at hand, which may or may not be the same as the variables of the interests, Y_i and D_i.

[13] As Kern et al. (2016) mention in a footnote, the PATE and TATE are often used interchangeably in the literature. In this article, however, we distinguish the PATE and TATE for the purpose of conceptual clarity.

is not a subset of P ($S \not\subseteq P$), the inference relates to transportability, and the quantity of interest is the TATE.[14]

The difference-in-means estimator, however, does not always yield the PATE or TATE. The internal validity biases, b_{S1} and b_{S2}, are well known in the causal inference literature (e.g., Angrist & Pischke 2008). The assignment selection bias, b_{S1}, is the bias due to the lack of random assignment. Returning to our running example, one may nonrandomly assign the aid projects to villages with lower levels of social trust. In such a case, if we simply compare the social trust of the treated villages with that of the control villages, the former can have a lower level of social trust. This, however, does not mean that the aid program has a perverse impact.

The second internal validity bias, b_{S2}, corresponds to treatment-effect heterogeneity—that is, the differential effect of the treatment in the treatment and control groups (Heckman & Vytlacil 2005, Pritchett & Sandefur 2013, Kern et al. 2016). Continuing with our example, the villages with low social trust may have different responses to the aid program due to other factors. They might, for instance, have nonrepresentative political institutions that hinder social trust. The lack of representative political institutions may also attenuate the effects of the foreign aid. If we simply compare the treated and control villages, we may understate the effect of the aid program. Even though the effect averaged over all villages can be positive and much larger, the difference-in-means estimation may indicate only marginal improvement in social trust.

Although these internal validity biases are well known, less acknowledged is the bias due to the lack of external validity—henceforth, external validity bias (see Andrews & Oster 2019, Westreich et al. 2019). With a randomized trial or equivalent designs, it is possible to unbiasedly estimate the SATE. The latter, however, does not perfectly correspond to our goal, the PATE or TATE. The gap between the SATE and PATE/TATE is represented by the external validity biases, b_P and b_V. The sample selection bias (b_P) is the difference in the treatment effects between those included and excluded from a sample weighted by the proportion of the excluded units in a population (see also Hotz et al. 2005, Allcott 2015). As we show in Section 7, the bias becomes zero if all units in a population are included in a sample[15] or the average effects are the same for the included and excluded units.[16] With reference to M–STOUT, the settings (S), units (U), and time (T) terms are those for which sample selection bias is a core concern.

External validity is also compromised when the population outcome and treatment of interest are different from the sample variables at hand. The b_V term captures the biases due to the heterogeneities in these variables—i.e., the variable selection bias. Mathematically, the variable selection bias is the PATE/TATE with variables of interest minus the PATE/TATE with variables at hand. The bias arises from wanting construct validity, including measurement error and lack of consonance between the variable operationalizations and theoretical targets. In the M–STOUT framework, the last two terms—treatments and outcomes—are those for which variable selection bias is a core concern.

3.2. External Validity Bias Can Be as Harmful as Internal Validity Bias

An interesting insight from the above discussion is that there is no a priori reason to prefer a randomized experiment with an unrepresentative sample over an observational study with a repre-

[14]Mathematically, the PATE and TATE differ based on whether a sample is a subset of a target population (generalizability and PATE) or not (transportability and TATE). When one would like to make an inference about a target sample S', the target population is identical to the target sample (that is, $P = S'$), and hence is reduced into a special case of transportability (i.e., $S \not\subseteq P = S'$). See also Footnote 13.

[15]In transportability inference, this can never be the case because of $S \not\subseteq P$.

[16]Breskin et al. (2019) show the same using a bounds analysis.

sentative sample (Breskin et al. 2019, Gisselquist 2020). Even though the randomized experiment usually ensures that the internal validity biases, b_{S1} and b_{S2}, are zero, the experiment still suffers potentially large bias due to nonrandom sample selection, b_P, and variable selection bias, b_V. By contrast, an observational study with random selection of units and relevant variables makes b_P and b_V zero, even though there may still be internal validity biases b_{S1} and b_{S2}. Importantly, there is no a priori reason why biases on certain terms are more or less consequential than biases on other terms. Although experiments provide an unbiased estimate of the SATE, if the SATE is very different from the PATE or TATE of interest, the results can potentially be highly misleading (Hartman et al. 2015). In principle, an observational study with a representative sample can suffer some biases due to the lack of random assignment, but it can still yield an estimate that is closer to the PATE or TATE than an experiment with an unrepresentative sample (Breskin et al. 2019).

To illustrate, let us consider stylized examples of two extreme cases. In one setup, a researcher working together with an aid agency randomly assigns foreign aid to villages, but the researcher intentionally selects a sample so that the aid has the maximum positive effect on social trust. By contrast, in another setup, a researcher conducts a correlational study (e.g., regression-based) but with a representative sample. Both studies find a positive relationship between the aid program and social trust and thus recommend broader applications of the aid program. Which recommendation would be more harmful? The evident answer is "both." The first recommendation is harmful because the aid program should have weaker—or even zero or negative—effects outside of the sample. The second recommendation is also harmful because it conflates correlation and causality. There is no reason to believe that either is harmless. The point should be clear: The lack of external validity is potentially as harmful as the lack of internal validity.

Certainly, this conclusion does not mean that representative observational studies would be better than experiments with unrepresentative samples. By the same token, experiments are not a priori superior. Our point is simply that bias in the estimation of the PATE or TATE enters in a variety of ways, none of which is inherently better or worse. Our contention is not that we should compromise on internal validity, but rather that social science needs to take external validity seriously, because it too affects the biases in our inferences. Indeed, there are eminently reasonable ways to do this, as we discuss in the next section.

4. TOWARD EVALUATIVE CRITERIA FOR EXTERNAL VALIDITY

In the rapidly growing literature on external validity, many scattered but useful ideas are emerging, but the social sciences lack clear evaluative criteria for external validity. In this section, we organize these insights to articulate three key themes that constitute the basis for better evaluation of external validity: model utility, scope plausibility, and specification credibility. The first two are mostly separate criteria, relating to the mechanisms and context of the external validity inference. The third criterion characterizes the credibility of specification of model utility and scope plausibility.

It bears repeating that not every result that emerges from a given sample needs to apply universally, but it must apply in some way outside the sample. More precisely, scholars need to strive for some level of external validity and accurately characterize its extent. When doing so, the challenges for external validity are not altogether different from those of internal validity. Within the potential outcomes framework, randomization of units to treatment and control groups ensures the internal validity of the inference (Imbens & Rubin 2015). By the same token, even gold-standard, randomized experiments face challenges such as attrition, noncompliance, and spillover, so inferential challenges that limit the applicability of an inference often exist. A similar dynamic characterizes external validity; therefore, the task is to make credible, rather than universally applicable, inferences about external validity.

4.1. Model Utility

A first evaluative criterion for external validity is model utility. It refers to the utility of a model that organizes the inference(s) from a sample or research synthesis (Lucas 2003, Rodrik 2009, Clarke & Primo 2012, Bates & Glennerster 2017, Deaton 2019).[17] External validity inferences need useful models underpinning them for the effects, findings, and inferences derived from study samples to apply to broader population dimensions. We identify three components of model utility.

Model Utility Component 1 External validity inferences are tied to mechanisms rather than specific point estimates.

A specific estimate from a study need not be judged for having some precise truth value (Lieberson 1985, McIntyre 2019). By definition, point estimates are bounded, so they are unlikely to generalize to a broader population or transport to others given potentially diverse scales across M–STOUT. Although Vivalt (2020) finds some evidence that the point estimates from many randomized control trials generalize using meta analysis based on Bayesian hierarchical models, Vivalt's (2020) approach cannot account for sample/site selection bias or potential construct validity challenges that underpin variable selection bias (see Section 3). Accordingly, a point estimate–centered approach places unrealistic and misguided demands on the precision of effect size estimates. Instead, a model's usefulness captures the extent to which it characterizes the mechanism(s) and therefore can provide insight across a broader set of cases, in line with model utility components 2 and 3.

Model Utility Component 2 The mechanism is clearly specified, which entails an articulation of the causal principles.

