Esterling, Kevin M.; Brady, David; Schwitzgebel, Eric

Working Paper

The Necessity of Construct and External Validity for Generalized Causal Claims

I4R Discussion Paper Series, No. 18

Provided in Cooperation with:
The Institute for Replication (I4R)

Suggested Citation: Esterling, Kevin M.; Brady, David; Schwitzgebel, Eric (2023) : The Necessity of Construct and External Validity for Generalized Causal Claims, I4R Discussion Paper Series, No. 18, Institute for Replication (I4R), s.l.

This Version is available at:
https://hdl.handle.net/10419/268605

INSTITUTE for REPLICATION

No. 18
I4R DISCUSSION PAPER SERIES

The Necessity of Construct and External Validity for Generalized Causal Claims

Kevin M. Esterling

David Brady

Eric Schwitzgebel

February 2023

INSTITUTE for REPLICATION

I4R DISCUSSION PAPER SERIES

I4R DP No. 18

The Necessity of Construct and External Validity for Generalized Causal Claims

Kevin M. Esterling[1], David Brady[2], Eric Schwitzgebel[3]

[1]School of Public Policy and UC-Riverside, Dept. of Political Science, Riverside/USA
[2]School of Public Policy UC-Riverside/USA, WZB Berlin Social Science Center/Germany
[3]UC-Riverside, Dept. of Philosophy, Riverside/USA

FEBRUARY 2023

E-Mail: joerg.peters@rwi-essen.de
RWI – Leibniz Institute for Economic Research
Hohenzollernstraße 1-3
45128 Essen/Germany
www.i4replication.org

ISSN: 2752-1931

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

Kevin M. Esterling
(Corresponding Author)
Professor
School of Public Policy and
Department of Political Science
UC–Riverside
900 University Ave.
Riverside, CA 92506 USA
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 6, 2023

*Presented at the 2021 Annual Summer Meeting of the Society for Political Methodology and at the 2021 Annual Meeting of the American Political Science Association. An earlier version was presented in the WZB Talks Series, July 2020. We thank Elias Bareinboim, Michael Bates, Shaun Bowler, Nancy Cartwright, Carlos Cinelli, Justin Esarey, Uljana Feest, Diogo Ferrari, Christian Fong, Don Green, Justin Grimmer, Francesco Guala, Steffen Huck, Macartan Humphreys, Robert Kaestner, Sampada KC, Jon Krosnick, Dorothea Kubler, Doug Lauen, Joscha Legewie, Michael Neblo, Judea Pearl, Jörg Peters, Alex Rosenberg, Tara Slough, Heike Solga, Jacqueline Sullivan, Nicholas Weller, Bernhard Wessels, Ang Yu and the participants in the MAMA workshop in UCR Psychology for comments.

Abstract

The Credibility Revolution advances quantitative research designs intended to *identify* causal effects from observed data. The ensuing emphasis on internal validity however has enabled the neglect of construct and external validity. This article develops a framework we call *causal specification*. The framework formally demonstrates the joint necessity of assumptions regarding internal, construct and external validity for causal generalization. Indeed, the lack of any of the three types of validity undermines the Credibility Revolution's own goal to understand causality deductively. Without assumptions regarding construct validity, one cannot accurately label the cause or outcome. Without assumptions regarding external validity, one cannot label the conditions enabling the cause to have an effect. These assumptions ultimately are founded on qualitative and theoretical understandings of a causal process. As a result, causal specification clarifies the central role of qualitative research in underwriting deductive understandings of causality in quantitative research.

Keywords: Causality, Construct Validity, Deduction, External Validity, Generalization, Identification

Over the past few decades, the social sciences have seen the rise of the "Credibility Revolution" in quantitative causal inference (Angrist & Pischke 2010; Banerjee & Duflo 2009; Card 2022; Pearl 2000; Rubin 1974). Researchers influenced by the Credibility Revolution advocate for a design-based approach to evaluating causal effects. Arguably, this approach has come to dominate quantitative research in political science and the social sciences (Keele 2015; Samii 2016). The design-based approach is at the core of popular textbooks that political scientists use to teach quantitative research design to PhD students (Gerber & Green 2012; Morgan & Winship 2015). Twenty out of 49 papers presented at the 2022 summer meetings of the Society for Political Methodology, including the keynote speech from economist Susan Athey, advance designs for causal inference. To recognize the impact of this approach on quantitative social science, two of the key innovators of the Credibility Revolution, Joshua Angrist and Guido Imbens, were awarded the Nobel Prize in economics in 2021. In its announcement, the Nobel Committee stated the Credibility Revolution "has transformed applied work and improved researchers' ability to answer causal questions of great importance for economic and social policy . . . "

The Credibility Revolution has contributed to the study of causality in at least two important ways (Keele 2015). First, the field promotes a *deductive* approach to discovering causal effects by clarifying that explicit assumptions are necessary to make a causal claim "credible." In this tradition, a causal claim is credible when the connection between a statistical result and a causal effect of interest deductively follows from transparently-stated assumptions (Gelman & Imbens 2013; Keele 2015; Lundberg *et al.* 2021). In a deductive view, causality is not a property of the data and analysis. Instead, a claim regarding causality follows from the premises encoded within stated assumptions.

