The Necessity of Construct and External Validity for Generalized Causal Claims[*]

Kevin M. Esterling
Professor
School of Public Policy and
Department of Political Science
UC–Riverside
kevin.esterling@ucr.edu

David Brady
Professor
School of Public Policy
UC–Riverside and
WZB Berlin Social Science Center
david.brady@ucr.edu

Eric Schwitzgebel
Professor
Department of Philosophy
UC–Riverside
eric.schwitzgebel@ucr.edu

February 16, 2021

[*] An earlier version was presented as part of the WZB Talks series. We thank Michael Bates, Shaun Bowler, Uljana Feest, Diogo Ferrari, Macartan Humphreys, Sampada KC, Dorothea Kubler, Joscha Legewie, Michael Neblo, Alex Rosenberg, Heike Solga, Jacqueline Sullivan, Nicholas Weller, Bernhard Wessels, Ang Yu and the participants in the MAMA workshop in UCR Psychology for comments. Word count: 12,451.

Abstract

The credibility revolution has facilitated tremendous progress in the social sciences by advancing design-based strategies that rely on internal validity to deductively identify causal effects. We demonstrate that prioritizing internal validity while neglecting construct and external validity undermines causal generalization and misleadingly converts a deductive claim of causality into a claim based on speculation and exploration – undermining the very goals of the credibility revolution. We develop a formal framework of *causal specification* to demonstrate that internal, external and construct validity are jointly necessary for generalized claims regarding a causal effect. If one lacks construct validity, one cannot assign meaningful labels to the cause or to the outcome. If one lacks external validity, one cannot make statements about the conditions required for the cause to occur. Re-balancing considerations of internal, construct and external validity via causal specification preserves and advances the intent of the credibility revolution to understand causal effects.

1 Introduction

The Gold Standard Lab (GSL) undertakes a study of an intervention intended to increase juror turnout.[1] Inspired by get out the vote (GOTV) studies (e.g., Abrajano and Panagopoulos, 2011; Arceneaux and Nickerson, 2010; Gerber and Green, 2000; Green and Gerber, 2008; Green et al., 2013; Trivedi, 2005), GSL exposes residents to different messages using jury summons reminder postcards (partially replicating Bowler et al., 2014). Fortunately, the researchers in GSL are well-trained in design-based causal inference and so they conduct a gold-plated randomized controlled trial (RCT). GSL ask the Riverside (CA) County Superior Court to mail official government postcards to residents who recently received a jury summons, randomizing so that half receive a standard reminder postcard and the other half receive a postcard indicating that failure to appear could result in fines or imprisonment. The "enforcement" condition results in a statistically significant 10 percent increase in turnout relative to the "control" condition – an effect size more than 20 times that typically found in GOTV postcard experiments.

Given these strong results, GSL recommends courts adopt postcards with the enforcement message as a policy, and they publish an article containing the causal generalization: "Enforcement messages increase juror turnout." Eager to demonstrate the efficacy of the enforcement message in other jurisdictions, GSL next collaborates with the superior court in Orange County, California – a more affluent county adjacent to Riverside – to implement the identical gold-plated evaluation. Much to their surprise, the enforcement postcard shows no treatment effect. Discouraged, GSL returns to Riverside to replicate the initial results. This time the jury administrator is too busy to collaborate so GSL use postcards that do not have the official court seal and return address. They believe the different postcards will not matter because the enforcement message – the causal variable

[1] Jury service is a form of democratic participation (Amar, 1995), which is a political right that governments can coerce (Rose, 2005). Courts routinely seek low-cost methods to increase the yield for jury summonses (Boatright, 1999).

in their understanding – is the same as in the original study. Unfortunately, they find no treatment effect in this replication.

GSL understands research best practices in light of the “credibility revolution,” which focuses on the necessity of internal validity to enable claims of causality (Angrist and Pischke, 2010). However, the GSL vignette illustrates a pervasive problem in the social sciences guided by the credibility revolution. Routinely, there is a substantial gap between identifying an internally valid relationship between measured variables and arriving at a correct generalization about what causes what. The current approaches advanced by the credibility revolution for identifying causal effects from data fail to close this gap and as a result, confront fundamental problems for warranting causal generalizations (for a thorough treatment, see Shadish et al., 2002).

When a researcher claims to have established causality with data – such as when GSL claimed an effect of enforcement messages after the first Riverside trial – the researcher must undertake potentially risky inferences beyond what is implied by the data, and internal validity can address only a subset of the inferential risks. One risk concerns the proper specification of the causes and the effects (e.g., that the *enforcement message* increased turnout, as opposed to some other feature of the intervention). Another risk concerns properly specifying the scope conditions of the generalization (e.g., GSL’s erroneous expectation that the intervention would also work in Orange County). We argue that the first of these two risks can be understood as the risk of a failure of construct validity and the second can be understood as the risk of a failure of external validity.

The methods proposed by the credibility revolution identify causal effects in terms of measured variables, rather than in terms of the actual causes and effects, and often leave the scope of the generalization as implicitly local to the experimental setting. However, researchers virtually never interpret their (implicitly local) findings in terms of the measured variables themselves. Instead, they typically state a generalization that posits the actual causes and effects. When researchers make causal generalizations in ordinary language based on measured variables, those generalizations are only deductively sound given a set of background assumptions, including assumptions about construct and external validity. While practitioners may intend their causal claims to be deductive, neglecting considerations of construct and external validity instead makes those claims speculative and exploratory (Banerjee et al., 2017). The claims instead become inductions based on possibly misconceptualized categories and conditions supported by verbal assurances and intuitive plausibility – contrary to the animating ideas of the credibility revolution.

In this sense, the identification assumptions advanced by the credibility revolution are inadequate to support generalized causal claims. By focusing excessively on internal validity, such methods attend too little to the inferential risks in causal generalization that are addressed by considerations of construct and external validity. Thus, even when the target estimand is correctly "identified" (in the sense that term is ordinarily used in literature inspired by the credibility revolution), the causal elements remain speculative or unknown. Our framework, which we call *causal specification*, demonstrates the joint necessity of internal, construct, and external validity in causal generalization. We show that even a minimal claim about what causes what cannot be correct if it has only internal validity and lacks construct and external validity. We show instead that augmenting identification with considerations of construct and external validity can preserve the deductive nature of such claims. Doing so is crucial to preserving the intent of the credibility revolution to focus on understanding causal effects.

2 The Credibility Revolution

The GSL team published a paper reporting the successful results from the first Riverside trial because their training in design-based causal inference led them to believe that the RCT identified a causal effect. Their training was built on the "credibility revolution," which promotes quantitative research designs for identifying *causal effects* from observed data (Angrist and Pischke, 2010; Imbens, 2019; Pearl, 2000; Rubin, 1974). In this tradition, causal claims are "credible" because the connection between a statistical result and a causal effect of interest deductively follows from transparently-stated assumptions (Gelman and Imbens, 2013; Lundberg et al., 2021), such as an assumption of internal validity that posits the absence of confounding under a randomization. While we highlight RCTs throughout, this tradition also advances natural-experimental designs such as instrumental variables, matching, difference-in-differences and regression discontinuity that posit different assumptions (Angrist and Pischke, 2015; Keele and Minozzi, 2013).

Many of the advances in the credibility revolution have been governed by one of two comprehensive frameworks for causal inference, the *potential outcomes* framework, also known as the "Rubin causal model" (RCM) (Holland, 1986; Imbens, 2019), and the *structural causal model* (SCM) framework developed by Pearl (Pearl, 2000). Each of these frameworks is rooted in the counterfactual notion of causality (Lewis, 1973; Neyman, 1935; Woodward, 2004), although in different ways.

In either framework, a causal effect is defined by comparing the counterfactual outcomes – that is, what *would have happened if* the cause had been present versus absent, while everything else had remained the same. Because it is not possible to observe events that don't actually occur, at least half of the relevant cases remain unobserved. A causal effect is said to be *identified* only if the effect described in counterfactual terms can be uniquely expressed as a function of the observed data alone (Petersen and Laan, 2014). Identification, because it involves inference from observed data to a general causal pattern expressed partly in terms of unobserved counterfactuals, requires a set of assumptions about the relation between the observed data and the causal effect of interest.

We illustrate identification using the RCT design. To simplify, we assume a binary representation of the causal process, where the random variable A indicates the treatment state and $\neg A$ indicates the control state ($\neg$ is a symbol meaning "not" or "false"), and a binary outcome variable B. We also assume that the intervention A contains causally efficacious elements that monotonically increase the probability of their effects, including in the presence of other causes. In an RCT, we say that internal validity is present if

$$\{(B_i^A, B_i^{\neg A})\} \perp A_i. \tag{1}$$

Subscript i indicates an event or realization of a random variable for unit i.[2] The symbol $\perp$ means "is independent of."[3] Here, B_i^A represents the counterfactual outcome B conditionally upon unit i's being exposed to A, regardless of whether that unit was actually exposed to A. Likewise $B_i^{\neg A}$ represents the counterfactual conditionally upon that unit's not being exposed to A. For every unit of analysis or trial in an experiment, these terms represent the event B that *would* have occurred *had* A occurred (B^A), or respectively not occurred ($B^{\neg A}$).

In an RCT, the assumption of internal validity in 1 generally holds that the units in treatment and control have identical distributions of potential outcomes, and hence each group can supply the missing counterfactuals for the other. If assumption 1 is true, then it follows that the causal effect estimand τ_i is identified using the observed data:[4]

$$\tau = p(B|A) - p(B|\neg A), \tag{2}$$

since the only systematic difference between those in the A condition and those in the $\neg A$ condition is their exposure to A.[5] The difference between the conditional probabilities

[2] Throughout we assume that each unit i is an element of a set S, where S is the set of units either selected into, or potentially selected into, the RCT. For example S could be a convenience sample, or it could be a sampling frame.

[3] A weaker version of 1 only requires the claim to be true within strata of covariates.