To articulate the causal mechanism appropriately, a theory or substantive argument needs to specifically articulate the causal principles in the sample and in the target population (Russell 1912, Cartwright & Hardie 2012). Causal principles describe the underlying structure of how a cause and effect are related and, in the process, characterize the mechanisms. Most basically, causal principles identify whether a cause is necessary, sufficient, necessary and sufficient, or some probabilistic condition thereof. Causal principles also identify underlying causal assumptions, including whether equifinality characterizes the causal process, whether cause–effect relationships are subject to ceteris paribus considerations, and so forth.

For the purposes of external validity, a particularly important causal principle is the INUS condition (insufficient but necessary part of an unnecessary but sufficient condition) (Mackie 1965). When an INUS condition is present, a causal factor is important for producing an outcome but only when coupled with other factors.[18] In simpler terms, INUS conditions capture the idea that context or structural factors matter for how a cause produces an effect. For a treatment to affect an outcome, the treatment depends on other factors and, likewise, the outcome can be produced by an entirely different combination of factors.

Critically, the point here is that some causal principles characterize interactions of a cause with the underlying context. In the social sciences, various contextual or support factors, such as institutions, structure whether and how a cause is related to an effect (Acemoglu 2010, Cartwright & Hardie 2012). Returning to our running example, the success of a CDD project in Liberia may

[17] We take the idea of model utility from all of these citations, but perhaps the greatest inspirational source is Clarke & Primo's (2012) discussion of organizational models.

[18] Much social science research mixes across principles, for example, positing sufficiency with possible INUS, subject to ceteris paribus (Ashworth et al. 2014).

be realized only when it emphasizes women's, rather than youth, empowerment. Alternatively, perhaps the CDD project will be successful only when it operates at a village level rather than a county level. A useful model shows that the mechanism M does, in fact, play a causal role in a specific study's STOUT. It also makes the case that the mechanism needs to operate similarly (based on similar causal principles) in the population STOUT.

Model Utility Component 3 The level of abstraction of the mechanism, which is the subject of the external validity inference, is well conceptualized and articulated.

Treatments are rarely identical across contexts but, as argued above, they often have a similar mechanism underpinning the causal principles in play. Theorizing about a useful level of abstraction, and designing to check that abstraction, is key to understanding how the mechanisms travel to other contexts (Sartori 1970, Garcia & Wantchekon 2010, Cartwright & Hardie 2012, Pearl & Mackenzie 2018). In particular, a useful model of the underlying mechanism makes clear what can be grouped for generalization or transportation and what cannot.

The search for a useful level of abstraction gives rise to a related question of scale (Banerjee et al. 2017a, Bold et al. 2018, Grossman et al. 2020). Many studies are constrained by design, meaning that they focus only on partial equilibrium,[19] whereas a theoretical mechanism and context under investigation could usefully be analyzed at a higher, general equilibrium level (Acemoglu 2010, Deaton 2010).[20] Studies that account for general equilibrium incorporate more of the M–STOUT space, which can be beneficial. A general equilibrium approach, however, runs the risk of stretching the model's mechanism(s) beyond its true domain and, thereby, producing inaccurate predictions (Sartori 1970, Deaton & Cartwright 2018).

4.2. Scope Plausibility

A second evaluative criterion is the plausibility of the scope for external validity inferences. Arguments and inferences are always bounded in their applicability, both by theory and by design (Walker & Cohen 1985, Clarke & Primo 2012, Neumayer & Plumper 2017). We identify four core components of scope plausibility.

Scope Plausibility Component 1 Both the theoretical and the accessible populations for all STOUT dimensions are identified and articulated.

Plausibly establishing scope begins with concretely defining, at the theoretical and design stages, all of the theoretical populations of STOUT that a mechanism informs. Of course, scholars can make clarifications ex post, but it is first necessary to specify a theoretical population. Then, only after a theoretical population is clear, researchers need to specify an accessible population (or sampling frame) from which to construct the sample.

Unfortunately, common practice is often backwards. Only after studying a sample from an accessible population (generalizability inferences), or a sampling with no predefined population (transportability inferences), do scholars then retrofit results from the accessible population back to a theoretical population. Moreover, existing scholarship often makes sample-specific inferences (Imbens 2010), and set-theoretic methods sometimes define the population inductively (Ragin 2000). If scholars have too much latitude to define the scope and population ex post, then they

[19] Partial equilibrium models aim to make inferences about the effects of a treatment on an outcome while holding fixed other macro- or meso-level factors (e.g., technology, institutions) (Acemoglu 2010).

[20] General equilibrium models, by contrast, begin with a higher-level theoretical mechanism and account for relevant macro- or meso-level factors (Acemoglu 2010).

may pick a population that confirms his/her theory. As with internal validity, scholars do not have a "get out of jail free card" on external validity. Plausible scope and population definitions need to occur ahead of time.

Scope Plausibility Component 2 Causal interaction between mechanisms (M) and all STOUT are articulated with specific reference to contextual dependencies or irrelevancies.

Scholarship mostly posits simple causal relationships, only occasionally exploring causal interactions between mechanisms (M) and other STOUT dimensions as an afterthought. Ideally, scholars need to determine whether causal interaction is theoretically plausible ahead of time and, if so, build such interaction into all relevant research design decisions (Falleti & Lynch 2009, Cartwright & Hardie 2012, Muller 2015). As Nobel laureate Angus Deaton explains, demonstrating why a treatment works entails a detailed examination of the context that supports the mechanism in play (Deaton 2010, p. 448). Studies thus need to be explicit about the STOUT context and articulate how the mechanism travels across that context, whether in generalizability or transportability inferences.

The narrowest view of causal interaction is that no single process across two different contexts is the same, but that view is extremely restrictive. If appropriately abstracted, it is possible to make external validity inferences about how a causal mechanism interacts across STOUT dimensions.[21] Notably, plausible scope depends on the appropriate abstraction of causal mechanisms (M) and the STOUT contexts across which M travel. In any event, scholars need to devote attention to the articulation of possible causal interaction and the ex post analytical investigation of such interaction. Not all context matters, however, and in some cases inferences are fairly constant or homogeneous, so it is necessary to explicitly report the irrelevancies (e.g., Berinsky et al. 2012, Mullinix et al. 2015, Coppock 2018). Given that causal interaction is theoretically critical, and yet not always shown to be empirically operative, scholars need to devote attention to establishing the threshold at which causal interaction matters. Moreover, the social sciences need to establish clear criteria for abstraction such that inferences relying on causal interaction can be appropriately generalized or transported.

Scope Plausibility Component 3 Samples for all STOUT dimensions are selected at random, as-if random, or stratified random (as useful). When samples are not selected as-if randomly, prespecified weighting and poststratification can improve representativeness.

Random sampling provides a powerful solution to external validity because it ensures representativeness of the sample on observable and unobservable dimensions of a population. Representativeness can be defined in various ways (Kruskal & Mosteller 1979), but here we use it to characterize a sample that unbiasedly represents a population. A well-known strategy for achieving representativeness is formal random sampling, but it is almost always discarded as practically impossible (Shadish et al. 2002, Goertz & Mahoney 2012). Recently, though, randomly sampling from defined populations, such as through population-based survey experiments (Mutz 2011), has become much more feasible. Advances in new and big data collection also allow for better definitions of the populations of STOUT, which can enable creative approaches to random sampling. For example, census records, polling stations, and Google Earth street maps have all been used (e.g., Findley et al. 2017a, Dunning et al. 2019).

[21] For a related, more technical discussion, see Cartwright's (1999) discussion of causal systems and nomological machines.

Because random sampling is such a powerful principle, it serves as a benchmark. Accordingly, we propose that studies without random sampling be evaluated based on the principle of as-if random sampling. Similar to the experiment/natural experiment distinction,[22] sometimes the researcher possesses the control over sampling. Other times, a researcher does not have the control but can find as-if random sampling in an observational setup. To the extent that a sample can be considered as-if random, the benefits of representative sampling provide substantial leverage. As a sample under study approaches an as-if randomly sampled state, the sample will be sufficiently similar to make inferences about the target population.