Second, tightly related to the first contribution, the Credibility Revolution advances designs that state sufficient assumptions to "identify" causal effects from observed data (Banerjee & Duflo 2009; Card 2022; Pearl 2000; Rubin 1974). In particular, the field advances designs that warrant an assumption of *internal validity*. While we will highlight

randomized controlled trials, this tradition also advances natural-experimental designs such as instrumental variables, matching, difference-in-differences and regression discontinuity that each posit different assumptions (Keele & Minozzi 2013). Each of these designs seeks to ensure that the contrast of interest to the research question is not confounded and hence is internally valid. Given the design assumptions, the quantitative analyst can compare aggregate outcomes between counterfactual states of the world in order to detect the presence of a cause.

While casual identification strategies primarily focus on internal validity, they routinely omit considerations of *construct* and *external validity*.[1] This article explains that this neglect results in a substantial inferential gap. Identifying a causal effect among measured variables in an internally valid research design is insufficient on its own to justify a generalization about what actually causes what. Valid generalizations about causes and effects require not only internal validity but also construct validity and external validity. Without construct validity, one risks mislabeling the cause or outcome (Shadish *et al.* 2002). Without external validity, one risks misunderstanding the conditions enabling the cause to have an effect (Falleti & Lynch 2009). Thus, assumptions regarding construct and external validity are necessary to move beyond measured variables to a deductive, generalized claim about a cause and an effect.

As even Angrist & Pischke (2010, 23) acknowledge, the most that internal validity allows the analyst to deduce is the simple fact that some cause occurred in the course of an experiment. But internal validity provides no guidance on how such facts from a given study or across studies can accumulate into knowledge. Instead, in their view,

[1]The influential textbooks from Morgan & Winship (2015), Imbens & Rubin (2015), and Gerber & Green (2012) have no index entries for either construct or external validity. The assertion that causal claims can be identified from internal validity alone is pervasive; for a formal example, see Keele (2015, 316). The textbook by Morton & Williams (2010, chapter 7) has a sophisticated discussion of both construct and external validity, drawing largely from Shadish *et al.* (2002), but neither of these works discusses the necessity of all three validities for deductive causal inference. Findley *et al.* (2021) and Egami & Hartman (2022) formalize external validity but only in terms of variables, so they do not demonstrate the core argument of our paper, which is the necessity of external validity (and construct validity) to preserve the deductiveness of causal claims.

any substantive claims about what was the cause, what was the effect and what were the enabling conditions can at best be – in their words – "speculative" (see also Holland 1986, 959). However, without explicit assumptions about construct and external validity, such speculations remain tentative in a way that undermines the deductive "credibility" of a causal claim.

In practice, applied researchers – even leading proponents of the Credibility Revolution – routinely make general causal claims without explicitly declaring what is speculation. For example, Gerber *et al.* (2008, 33) claim that mailing postcards that reveal one's voting history to neighbors "demonstrate[s] the profound importance of social pressure as an inducement to political participation." Likewise, Angrist *et al.* (2012, 858) claim that a specific charter school design called KIPP "generated substantial score gains," attributing causality to the KIPP design itself. Based on RCTs of an anti-poverty program, Banerjee *et al.* (2015) claim: "It is possible to make sustainable improvements in the economic status of the poor with a relatively short-term intervention." These authors make claims about actual causes and actual effects without sufficiently clarifying that they had intended these internally valid but not necessarily construct or externally valid claims to rely on – and knowledge in their field to accumulate from – "speculation."

Building on the philosophy of science and causality, this article provides a critique that aims to augment the Credibility Revolution. We explain how, whether the analyst recognizes it or not, any "credible" generalized causal claim requires assumptions of construct and external validity – assumptions that ultimately derive from theory (see Slough 2022; Slough & Tyson 2022) and qualitative knowledge (see Kocher & Monteiro 2016). Thus, deduction as a scientific enterprise in quantitative causal inference fundamentally depends on a combination of theory and qualitative knowledge along with any design-based statistical evidence.

To address this neglect of construct and external validity, we develop a framework we call *causal specification*. This framework formalizes assumptions regarding internal,

construct and external validity within a single causal expression, and shows that causal generalization requires a rebalancing that *equally* values all three validities. Within a deductive understanding of causality, internal validity has no special status or lexical priority. The causal specification framework explicitly recognizes the contributions of theory and qualitative knowledge to quantitative causal inference, and charts a way forward for social scientists who aim to make deductive causal generalizations.

The Credibility Revolution

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), and the *structural causal model* (SCM) framework developed by Pearl (Pearl 2000). 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 (Lewis 1973; Neyman 1935; Woodward 2004). Because it is not possible to observe events that do not actually occur, at least half of the relevant cases remain unobserved post intervention. A causal effect is said to be *identified* only if the effect described in counterfactual terms can be deduced from the data in the ideal circumstances where the analyst hypothetically had an infinite amount of data (Keele 2015; Manski 1995; Petersen & van der Laan 2014). Because it involves inference from observed data to unobserved counterfactuals, identification requires a set of assumptions about the relation between the observed data and the causal effect of interest.

We illustrate identification within the RCT design using a fictional vignette in which The Gold Standard Lab (GSL) undertakes a quantitative test of the causal effect of an intervention aiming to increase juror turnout. Jury service is a form of democratic participation, and is a political right that governments can coerce (Rose 2005). Of course,

not everyone who receives a summons actually shows up at the courthouse for jury service, so courts routinely seek low-cost methods to increase the yield for jury summonses (Boatright 1999). Inspired by get out the vote (GOTV) studies (e.g., Arceneaux & Nickerson 2010; Gerber & Green 2000), GSL exposes residents in the court's jurisdiction to different messages using jury summons reminder postcards (partially replicating Bowler *et al.* 2014) to evaluate the messages' causal effect.