[4] Note that in 2 we are conditioning on the observed data, $B_i = A_i B_i^A + \neg A_i B_i^{\neg A}$.

[5] Here and below we suppress subscripts for simplicity. One could write 2 as $\tau = p_{i \in S|A}(B_i|A_i) - p_{i \in S|\neg A}(B_i|\neg A_i)$. This is the "intention to treat" effect and it also requires the *stable unit treatment value assumption* (SUTVA). Identification of an intervention effect requires a third assumption called the *exclusion restriction*. We examine the exclusion restriction and SUTVA in the section on construct validity below. It is not relevant to the argument whether the probabilities in 2 are frequentist or subjective.

on the right-hand side of 2 is treated as an estimate of the causal effect τ, since under assumption 1 the magnitude of the difference is not driven by bias that would otherwise occur from confounding; see Gerber and Green (2012, 38).[6] Researchers can move from the estimated statistical relationship between A and B to a counterfactual claim with a causal interpretation, saying "A causes B" if the estimated τ is different from zero.

In this paper, we focus on RCTs where the assumption of internal validity is well-justified by the randomization of unit assignments. That is, randomization renders the assumption of internal validity relatively weak and plausible. The claim to have identified a causal effect, however, in no way depends on the strength of the assumptions associated with any specific research design. The conclusion that A causes B deductively follows from the premises encoded in the assumptions laid out in the formal apparatus of identification, such as selection on observables for matching and regression or the parallel path assumption for difference-in-differences (Keele and Minozzi, 2013; Pearl, 2000). Once the assumptions are made, the conclusion follows.

The credibility revolution was motivated by previous generations' naïve reliance on regression models of observational data to test for causality. In that context, applied researchers invoked verbal assurances that they had knowledge of which variables needed to be included, and access to measures of those variables. These assurances typically strained credulity (Leamer, 1983). In turn, many researchers developed what Stokes (2014) refers to as "radical skepticism" about unobserved confounders. We develop our arguments with the RCT design in which the internal validity assumption is relatively weak, so that we can focus on what even perfect internal validity does not accomplish. Radical skepticism was an excess, but the concerns of the radical skeptics, which helped motivate the current prioritization of internal validity, recur also for construct validity and external validity, as will become evident in our notation and discussion below.

[6] In addition, researchers can use departures from the internal validity assumption in 1 to explore the sensitivity of τ to such departures, to assess the strength of the assumption.

3 Validity and Causal Generalization

The credibility revolution has made enormous contributions by emphasizing the critical role of assumptions for deductive causal claims (Gelman and Imbens, 2013; Keele and Minozzi, 2013). In current practice, however, identification of causal effects fails to incorporate assumptions regarding construct and external validity that also are necessary for supporting generalized causal claims, and this is true whether one relies on the RCM or the SCM for identification. In what follows, we explain how our causal specification framework clarifies these essential quantities that must be present in either framework for generalized causal claims to have validity. We also explain how the neglect of construct and external validity undermines the capacity of scholars in the credibility revolution to make the deductive claims they intend to make. In the discussion we show how to modify the RCM and SCM – varying from revisions and extensions to altering fundamental principles (Edwards et al., 2015) – in order to accommodate our definitions of construct and external validity.

As Holland (1986) notes, the variables that actually are measured in a scientific procedure – such as A and B – are "primitive" to the potential outcomes framework, and hence counterfactuals and identification are each with respect to the measured variables.[7] Shadish et al. (2002) forcefully argue, however, that researchers' semantic statements of causal effects are virtually always stated at the level of cause and effect – what we call the causal "relata" – and virtually never in terms of the measured variables themselves (see also Kim, 1971). That is, in most applications, *only some aspects of the measured*

[7] In practice, users of the SCM framework depict measured variables as causal nodes, and so also take measured variables as primitive. As Pearl (2010, 872) explains, under the "consistency rule" of SCM, all relevant causal features must be labeled and depicted in a graph in order for the model to be consistent with a counterfactual representation of causality. However, when positing measured variables as causes, the graph only implicitly encodes assumptions of how the counterfactuals connect to the measured variables (Edwards et al., 2015), which violates the consistency rule. The appendix shows that conditioning on measured variables as causal nodes in a graph can never depict a valid representation of a causal claim in that each measured variable only serves as a collider.

variables are causally relevant. Accurately specifying the causal relata – specifying which aspects of the measured intervention had which relevant effects – is essential to stating a causal claim that is useful for understanding the world. In addition, the experiment is always embedded in a set of conditions that also matter for the presence of causality.

Consider a typical causal process about which researchers wish to make a causal inference (Mackie, 1965; Paul and Hall, 2013; Rothman, 1976). In any experimental design, the intervention will contain a (potentially null) set of causes, which we label "active ingredients," along with other elements that are not causal, or "inert ingredients" (see Cook et al., 2014). Likewise, the outcome will contain elements that are of interest ("the disease") and that are not of direct interest ("symptoms"). We refer to the intervention's active ingredient and the outcome of interest as the causal "relata." In our example, GSL described the manipulated active ingredient as the "enforcement message" and "juror turnout" as the outcome of interest. The active ingredients in the intervention will be catalyzed by conditions that are present in the setting or field (Cartwright, 2011; VanderWeele and Hernán, 2006) without which the cause will not occur – ingredients perhaps present in Riverside but not in Orange County.

We formalize the relata and conditions using the following simple causal claim,

$$\text{“}\alpha \text{ causes } \beta \text{ in } \gamma\text{,”} \tag{3}$$

where α and β are types or classes of events, such as random variables, and γ is a set of contexts or conditions. We assume that causation is a relation between individual concrete events (see Schaffer, 2016, for a review of other metaphysical alternatives), indicated with subscript i. For example, α is the class of events in which summoned jurors receive a postcard with such-and-such a content, and α_i is an individual event of a particular summoned juror receiving a particular postcard. Following convention in philosophy, we use quotation marks to designate a *claim* (e.g. as opposed to a fact in nature). A particular causal claim is a type of historical, hypothetical, or predictive statement about the relationship between two individual events: "α_i caused (or would have caused or will cause) β_i." A general causal claim or *causal generalization* references a pattern among the relata, that "events of type α cause events of type β."

Often, as we have just done, the conditions or settings in which the generalization holds are left implicit or unstated. Some causal generalizations in physics might be intended to be truly universal (Holland 1986, 947, though see Cartwright 1983 for concerns): α causes β (or tends to cause β) whenever and wherever α occurs. However, in social science, causal generalizations are nearly always implicitly restricted to settings with relevant local conditions, γ, that are often vaguely stated or understood. For example, the GSL could only reasonably expect postcards to work in functioning democracies and among literate participants who know English, even if they did not explicitly say so.

The statement "α causes β in γ" is a generalization: it is a claim that one thing generally causes another in a certain range of conditions (Kruglanski and Kroy, 1976). A causal claim is *valid* if the claim is true, that is, if the purported cause and purported effect are the actual cause and actual effect.[8] In our terminology, the general causal claim that "α causes β in γ" is valid if it is indeed the case that α causes β in γ. In this conception, validity is at its core a *relationship between a claim and the world* – the relationship that holds if and only if the claim correctly reflects reality (Shadish et al., 2002, 35).

There are exactly four ways in which the causal generalization "α causes β in γ" can be invalid, corresponding to the four parts of the claim:

(i) α

(ii) causes

(iii) β

[8] This definition comes from measurement theory originating in Kelly (1927), "the problem of validity is whether a test really measures what it purports to measure" and is consistent with Borsboom et al. (2004). On inconsistencies and conceptual difficulties in the concept of validity, see Feest (2020), Jiménez-Buedo (2011) and Sullivan (2009).

(iv) in γ

Something might cause β in γ, but that something might not be events of type α (falsity in part i). Events of type α might cause something in γ, but that something might not be events of type β (falsity in part iii). Events of type α might cause events of type β across some range of conditions, but not across the range γ (falsity in part iv). Or events of type α might be related to events of type β across conditions γ but the relationship might not be a directional causal relationship of the sort claimed (falsity in part ii).

To illustrate, consider how GSL's causal generalization, "Enforcement messages increase juror turnout," might fail:

(i) Their claim might fail because the researchers have misconstrued the nature of the cause, assigning an incorrect semantic label. The postcards might have had their effect not because of the specific words in the enforcement message, but rather because that message was contained on an official postcard.

(ii) Their claim might fail internally, due to chance or poor experimental design. Maybe jurors who were already planning to appear at court disproportionately received the threatening postcards.

(iii) Their claim might fail because the researchers have misconstrued the nature of the outcome, assigning an incorrect semantic label. Maybe juror turnout was mismeasured – for example, if the jury administrator classifies respondents who request an excuse from jury service as a successful recruitment.

(iv) Their claim might fail because they have implicitly mischaracterized how broadly their claim generalizes. The claim invites the reader to generalize to most normal U.S. jury-summons contexts (though unfortunately this remains vague); but perhaps it only works in certain communities.

Generalizations always go beyond the scientific evidence. Researchers will have witnessed only a finite number of events in a specific time and place. In order to make general causal claims that are meaningful to others, researchers must make a *causal inference*, moving from the evidence to a causal claim on the basis of theory, common sense, and other considerations – that is, assumptions combined with the evidence (see Lundberg et al., 2021). A *causal generalization* results from an inferential leap to the conclusion that in general, under conditions γ, α-type events cause β-type events. Such an inference may or may not be warranted, but without an inference, even if a study induced causality it could not support a general causal claim.