Theoretically, prespecified weighting offers a path toward plausible representativeness in the absence of random sampling (Olsen et al. 2013, Hartman et al. 2015, Kern et al. 2016, Franco et al. 2017, Nguyen et al. 2017, Buchanan et al. 2018, Miratrix et al. 2018). Weighting part(s) of observations over others generally makes most intuitive sense when some units in a sample are underrepresented relative to the population counterpart. Poststratification—or the analysis of unstratified data with weights and strata that mimic how the data would have been if they were collected through representative stratification (Gelman & Hill 2007, p. 181)—follows the principle of approximating random, representative sampling. To the extent that biases are minimized by either design, covariates, or priors, the sample becomes representative (Little & Pepinsky 2021). Given the concerns about p-hacking or fishing, weighting approaches must be guided by theory and design (Franco et al. 2017).

By the same token, the benefits of random (or as-if random) sampling can unravel in numerous ways. If causal interaction between the mechanisms M and STOUT exists, for example, then simple random sampling may not provide the needed representativeness. Even with a representative sample, choices made by a researcher—such as listwise deletion (Rubin 2004) and the exclusion of noncompliers from the analysis of experiments (Berinsky et al. 2014)—can compromise external validity as well. Moreover, pooling observations is not the panacea that many scholars believe it to be. In fact, adding irrelevant observations can decrease the representativeness of a sample. Even if properly guided by theory, pooling problems can compound as researchers use methods such a linear regression, which can often weight units in ways that undo any the representativeness gains from using complete TSCS data sets (Aronow & Samii 2016). To guard against these pitfalls, researchers must carefully evaluate and justify the choices they make in an analysis.

Scope Plausibility Component 4 Theoretically guided, nonrandom sample selection facilitates principled extrapolation.

Although seeking plausibly random samples is useful, there are nonetheless principled uses of nonrandom selection that facilitate learning about specific M–STOUT dimensions. Selection of treatments and outcomes cannot occur at random. Instead, it requires theory to establish the match between sample operationalization and the population construct for which certain inferences need to travel (Campbell 1986, Wells & Windschilt 1999). Unfortunately, the norm is to select operationalizations based on convenience, which neglects how the sample and population connect.

Due to the vast parameter space of all possible M–STOUT, nonrandom selection holding constant (or excluding) a single dimension may be required to make definitive inferences about one M–STOUT dimension relative to another (McFadden et al. 1977, Keane & Wolpin 2007, Wolpin 2007, Morton & Williams 2010). That is, when all dimensions are simultaneously varied, investigating dimension-specific contributions or constraints becomes much more difficult. Holding one or more dimensions constant, conversely, enables isolation of the effects of other

[22] As Dunning (2012) explains, a researcher does not control the randomization in a natural experiment, but the researcher does control the randomization in a field, laboratory, or survey experiment.

dimensions. The Metaketa Initiative, for example, pegged a common treatment arm and outcome while varying sampling units and, to some extent, settings.

In a similar vein, purposively choosing certain cases allows the researcher to show that if a result holds in a single case then it should hold in all other cases along that dimension. If a result holds in a least likely (influential) case, for example, then it should hold in a much larger set of cases. Purposive case selection could attempt to satisfy other criteria such as to identify typical or ideal cases, which may be representative in a sense and provide inferences about other cases (Kruskal & Mosteller 1979). Although the justification for case selection strategies has largely turned on questions of the identification of causal effects or mechanisms, their value in establishing external validity should not be overlooked.

Explicitly modeling self-selection (Gaines & Kuklinski 2011) or heterogeneous characteristics (Huff & Tingley 2015) are other critical steps toward extrapolation across M–STOUT. We caution, however, that the spate of recent studies suggesting similarities between nonrandom and random samples is confined largely to the United States and many of them rely on MTurk in WEIRD[23] societies (Henrich et al. 2010). At issue, self-selection processes in other contexts are still poorly understood. Even in those cases, however, well-specified theories, such as formal models, can help model the selection processes.

4.3. Specification Credibility

If scholars begin to take external validity more seriously, it will require credible specification of theory, design, and synthesis. We identify four critical components.

Specification Credibility Component 1 Making an external validity inference requires a theory and research design, which ensures that the external validity inferences are falsifiable.

In contrast to the current norms of drawing out post hoc similarities, scholars need to theorize and design for external validity and then evaluate it rigorously. A single test is an instantiation, and then a test from another sample or population is a different instantiation. Those other instances have a distribution, and the task for scholars is to appropriately characterize that distribution. For external validity inferences, the core question becomes: What does one mean by X causes Y in this instance—meaning in this instantiation of M across STOUT? Once one specifies a theory and design, external validity inferences become falsifiable. As discussed above, this task is not reducible to whether point estimates transfer from here to there. Instead, it requires careful theorizing about mechanisms and design for scope considerations. Once a theory and design are in place, falsifiable standards must then guide the study's conclusions with respect to external validity.

Specification Credibility Component 2 The assumptions and features of the external validity model are defensible.

Various approaches can be used to specify model utility and scope plausibility, and they range in the credibility of the underlying assumptions and the framework features. For instance, when a researcher uses a method relying on random sampling, the researcher needs to provide detailed explanations about the sampling procedures to defend the assumption. Similarly, when a researcher exploits as-if random sampling in an observational setup, they must provide explanations about the source of as-if randomness and compellingly defend the assumption. In other cases, when (as-if) random sampling is unattainable, a researcher may wish to use prespecified weighting or poststratification. These methods usually require the assumption that, conditional on the observed

[23] Western, educated, industrialized, rich, democratic (Henrich et al. 2010).

covariates, the sample selection is independent of treatment assignment.[24] Thus, external validity critically hinges on how plausibly the researcher can defend the conditional-on-observables assumptions.

Aside from those empirical approaches, researchers may use more theory-oriented approaches to ensure external validity. In these cases, it is even more crucial to defend the underlying assumptions. DAGs offer theoretical solutions to external validity (Bareinboim & Pearl 2013). DAG approaches explicitly model the transportability problem with emphasis on the core components that must transport and the factors that support the transportation. Although highly sophisticated, causal graphs rest on assumptions, most notably, that a well-defined causal model is specified (Aronow & Sävje 2020). The causal model therefore must be justified based on substantive knowledge. Similarly, structural models, which rely on modeling effect heterogeneity across contexts, as well as synthetic approaches including Bayesian model averaging or stacking, often depend on assumptions about distributions and priors (Yao et al. 2018, Dehejia et al. 2021, Dunning et al. 2019, Hollenbach & Montgomery 2020). Instead of uncritically accepting the common practices in the literature, researchers need to justify why the assumptions in their applications are plausible.

Finally, game theory, computational models, and structural inference provide ways to theorize about the processes of sample selections but, again, with assumptions. These theoretical models are precise about the mechanisms M and often the STOUT context. Well-designed structural models, in particular, are precise about the dynamics, functional form, and the extent to which parameters are separable (Low & Meghir 2017). However, most theoretical models rest also on assumptions about players, their interactions, utility functions, and information sets—though there are some exceptions (e.g., Fey & Ramsay 2011). Even though the assumptions need not be true or even plausible, when a researcher applies theoretical models, they must explain how the underlying assumptions match the case. Most often, researchers do not do so.

Of course, much social science scholarship employs nonformal theoretical frameworks, too. Although they are typically not as explicit about underlying assumptions, informal theoretical frameworks make assumptions nonetheless. Making an external validity inference on the basis of a most-similar case design, for example, carries with it assumptions about the definition of the case and the extent to which others are similar. In turn, regardless of whether the theoretical approach is formal or nonformal, scholars need to articulate relevant parameters and context as well as the assumptions on which the model depends. Following such a course of action allows for credible evaluation of the specification that produces an external validity inference.

Specification Credibility Component 3 The study's estimand preserves the integrity of the theoretical target population of interest.

Most social science research aims to recover the average treatment effect in a target population (PATE or TATE) through unbiased estimation of the SATE, but that is often neither possible nor desirable from the perspective of internal validity. Accordingly, scholars must consider other research designs that employ different estimands. However, using other estimands changes the composition of the M–STOUT of the inference, which means inferences about external validity must be tailored to the altered M–STOUT.