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 "reminder" condition – an effect size more than 20 times that typically found in GOTV postcard experiments.

Given these strong results, GSL recommends courts adopt the enforcement message as a policy, and they publish an article containing the causal generalization: "Enforcement messages cause juror turnout." Eager to demonstrate the efficacy of the enforcement message in other contexts, GSL next collaborates with the superior court in Orange County, California – a more affluent adjacent county – to implement the identical gold-plated evaluation. Much to their surprise, the enforcement postcard shows no treatment effect, from which they reason that affluence suppresses the enforcement message effect.

The GSL design very much adheres to the identification strategies of the Credibility Revolution. We can formalize the GSL study design and analysis as follows. There are n units indexed by $i \in \{1, \ldots, n\}$. Throughout we assume that each unit i is an element of $\mathcal{S}$, which is the sampling frame (on the role of sampling in RCTs, see Imai *et al.* 2008). We assume a binary representation of the causal variable, where $A = 1$ indicates the treatment state (having received a postcard with the enforcement message) and $A = 0$

indicates the control state (having received a postcard with the reminder message). B is a binary outcome variable where $B_i = 1$ if the unit reported for jury duty and $B_i = 0$ if they did not. $B_i(A = 1)$ represents the counterfactual outcome B conditioned on unit i's being in the treatment state, regardless of which postcard the unit actually received. Likewise $B_i(A = 0)$ represents the counterfactual outcome conditionally upon that unit being in the control state. For every unit of analysis i, these terms represent the potential outcome B that *would* have occurred *had* $A = 1$ occurred or respectively not occurred.

In an RCT, we say that internal validity is present if the experimental units' assignment to treatment or control is unrelated to their outcomes in either state, represented formally as:

$$[B_i(A = 1), B_i(A = 0)] \perp\!\!\!\perp A_i. \tag{1}$$

The symbol $\perp\!\!\!\perp$ means "is independent of." A weaker version of 1 only requires the claim to be true within strata of covariates. The assumption in 1 implies that the units in treatment and control have identical distributions of potential outcomes in expectation, and hence each group can supply the missing counterfactuals for the other. It follows that the "average treatment effect" *estimand* (ATE) is identified and can be estimated using the observed data:

$$ATE = \mathbb{E}_{\mathcal{S}}[B_i(A_i = 1)] - \mathbb{E}_{\mathcal{S}}[B_i(A_i = 0)]. \tag{2}$$

Under assumption 1 the only systematic difference between those in the $A = 1$ condition and those in the $A = 0$ condition is their exposure to A. $\mathbb{E}_{\mathcal{S}}$ is the expectation over $\mathcal{S}$ and indexing A indicates estimation is over the observed data.

The right-hand side of 2 is the estimand of the ATE since, under assumption 1, the expected difference is not driven by bias that would otherwise occur from confounding (Gerber & Green 2012, 38). This is the "intention to treat" effect and it also requires the *stable unit treatment value assumption* (SUTVA). Identification of the average treatment effect requires a third assumption called the *exclusion restriction* (Angrist *et al.* 1996). We

return to those two assumptions in the section on construct validity below. Researchers can interpret the estimated statistical relationship between A and B as a counterfactual causal claim of identifying the ATE if these ancillary assumptions and assumption 1, the core assumption of internal validity, are true.

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. 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 design identified a causal effect 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, continuity for regression discontinuity, or the parallel path assumption for difference-in-differences (Keele & Minozzi 2013; Pearl 2000). Once the assumptions are made, the conclusion follows deductively.

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 included all needed control variables. These assurances typically strained credulity (Samii 2016). In turn, many researchers developed what Stokes (2014) refers to as "radical skepticism" about unobserved confounders. We develop our arguments with the RCT design so that we can focus on what even perfect internal validity does not accomplish. Radical skepticism was an excess. However, the skeptical thinking that helped motivate the focus on internal validity equally justifies concerns about construct validity and external validity, as will become evident in our notation and discussion below.

Validity and Causal Generalization

The methods inspired by the Credibility Revolution take the variables that are actually measured as fundamental for understanding causality. As Holland (1986) notes, the variables that actually are measured in a scientific procedure – such as A and B – are "primitive" in the Rubin causal model, and hence counterfactuals and identification are each with respect to the measured variables. Researchers can identify that some cause occurred in a given experiment when the expected value of the variable B differs between counterfactual states characterized by A. Equally importantly, identification strategies often leave the scope of generalization vague or implicitly local to the experimental setting.

Nevertheless, researchers rarely interpret their findings in terms of only the measured variables and are often not explicit about their limited scope of generalization. Instead, scholars typically interpret their findings in terms of general causes ("enforcement messages") and general effects ("juror turnout") that generalize beyond their study's setting. As Shadish *et al.* (2002) forcefully argue, researchers' semantic statements are virtually always in terms of underlying causes and effects and virtually never in terms of the measured variables themselves (see also Kim 1971). Furthermore, all experiments are embedded in a set of conditions that also matter to causality (Falleti & Lynch 2009). Causes, effects, and the conditions in settings are referents in nature, and claims regarding their relationships are ontological. As a result, a correct generalized causal claim requires not only the identification of a causal effect between measured variables (i.e. internal validity) but also the correct specification of the actual cause and of the actual effect (i.e. construct validity) and correct specification of the scope of the generalization (i.e. external validity).