Now consider the evidence that is created in an experiment. Our framework distinguishes the active causes in the relata and conditions from the inert ingredients and other events that are bundled with them in the intervention and outcome measurement; that is, we do not take variables as the "primitives" of the analysis. To clarify this distinction, we label the elements of the bundles (the causes, effects and conditions of interest, plus inert ingredients and other events) with lower-case Greek letters, and we label the measured bundles with upper case Latin letters. In the fully binary case

$$A \triangleq \{\alpha \wedge \theta_\alpha\} \tag{4a}$$

$$B \triangleq \{\beta \vee \theta_\beta\} \tag{4b}$$

$$C \triangleq \{\gamma \wedge \theta_\gamma\} \tag{4c}$$

Note that $\triangleq$ means "is definitionally equal to," $\wedge$ logically means "and," (requiring both elements to be true) and $\vee$ logically means "or" (requiring one or both elements to be true). In the binary case, each variable can be set to either *true* or *false*.

The upper case variables A, B, C are observed measures of the intervention, outcome, and conditions, respectively.[9] Greek letter variables represent hypothesized causes, conditions and effects, inert elements, and elements that are not of direct interest, which combine into informationally-equivalent sets (Dafoe et al., 2018). The elements $\{\alpha, \gamma\}$

[9] We omit measurement error in the notation (see Edwards et al., 2015).

are “active ingredients” that have a causal effect on the outcome of interest β. The elements $\{\theta_\alpha, \theta_\gamma\}$ are “inert ingredients” and θ_β is a related outcome that is not of interest.[10]

This notation makes clear that measured variables are inherently bundles. In particular, removing the inert ingredient θ_α in equation 4a would make A false. This is because active and inert ingredients $\{\alpha, \theta_\alpha\}$ are bundled together in the intervention. For example, in the first GSL trial, the postcards bundled the enforcement message with an implied threat, an emotional tone, numerals related to the relevant statute, amount of ink, and sentence complexity (all of which vary at least slightly between treatment and control). Likewise, C bundles all of the conditions $\{\gamma, \theta_\gamma\}$ that remain constant. These include design elements that are identical for both the treatment and control condition (e.g., the cardstock and court seal), the attributes of the experimental units that are assumed to be balanced through randomization (e.g., employment status of the recipient and time the postcard was received), and the conditions in the setting (e.g., Riverside during the rainy season). Typically, neither these conditions nor C itself is literally “measured” beyond disclosures of experimental procedures and the setting of the RCT.

The disjunction in 4b represents that the measurement of an outcome might reflect the real outcome of interest, β, *or* instead a related event not of direct interest, θ_β (e.g. a legally valid request for excuse). In most studies, β itself cannot be measured in isolation but must be inferred from self-reports, records, or events assumed to stand in some felicitous causal relation to β. For simplicity, our notation omits cases in which β occurs but remains unmeasured: It models the risk of “false positives” while leaving the risk of “false negatives” unmodeled.[11]

In an RCT, the evidence is limited to the measured variables. Typically, however, researchers’ causal claims make reference to the events here represented as Greek letters, not to the measured entities characterized by Latin letters. The definition in 4 makes explicit that the Greek-letter reality behind the Latin-letter measures remains a matter

[10] For simplicity, we omit interactions between elements of the information set.

[11] False positives would require a disjunction of conjunctions in the definition of B.

of inference, and this is true even when the causal estimand is identified. Under our simplifying assumptions, identification establishes the following specific causal generalization about what has actually been manipulated and measured:

$$\text{“}p(B|A \wedge C) > p(B|\neg A \wedge C),\text{”} \tag{5}$$

where C is the implicit local situation.[12] Note that throughout we use the textbook definition of "identification" as the result of the relevant design-based assumptions. This means that identification is only with respect to the Latin-letter variables – taken to be primitive – not the Greek-letter causal relata or conditions. It follows from definition 4 that claim 5 is not the same as the (generalized) causal process of interest, which is expressed as

$$\text{“}p(\beta|\alpha \wedge \gamma) > p(\beta|\neg\alpha \wedge \gamma).\text{”} \tag{6}$$

Of course, if the counterfactuals are literally defined only in terms of the measured variables A and B, then considerations of construct validity are irrelevant; "A" is A and "B" is B. Likewise, if the causal effect is only local to the conditions and setting C, then considerations of external validity are largely irrelevant. But absent construct and external validity, one cannot communicate the meaning of the results in claim 5 beyond the statement "Whatever it was we did, at that one time and place, had an effect on whatever it was we measured" (Cronbach, 1982). Since the identification strategy of an RCT is only with respect to the measured variables, identification can only license a claim such as 5 – a claim that would never be published. Claim 5 does not license the meaningful semantic statement represented by claim 6 (Cook et al., 2014).

Note that a claim such as 5 is in the form of a deductive test; it is what Gelman and Imbens (2013) call a "what if"- rather than a "why"-type assessment. In moving implicitly

[12] When stating claims regarding probability relations among event types (rather than specific event sets, as in Section 2), the researcher can choose between frequentist versus counterfactual versus subjective interpretations of probability.

from 5 to 6, however, the researcher inadvertently transforms a "what if" question among the measured variables into a speculative or exploratory "why" accounting of the relata and conditions driving the statistical patterns that support claim 5. This is the case even when claim 5 is said to be identified (Banerjee et al., 2017). When researchers neglect construct and external validity, they are simply hazarding an exploratory guess about causality. This is just like when earlier generations of social scientists offered exploratory guesses about included covariates in hope of achieving internal validity.

Because identification is only with respect to the measured variables, and to local conditions, the identification of potential outcomes implies nothing about parts (i), (iii), or (iv) of a valid generalized causal claim, that is, the labeling of the relata and the conditions under which the cause will occur.[13] However, a causal generalization is not valid unless all four elements are present: construct validity of the cause (i), internal validity (ii), construct validity of the outcome (iii), and external validity (iv), which together capture four distinct inferential risks in moving from scientific evidence to a general causal claim. Hence, design-based identification does not provide sufficient assumptions to justify or deduce a generalized causal claim.

Shadish et al. (2002) provides the classic statement of the existence of this gap between measured variables and the constructs the variables are intended to represent. They propose a theory of causal generalization that states, even when internal validity is achieved, moving from a claim such as 5 to a claim such as 6 requires construct validity to understand how variables map onto underlying constructs, and external validity to understand how causal effects can transport to new situations.[14]

[13] RCTs require two other assumptions beyond internal validity for identification, the exclusion restriction and SUTVA, but as we explain below, neither of these assumptions address considerations relevant to either construct or external validity.

[14] Shadish et al. (2002) assert each element of an experiment – the units, treatments, outcomes and settings – represents a construct, and so construct validity is relevant to each. In contrast, our definition of construct validity only considers claims regarding the relata, while our definition of external validity only considers claims to generalization to other units and settings.

Next we develop the framework of *causal specification* to clarify the necessity of internal, construct and external validity for preserving the deductive nature of "what if" causal questions. Our framework formalizes the insights of Shadish et al. (2002) into a single expression that explicitly demonstrates the parity of each type of validity when supporting a causal claim. Our framework provides the necessary assumptions regarding each type of validity – internal, construct and external – that can ensure the deductive conclusions operate both at the level of measured variables and at the level of causal relata and conditions. As we detail below, *construct validity* is present when α and β are correctly specified, and *external validity* is present when γ is correctly specified. Construct and external validity, including correctly specified conditions and causal relata, are no less necessary for causal generalization than internal validity.

4 Causal Specification for Internal Validity

Ever since Campbell introduced the definitions of internal and external validity in the 1950s, generations of scientists have been tutored to understand that internal validity can warrant a "local" and "molar" causal claim (Campbell and Stanley, 1963; Cook, 2012; Shadish et al., 2002), that is, a reportable finding of a causal effect for a (local) sample.[15] Because internal validity establishes the presence of causality, it has had a lexical scientific priority over external validity (e.g. Imbens and Rubin, 2015, 359).

Recall that we stated our general causal claim as the sentence: "α causes β in γ." In an RCT, identification requires internal validity, that is, that the probability of β under the counterfactual of α being either true or false is in fact unrelated to the value of α

[15] Cook (2012) and Julnes (2004) document how Campbell's later work distinguishes "molar" interventions, which are complex packages of elements within the manipulated variable, and "molecular" interventions which are the causal elements (see also Rothman, 1976). We build on this distinction below by labeling the causal elements "active ingredients" and the remaining elements "inert ingredients."

within an experiment, or

$$\text{“}\{p(\beta^{\alpha}), p(\beta^{\neg\alpha)}\} \perp \alpha.\text{”} \tag{7}$$

which is analogous to equation 1, except it is stated at the level of the causal relata, and it is intended as a claim rather than a truth.

Establishing internal validity for a causal generalization requires specifying 7. That is, it requires specifying the absence of a certain type of unbalanced distribution of confounding causes.[16] Of course, no research design can ensure that claim 7 is true. A randomized design makes a claim of unconfoundedness more plausible. Yet, even in a perfectly designed RCT with a large sample, claim 7 might still be false. As a result, all standard treatments of RCT design also require an assumption of independence (see Holland, 1986) or randomization (Angrist et al., 1996). For generalization, the analogous assumption must be made among the relata-level potential outcomes, which in turn requires modifying the potential outcomes framework to relax the requirement that measured variables are primitive (Edwards et al., 2015).[17]

5 Causal Specification for Construct Validity

Traditionally, construct validity centers on considerations of the quality of observed measures when a criterion measure does not exist, to ensure that the outcome that is measured in fact corresponds with the concept of interest (Adcock and Collier, 2001, 529). In causal analysis, this means the semantic labels assigned to the causal relata are correct (Cook et al., 2014).[18] The notion originates in Cronbach and Meehl (1955) who proposed assessing whether the pattern of convergences and divergences in a set of correlations meets the theoretical expectations of a “nomological network.” As Borsboom et al. (2004) explains,

[16] We ignore inert properties that remain unbalanced.

[17] The appendix shows how the SCM can incorporate latent causes directly into a causal graph.