Both experiments and natural experiments are particularly likely to use estimands other than the SATE. A challenge for experiments, particularly field experiments, is that attrition,

[24]Covariates make possible other approaches, such as matching, trend similarity, latency equivalence, and mundane realism (see, for examples, Aronson & Carlsmith 1968; Aronson et al. 1994; Guala 2005, 2010; Abadie et al. 2010; Tipton et al. 2014; Angrist & Rokkanen 2015; van Eersel et al. 2019).

noncompliance, and spillover alter the sample in critical ways. In turn, they force scholars to estimate intent-to-treat effects instead of the SATE or to correct for problems by examining only compliers such as in complier average causal effect analysis (Gerber & Green 2012). Rarely, though, do scholars make the necessary qualifications for explaining the external validity of their experiments after attrition, noncompliance, or spillover.

Natural experiments suffer from similar challenges. Instrumental variable and regression discontinuity designs, for example, estimate different local average treatment effects (LATE),[25] and synthetic control models estimate the effect on a particular treated unit.[26] None of these estimands translate neatly to the SATE, and not much at all to the PATE or TATE (Deaton 2010, Heckman & Urzúa 2010). However, scholars can employ various techniques to increase the external validity of natural experiments (Imbens 2010, Angrist & Rokkanen 2015, Bisbee et al. 2017, Wing & Bello-Gomez 2018, Bertanha & Imbens 2020). Although these techniques usually require stronger assumptions than those needed for causal inference, as far as the assumptions are substantively defended, the methods can enhance the external validity.

Finally, observational research is not necessarily superior to experimental research with respect to estimating the PATE and TATE, though it can sometimes produce better lessons than transporting experimental results or synthesizing through meta-analysis (Pritchett & Sandefur 2015). Although covering all observational methods is outside the scope of this review, for the case of a linear regression, Aronow & Samii (2016) show that linear regression often uses problematic weighting and thus biases estimates. The authors also provide a way to reweight and hence recover the unbiased estimates.

Specification Credibility Component 4 Theoretically guided research synthesis substantiates the external validity of research programs.

Although repeated measurement and analysis can be useful, simply amassing more and more data is unlikely to solve external validity. Notably, many variables of interest elude valid measurement, and much of what is measurable is not germane. Moreover, research emerges in a decentralized, nonrandom fashion, so even meta-analytic studies that include a full universe of studies on a given phenomenon may still be incomplete if they do not capture sufficient heterogeneity (Allcott 2015). Useful meta-analyses—including Dunning et al. (2019) on information and political accountability and Banerjee et al. (2015) on livelihoods—have common treatment arms, operate on the same lower-level unit (individuals), and employ settings that are comparable (field experiments). When comparing the core characteristics of countries in the syntheses relative to other countries, and subnational sites relative to all possible subnational sites, however, the studies exhibit extremely little variation on multiple dimensions. Ideally, studies need to encompass some random selection and substantial coverage of the M–STOUT parameter space—with theory guiding the type of variation on M–STOUT to explore.

5. REPORTING

Several decades into the credibility revolution, scholars have a good sense of how to gauge a study's internal validity. Indeed, nearly all studies report on it explicitly. However, outside of statistical

[25] The canonical regression discontinuity design using the continuity-based framework estimates a LATE at the cutoff (Sekhon & Titiunik 2017). Instrumental variable models only estimate a LATE for compliers. In an experiment, a complier is a unit that, if assigned to treatment (control), takes the treatment (control). In an observational study, a complier is a unit for which outcomes are shifted in the theoretically hypothesized direction (Imbens & Angrist 1994, Angrist et al. 1996).

[26] Similarly, matching often "prunes" observations to obtain better matches, resulting in different M-STOUT dimensions (Ho et al. 2007) and, in turn, external validity inferences.

sampling concerns, scholars rarely make more than superficial attempts to report on external validity accurately—if they report on it at all. It is curious that reporting standards in political science only cover external validity in cursory form (Gerber et al. 2014), especially given that other social science fields take it far more seriously in both quantitative (Appelbaum et al. 2018) and qualitative research (Levitt et al. 2018).

Our practical ambition is that every published social science study include a dedicated discussion of external validity. For it to be taken seriously, authors need to report on external validity, and reviewers and editors need to insist on reporting as a matter of course. Along these lines, we fully agree with (Rodrik 2009, p. 39) who, in discussing randomized experiments, said: "It is incumbent on the authors to convince the reader that the results are reasonably general and also to address circumstances under which they may not be. This is as important as justifying causal identification in other types of empirical work."

When transparently reporting on external validity, the three evaluative criteria provide a guide for what authors should discuss and reviewers/editors should evaluate. Even when studies do not maximize on these criteria, it is still incumbent upon scholars to characterize levels of external validity accurately. The most credible studies from the perspective of external validity not only report on the S or the U but also characterize all M–STOUT dimensions; clearly define the theoretical populations, accessible populations, and samples; and, if applicable, are explicit on the extent to which the inferences from the sample are meant to generalize or transport (or both). If all authors follow such guidance as a matter of course, then the social sciences can collectively build more generalized knowledge by gradually moving scientific priors (Rodrik 2009).

6. CONCLUSION

According to existing epistemological and methodological standards, external validity is fundamental, not incidental, in social science (e.g., King et al. 1994, Shadish et al. 2002, Gerring 2011).[27] Despite McDermott's (2011) claim to the contrary, internal validity has been much more prominent than external validity in recent scientific inquiry. In this article, we organized insights from political science and other social science fields to articulate emerging themes, which we also formalized to engage directly with rigorous internal validity approaches. In so doing, we have attempted to show that the fixation on internal validity is not only misplaced but also out of step with many recent arguments across the social sciences.

Both science and public policy demand greater attention to external validity. Scholars appear committed to the ideal of producing general knowledge, but they have not yet implemented more rigorous external validity standards in their day-to-day research. For all the work of a number of organizations, such as the Empirical Implications of Theoretical Models, EGAP, the Open Science Foundation, the Berkeley Institute for Transparency in the Social Sciences, Empirical Studies of Conflict, and the Abdul Latif Jameel Poverty Action Lab (J-PAL), among others, such organizations have given very little systematic attention to external validity concerns. One reaction may be that external validity is a noble but unattainable goal. We have attempted to provide evidence to the contrary.

Applied researchers also appear to have a false sense of security about the level of external validity in their studies. In political science, for example, many observational researchers studying

[27] Of course, the production of general knowledge is obviously a broad endeavor with many components (see, for example, Elman et al. 2019).

large samples of countries or states seem content that they are studying the "real world" and therefore must necessarily have external validity (McDermott 2011). With the advent of big data, the problem may be compounded. As Nagler & Tucker (2015, p. 85) argue, "With big data comes the illusion of big precision." This hope currently rests on a precarious foundation of brittle assumptions about causal processes, sampling, and representativeness. But it does not have to be this way; indeed, much can be done.

In our analysis of external validity, we have built upon the previous literature to provide an improved conceptualization and relevant criteria that scholars can use to evaluate the external validity of studies. The predominant current practices of neglect or merely thinking about external validity as an afterthought must change. The "ultimate goal of all good social science" is to make inferences beyond the data at hand (King et al. 1994, pp. 8, 34), and that will only become feasible as authors, reviewers, and editors heed our call for a dedicated discussion of external validity in every research product. In the process, political scientists will become more systematic in how they assess the extent to which results generalize to broader populations or transport to other populations—and, in turn, how knowledge accumulates.

7. APPENDIX: FORMALIZATION OF EXTERNAL VALIDITY

7.1. Internal Validity Bias

Consider that one would like to estimate the causal effect of a treatment D on an outcome variable Y with a sample $S = \{0, 1, \ldots, n\}$. By convention, assume that the treatment variable is dichotomous $D \in \{0, 1\}$. The sample average treatment effect (SATE) is $\delta_S = E_S[Y_j(1) - Y_j(0)]$ for $j \in S$, which captures the difference in potential outcomes due to treatment status. The corresponding difference-in-means estimator is $\hat{\delta}_S = E_S[Y_j(1)|D_j = 1] - E_S[Y_j(0)|D_j = 0]$ for $j \in S$, which captures the observable difference in the sample averages. **Table 2** summarizes the notation.