Accurately specifying which aspects of the measured intervention had which relevant effects is essential to deducing a generalized causal claim. Consider a typical causal process about which researchers wish to make a causal inference and claim (Heckman 2005; Mackie 1965; Paul & Hall 2013; Rothman 1976). In any experiment, 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 not of direct interest ("symptoms"). We refer to the intervention's active ingredients and the outcome of interest as the causal "relata." 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 supported or countered by conditions in the setting (Cartwright 2011; VanderWeele & Hernán 2006) – ingredients such as affluence perhaps present in coastal Orange County but not in inland Riverside County.

Using a classic notation from philosophy, 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 ontological events, such as random variables, and γ is a set of supporting or countering conditions. We assume that causation is a relation between individual concrete events (see Schaffer 2016, for a review of 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* (as opposed to a fact in nature). A particular causal claim is a 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* asserts 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. Although some causal generalizations in physics might be truly universal (α causes β whenever and wherever α occurs), in social science, causal

generalizations are nearly always realistically restricted to settings with relevant local conditions, γ, that are often vaguely stated or understood (Falleti & Lynch 2009). For example, the GSL could only reasonably expect postcards to work in functioning democracies and among literate participants, 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 certain conditions (Kruglanski & Kroy 1976). We say that 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. Our definition of validity comes from measurement theory originating in Kelly (1927) 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)). In our framework, the general causal claim that "α causes β in γ" is valid if it is indeed the case that α causes β in γ. Hence, validity is a *relationship between a claim and the world* – the relationship that holds if and only if the claim correctly reflects reality (cf. 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

(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 in some conditions, but not in 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 initial causal generalization, "Enforcement messages cause juror turnout," might fail:

(i) Their claim might fail because the researchers misconstrued the nature of the cause, assigning an incorrect semantic label. The postcards' effect might not be due to the specific words but rather because the text was longer in the enforcement compared to reminder message.

(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 misconstrued the nature of the outcome, assigning an incorrect semantic label. Maybe juror turnout was mismeasured – for example, if excused absences were classified as successful recruitments.

(iv) Their claim might fail because the researchers mischaracterized how broadly their claim generalizes. The claim invites the reader to generalize across the U.S. (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 specific times and places. 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, qualitative knowledge, and even common sense – 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 (or often enough cause, or cause absent interference) 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 from an experiment. Our framework distinguishes active causes from inert ingredients as events that are bundled within the measures of the intervention and outcome. That is, we do not take measured 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) 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 for the expression to be true) and $\vee$ logically means "or" (requiring one or both elements to be true for the expression to be true). In the binary case, each variable can be either *true* or *false*.

The Latin letters A, B, C are observed measures of the intervention, outcome, and conditions in settings, respectively, here assumed to be measured without error (see Edwards *et al.* 2015). The Greek letters represent hypothesized causes, effects, causal conditions, and inert elements which combine into informationally-equivalent sets (Dafoe *et al.* 2018). The elements $\{\alpha, \gamma\}$ 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. For simplicity, we omit interactions between elements of the sets.

This notation makes clear that measured variables are inherently bundles (Heckman 2005). In particular, since active and inert ingredients $\{\alpha, \theta_\alpha\}$ are bundled in the intervention, removing the inert ingredient θ_α in equation 4a would make A false. For example, in the first GSL trial, the postcards bundled the enforcement message with an emotional tone, numerals related to the relevant statute, amount of ink, and sentence length and complexity (all varying at least slightly between treatment and control). Likewise, C bundles all of the conditions in the setting $\{\gamma, \theta_\gamma\}$ that remain constant. These include design elements that are identical for treatment and control (e.g., the cardstock and court seal), attributes of the units that are assumed to be balanced through randomization (e.g., employment status of the recipient), and features of the context (e.g., Riverside during the rainy season). Typically, neither these conditions nor C itself is literally "measured"

beyond disclosures of experimental procedures, units, 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 or records (B) assumed to stand in some felicitous causal relation to β. For simplicity, our notation omits cases in which β occurs but remains unmeasured, modeling the risk of false positives but not false negatives.

In an RCT, the evidence is limited to the measured variable bundles. Typically, however, researchers' causal claims reference the ontological events here represented as Greek letters ("enforcement message" and "turnout"), not the measured bundles characterized by Latin letters ("A" and "B"). The definition in 4 makes explicit that the Greek-letter reality behind the Latin-letter measures remains a matter of inference, and this is true even when the causal estimand is "identified" (as in Keele & Minozzi 2013). Under our assumptions, textbook identification establishes the following specific causal claim as a fact about what has actually been manipulated and measured:

$$\text{“}ATE = \mathbb{E}_{\mathcal{S}}[B_i(A_i = 1, C_i)] - \mathbb{E}_{\mathcal{S}}[B_i(A_i = 0, C_i)],\text{”} \tag{5}$$

where C bundles the conditions in the setting, which are either balanced or constant across units. The quotes make it explicit that 5 is a claim. Thus, causal *identification* only warrants claims with respect to the Latin-letter variables – in words, "the average treatment effect is a causal relationship between variables A and B in setting C" – not the Greek-letter causal relata and causal conditions that are present in the world. It follows from definition 4 that the fact of claim 5 is not the same as the (generalized) causal process of interest τ, which is expressed as counterfactuals regarding states of α and γ,

$$\text{“}\tau = \mathbb{E}_{\mathcal{S}}[\beta_i(\alpha_i = 1, \gamma_i)] - \mathbb{E}_{\mathcal{S}}[\beta_i(\alpha_i = 0, \gamma_i)],\text{”} \tag{6}$$