[18] In this paper we set aside the issue of reliability. Reliability is related to validity in that if the measurement is unreliable it may not be clear what it measures (Hood, 2009).

such an analysis of correlations can never fully serve to match a measure with a concept that is best understood as an ontological referent; such an approach would mistake empirical validation procedures for validity (Alexandrova, 2017).[19]

According to Borsboom et al. (2004), a measure is construct valid if measured observations are themselves caused by the underlying (ontological) referent of interest. Referring back to our definition 4, a causal generalization is construct valid only if A accurately tracks α and B accurately tracks β. As ontological referents, α and β are not normally measurable in isolation by the researcher. Instead, the correspondence between the measured variables and the intended relata is a (possibly warranted) assumption, governed by considerations of construct validity, just as the presence of internal validity is an assumption. Assigning correct semantic labels "α" and "β" to the causal relata thus stands as one of the core inferential risks when making causal claims based on the statistical relationship between A and B.[20] Without an explicit assumption and justification for their semantic labels, researchers cannot properly claim to have stated a generalized causal effect. One knows only that something caused something, not what causes what.

Construct Validity of the Cause. Construct validity is essential for understanding the role of the intervention as a possible causal agent, and so we first consider construct validity of the cause. Generally, analysts claim the specific physical properties of an intervention stand as an instance of an underlying, latent causal construct (Sartori, 1970). For example, Gerber et al. (2008) takes the text statement on a postcard promising to reveal one's voting behavior to one's neighbors as an instance of "social pressure," much like

[19] On balance, a close reading of Cronbach and Meehl (1955) need not support the Borsboom et al. (2004) allegation that an investigation of a nomological network only amounts to an atheoretical search for "meaning without referent" (see Westen and Rosenthal, 2003, 609). Likewise a close reading of Borsboom et al. (2004) need not support the allegation in Hood (2009) that the authors over-correct by simply sidestepping epistemological considerations. Neither of these camps, however, is as explicit as us that validity is a relationship between a semantic claim and the world.

[20] As Kim (1971) writes, "the orthographic features of an event description are not a reliable guide to the ontological structure of the event it describes . . . "

GSL took their text to be an instance of "enforcement." The correspondence between the observed physical intervention and the underlying construct is necessarily imperfect, however (Adcock and Collier, 2001, 534). For example, different physical manifestations can correspond to the same referent depending on the context, such as when Dunning (2008, 43) devises different informational voting interventions to match across implementations in Latin America, South Asia and Africa (see Gilbert et al., 2016).[21]

In a proposed empirical test of the causal process, that is, whether A causes B, the manipulated intervention variable A is presumed to contain at least one necessary component (active ingredient) for the cause to occur (Mackie, 1965; Rothman, 1976). Every intervention must be a bundle of components, however, some of which are active (α) and some of which are inert (θ_α). Internal validity itself cannot warrant assigning the label "active ingredient" to any of the elements in the intervention because the manipulation itself is always potentially confounded with active ingredients not explicitly labeled by the researchers (Cook et al. 2014, 379,382; Fong and Grimmer 2019). This is the problem Dafoe et al. (2018) identify as "informational equivalence." Instead, as a minimum requirement, a valid causal claim must assume and specify the active ingredient α and assign to it a construct valid, semantically-meaningful label.

The active ingredients in social science interventions typically are not as easily identified as in the case of drug trials. For the GSL example, the manipulation is not only the enforcement message but everything else bundled with the intervention, including the level of threat, the presence of the numerals indicating the statute, sentence complexity, and so on (see Fong and Grimmer, 2019). Because the inert and active ingredients perfectly covary within an internally valid design, internal validity alone cannot distinguish active from inert ingredients. Some elements might not be entirely ontologically distinct, such as "enforcement" and "threat" in the GSL example. If the elements are sufficiently

[21] Gilbert et al. (2016) considers whether, to an Israeli, certain employment consequences that follow taking leave for mandatory military service are the same, to an American, as those that follow from taking leave from work to get married and go on a honeymoon.

distinguishable, conceptually and empirically, which ingredient best characterizes what is actually driving outcomes remains an open question.[22]

The credibility revolution understands aspects of this problem of confounding in the intervention, although they address the problem by stipulating ancillary assumptions rather than treating it as a core element of validity. In particular, when there is full compliance with the protocol, RCT designs rely on two assumptions in addition to the assumption of randomization, known as the "exclusion restriction" and the "stable unit treatment value assumption" (SUTVA) (Angrist et al., 1996; Gerber and Green, 2012). The exclusion restriction assumes the assignment has no effect on the outcome other than its effect in changing a unit's exposure to the intervention. SUTVA assumes that the treatment each unit receives is not affected by other units, irrespective of whether the other units were assigned to treatment or control.[23] The substantive purpose of these two assumptions is to rule out certain, but not all, aspects of the confounding within the intervention that can remain even when internal validity is perfect (see Julnes, 2004).

First consider the exclusion restriction. Absent blinding, random assignment itself can create confounds such as John Henry and Hawthorne effects that occur simply because the unit is aware of assignment. To assume the assignment itself is not causal under the exclusion restriction is to assume that the assignment is not among the active ingredients. The purpose of the exclusion restriction is to label the assignment process as an inert component, but the assumption itself does not identify the active component of the intervention required by construct validity (Julnes, 2004, 176).[24]

[22] Identifying an unconfounded molar package of elements is potentially useful for prevention. If one element of the package is necessary, withholding the full package would disable the cause (Rothman, 1976, 558). Often though analysts wish to identify the cause that brings about an outcome, rather than how to prevent outcomes.

[23] Technically, since the exclusion restriction and SUTVA allow the analyst to ignore each unit's assignment as well as the assignment and exposure vector of all other units, the two assumptions greatly reduce the number of potential outcomes that need to be considered (Angrist et al., 1996).

[24] Dafoe et al. (2018) recommend evoking a stronger version of the exclusion restriction to assume that all necessary latent components that are not of interest to the analyst are

Second, SUTVA is intended to identify the active ingredients in the intervention. For example, SUTVA rules out the presence of spillover from the treatment units to the control units, such as when someone in GSL's treatment group is friends with someone in the control group and so shares the postcard message. Randomization in an RCT does not rule out this scenario and hence the analyst must assume the states that define treatment and control are the ones that the analyst had intended. Construct validity however requires semantic labels to match the referents. Merely *assuming* the label matches the underlying real cause, via the exclusion restriction and SUTVA, cannot adequately substitute for specification of the labels.

Formalizing Construct Validity of the Cause. Under our notation, a claim for *weak construct validity of the cause* takes the form,

$$\text{``}p[B|(\alpha \wedge \theta_\alpha) \wedge C)] > p[(B|(\neg\alpha \wedge \theta_\alpha) \wedge C] \;\forall\; \theta_\alpha\text{,''} \tag{8}$$

where the $\forall$ symbol means "for each" – that is, the cases where θ_α is either true or false.[25] Under this claim, when α is present the probability of the outcome is higher than when α is absent, and that inequality holds both when the potential confound θ_α is present and when it is not. In other words, the cause α must be *specified.* If Claim 8 is false (given our background assumptions), the claimed cause "α" is not a real cause α and construct validity is absent.

Our definition of construct validity cannot be accommodated in either the RCM or the SCM whenever practitioners in either framework take measured variables as primitive. To enable valid generalized causal claims, the RCM would need to relax the requirement

uncorrelated with the latent component of interest (see also Fong and Grimmer, 2019). This stronger assumption implies that the analyst correctly understands *a priori* which components of the manipulation are active and which are inert.

[25] For simplicity we omit cases in which θ_α is true on one side of the statement and false on the other side. This is analogous to the SUTVA assumption evoked in RCTs to simplify the number of counterfactuals an experiment has to consider.

that potential outcomes are defined over measured variables only (Edwards et al., 2015). The SCM would need to encode assumptions regarding the active ingredients as latent causal nodes in a DAG (see the appendix). Claim 8 demonstrates the inadequacy of the exclusion restriction and SUTVA as a substitute for construct validity, in that each of these is only a special case of assumptions regarding inert ingredients, for example, that θ_α characterizes the assignment process or non-causal components of intervention itself. Instead, the claim in 8 holds that the cause the researcher postulates to be the actual cause is in fact an actual cause – in other words, that it matters specifically whether α is present, not just whether θ_α is absent.

A claim of *strong construct validity of the cause* would add the following to claim 8:

$$\text{“}p[B|(\alpha \wedge \theta_\alpha) \wedge C)] \approx p[(B|(\alpha \wedge \neg\theta_\alpha) \wedge C] \;\forall\; \alpha.\text{”} \tag{9}$$

When α is present, the probability of B is not affected by θ_α, and likewise when α is absent. Under our background assumptions, unless this is true the claimed inert ingredient “θ_α” is not the real inert ingredient θ_α, and hence construct validity does not occur. The difference between the weak and strong version is that in the weak version α is relevant to the outcome regardless of whether θ_α is present. By contrast, the strong version adds that θ_α's presence or absence is irrelevant to the outcome.

Construct Validity of the Outcome. Correctly identifying and measuring the outcome of interest is also essential for causal identification. In a clinical trial, for example, one might relieve the symptoms and mistakenly conclude one has cured the underlying disease (e.g., using fever as the measure of disease then applying ice to the patient and claiming the disease cured). In the GSL example, the intervention aims to increase juror turnout, but the jury administrator might record an excuse from service as also having fulfilled the legal requirements.

Construct validity in the outcome is present when the outcome is correctly labeled and conceptualized. The directly measured outcome B might stand in a variety of relationships to the outcome of interest β. In some cases, β itself might be directly measurable (e.g., response time) in which case B=β. More commonly, B and β stand in some causal relationship, where B is a presumed cause or effect of β or the two are related to a common cause. For example, if β is juror turnout, B might be the clerk's record of which residents reported on the assigned day, which could be entirely accurate or contain false positives or negatives. Generally, the tighter the causal relationship between β and B, the better the warrant for inferring from the directly observed B to the claimed β.