Table 2 Mathematical notation

Term	Definition
Y	Outcome of interest
$\tilde{Y}$	Outcome at hand
D	Treatment of interest
$\tilde{D}$	Treatment at hand
W	Indicator of sample selection
S	Sample
P	Population
i	Unit in a population
j	Unit in a sample
$\hat{\delta}_S$	Difference-in-mean estimate
δ_S	Sample average treatment effect
δ_P	Population average treatment effect of D on Y
$\tilde{\delta}_P$	Population average treatment effect of $\tilde{D}$ on $\tilde{Y}$
b_{S1}	Assignment selection bias
b_{S2}	Within-sample effect heterogeneity
b_P	Sample selection bias
b_V	Variable selection bias

As is commonly known, the difference-in-means estimator is decomposed to the SATE with a bias term:

$$\hat{\delta}_S = \delta_S + b_{S1} + b_{S2}. \tag{2.}$$

That is, the empirical estimator is the SATE plus the assignment selection bias, $b_{S1} = E_S[Y_j(0)|D_j = 1] - E_S[Y_j(0)|D_j = 0]$, and the within-sample effect heterogeneity, $b_{S2} = \Pr(D_j = 0)\{E_S[Y_j(1) - Y_j(0)|D_j = 1] - E_S[Y_j(1) - Y_j(0)|D_j = 0]\}$. The selection bias is the difference in the potential outcomes between the treated and control groups, whereas the within-sample effect heterogeneity is the difference in the causal effect between those two groups. Experimental design and random assignment ensure that the treated and control units are equivalent in expectation and, therefore, that these two bias terms are zero. The absence of biases allows us to obtain internally valid estimates of the SATE.

7.2. External Validity Bias

Although the SATE itself can be of interest in some cases (Wilke & Humphreys 2020), we usually would like to estimate the causal effect of a target population P. The population average treatment effect (PATE) is defined as $\delta_P = E_P[Y_i(1) - Y_i(0)]$ for $i \in P$. However, the PATE can be systematically different from the SATE. Let $W_i \in \{0, 1\}$ be a sample selection variable that takes 1 if a unit i is selected into a sample S and otherwise 0 (that is, $W_i = 1$ if $i \in S$ for $i \in P$, and otherwise $W_i = 0$). Then, the SATE is equal to the PATE plus a sample selection bias:

$$\delta_S = \delta_P + b_P. \tag{3.}$$

By definition, $\mathrm{E}_S[Y_j(1) - Y_j(0)] = \mathrm{E}_P[Y_i(1) - Y_i(0) \mid W_i = 1]$. Note that by the property of expectation,

$$\begin{aligned} &\mathrm{E}_P[Y_i(1) - Y_i(0)] \\ &\quad = \mathrm{E}_P[Y_i(1) - Y_i(0) \mid W_i = 1]\, P(W_i = 1) + \mathrm{E}_P[Y_i(1) - Y_i(0) \mid W_i = 0]\, \Pr(W_i = 0). \end{aligned} \tag{4.}$$

This is equivalent to

$$\begin{aligned} &\mathrm{E}_P[Y_i(1) - Y_i(0) \mid W_i = 1] \\ &\quad = \frac{1}{\Pr(W_i = 1)}\{\mathrm{E}_P[Y_i(1) - Y_i(0)] - \mathrm{E}_P[Y_i(1) - Y_i(0) \mid W_i = 0]\, \Pr(W_i = 0)\} \\ &\quad = \mathrm{E}_P[Y_i(1) - Y_i(0)] + \frac{\Pr(W_i = 0)}{\Pr(W_i = 1)}\{\mathrm{E}_P[Y_i(1) - Y_i(0)] - \mathrm{E}_P[Y_i(1) - Y_i(0) \mid W_i = 0]\}. \end{aligned} \tag{5.}$$

Note that

$$\begin{aligned} &\mathrm{E}_P[Y_i(1) - Y_i(0)] - \mathrm{E}_P[Y_i(1) - Y_i(0) \mid W_i = 0] \\ &\quad = \mathrm{E}_P[Y_i(1) - Y_i(0) \mid W_i = 1]\, P(W_i = 1) + \mathrm{E}_P[Y_i(1) - Y_i(0) \mid W_i = 0]\, \Pr(W_i = 0) \\ &\qquad - \mathrm{E}_P[Y_i(1) - Y_i(0) \mid W_i = 0] \\ &\quad = \Pr(W_i = 1)\{\mathrm{E}_P[Y_i(1) - Y_i(0) \mid W_i = 1] - \mathrm{E}_P[Y_i(1) - Y_i(0) \mid W_i = 0]\}. \end{aligned} \tag{6.}$$

By inserting this and simplifying it,

$$\begin{aligned} &\mathrm{E}_P[Y_i(1) - Y_i(0) \mid W_i = 1] \\ &\quad = \mathrm{E}_P[Y_i(1) - Y_i(0)] + \Pr(W_i = 0)\{\mathrm{E}_P[Y_i(1) - Y_i(0) \mid W_i = 1] - \mathrm{E}_P[Y_i(1) - Y_i(0) \mid W_i = 0]\} \\ &\quad = \delta_P + \Pr(W_i = 0)\{\mathrm{E}_P[Y_i(1) - Y_i(0) \mid W_i = 1] - \mathrm{E}_P[Y_i(1) - Y_i(0) \mid W_i = 0]\}. \end{aligned} \tag{7.}$$

The sample selection bias, $b_P = \Pr(W_i = 0)\{E_P[Y_i(1) - Y_i(0) \mid W_i = 1] - E_P[Y_i(1) - Y_i(0) \mid W_i = 0]\}$, is the difference in the average causal effects between those included in the

sample and those excluded from the sample, weighted by the the proportion of missing units. In general, the larger the difference is, and the more missing units exist, the larger the bias is. The bias becomes zero either (*a*) when all units in the population are included in the sample or (*b*) when the average causal effects are the same between the included and excluded units. In the case of generalizability, the first condition would not be satisfied unless one is interested in the SATE itself. The second condition usually requires random sampling, even though random sampling can be something difficult to implement in practice. The problems become even more acute in the case of transportability, since the target population is different from the population that a sample represents. Thus, a sample at hand is unlikely to be a full or random sample of a target population.

The above discussion so far focuses on the cases of generalizing or transporting estimates with certain units to a broader or different population. In other cases, however, one may be interested in transporting estimates with certain treatment or outcome variables to different versions of those variables. This raises questions about variable external validity. Consider that an outcome Y and treatment D of interest are different from the outcome and treatment at hand, $\tilde{Y}$ and $\tilde{D}$. Then the PATE at hand becomes $\tilde{\delta}_P = E_P[\tilde{Y}_i(\tilde{D}_i = 1) - \tilde{Y}_i(\tilde{D}_i = 0)]$ for $i \in P$. In presence of variable heterogeneities, this PATE at hand can be systematically different from the PATE of interest δ_P:

$$\tilde{\delta}_P = \delta_P + b_V. \qquad 8.$$

Variable selection bias, $b_V = E_P[\tilde{Y}_i(\tilde{D}_i = 1) - \tilde{Y}_i(\tilde{D}_i = 0)] - E_P[Y_i(D_i = 1) - Y_i(D_i = 0)]$, represents the heterogeneity in the different outcomes responding to different treatments.[28]

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

For helpful research assistance, we thank Megan Farrell, Klara Fredriksson, Kai Yue (Theodore) Charm, Juan Lozano, Amila Lulo, Xin Nong, Ayu Sofyan, Aditya Tantravahi, and Alex Wais as well as Vanessa Gonzalez, Jonah Isaac, Samiya Javed, Sara Lowe, and Jenny Rodriguez. For helpful feedback, we thank Bethany Albertson, Trey Billing, Susanna Campbell, Jiseon Chang, Paul Diehl, Colin Elman, Chris Fariss, John Gerring, Gary Goertz, Ben Graham, Guy Grossman, Alan Jacobs, Diana Kapiszewski, Jim Kuklinski, Jenn Larson, James Mahoney, Tetsuya Matsubayashi, Rich Nielsen, Dan Nielson, Xin Nong, Raul Pacheco-Vega, Jan Pierskalla, Raul Roman, Cyrus Samii, Jason Seawright, Renard Sexton, Jason Sharman, James Stevenson, Dustin Tingley, Stella Wancke, Nick Weller, Steven Wilson, Rebecca Wolfe, and Cornelia Woll. We also benefited immensely from the 2019 Institute for Qualitative and Multi-Method Research authors' workshop in Syracuse, New York.