that is, the semantic claim about the ontological process of interest – "enforcement messages cause turnout in the absence of affluence." As our notation makes plain, even if an identification strategy justifies claim 5, that by itself does not justify making the claim about the causal process meaningfully expressed in claim 6. To see this, we expand claim 5 using definition 4 to the equivalent statement,

$$\text{“}ATE = \mathbb{E}_{\mathcal{S}}\{[\beta_i \vee \theta_{\beta_i}]([\alpha_i \wedge \theta_{\alpha_i}] = 1, [\gamma_i \wedge \theta_{\gamma_i}])\} - \mathbb{E}_{\mathcal{S}}\{[\beta_i \vee \theta_{\beta_i}]([\alpha_i \wedge \theta_{\alpha_i}] = 0, [\gamma_i \wedge \theta_{\gamma_i}])\}.\text{”} \quad (7)$$

Plainly, moving from identifying the ATE in 5 to the generalized causal effect τ in 6 requires an extensive set of assumptions about claim 7 that go beyond the assumption of internal validity. Comparing claim 5 to claim 7, expanding A problematizes construct validity of the cause; expanding B problematizes construct validity of the outcome; and expanding C problematizes external validity.[2]

Of course, if researchers literally only care about the measured variables A and B as manifested in the exact setting C, they need not make assumptions about the relationships between A and α, B and β, and C and γ. However, they would then be unable to communicate the meaning of their results beyond saying "Whatever it was we did (designated by the symbol A), at that one time and place (designated by the symbol C), had an effect on whatever it was we measured (designated by the symbol B)" – a claim that would never be published and indeed is not even a generalization (see Cronbach 1982). Since textbook identification strategies are only with respect to the measured variables, identification can only license a claim such as 5. Claim 5 does not license the meaningful semantic statement represented by claim 6 (Cook *et al.* 2014).

Note that a claim such as 5 can be the result of a deductive test, under assumptions

[2]Cronbach (1982) and Shadish *et al.* (2002) propose the UTOS framework for understanding validity regarding the elements of a research design (for extensions, see Findley *et al.* 2021). UTOS is an acronym for "units, treatments, outcomes and settings." While UTOS is helpful in organizing the elements of a research design, it does not serve our purpose in distinguishing the different types of validity since the framework holds that each of the UTOS elements is a construct, *and* it holds that each element is a consideration for evaluating external validity. Thus we cannot rely on UTOS to develop our argument.

of internal validity, using standard procedures of causal identification. It can be what Gelman & Imbens (2013) call a "what if" rather than a "why"-type assessment. In moving *implicitly* from 5 to 6, however, the researcher transforms a "what if" question among 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 identified. When researchers make a generalized causal claim without addressing the relationship between the Latin-letter measured variables and the Greek-letter causal relata and conditions, they are simply hazarding an exploratory guess or "structured speculation" about the underlying causality (Banerjee *et al.* 2017), with the hope that causal knowledge will somehow accumulate coherently from a sequence of results such as 5 (Angrist & Pischke 2010, 23). This hope for accumulation of knowledge under speculation is similar to when earlier generations of applied statisticians made exploratory guesses and offered verbal assurances about control variables achieving internal validity.

The authors of the Credibility Revolution partially acknowledge this limitation, and attempt to place boundaries on claims that are "credibly" established under internal validity through a familiar saying that identification can deductively establish the *effects of a cause* – that is to establish a fact that a cause occurred by comparing outcomes across counterfactual states of the world – but not the *causes of effects* which would require naming the actual causes (Holland 1986). This distinction fails to establish these bounds, however, in that communicating *any* description of the counterfactual states over which a cause and effect are detected requires labeling those states in some way, which, even if those labels are abstract or generic, is still to require an assumption of construct validity. And to assume the cause will occur in any other time and place, or even a generic assumption that causal effects are homogeneous, requires an assumption of external validity. Even if one were to try to communicate a finding modestly as an "effect of cause"-type claim, that would not relieve the analyst from considering these aspects of validity.

Because identification is only with respect to measured variables in one setting, identification does not address parts (i), (iii), or (iv) of a generalized causal claim, that is, the labeling of the relata and the conditions under which the cause will occur. Since a causal generalization is not valid unless all four aspects of the claim are correct, causal identification does not provide sufficient assumptions to deduce a generalized causal claim. Thus we augment causal identification with what we call *causal specification*, which formalizes the insights of Shadish *et al.* (2002) into the single expression of claim 7 that shows the equal importance of each type of validity in deductions regarding causal relata and conditions. In causal specification, construct validity is present when α and β are correctly specified, and external validity is present when γ is correctly specified.

In the remainder of the paper, we expand on the concepts of construct and external validity as they fit within the framework of causal specification, showing the necessity of each type of assumption to preserve the deductivenss of causal claims. The formal approach using the potential outcomes framework, as we have shown here, can be extended for each of internal, construct and external validity. However, we confine those formalizations to appendix A. In addition, for interested readers, appendix B shows our arguments in a structural causal models framework. These appendices can be skipped without loss of continuity.

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 & Collier 2001, 529). In causal analysis, this means the semantic labels assigned to the causal relata are correct (Cook *et al.* 2014). The notion originates in Cronbach & 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,

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).

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, in our framework a causal generalization is construct valid only if α and β, inferred from observations of A and B, are the real underlying cause and effect. As ontological referents, α and β are latent and so 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 (see Kim 1971). Without an explicit assumption and justification for their semantic labels, researchers cannot properly claim to have deduced 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 causal referent (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 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 & 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.