The details of causal modeling of the relationship between B and β elude our simple probabilistic notation. We note that when B is observed, it *might* be true that the outcome of interest β occurred or (in false positive cases) it might be true that only a related outcome not of interest θ_β might have occurred. For example if the jury administrator recorded a request for a deferment or exemption as satisfying the juror's legal obligation, even though GSL might not have intended such behavior to count as jury service. Internal validity does not establish the existence of the required relationship between β and B. Absent specification that B captures β, one cannot properly claim to have specified the real outcome. Hence construct validity of the outcome would be lacking.

6 Causal Specification for External Validity

The traditional definition of external validity focuses on whether an identified causal effect extrapolates or is "generalizable" to other settings (Cook, 2014b; Findley et al., 2021; Guala, 2005; Julnes, 2004; Shadish et al., 2002). "Settings" include different countries, time periods, populations, contexts and laboratories. Although this definition is standard, in the social sciences it is often seen as an unattainable ideal (Deaton and Cartwright, 2018; Findley et al., 2021) as there are very few social science studies that yield the same results across any and all settings of human existence (Cook, 2014b; Julnes, 2004). Hence, external validity has often been one ideal that social scientists feel comfortable failing to attain.

The GSL vignette exemplifies the pervasive lack of traditionally-defined external validity in studies with strong internal validity. Indeed, RCTs usually yield widely varying results across settings (Allcott, 2015; Deaton, 2010, 2019; Deaton and Cartwright, 2018; Olsen et al., 2013; Peters et al., 2018; Pritchett and Sandefur, 2013, 2015; Vivalt, 2020; Weiss et al., 2014), and results from psychological experiments vary dramatically across societies and "WEIRD (Western Educated Industrialized Rich Democracies) subjects are particularly unusual compared with the rest of the species" (Henrich et al., 2010).[26] As Banerjee and Duflo (2009, 160) acknowledge, "Without assumptions, results from experiments cannot be generalized beyond their context."

Clarifying the Definition of External Validity. Partly because of the pervasive lack of traditionally-defined external validity, we propose a clarification of the definition of external validity. Our definition is less a revision and more a clarification that aims to unify recent advances.

First, we define *conditions* as any features that are balanced across treatment and control groups or constant in a setting. Like Cartwright's (2011) "helping factors" and "countering causes," Deaton and Cartwright's (2018) "support factors," and Findley and colleagues' (2021) "context or structural factors," conditions can augment or undermine a cause. Conditions include quintessential characteristics of settings like culture, history, and institutions. The conditions also include all elements that are constant in the experimental design, attributes of the experimental units that are balanced between treatment and control, and aspects of the causal field (Mackie, 1965) that are all constant relative to the intervention. In the classic example, oxygen is an active condition that is necessary for the treatment effect of striking a match to result in the outcome of fire (Pearl, 2019).

[26] Many highlight how imperfect implementation triggers variation in effects across settings (Allcott, 2015; Deaton and Cartwright, 2018; Olsen et al., 2013; Peters et al., 2018; Ravallion, 2012; Vivalt, 2020; Weiss et al., 2014). Yet, low external validity is common even with ideal implementation.

As we explain below, other conditions are *inert* (e.g. nitrogen in the air).

Second, we clarify that external validity is the correct *specification* of the conditions that define the causal generalization. Put differently, external validity requires correctly labeling the conditions that enable the treatment to produce its effects. According to the traditional definition, a claim is externally valid if it generalizes across settings. We say a claim is externally valid to the extent that the researchers have accurately specified the conditions defining the range of settings across which the effect generalizes. Rather than characterizing the extent to which findings generalize (Findley et al., 2021), we propose external validity specifies exactly when a causal effect does or does not generalize. Hence, external validity means the researcher has accurately specified the conditions that enable or disable a causal effect.

This revised definition is more useful for the realities of social science. Unlike the traditional definition, we embrace that treatment effects will vary across settings because of the inescapable role of conditions that also matter for the outcome.[27] It is not as helpful to simply say a causal claim lacks external validity because it is constrained to a narrow range of settings. We say it is more helpful to specify how a causal claim is contingent on specific conditions. The heart of external validity then becomes the correctness of one's causal specification based on the relevant conditions.

While GSL lacked traditionally-defined external validity, the actual problem is that GSL does not even know why the treatment worked in Riverside and not Orange County. GSL could attain external validity however by correctly specifying which conditions moderate their causal effect and define the range of settings. For example, GSL might demonstrate the intervention works in the setting of Riverside but not in Orange County because

[27] Several others make this point about RCTs (e.g. Deaton and Cartwright 2018; Guala 2005; Peters et al. 2018; Weiss et al. 2014). For example, Pritchett and Sandefur (2015, 474) write, "Social programs, in contrast, are embedded in contexts which encompass a long list of unknown factors which interact in often unknown ways." Ravallion (2012, 110) refers to "the institutional-implementation factors that might make a given program a success in one place, or at one scale, but not another."

of the condition of affluence. This clarifies the essential role of a particular condition across settings for causal specification. Traditionally, one would say GSL lacks external validity because the treatment worked in Riverside but not Orange County. We propose GSL has external validity if they can say the treatment works in Riverside because of an absence of affluence and fails in Orange because of a presence of affluence.

Our revision unifies recent efforts to incorporate external validity into causal frameworks. First, in Pearl and colleagues' transportability approach (Bareinboim and Pearl, 2016), one needs to know which conditions are modifying the causal effect (Humphreys and Scacco, 2020). Knowing "where" in the directed acyclic graph that effect moderation is occurring requires knowledge of what conditions in settings matter. Second, Banerjee et al. (2017) argue for an "inherently subjective" "structured speculation" to use theory and knowledge of conditions to generalize treatment effects across settings. Third, Stuart et al. (2011) use propensity scores and weighting to assess if RCT samples generalize and have external validity to a target population. This requires the researcher know which conditions to include in the propensity score model and on what conditions to compare sample and target population (Pritchett and Sandefur, 2013).

Despite their contributions, it is important to note that these recent efforts are vulnerable to the same criticisms of unknown confounding that motivated the "radical skeptic" camp of the credibility revolution. After all, it is not clear why one should have "radical skepticism" about unknown confounders within one setting while maintaining even modest optimism about knowing what unobserved conditions enable generalization across settings. In all three recent approaches, one relies on the same sorts of verbal assurances that the radical skeptics sought to marginalize in claims about internal validity.

For instance, Deaton and Cartwright (2018, 13) explain propensity scores and weighting only work when observables perfectly align with the relevant conditions, the conditions are present both in the initial sample and population, and the effects of conditions are the same in both the initial sample and population. In the GSL vignette, suppose that in the first Riverside trial the court provided GSL with zip codes for each of the prospective jurors but not other covariate information (as in Bowler et al., 2014). Say GSL determined that the causal effect of the enforcement postcard did not vary significantly across zip codes within Riverside and tried to extrapolate from that. But if the affluence in Orange County is not reflected across the zip codes in Riverside County, such subgroup analyses could not enable GSL to generalize. Hence, GSL would fail to have external validity.

Why External Validity Matters. The social sciences have greatly prioritized internal validity over external validity for at least several decades (Cook, 2014a; Findley et al., 2021; Julnes, 2004; Pirog, 2014; Pritchett and Sandefur, 2013). For instance, the justly influential causal inference textbooks by Morgan and Winship (2015) and by Imbens and Rubin (2015) have zero mentions of external validity in their indexes.

The credibility revolution often sidesteps external validity by simply acknowledging a causal estimate is "local." However, if one has identified only a local causal effect, with no considerations of external validity, one can only accomplish a very limited kind of knowledge (Cronbach, 1982). A local internally valid causal effect is circumscribed to a specific set of units exposed to a specific event in a specific time and specific place and is only knowable retrospectively (Cook, 2012; Deaton, 2010, 2019; Guala, 2005; Rothman, 1976; VanderWeele and Hernán, 2006; Vivalt, 2020; Westreich et al., 2019). As Cartwright (2011) explains, internal validity only shows "it works somewhere." Actually, internal validity only shows it historically worked (in the past tense) somewhere (Nosek and Errington, 2020). This does not yield the sort of knowledge or inference social scientists typically want (Findley et al., 2021). Cartwright (2011) explains, we want to know: "it works widely" or "it will work for us" (Cook, 2014b; Deaton, 2010; Deaton and Cartwright, 2018; Pritchett and Sandefur, 2013, 2015).[28]

[28] For these reasons, Rubin (1974) stressed the need for "subjective random sampling" of settings to ensure a study was of "practical interest," "representative" and "useful." Cronbach (1982, 137) similarly argues that internal validity alone is "trivial, past-tense, and local."

Unfortunately, internal validity alone does not warrant making any claims beyond a historical claim about what happened in the one very specific setting in which the research was conducted. Confronted with the lack of external validity, scholars often respond one of two ways. First, many claim that identifying an effect in one setting is sufficient and they have no intention to produce general knowledge. However, by claiming they only intend to make an extremely specific historical claim about the effect of something in only one setting, they are retreating to what we call *the historicist's refuge*.

Historians' idiographic causal narratives of specific events are certainly valuable. Nevertheless, we doubt that social scientists – if pressed on the matter – would concede they have no desire to be different from historians (Findley et al., 2021; Henrich et al., 2010). As Nosek and Errington (2020, 3) explain, researchers rarely limit their inferences to a "particular climate, at particular times of day, at a particular point in history, with a particular measurement method, using particular assessments, with a particular sample." Indeed, there is an implicit generalization even in topic and setting selection – as researchers choose topics and settings for the very purpose that they are likely to illustrate some general phenomenon.

If it is truly the case that a social scientist only wants to make a "local" and non-generalizable claim, it is essential to be explicit. Just as the credibility revolution has transformed scientific discourse by requiring causal identification for any language of causal effects, scholars should declare their lack of any intent to generalize *to any setting other than the experimental setting that already occurred.*[29] This would require a substantial change to prevailing practices as readers would need to police against any language of generalization in the absence of external validity just like readers currently police against causal language in the absence of internal validity.