LITERATURE CITED

Abadie A, Diamond A, Hainmueller J. 2010. Synthetic control methods for comparative case studies: estimating the effect of California's tobacco control program. *J. Am. Stat. Assoc.* 105:493–505

[28]Decomposing the variable selection bias to those due to the heterogeneities in outcome and treatment is not immediately clear.

Acemoglu D. 2010. Theory, general equilibrium, and political economy in development economics. *J. Econ. Perspect.* 24:17–32

Allcott H. 2015. Site selection bias in program evaluation. *Q. J. Econ.* 130:1117–65

Andrews I, Oster E. 2019. A simple approximation for evaluating external validity bias. *Econ. Lett.* 178:58–62

Angrist JD, Imbens GW, Rubin DB. 1996. Identification of causal effects using instrumental variables. *J. Am. Stat. Assoc.* 91:456–58

Angrist JD, Pischke JS. 2008. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton, NJ: Princeton Univ. Press

Angrist JD, Pischke JS. 2010. The credibility revolution in empirical economics: how better research design is taking the con out of econometrics. *J. Econ. Perspect.* 24:3–30

Angrist JD, Rokkanen M. 2015. Wanna get away? Regression discontinuity estimation of exam school effects away from the cutoff. *J. Am. Stat. Assoc.* 110:1331–44

Appelbaum M, Cooper H, Kline RB, Mayo-Wilson E, Nezu AM, Rao SM. 2018. Journal article reporting standards for quantitative research in psychology: the APA Publications and Communications Board Task Force report. *Am. Psychol.* 73:3–25

Aronow PM, Samii C. 2016. Does regression produce representative estimates of causal effects? *Am. J. Political Sci.* 60:250–67

Aronow PM, Sävje F. 2020. The book of why: the new science of cause and effect. *J. Am. Stat. Assoc.* 1459:482–85

Aronson E, Carlsmith JM. 1968. Experimentation in social psychology. *Handb. Soc. Psychol.* 2:1–79

Aronson E, Wilson TD, Akert RM. 1994. *Social Psychology: The Heart and the Mind.* New York: Harper Collins

Ashworth S, Berry CR, Mesquita EBD. 2014. All else equal in theory and data (big or small). *PS: Political Sci. Politics* 48:89–94

Banerjee A, Banerji R, Berry J, Duflo E, Kannan H, et al. 2017a. From proof of concept to scalable policies: challenges and solutions, with an application. *J. Econ. Perspect.* 31:73–102

Banerjee A, Chassang S, Snowberg E. 2017b. Decision theoretic approaches to experiment design and external validity. In *Handbook of Economic Field Experiments*, Vol. 2, ed. Duflo E, Banerjee A, pp. 141–74. Oxford, UK: Elsevier

Banerjee A, Duflo E, Goldberg N, Karlan D, Osei R, et al. 2015. A multifaceted program causes lasting progress for the very poor: evidence from six countries. *Science* 348:1260799

Bareinboim E, Pearl J. 2013. A general algorithm for deciding transportability of experimental results. *J. Causal Inference* 1:107–134

Bareinboim E, Pearl J. 2016. Causal inference and the data-fusion problem. *PNAS* 113:7345–52

Bates MA, Glennerster R. 2017. The generalizability puzzle. *Stanford Soc. Innov. Rev.* 201:50–54

Berinsky AJ, Huber GA, Lenz GS. 2012. Evaluating online labor markets for experimental research: Amazon.com's Mechanical Turk. *Political Anal.* 20:351–68

Berinsky AJ, Margolis MF, Sances MW. 2014. Separating the shirkers from the workers? Making sure respondents pay attention on self-administered surveys. *Am. J. Political Sci.* 58:739–53

Bertanha M, Imbens GW. 2020. External validity in fuzzy regression discontinuity designs. *J. Bus. Econ. Stat.* 38:593–612

Biglaiser G, Staats JL. 2010. Do political institutions affect foreign direct investment? A survey of U.S. corporations in Latin America. *Political Res. Q.* 63:508–22

Bisbee J, Dehejia R, Pop-Eleches C, Samii C. 2017. Local instruments, global extrapolation: external validity of the labor supply–fertility local average treatment effect. *J. Labor Econ.* 35:S99–147

Bold T, Kimenyi M, Mwabu G, Ng A, Sandefur J. 2018. Experimental evidence on scaling up education reforms in Kenya. *J. Public Econ.* 168:1–20

Breskin A, Westreich D, Cole SR, Edwards JK. 2019. Using bounds to compare the strength of exchangeability assumptions for internal and external validity. *Am. J. Epidemiol.* 188:1355–60

Brunswik E. 1947. Systematic and representative design of psychological experiments. In *Proceedings of the Berkeley Symposium on Mathematical Statistics and Probability*, pp. 143–202. Berkeley: Univ. Calif. Press

Buchanan AL, Hudgens MG, Cole SR, Mollan KR, Sax PE, et al. 2018. Generalizing evidence from randomized trials using inverse probability of sampling weights. *J. R. Stat. Soc. Ser. A: Stat. Soc.* 181:1193–209

Camerer C. 2015. The promise and success of lab–field generalizability in experimental economics: a critical reply to Levitt and List. In *Handbook of Experimental Economic Methodology*, ed. GR Fréchette, A Schotter, pp. 249–95. Oxford, UK: Oxford Univ. Press

Campbell DT. 1957. Factors relevant to the validity of experiments in social settings. *Psychol. Bull.* 54:297–312

Campbell DT. 1986. Relabeling internal and external validity for applied social scientists. In *Advances in Quasi-Experimental Design and Analysis*, ed. WMK Trochim, pp. 67–77. San Francisco: Jossey-Bass

Carroll L. 1865. *Alice's Adventures in Wonderland*. London: Macmillan

Cartwright N. 1999. *The Dappled World: A Study of the Boundaries of Science*. Cambridge, UK: Cambridge Univ. Press

Cartwright N. 2020. Middle-range theory: Without it what could anyone do? *Theoria* 35:269–323

Cartwright N, Hardie J. 2012. *Evidence-Based Policy: A Practical Guide to Doing It Better*. Oxford, UK: Oxford Univ. Press

Clarke KA, Primo DM. 2012. *A Model Discipline: Political Science and the Logic of Representations*. Oxford, UK: Oxford Univ. Press

Cole SR, Stuart EA. 2010. Generalizing evidence from randomized clinical trials to target populations: the ACTG 320 Trial. *Am. J. Epidemiol.* 172:107–15

Cook TD, Campbell DT, eds. 1979. *Quasi-Experimentation: Design and Analysis for Field Settings*, Vol. 3. Chicago: Rand McNally

Coppock A. 2018. Generalizing from survey experiments conducted on Mechanical Turk: a replication approach. *Political Sci. Res. Methods* 7:613–28

Coppock A, Green DP. 2015. Assessing the correspondence between experimental results obtained in the lab and field: a review of recent social science research. *Political Sci. Res. Methods* 3:113–31

Coppock A, Hill SJ, Vavreck L. 2020. The small effects of political advertising are small regardless of context, message, sender, or receiver: evidence from 59 real-time randomized experiments. *Sci. Adv.* 6:eabc4046

Coppock A, Leeper TJ, Mullinix KJ. 2018. The generalizability of heterogeneous treatment effect estimates across samples. *PNAS* 115:12441–46

Cronbach LJ, Shapiro K. 1982. *Designing Evaluations of Educational and Social Programs*. San Francisco: Jossey-Bass

Dahabreh IJ, Hayward R, Kent DM. 2016. Using group data to treat individuals: understanding heterogeneous treatment effects in the age of precision medicine and patient-centred evidence. *Int. J. Epidemiol.* 45:2184–93

Deaton A. 2010. Instruments, randomization, and learning about development. *J. Econ. Lit.* 48:424–55

Deaton A. 2019. *Randomization in the tropics revisited: a theme and eleven variations*. NBER Work. Pap. 27600

Deaton A, Cartwright N. 2018. Understanding and misunderstanding randomized controlled trials. *Soc. Sci. Med.* 210:2–21

Dehejia R, Pop-Eleches C, Samii C. 2021. From local to global: external validity in a fertility natural experiment. *J. Bus. Econ. Stat.* 39:217–43