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). Establishing internal validity alone 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 & Grimmer 2019). This is the problem Dafoe *et al.* (2018) identify as "informational equivalence." Instead, as a minimum requirement, a valid generalized 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 & Grimmer 2019). Because the inert and active ingredients perfectly covary within a well-designed RCT, a well-designed RCT cannot by itself distinguish the active from the inert ingredients. Furthermore, some elements might not be entirely ontologically distinct, such as "enforcement" and "threat" in the GSL example. Even if the elements are sufficiently distinguishable, conceptually and empirically, which ingredient best characterizes what is actually driving outcomes remains an open question.

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 & Green 2012). The exclusion restriction and SUTVA allow one to ignore each unit's assignment and the assignment and exposure vector of all other units, and so the two assumptions greatly reduce the number of potential outcomes to consider (Angrist *et al.* 1996). Substantively, these two assumptions rule out certain, but not all, aspects of confounding within the intervention that can remain even when internal validity is perfect (see Julnes 2004).

First consider the exclusion restriction, which assumes that the assignment itself has no direct or indirect effect on the outcome other than through the treatment. Absent blinding, random assignment 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. Although the exclusion restriction labels the assignment process as an inert component, it does not label the active component of the intervention (Julnes 2004, 176).

Second, 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. 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. Neither the exclusion restriction nor SUTVA adequately substitutes for a justified specification of the labels.

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, for example, 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 of 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 $B{=}\beta$. 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 basic notation. In short, 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. Establishing 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.

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 2014; 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,

it is often viewed as an unattainable ideal (Deaton & Cartwright 2018; Findley *et al.* 2021). Very few social science studies yield the same results across *all* settings of human existence (Cook 2014; Julnes 2004). Indeed, despite high internal validity, RCTs usually yield substantially varying results across settings (Deaton 2019; Deaton & Cartwright 2018; Peters *et al.* 2018; Pritchett & Sandefur 2013; Vivalt 2020; Weiss *et al.* 2014). While, of course, inconsistent execution of RCTs also triggers variation in treatment effect estimates, low external validity in the traditional sense is quite common in RCTs even when executed consistently (Deaton & Cartwright 2018; Henrich *et al.* 2010; Ravallion 2012).

Clarifying the Definition of External Validity. Partly because of the pervasive lack of traditionally-defined external validity, we propose a clarification and modest elaboration of the definition of external validity for use in causal specification.

First, we define *causal conditions* (γ) as any active ingredients that are balanced or constant 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," causal conditions can augment or undermine a cause. These conditions include quintessential characteristics of settings like culture and institutions (Cartwright & Hardie 2012; Falleti & Lynch 2009). They 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 (e.g., Pearl 2019), oxygen is an active condition that is necessary for the treatment effect of striking a match to result in the outcome of fire. As we explain below, other conditions are *inert* (e.g., nitrogen in the air).

Second, we clarify that external validity requires the correct *specification* of the conditions that define the causal generalization. Thus, external validity requires evidence or

assumptions about what conditions are needed for 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 one has accurately specified *why* or *how* the effect generalizes across settings. This means specifying the causal conditions defining the range of settings across which the effect generalizes. Hence, we define external validity as the correct specification of the conditions that enable or disable a causal effect (for a similar definition, see Egami & Hartman 2022). External validity is present when that specification is true.

This revised definition is more general than and subsumes the traditional definition. The traditional definition of external validity is a special case of our definition. The traditional definition only considers the case where the conditions γ are widespread.

Our definition is also more attainable. Unlike the traditional definition, we embrace the reality that treatment effects will vary across settings because of the inescapable role of conditions that generally also matter for the outcome (Deaton & Cartwright 2018; Guala 2005; Peters *et al.* 2018; Ravallion 2012, 110; Weiss *et al.* 2014). Our approach reveals that it is as important to know the settings where a cause *will not* occur as to know the settings where it *will* occur. Thus, a deductive and general understanding of causality requires specifying how a causal claim is contingent on specific conditions. While GSL lacked traditionally-defined external validity, the actual problem is that GSL does not understand why the treatment worked in Riverside and not Orange County. However, GSL can attain external validity by correctly specifying which conditions moderate the treatment and define the range of settings. For example, GSL might successfully justify the assumption that the intervention works in the setting of Riverside but not in Orange County by specification of the condition of affluence.

In many ways, our revision embraces and unifies recent efforts to incorporate external validity into causal frameworks. Pearl and colleagues' transportability approach (Bareinboim & Pearl 2016) specifies which conditions are modifying the causal effect (Humphreys

& Scacco 2020). Knowing "where" and "why" in the directed acyclic graph that effect moderation is occurring requires knowledge of conditions in settings. The "inherently subjective," "structured speculation" of Banerjee *et al.* (2017) uses theory and knowledge of conditions to generalize treatment effects across settings. Propensity score approaches require knowing which conditions to include in the propensity score model and on what conditions to compare sample and target population (Pritchett & Sandefur 2013; Stuart *et al.* 2011). Egami & Hartman (2022) and Findley *et al.* (2021) each provide a general framework for external validity, but these frameworks differ from causal specification because they take the measured variables rather than the actual causes and conditions as the objects of interest for causal analysis (Slough & Tyson 2022).

Why External Validity Matters. For several decades, the social sciences have prioritized internal validity over external validity (Findley *et al.* 2021; Julnes 2004; Pritchett & Sandefur 2013). As result, the credibility revolution often sidesteps external validity by simply declaring a causal effect is "local."