[29] Notably, studies of the U.S. rarely justify selecting the unusual U.S. case, specify what conditions make the U.S. unusual, or qualify how conclusions might not generalize (Findley et al., 2021). By contrast, studies of non-U.S. contexts are routinely required to justify precisely how their context is unique and/or generalizes.

Ultimately, in the historicist's refuge, one takes the position of claiming to identify a causal effect while having no understanding of the conditions in the setting enabling that effect. The researcher does not know how much of the effect is due to the treatment, the conditions in the setting, or some complex interaction of treatment and conditions. Nor does the researcher even know if the conditions are common or unusual and hence whether the treatment will have an effect in any other setting. If the conditions in the setting are unusual, then the causal effect will be unusually large or small – a form of selection bias that is similar to the unknown confounding bias that concerned the credibility revolution (Findley et al., 2021). In the GSL vignette, GSL does not know if Riverside or Orange County reveals the true causal effect and it could equally be because of either helping factors in Riverside or countering causes in Orange (or both). There could even be unknown conditions that reveal both Riverside and Orange are unusual.

Second, some admit to a lack of external validity and say the "next step" is to go forth across a range of settings. For instance, Banerjee and Duflo (2009, 162) write: "If we were prepared to carry out enough experiments in varied enough locations, we could learn as much as we want to know about the distribution of the treatment effects across sites." This is not possible however without external validity and causal specification. Sampling a "range of settings" or "similar settings" presumes one knows what defines the range or similarity. Sampling from a part of the population does not represent the population, and the law of large numbers does not remedy sampling from a corner of the sample space.[30] Defining the sample space can only be a theoretical speculation about what moderates the treatment (Banerjee and Duflo, 2009; Deaton and Cartwright, 2018). In the GSL vignette, the method of choosing Orange County as the next step does not establish relevant differences across the range of settings because those differences are

[30] This leads to problematic "simple enumerative induction." Cartwright (2011) and Deaton and Cartwright (2018) review several examples of why one cannot infer: RCT 1 worked, RCT 2 worked, therefore RCT 3 or all RCTs will work. Cartwright (2011) explains that this is an unwarranted "power or capacity claim" that the treatment "reliably promotes" outcomes.

unknown just like unknown confounders.

Again, this is hard to square with the "radical skeptics"' concerns about unknown confounding. Unfortunately, there is no clear methodology for discovering the unknown conditions across the range of settings (Cartwright, 2011; Deaton, 2010; Deaton and Cartwright, 2018). Hence, when a researcher pursues internal validity without external validity, they appear to have "radical skepticism" about unknown confounders but more than modest optimism about knowing the relevant conditions across settings. Absent clear, epistemically warranted specification of relevant conditions, any sampling of the range of settings remains speculative and exploratory (e.g., Banerjee and Duflo, 2009). One simply does not know that well-known characteristics of settings are what actually moderates the treatment and should be the basis of sampling across settings (Pritchett and Sandefur, 2013; Ravallion, 2012). This is just like pre-credibility revolution regression analyses that fell back on verbal assurances because they had no way of knowing if all confounders had been controlled (Deaton and Cartwright, 2018; Muller, 2015).

Formalizing External Validity. We now offer a formal definition of external validity. Our definition of external validity relies on understanding conditions as all features of the setting, units, or intervention that are constant or balanced between values of α. Recall we define the setting and conditions in the equation, $C \triangleq (\gamma \wedge \theta_\gamma)$, that is, C is true if both γ and θ_γ are true; γ are the active ingredients in the setting that moderate the manipulated cause α, and θ_γ are the inert ingredients that are also in the setting.

Under our definition of external validity, the causal conditions γ must be correctly specified to make a valid causal generalization. The presence of θ_γ clarifies there are lots of other ways that settings vary and many of those are ignorable. A claim of *weak external validity* takes the form

$$\text{“}p[B|A \wedge (\gamma \wedge \theta_\gamma)] > p[B|A \wedge (\neg\gamma \wedge \theta_\gamma)] \ \forall\ \theta_\gamma,\text{”} \tag{10}$$

Note the close parallel with weak construct validity of the cause in claim 8. The outcome has a higher probability when γ is present versus absent, regardless of whether θ_γ is present or absent. The RCM currently is not well-equipped to handle considerations of external validity, given that its focus is on identifying local effects. The SCM addresses considerations of external validity using notions of the notions of "transportability" described in Bareinboim and Pearl (2016), although as the appendix shows, claims of transportability must be over latent conditions γ rather than measured contextual variables C.

A claim of *strong external validity* adds the following to claim 10:

$$\text{“}p[B|A \wedge (\gamma \wedge \theta_\gamma)] \approx p[B|A \wedge (\gamma \wedge \neg\theta_\gamma)] \; \forall \; \gamma.\text{”} \tag{11}$$

In the strong external validity claim, which θ_γ occurs is irrelevant to the probability of the outcome, provided that A and γ are constant. Under our background assumptions, unless both equations are true the claimed causal condition "γ" is not the real causal condition γ and the claimed inert condition "θ_γ" is not the real inert condition θ_γ.

7 A Framework for Causal Specification

The tight linkage between the concept of internal validity and the concept of causality is encoded in the causal frameworks that governed the credibility revolution. The *potential outcomes* framework (Holland, 1986; Rubin, 1974) and the *structural causal models* framework (Pearl, 2000) are each centered on the problem of unconfoundedness, but with little consideration of the necessity of external validity or construct validity. As a result, neither serves as an adequate general framework for validity.

In the potential outcomes framework, measured variables are primitive when defining counterfactuals: "A" is simply A, "B" is simply B, and the results typically are implicitly local to the study conditions C, each without regard to the causal relata and conditions that are the actual components of the causal process. In this respect, the RCM requires some modification of its fundamental principles to accommodate these more general notions of validity (Edwards et al., 2015). In the SCM, while it is permissible to include latent causal nodes in a directed graph, such as α or γ, the appendix shows that a graph that includes measured variables as causal nodes violates the consistency rule (Pearl, 2010, 872) in that such a graph would mistakenly condition on a non-causal entity.

More importantly, causal specification informs users of either framework the assumptions that must be added to the common identification strategies in order to support deductive causal claims. A strong claim of *validity* requires causal specification of all of claims 7 to 11. No one of them is sufficient for specifying a causal generalization. A weaker claim omits 9 and 11. When researchers focus on internal validity, they can often establish claim 5 (“$p(B|A \wedge C) > p(B|\neg A \wedge C)$”) but they risk inferring without adequate justification from claim 5 to claim 6 (“$p(\beta|\alpha \wedge \gamma) > p(\beta|\neg\alpha \wedge \gamma)$”). In the absence of construct and external validity, the researcher mistakenly converts a deductive “what if” question to an exploratory “why” question. One merely posits that α is the active ingredient in A, that B accurately tracks β, and that the causal relationship holds across a range of (only implicitly specified) settings. To make these assumptions without sufficient warrant is to assume away most problems at the core of validity, substituting informal speculation for rigorous clarification of the nature of the cause, the nature of the effect, and the scope of the generalization.

As our notation makes plain, even if a formal identification strategy justifies 5, that by itself in no way justifies making the claim about the causal process expressed in equation 6. To see this, we expand claim 5 using definition 4 to the equivalent statement,

$$\text{“}p[\beta \vee \theta_\beta|(\alpha \wedge \theta_\alpha) \wedge (\gamma \wedge \theta_\gamma)] > p[\beta \vee \theta_\beta|\neg(\alpha \wedge \theta_\alpha) \wedge (\gamma \wedge \theta_\gamma)].\text{”} \tag{12}$$

Comparing claim 5 to claim 12, expanding A problematizes construct validity of the cause; expanding B problematizes construct validity of the outcome; and expanding C problematizes external validity. We depict claims 12 as a causal graph (DAG) in the appendix. The limits of internal validity are illustrated by the GSL example. Solely focusing on internal validity, GSL does not know what conditions explain why the treatment worked at all, much less why it worked in the first setting and not subsequent settings. Internal validity is of no assistance in labeling the active ingredients in *either* the manipulation or in the setting or conditions.

8 Conclusion

Social scientists typically aim to contribute to general knowledge about what causes what in what conditions, and not merely historical knowledge that something caused something one time in the past. That is, social scientists aim to arrive at valid causal generalizations.

In our framework for understanding validity, a causal generalization of the form "α causes β in γ" is valid if and only if it is true that α causes β in γ. The challenge of causal specification is not only the challenge of confirming that in fact something caused something in one setting (i.e. the focus of internal validity) but equally the challenge of specifying – correctly labeling and conceptualizing – the nature of the cause, the nature of the effect, and the conditions under which the generalization holds. By itself, even the most rigorous proof of internal validity shows only that some aspect of the manipulation (A but not necessarily α) caused some measured outcome (B but not necessarily β) in one setting (C, typically leaving γ implicit). Construct validity is achieved when the semantically asserted cause and effect are the actual cause and effect. External validity is achieved when the scope of the generalization is correctly specified. Unless all three types of validity are present, a claim that α causes β in γ is false. All three types of validity are required; none have priority.

We show that the formal identification assumptions within "credible designs" are insufficient for supporting generalized causal claims, irrespective of whether one is working in the SCM or the RCM. The commonly-cited identification assumptions ensure internal validity but fail to ensure construct and external validity. We encourage social scientists attend equally to internal, construct, and external validity. As our framework of causal specification makes clear, all three are equally necessary for generalized causal claims and hence supply the additional assumptions that must augment current approaches to identification in order to support the deductive justification of causal claims. These assumptions inevitably rely partly on subjective, theoretically-grounded, and verbally-justified labeling of the relata for construct validity, and of the conditions for external validity. These additional assumptions regarding the relata and conditions are as necessary for deriving a causal claim as is the assumption of internal validity.