Druckman JN, Green DP, Kuklinski JH, Lupia A. 2006. The growth and development of experimental research in political science. *Am. Political Sci. Rev.* 100:627–35

Druckman JN, Kam CD. 2011. Students as experimental participants. In *Cambridge Handbook of Experimental Political Science*, ed. JN Druckman, DP Greene, JH Kuklinski, A Lupia, pp. 41–57. New York: Cambridge Univ. Press

Dunning T. 2012. *Natural Experiments in the Social Sciences: A Design-Based Approach*. Cambridge, UK: Cambridge Univ. Press

Dunning T, Grossman G, Humphreys M, Hyde S, McIntosh C, Nellis G, eds. 2019. *Information, Accountability, and Cumulative Learning: Lessons from Metaketa I.* Cambridge, UK: Cambridge Univ. Press

Elman C, Gerring J, Mahoney J. 2019. *The Production of Knowledge: Enhancing Progress in Social Science*. Cambridge, UK: Cambridge Univ. Press

Falk A, Heckman JJ. 2009. Lab experiments are a major source of knowledge in the social sciences. *Science* 326:535–38

Falleti TG, Lynch JF. 2009. Context and causal analysis. *Comp. Political Stud.* 42:1143–66

Fey M, Ramsay KW. 2011. Uncertainty and incentives in crisis bargaining: Game-free analysis of international conflict. *Am. J. Political Sci.* 55:149–69
Findley MG, Denly M, Kikuta K. 2022. *External Validity in the Social Sciences: An Integrated Approach*. Cambridge, UK: Cambridge Univ. Press. Manuscript under contract
Findley MG, Harris AS, Milner HV, Nielson DL. 2017a. Who controls foreign aid? Elite versus public perceptions of donor influence in aid-dependent Uganda. *Int. Organ.* 71:633–63
Findley MG, Laney B, Nielson DL, Sharman JC. 2017b. External validity in parallel global field and survey experiments on anonymous incorporation. *J. Politics* 79:856–72
Franco A, Malhotra N, Simonovits G, Zigerell LJ. 2017. Developing standards for post-hoc weighting in population-based survey experiments. *J. Exp. Political Sci.* 4:161–72
Gaines BJ, Kuklinski JH. 2011. Experimental estimation of heterogeneous treatment effects related to self-selection. *Am. J. Political Sci.* 55:724–36
Garcia FM, Wantchekon L. 2010. Theory, external validity, and experimental inference: some conjectures. *Ann. Am. Acad. Political Soc. Sci.* 628:132–47
Gelman A, Hill J. 2007. *Data Analysis Using Regression and Multilevel/Hierarchical Modelling*. Cambridge, UK: Cambridge Univ. Press
Gerber A, Arceneaux K, Boudreau C, Dowling C, Hillygus S, et al. 2014. Reporting guidelines for experimental research: a report from the experimental research section standards committee. *J. Exp. Political Sci.* 1:81–98
Gerber AS, Green DP. 2012. *Field Experiments: Design, Analysis, and Interpretation*. New York: WW Norton
Gerring J. 2008. The mechanismic worldview: thinking inside the box. *Br. J. Political Sci.* 38:161–79
Gerring J. 2011. *Social Science Methodology: A Unified Framework*. Cambridge, UK: Cambridge Univ. Press
Gisselquist RM. 2020. How the cases you choose affect the answers you get, revisited. *World Dev*. 127:104800
Goertz G. 2017. *Multimethod Research, Causal Mechanisms, and Case Studies: An Integrated Approach*. Princeton, NJ: Princeton Univ. Press
Goertz G, Mahoney J. 2012. *A Tale of Two Cultures: Qualitative and Quantitative Research in the Social Sciences*. Princeton, NJ: Princeton Univ. Press
Grossman G, Humphreys M, Sacramone-Lutz G. 2020. Information technology and political engagement: mixed evidence from Uganda. *J. Politics* 82:1321–36
Grzymala-Busse A. 2011. Time will tell? Temporality and the analysis of causal mechanisms and processes. *Comp. Political Stud.* 44:1267–97
Guala F. 2005. *The Methodology of Experimental Economics*. Cambridge, UK: Cambridge Univ. Press
Guala F. 2010. Extrapolation, analogy, and comparative process tracing. *Philos. Sci.* 77:1070–82
Guardado J, Wantchékon L. 2018. Do electoral handouts affect voting behavior? *Electoral Stud*. 53:139–49
Hartman E, Grieve R, Ramsahai R, Sekhon J. 2015. From sample average treatment effect to population average treatment effect on the treated: combining experimental with observational studies to estimate population treatment effects. *J. R. Stat. Soc. Ser. A: Stat. Soc.* 178:757–78
Heckman JJ, Urzúa S. 2010. Comparing IV with structural models: what simple IV can and cannot identify. *J. Econ.* 156:27–37
Heckman JJ, Vytlacil E. 2001. Policy-relevant treatment effects. *Am. Econ. Rev. Pap. Proc.* 91:107–11
Heckman JJ, Vytlacil EJ. 2005. Structural equations, treatment effects, and econometric policy evaluation. *Econometrica* 73:669–738
Henrich J, Heine SJ, Norenzayan A. 2010. The weirdest people in the world? *Behav. Brain Sci.* 33:61–83
Ho DE, Imai K, King G, Stuart EA. 2007. Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Anal.* 15:199–236
Holland PW. 1986. Statistics and causal inference. *J. Am. Stat. Assoc.* 81:945–60
Hollenbach FM, Montgomery JM. 2020. Bayesian model selection, model comparison, and model averaging. In *SAGE Handbook of Research Methods in Political Science and International Relations*, ed. L Curini, RJ Franzese, pp. 937–60. London: SAGE
Hotz VJ, Imbens GW, Mortimer JH. 2005. Predicting the efficacy of future training programs using past experiences at other locations. *J. Econ.* 125:241–70
Huff C, Tingley D. 2015. "Who are these people?" Evaluating the demographic characteristics and political preferences of MTurk survey respondents. *Res. Politics* 2:1–12