However, if one has identified only a local causal effect without external validity, one is limited to the hope that causal knowledge can accumulate, but with no guidance on how the accumulation can occur. Absent external validity, any claim must be 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 2019; Gailmard 2021; Guala 2005; Rothman 1976; VanderWeele & Hernán 2006; Vivalt 2020). Cartwright (2011) explains internal validity only shows "it works somewhere." Actually, internal validity only shows it historically worked (in the past tense) somewhere (Nosek & Errington 2020). As Cronbach (1982, 137) explains, internal validity alone is "trivial, past-tense, and local." This is not the sort of general causal knowledge social scientists typically want to accumulate (Findley *et al.* 2021). Cartwright (2011) explains, we want to know: "it works widely" or "it will work for us" (Cartwright & Hardie 2012; Cook 2014; Deaton & Cartwright 2018; Pritchett

& Sandefur 2013). For this exact reason, Rubin (1974) originally stressed the need for "subjective random sampling" of settings to ensure a study was of "practical interest," "representative" and "useful."

Confronted with this, some might claim that identifying an effect in one setting is sufficient and they have no intention to produce "universal" knowledge. At least implicitly, such a claim asserts general knowledge can accumulate as if the conditions in settings are irrelevant (see Angrist & Pischke 2010, 23). They may even say that one setting at one time defines their "population" and their sample generalizes to that population. We propose that by claiming they only intend to make a specific historical claim about the effect of something in only one setting – which is actually a time and place with certain very specific conditions – they are retreating to what we call *the historicist's refuge*.

Historians' idiographic causal narratives of specific events are certainly valuable, and indeed, as Kocher & Monteiro (2016) note, even essential for developing and justifying research designs for natural experiments. Nevertheless, we doubt that social scientists truly have no desire to be different from historians (Findley *et al.* 2021; Henrich *et al.* 2010). As Nosek & Errington (2020, 3) explain, social scientists 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, we choose topics to study because they are instances of some generalization. For example, GSL's study was an instance of the general phenomena of jury service or democratic participation. When researchers intentionally choose topics to understand general phenomena, it is not credible to back out of the generalization by declaring post-hoc that causal effects are "local."

However, if one is truly only making a "local" claim in the historicist's refuge, this would require explicit language. Just as the credibility revolution requires causal identification for any language of causal effects, historicists should declare their inability to generalize *to any setting other than the one experimental setting at the one time when the*

experiment was actually conducted. Readers would need to police against any language of generalization just like readers currently police against causal language for descriptive research. It seems fair to note that such a practice would probably require fairly dramatic changes to studies of the U.S., which are rarely forced to justify selecting the highly unusual U.S. case (Findley *et al.* 2021; Henrich *et al.* 2010).

Even with careful language however, the historicist's refuge cannot lead to a coherent, deductive understanding of causality (Cartwright & Hardie 2012). Historicists in their refuge claim to identify a causal effect while having no understanding of the conditions in the setting enabling that effect. One does not know how much of the effect is due to the treatment or some complex interaction between the treatment and conditions in the setting. One does not even know what the relevant conditions might be. Hence, the lack of external validity reveals a lack of understanding about what really causes what.

Further, any commitment to replication forces one to inevitably abandon the historicist's refuge. If even one other setting yields a different result, beyond sampling variability, this proves conditions in settings matter. Once GSL found a different result in Orange County, they had no way of knowing if the treatment effect is aided by helping factors in Riverside County or suppressed by countering causes in Orange County (or both). Unknown conditions might even make both Riverside and Orange unusual. If the conditions are unusual, then the causal effect could be unusually large or small. Any broader generalization would suffer from a selection bias just like any sample selection bias (Findley *et al.* 2021).

In response to these challenges, some might admit a lack of external validity and say the "next step" is to go forth inductively across a range of settings. For instance, Banerjee & 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 feasible however without causal specification (Cartwright & Hardie 2012). Sampling a "range of settings" or "similar settings" presumes

one knows what defines the range or similarity. The law of large numbers does not ensure representativeness if one is sampling from a corner of the sample space, and "simple enumerative induction" does not warrant claiming the treatment "reliably promotes" outcomes (Cartwright 2011). Generalizing beyond one case requires causal specification about what enables the cause (Banerjee & Duflo 2009; Deaton & Cartwright 2018). In the GSL vignette, choosing Orange County as the next step is merely haphazard without some understanding of what conditions might be relevant.

Ultimately, neglecting external validity raises concerns very similar to the "radical skepticism" about unknown confounding (Stokes 2014). Verbal assurances of external validity strain credibility just like verbal assurances about control variables (Deaton & Cartwright 2018). Radical skeptics about unknown confounding should be similarly radically skeptical about any generalization based on unknown conditions.

The Role of Theory and Qualitative Knowledge in Quantitative Causal Inference

Our framework for causal specification clarifies the assumptions that must be added to prevailing causal frameworks and textbook identification strategies in order to support credible deductive claims about causal effects. In the absence of construct and external validity, the researcher converts a deductive "what if" question to an exploratory "why" question. Researchers might implicitly speculate that α is the active ingredient in A, that B accurately tracks β, and that the causal conditions γ in C operate similarly elsewhere. These however are key assumptions that underwrite causal claims of the nature of the cause, the nature of the effect, and the scope of the generalization. *Construct and external validity are as necessary as internal validity for a deductive understanding of causality*, and all three are equally essential for the credibility of causal claims.