If applied researchers ignore construct and external validity when stating deductive causal claims, they mistakenly convert the intended deductive claim into a claim based on exploration and speculation – contrary to the goals of the credibility revolution. Our framework for causal specification corrects this, and offers a means for applied researchers to preserve the deductive nature of their claims not only at the level of measured variables but also – and more importantly – at the level of relata and conditions. In this way, causal specification clarifies the additional assumptions that are required for the credibility revolution to achieve its aspirations of understanding causal effects.

References

Abrajano, M. and C. Panagopoulos (2011). Does Language Matter? The Impact of Spanish Versus English-Language GOTV Efforts on Latino Turnout. *American Politics Research 39*(July), 643–663.

Adcock, R. and D. Collier (2001). Measurement validity: A shared standard for qualitative and quantitative research. *American Political Science Review 95*(3), 529–546.

Alexandrova, A. (2017). *A Philosophy for the Science of Well-Being.* New York, N.Y.: Oxford University Press.

Allcott, H. (2015). Site Selection Bias in Program Evaluation. *Quarterly Journal of Economics 130*, 1117–1165.

Amar, V. D. (1995). Jury Service as Political Participation Akin to Voting. *Cornell Law Review 80*, 203–259.

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

Angrist, J. D. and J.-S. Pischke (2010). The Credibility Revolution in Empirical Economics: How Better Research Design is Taking the Con out of Econometrics. *Journal of Economic Perspectives 24* (2), 3–30.

Angrist, J. D. and J.-S. Pischke (2015). *Mastering 'Metrics: The Path from Cause to Effect.* Princeton, N.J.: Princeton University Press.

Arceneaux, K. and D. W. Nickerson (2010). Comparing Negative and Positive Campaign Messages: Evidence from Two Field Experiments. *American Politics Research 38* (Jan.), 54–83.

Banerjee, A., S. Chassang, and E. Snowberg (2017). Decision Theoretic Approaches to Experiment Design and External Validity. *Handbook of Economic Field Experiments 1*, 141–174.

Banerjee, A. V. and E. Duflo (2009). The Experimental Approach to Development Economics. *Annual Review of Economics 1* (1), 151–178.

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

Boatright, R. G. (1999). Why Citizens Don't Respond to Jury Summonses and What Courts Can Do About It. *Judicature 82*, 156–164.

Borsboom, D., G. J. Mellenbergh, and J. van Heerden (2004). The Concept of Validity. *Psychological Review 111* (4), 1061–1071.

Bowler, S., K. Esterling, and D. Holmes (2014). GOTJ: Get Out the Juror. *Political Behavior 36*, 515–533.

Campbell, D. T. and J. C. Stanley (1963). *Experimental and Quasi-Experimental Designs for Research.* Chicago, Ill.: Rand McNally & Company.

Cartwright, N. (1983). *How the Laws of Physics Lie.* New York, N.Y.: Oxford University Press.

Cartwright, N. (2011). The Art of Medicine: A philosopher's view of the long road from RCTs to effectiveness. *Lancet (London, England) 377* (9775), 1400–1.

Cook, T. (2012, 5). Causal Generalization: How Campbell and Cronbach Influenced My Theoretical Thinking on This Topic, Including in Shadish, Cook, and Campbell. In M. C. Alkin (Ed.), *Evaluation Roots*, pp. 89–112. SAGE Publications, Inc.

Cook, T. D. (2014a). "Big Data" in Research on Social Policy. *Journal of Policy Analysis and Management 33*(2), 544–547.

Cook, T. D. (2014b). Generalizing Causal Knowledge in the Policy Sciences: External Validity as a Task of Both Multiattribute Representation and Multiattribute Extrapolation. *Journal of Policy Analysis and Management 33*(2), 527–536.

Cook, T. D., Y. Tang, and S. Seidman Diamond (2014). Causally Valid Relationships That Invoke the Wrong Causal Agent: Construct Validity of the Cause in Policy Research. *Journal of the Society for Social Work and Research 5*(4), 379–414.

Cronbach, L. J. (1982). *Designing Evaluations of Educational and Social Programs.* San Francisco: Jossey-Bass Publishers.

Cronbach, L. J. and P. E. Meehl (1955). Construct Validity in Psychological Tests. *Psychological Bulletin 52*(4), 281–302.

Dafoe, A., B. Zhang, and D. Caughey (2018). Information Equivalence in Survey Experiments. *Political Analysis 26*(4), 399–416.

Deaton, A. (2010). Instruments, randomization, and learning about development. *Journal of Economic Literature 48*(2), 424–455.

Deaton, A. (2019). Randomization in the Tropics Revisited: A Theme and Eleven Variations. In F. Bedecarrats, I. Guerin, and F. Rouboud (Eds.), *Randomized Control Trials in the Field of Development.* New York, N.Y.: Oxford University Press.

Deaton, A. and N. Cartwright (2018). Understanding and Misunderstanding Randomized Control Trials. *Social Science and Medicine 210*, 2–21.

Dunning, T. (2008). Improving Causal Inference: Strengths and Limitations of Natural Experiments. *Political Research Quarterly 61*(June), 282–293.

Edwards, J. K., S. R. Cole, and D. Westreich (2015, 8). All your data are always missing: Incorporating bias due to measurement error into the potential outcomes framework. *International Journal of Epidemiology 44*(4), 1452–1459.

Feest, U. (2020, 1). Construct validity in psychological tests – the case of implicit social cognition. *European Journal for Philosophy of Science 10*(1), 1–24.

Findley, M. G., K. Kikuta, and M. Denly (2021). External Validity. *Annual Review of Political Science forthcomin*, 1–51.

Fong, C. and J. Grimmer (2019). Causal Inference with Latent Treatments.

Gelman, A. and G. Imbens (2013, 11). Why ask Why? Forward Causal Inference and Reverse Causal Questions. Technical Report 19614, National Bureau of Economic Research.

Gerber, A. S. and D. P. Green (2000). The Effects of Canvassing, Telephone Calls, and Direct Mail on Voter Turnout: A Field Experiment. *American Political Science Review 94*(Sept.), 653–663.

Gerber, A. S. and D. P. Green (2012). *Field Experiments: Design, Analysis and Interpretation.* New York, N.Y.: W.W. Norton.

Gerber, A. S., D. P. Green, and C. W. Larimer (2008). Social Pressure and Voter Turnout: Evidence from a Large-Scale Field Experiment. *American Political Science Review 102*(Feb.), 33–48.

Gilbert, D. T., G. King, S. Pettigrew, and T. D. Wilson (2016). Comment on "Estimating the reproducibility of psychological science".

Green, D. P. and A. S. Gerber (2008). *Get Out the Vote: How to Increase Voter Turnout* (2nd ed.). Washington, D.C.: Brookings Institution Press.

Green, D. P., M. C. McGrath, and P. M. Aranow (2013). Field Experiments and the Study of Voter Turnout. *Journal of Elections, Public Opinion and Parties 23*(1), 27–48.

Guala, F. (2005). *The Methodology of Experimental Economics.* New York, N.Y.: Cambridge University Press.

Henrich, J., S. J. Heine, and A. Norenzayan (2010). The weirdest people in the world?

Holland, P. W. (1986). Statistics and Causal Analysis. *Journal of the American Statistical Association 81*(Dec.), 945–960.

Hood, S. B. (2009). Validity in Psychological Testing and Scientific Realism. *Theory & Psychology 19*(4), 451–473.

Humphreys, M. and A. Scacco (2020). The Aggregation Challenge. *World Development 127.*

Imbens, G. W. (2019). Potential Outcome and Directed Acyclic Graph Approaches to Causality: Relevance for Empirical Practice in Economics.

Imbens, G. W. and D. B. Rubin (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction.* New York, N.Y.: Cambridge University Press.

Jiménez-Buedo, M. (2011, 9). Conceptual tools for assessing experiments: Some well-entrenched confusions regarding the internal/external validity distinction. *Journal of Economic Methodology 18*(3), 271–282.

Julnes, G. (2004). Review of Experimental and Quasi-Experimental Designs for Generalized Causal Inference. *Evaluation and Program Planning 27*, 173–185.

Keele, L. and W. Minozzi (2013). How Much is Minnesota Like Wisconsin? Assumptions and Counterfactuals in Causal Inference with Observational Data. *Political Analysis 21*(Spring), 193–216.

Kelly, T. L. (1927). *Interpretation of Educational Measurements.* Yonkers-on-Hudson, N.Y.: World Book.

Kim, J. (1971). Causes and Events: Mackie on Causation. *Journal of Philosophy 68*(14), 426–441.

Kruglanski, A. W. and M. Kroy (1976). Outcome Validity in Experimental Research: A Re-conceptualization. *Representative Research in Social Psychology 7*(2), 166–178.

Leamer, E. E. (1983). Let's Take the Con Out of Econometrics. *The American Economic Review 73*(1), 31–43.

Lewis, D. (1973). Causation. *Journal of Philosophy 70*, 556–567.

Lundberg, I., R. Johnson, and B. M. Stewart (2021). What is Your Estimand? Defining the Target Quantity Connects Statistical Evidence to Theory. *American Sociological Review forthcomin.*

Mackie, J. (1965). Causes and Conditions. *American Philosophical Quarterly 12*, 245–265.

Morgan, S. L. and C. Winship (2015). *Counterfactuals and Causal Inference: Methods and Principles for Social Research* (2nd ed.). New York, N.Y.: Cambridge University Press.

Muller, S. (2015). Causal Interaction and External Validity: Obstacles to the Policy Relevance of Randomized Evaluations. *World Bank Economic Review 29*(Supplement), S217–S225.

Neyman, J. (1935). Statistical Problems in Agricultural Experimentation. *Supplement of Journal of the Royal Statistical Society 2*, 107–180.

Nosek, B. A. and T. M. Errington (2020, 3). What is replication? *PLOS Biology 18*(3), e3000691.