Imai K, Keele LJ, Tingley D, Yamamoto T. 2011. Unpacking the black box of causality: learning about causal mechanisms from experimental and observational studies. *Am. Political Sci. Rev.* 105:765–89
Imai K, King G, Stuart EA. 2008. Misunderstandings between experimentalists and observationalists about causal inference. *J. R. Stat. Soc. Ser. A: Stat. Soc.* 171:481–502
Imai K, Ratkovic M. 2013. Estimating treatment effect heterogeneity in randomized program evaluation. *Ann. Appl. Stat.* 7:443–70
Imbens GW. 2010. Better LATE than nothing: some comments on Deaton (2009) and Heckman and Urzua (2009). *J. Econ. Lit.* 48:399–423
Imbens GW, Angrist JD. 1994. Identification and estimation of local average treatment effects. *Econometrica* 62:467–75
Imbens GW, Rubin DB. 2015. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. New York: Cambridge Univ. Press
Keane MP, Wolpin KI. 2007. Exploring the usefulness of a nonrandom holdout sample for model validation: welfare effects on female behavior. *Int. Econ. Rev.* 48:1351–78
Kern HL, Stuart EA, Hill J, Green DP. 2016. Assessing methods for generalizing experimental impact estimates to target populations. *J. Res. Educ. Eff.* 9:103–27
Kessler J, Lise V. 2015. The external validity of experiments: the misleading emphasis on quantitative effects. In *Handbook of Experimental Economic Methodology*, ed. GR Fréchette, A Schotter, pp. 391–406. Oxford, UK: Oxford Univ. Press
King G, Keohane RO, Verba S. 1994. *Designing Social Inquiry: Scientific Inference in Qualitative Research*. Princeton, NJ: Princeton Univ. Press
Klašnja M, Titiunik R. 2017. The incumbency curse: weak parties, term limits, and unfulfilled accountability. *Am. Political Sci. Rev.* 111:129–48
Kruskal W, Mosteller F. 1979. Representative sampling, III: the current statistical literature. *Int. Stat. Rev./Rev. Int. Stat.* 47:245–65
Leamer EE. 2010. Tantalus on the road to asymptopia. *J. Econ. Perspect.* 24:31–46
Lesko CR, Buchanan AL, Westreich D, Edwards JK, Hudgens MG, Cole SR. 2017. Generalizing study results. *Epidemiology* 28:553–61
Levitt HM, Creswell JW, Josselson R, Bamberg M, Frost DM, Suarez-Orozco C. 2018. Journal article reporting standards for qualitative research in psychology: the APA Publications and Communications Board task force report. *Am. Psychol.* 73:26–46
Lieberson S. 1985. *Making It Count: The Improvement of Social Research and Theory*. Berkeley: Univ. Calif. Press
Little A, Pepinsky TB. 2021. Learning from biased research designs. *J. Politics*. In press. **https://www.journals.uchicago.edu/doi/10.1086/710088**
Low H, Meghir C. 2017. The use of structural models in econometrics. *J. Econ. Perspect.* 31:33–58
Lucas JW. 2003. Theory-testing, generalization, and the problem of external validity. *Sociol. Theory* 21:236–53
Mackie JL. 1965. Causes and conditions. *Am. Philos. Q.* 2:245–64
Marcellesi A. 2015. External validity: Is there still a problem? *Philos. Sci.* 82:1308–17
McDermott R. 2011. Internal and external validity. In *Cambridge Handbook of Experimental Political Science*, ed. JN Druckman, DP Green DP, JH Kuklinski, A Lupia, pp. 27–40. New York: Cambridge Univ. Press
McFadden D, Talvitie AP, and associates. 1977. *Demand model estimation and validation. Urban travel demand forecasting project: phase 1 final report series, Vol. V*. Rep. UCB-ITS-SR-77-9, Inst. Transport. Stud., Univ. Calif., Berkeley and Irvine
McIntyre L. 2019. *The Scientific Attitude: Defending Science from Denial, Fraud, and Pseudoscience*. Cambridge, MA: MIT Press
Miratrix LW, Sekhon JS, Theodoridis AG, Campos LF. 2018. Worth weighting? How to think about and use weights in survey experiments. *Political Anal.* 26:275–91
Morton RB, Williams KC. 2010. *Experimental Political Science and the Study of Causality: From Nature to the Lab*. Cambridge, UK: Cambridge Univ. Press
Muller SM. 2015. Causal interaction and external validity: obstacles to the policy relevance of randomized evaluations. *World Bank Econ. Rev.* 29:S217–25
Mullinix KJ, Leeper TJ, Druckman JN, Freese J. 2015. The generalizability of survey experiments. *J. Exp. Political Sci.* 2:109–38

Muralidharan K, Niehaus P. 2017. Experimentation at scale. *J. Econ. Perspect.* 31:103–24
Mutz DC. 2011. *Population-Based Survey Experiments*. Princeton, NJ: Princeton Univ. Press
Nagler J, Tucker JA. 2015. Drawing inferences and testing theories with big data. *PS: Political Sci. Politics* 48:84–88
Neumayer E, Plumper T. 2017. *Robustness Tests for Quantitative Research*. Cambridge, UK: Cambridge Univ. Press
Nguyen TQ, Ebnesajjad C, Cole SR, Stuart EA. 2017. Sensitivity analysis for an unobserved moderator in RCT-to-target-population generalization of treatment effects. *Ann. Appl. Stat.* 11:225–47
Olsen R, Bell S, Orr L, Stuart EA. 2013. External validity in policy evaluations that choose sites purposively. *J. Policy Anal. Manag.* 32:107–21
Pearl J. 2009. *Causality: Models, Reasoning, and Inference*. Cambridge, UK: Cambridge Univ. Press
Pearl J, Bareinboim E. 2014. External validity: from do-calculus to transportability across populations. *Stat. Sci.* 29:579–95
Pearl J, Bareinboim E. 2019. Note on "generalizability of study results." *Epidemiology* 30:186–88
Pearl J, Mackenzie D. 2018. *The Book of Why: The New Science of Cause and Effect*. New York: Basic Books
Pierson P. 2000. Increasing returns, path dependence, and the study of politics. *Am. Political Sci. Rev.* 94:251–67
Pritchett L, Sandefur J. 2013. *Context matters for size: why external validity claims and development practice don't mix*. Work. Pap., Cent. Glob. Dev., Washington, DC
Pritchett L, Sandefur J. 2015. Learning from experiments when context matters. *Am. Econ. Rev.* 105:471–75
Ragin CC. 2000. *Fuzzy-Set Social Science*. Chicago: Univ. Chicago Press
Ravallion M. 2012. Fighting poverty one experiment at a time: *Poor Economics: A Radical Rethinking of the Way to Fight Global Poverty*: review essay. *J. Econ. Lit.* 50:103–14
Rodrik D. 2009. The new development economics: We shall experiment, but how shall we learn? In *What Works in Development? Thinking Big and Thinking Small*, ed. J Cohen, W Easterly, pp. 24–50. Washington, DC: Brookings Inst. Press
Rubin DB. 2004. *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley & Sons
Russell B. 1912. On the notion of cause. *Proc. Aristot. Soc.* 13:1–26
Samii C. 2016. Causal empiricism in quantitative research. *J. Politics* 78:941–55
Sartori G. 1970. Concept misformation in comparative politics. *Am. Political Sci. Rev.* 64:1033–53
Schulz K. 2015. The rabbit-hole rabbit hole. *New Yorker*, June 4. **https://www.newyorker.com/culture/cultural-comment/the-rabbit-hole-rabbit-hole**
Sekhon JS, Titiunik R. 2017. On interpreting the regression discontinuity design as a local experiment. In *Regression Discontinuity Designs: Theory and Applications*, Vol. 38, ed. MD Cattaneo, JC Escanciano, pp. 1–28. Bingley, UK: Emerald Publ.
Shadish W, Cook TD, Campbell DT. 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. *J. Educ. Behav. Stat.* 38:239–66
Tipton E, Hedges L, Vaden-Kiernan M, Borman G, Sullivan K, Caverly S. 2014. Sample selection in randomized experiments: a new method using propensity score stratified sampling. *J. Res. Educ. Eff.* 7:114–35
Trochim WMK, Donnelly JP. 2006. *The Research Methods Knowledge Base*. Cincinnati, OH: Atomic Dog. 3rd ed.
van Eersel GG, Koppenol-Gonzalez GV, Reiss J. 2019. Extrapolation of experimental results through analogical reasoning from latent classes. *Philos. Sci.* 86:219–35
Vivalt E. 2020. How much can we generalize from impact evaluations? *J. Eur. Econ. Assoc.* 18(6):3045–89
Walker HA, Cohen BP. 1985. Scope statements: imperatives for evaluating theory. *Am. Sociol. Rev.* 50:288–301
Weller N, Barnes J. 2014. *Finding Pathways: Mixed-Method Research for Studying Causal Mechanisms*. Cambridge, UK: Cambridge Univ. Press
Wells GL, Windschilt PD. 1999. Stimulus sampling and social psychology. *Personal. Soc. Psychol. Bull.* 25:1115–25
Wells HG. 1905. *A Modern Utopia*. London: Chapman & Hall
Westreich D, Edwards JK, Lesko CR, Cole SR, Stuart EA. 2019. Target validity and the hierarchy of study designs. *Am. J. Epidemiol.* 188:438–43

Westreich D, Edwards JK, Lesko CR, Stuart EA, Cole SR. 2017. Transportability of trial results using inverse odds of sampling weights. *Am. J. Epidemiol.* 186:1010–14

Wilke A, Humphreys M. 2020. Field experiments, theory, and external validity. In *SAGE Handbook of Research Methods in Political Science and International Relations*, ed. L Curini, RJ Franzese, pp. 1007–35. London: SAGE

Wilson MC, Knutsen CH. 2020. Geographical coverage in political science research. *Perspect. Politics*. **https://doi.org/10.1017/S1537592720002509**

Wing C, Bello-Gomez RA. 2018. Regression discontinuity and beyond: options for studying external validity in an internally valid design. *Am. J. Eval.* 39:91–108

Wolpin KI. 2007. Ex ante policy evaluation, structural estimation, and model selection. *Am. Econ. Rev.* 97:48–52

Yao Y, Vehtari A, Simpson D, Gelman A. 2018. Using stacking to average Bayesian predictive distributions (with discussion). *Bayesian Anal.* 13:917–44