Olsen, R. B., L. L. Orr, S. H. Bell, and E. A. Stuart (2013). External Validity in Policy Evaluations That Choose Sites Purposively. *Journal of Policy Analysis and Management 32*(1), 107–121.

Paul, L. and N. Hall (2013). *Causation: A User's Guide.* Oxford: Oxford University Press.

Pearl, J. (2000). *Causality: Models, Reasoning and Inference* (2 ed.). New York, N.Y.: Cambridge University Press.

Pearl, J. (2010). On the Consistency Rule in Causal Inference: Axiom, Definition, Assumption, or Theorem? *Epidemiology 21*(6), 872–875.

Pearl, J. (2019). Sufficient causes: On oxygen, matches, and fires. *Journal of Causal Inference 7*(2).

Peters, J., J. Langbein, and G. Roberts (2018, 2). Generalization in the Tropics – Development Policy, Randomized Controlled Trials, and External Validity. *The World Bank Research Observer 33*(1), 34–64.

Petersen, M. L. and M. J. v. d. Laan (2014). Causal Models and Learning from Data: Integrating Causal Modeling and Statistical Estimation. *Epidemiology 25*(3), 418–426.

Pirog, M. A. (2014). Internal versus External Validity: Where are Policy Analysts Going? *Journal of Policy Analysis and Management 33*(2), 548–550.

Pritchett, L. and J. Sandefur (2013). Context Matters for Size: Why External Validity Claims and Development Practice Do Not Mix. *Journal of Globalization and Development 4*(Dec.), 161–197.

Pritchett, L. and J. Sandefur (2015). Learning from experiments when context matters. In *American Economic Review*, Volume 105, pp. 471–475. American Economic Association.

Ravallion, M. (2012). Fighting Poverty One Experiment at a Time. *Journal of Economic Literature 50*, 103–114.

Rose, M. R. (2005). A Dutiful Voice: Justice in the Distribution of Jury Service. *Law and Society Review 39*(3), 601–634.

Rothman, K. (1976). Causes. *American Journal of Epidemiology 104*, 587–592.

Rubin, D. B. (1974). Estimating Casual Effects of Treatments in Randomized and Nonrandomized Studies. *Journal of Educational Psychology 66*(5), 688–701.

Sartori, G. (1970). Concept Misinformation in Comparative Politics. *American Political Science Review 64*(4), 1033–1053.

Schaffer, J. (2016). The Metaphysics of Causation. In *The Stanford Encyclopedia of Philosophy.* https://plato.stanford.edu/archives/fall2016/entries/causation-metaphysics.

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

Stokes, S. (2014). A Defense of Observational Research. In *Field Experiments and Their Critics.*

Stuart, E. A., S. R. Cole, C. P. Bradshaw, and P. J. Leaf (2011). The use of propensity scores to assess the generalizability of results from randomized trials. *Journal of the Royal Statistical Society. Series A: Statistics in Society 174*(2), 369–386.

Sullivan, J. A. (2009, 4). The multiplicity of experimental protocols: A challenge to reductionist and non-reductionist models of the unity of neuroscience. *Synthese 167*(3), 511–539.

Trivedi, N. (2005). The Effect of Identity-Based GOTV Direct Mail Appeals on the Turnout of Indian Americans. *The Annals of the American Academy of Political and Social Science 601*(Sept.), 115–122.

VanderWeele, T. J. and M. A. Hernán (2006). From counterfactuals to sufficient component causes and vice versa.

Vivalt, E. (2020). How Much Can We Generalize from Impact Evaluations? *Journal of the European Economic Association Forthcomin*.

Weiss, M. J., H. S. Bloom, and T. Brock (2014). A Conceptual Framework for Studying the Sources of Variation in Program Effects. *Journal of Policy Analysis and Management 33*(3), 778–808.

Westen, D. and R. Rosenthal (2003). Quantifying Construct Validity: Two Simple Measures. *Journal of Personality and Social Psychology 84*(3), 608–618.

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

Woodward, J. (2004). *Making Things Happen: A Theory of Causal Explanation*. New York, N.Y.: Oxford University Press.

A Appendix 1: DAG Representation

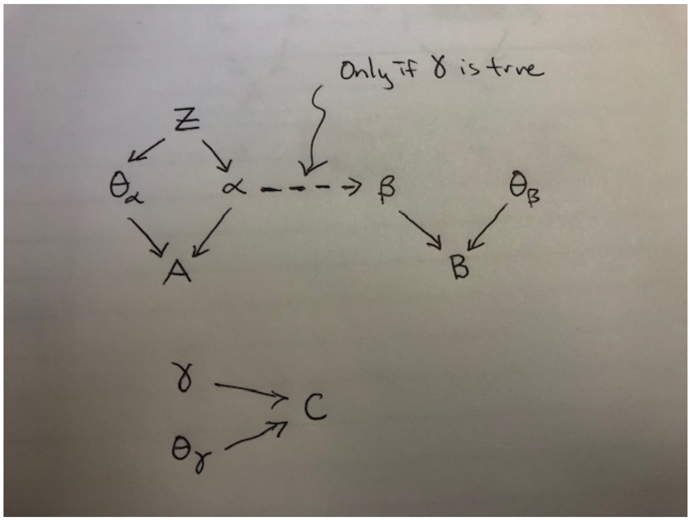

Figure 1

The DAG in this appendix section is a representation of the generalized causal claim "α causes β in γ" as defined in claim 12 of the main text; this representation is *valid* if it corresponds to nature. Z is an assignment mechanism. All of the other nodes are defined in the text. Nodes represented by Greek letters are latent or unobserved, and nodes represented by Latin letters are measured or observed. Among the latent nodes, the $\boldsymbol{\theta}$ vector contains "inert" ingredients in that the nodes do not have any effect, either direct or indirect, on the outcomes β or B. α and γ are "active" ingredients in that they cause the outcome of interest β. An ideal experiment would execute a $do(\alpha)$ procedure, in both the presence and absence of γ, but since α and γ are latent such a procedure is not possible.

The DAG offers a proof that a valid causal statement never conditions on the measured variables since doing so in all cases would condition on a collider variable. Instead, to support a strongly valid generalized causal claim based on an observed statistical relationship between A and B, all of the Greek letter nodes must be specified. To support a weakly valid causal claim, the $\boldsymbol{\theta}$ vector does not need to be specified.

B Appendix 2: Completeness of Validity Statements

Claims 8 and 9 show the claims regarding α and θ_α needed to meet our requirements for construct validity of the cause. Note that each claim has four arguments over these two parameters that are relevant to the claim, two on the left-hand side and two on the right-hand side. Here we show that the paper has already considered all of the relevant permutations for the binary case. As we state in footnote 25, for simplicity, we omit cases in which the universally quantified variable is true on one side of the statement and false on the other side. That is, we are not considering the case where counterfactuals over both α and θ_α can occur at the same time. This is analogous to the SUTVA assumption evoked in RCTs to simplify the number of counterfactuals an experiment has to consider.

Table 1 gives all of the possible permutations of the elements in the four arguments for construct validity of the cause, and we show that the permutations we ignore are equalities that are not relevant to construct validity or simply provide redundant information. In addition, since the conditions for external validity are exactly symmetric to those of construct validity of the cause, the same results apply to demonstrate that equations 10 and 11 for external validity are similarly comprehensive.

Notice first that rows 1 to 4 of table 1 simply state equalities that have no bearing on validity. Notice next that rows 11-16 are redundant statements that simply switch which argument set is on the left- or right-hand side, where the specific redundancies are listed in the final column, and so these rows provide no additional information for validity. Finally, notice that rows 5-6 allow both counterfacutuals to occur at the same time, which we have ruled out by assumption. And so the only rows that we need to consider are 7-10, which correspond to equations 8 and 9 that are in the paper. In particular, rows 7 and 8 correspond to the requirements for weak construct validity of the cause with θ_α set to true and false, respectively. Rows 9 and 10 correspond to the requirements for strong construct validity of the cause with α set to true and false, respectively.

This analysis does not consider the much more complex case where there are interac-

Table 1: Permutations for Construct Validity of the Cause

Row	Arg. 1	Arg. 2	Arg. 3	Arg. 4	Result
1	α	θ_α	α	θ_α	Equality
2	α	$\neg\theta_\alpha$	α	$\neg\theta_\alpha$	Equality
3	$\neg\alpha$	θ_α	$\neg\alpha$	θ_α	Equality
4	$\neg\alpha$	$\neg\theta_\alpha$	$\neg\alpha$	$\neg\theta_\alpha$	Equality
5	α	θ_α	$\neg\alpha$	$\neg\theta_\alpha$	Ruled Out by Assumption
6	α	$\neg\theta_\alpha$	$\neg\alpha$	θ_α	Ruled Out by Assumption
7	α	θ_α	$\neg\alpha$	θ_α	Claim 8, First Case
8	α	$\neg\theta_\alpha$	$\neg\alpha$	$\neg\theta_\alpha$	Claim 8, Second Case
9	α	θ_α	α	$\neg\theta_\alpha$	Claim 9, First Case
10	$\neg\alpha$	θ_α	$\neg\alpha$	$\neg\theta_\alpha$	Claim 9, Second Case
11	α	$\neg\theta_\alpha$	α	θ_α	Redundant to Row 9
12	$\neg\alpha$	θ_α	α	θ_α	Redundant to Row 7
13	$\neg\alpha$	θ_α	α	$\neg\theta_\alpha$	Redundant to Row 6
14	$\neg\alpha$	$\neg\theta_\alpha$	α	θ_α	Redundant to Row 5
15	$\neg\alpha$	$\neg\theta_\alpha$	α	$\neg\theta_\alpha$	Redundant to Row 8
16	$\neg\alpha$	$\neg\theta_\alpha$	$\neg\alpha$	θ_α	Redundant to Row 10

tions between construct and external validity, or for interactions between active and inert elements. These interactions do not exist by assumption, and so are outside the scope of this paper, but relaxing those assumptions is straightforward.