Since assumptions are not themselves testable (absent specifying additional assumptions) causal specification is not established by any statistical procedure, boundary con-

struction or sensitivity analysis. Of course, there exist design-based statistical procedures that can *warrant* assumptions regarding construct and external validity, similar to how randomization can warrant a claim of internal validity. For example, a conjoint experiment (Hainmueller & Hopkins 2014) can test for the causal effect of specific (active ingredient) components contained within an intervention, but setting up the test requires construct valid assumptions about the cause or possible causes. And multisite studies (Dunning *et al.* 2019) can test for variation in the treatment effect across settings, but to know which sites to choose requires externally valid assumptions about variation in the underlying conditions that enable or disable the cause.

Ultimately, there are no statistical procedures that can solve the ontological problems of how to assign semantic labels to causes, outcomes and conditions (Kim 1971). Instead, as Kocher & Monteiro (2016, 953) and Slough (2022) emphasize, the assumptions that underwrite a design for causal inference are not statistical but instead are derived from theory and qualitative or historical knowledge. For example, process tracing methods qualitatively depict a "snapshot" of a causal event at one point in time and an understanding of unfolding of events over time (Collier 2011; Collier *et al.* 2010). As Beach (2017, 9) writes, the purpose of process tracing is not to assess "the difference that changes in values of X make for values of Y. Instead, inferences are made using the correspondence between hypothetical and actual observable manifestations of the operation of mechanisms within a selected case." Similarly, Fenno (1978) advocates for a "soak and poke" approach to qualitative understanding, while Weller & Barnes (2014, 21) advocate for qualitative pathway analysis in mixed methods research in order to understand generalized causal links across settings.

In each of these qualitative methodological approaches, the analyst can develop a theoretical and substantive understanding of the underlying referents that correspond to the actual causal relata and the relevant conditions in settings. To say that quantitative causal inference relies on qualitative knowledge to produce claims is not to say however

that either approach somehow subsumes the other. Indeed, as Beach & Kaas (2020) note, in many ways the two modes of causal inquiry are incommensurable. For example, process tracing can only recover what was observed in a causal event, but not the event's counterfactual. And quantitative methods can only recover treatment effects averaged over units, not the causal effect for a given unit. Deductive approaches to quantitative causal inference *require* theory and qualitative knowledge, however, because theory and qualitative knowledge enable the researcher to choose and recognize relatively plausible versus implausible background assumptions concerning all three types of validity. Acknowledging this requires acknowledging the importance of qualitative research for scientific progress in quantitative research (Kocher & Monteiro 2016).

Further, to say that generalized quantitative causal inference relies on assumptions about the causal conditions and relata does not somehow make the approach we recommend unscientific. Indeed, as practitioners of the credibility revolution well know, internal validity also remains an assumption even after randomization and even after balance tests have been passed. These assumptions at the core of causal specification are simply and fundamentally necessary to preserve the deductiveness of (scientific) causal claims.

Conclusion

Social scientists typically aim to produce general knowledge about what causes what in what conditions, and not just historical knowledge that something caused something one time in one setting in the past. That is, social scientists aim for valid causal generalizations. The tight linkage between the concept of internal validity and the concept of causality is encoded in the causal frameworks that have governed the Credibility Revolution. The Rubin causal model (Holland 1986; Rubin 1974) and structural causal models (Pearl 2000) have made tremendous contributions while being centered on the problem of unconfoundedness and internal validity. However, the Credibility Revolution has hereto-

fore provided insufficient consideration of external validity or construct validity. As a result, it lacks an adequate general framework for validity and causal generalization.

In our causal specification framework, 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 (the focus of internal validity) but equally the challenge of correctly labeling 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 has priority.

We show that the textbook identification assumptions within "credible designs" are insufficient for deducing generalized causal claims, irrespective of whether one is working in the SCM or the RCM. Textbook identification focuses on internal validity but typically neglects assumptions regarding construct and external validity. Social scientists who wish to make deductive and generalized causal claims must 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 on qualitative, 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 generalized causal claim as are assumptions of internal validity.

We recommend that researchers who wish to make general contributions to our understanding of causal processes explicitly specify their αs, βs, and γs and defend the inertness of their θs. This requires quantitative scholars to attend to theory or to qualitative epistemologies such as process tracing, pathway analysis or soak and poke in order to justify assumptions, or to collaborate with scholars who specialize in qualitative methods. For researchers who already take construct and external validity explicitly into account, this might amount to only a more formal statement of their assumptions. For other researchers, like GSL, this might require confronting and clarifying causal assumptions that they would otherwise disregard and leave implicit in the background.

If applied researchers ignore construct and external validity when stating causal claims, they mistakenly convert an intended deductive claim into a claim based on exploration and speculation – contrary to the fundamental 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 – more importantly – at the level of relata and conditions. In this way, causal specification clarifies the additional assumptions required for the Credibility Revolution to achieve its aspirations of understanding causal effects.

References

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

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

Angrist, Joshua D, & Pischke, Jörn-Steffen. 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, Joshua D., Imbens, Guido W., & Rubin, Donald B. 1996. Identification of Causal Effects using Instrumental Variables. *Journal of the American Statistical Association*, **91**(June), 444–455.

Angrist, Joshua D., Dynarski, Susan M., Kane, Thomas J., Pathak, Parag A., & Walters, Christopher R. 2012. Who Benefits from KIPP? *Journal of Policy Analysis and Management*, **31**(4), 837–860.

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

Banerjee, Abhijit, Duflo, Esther, Goldberg, Nathanael, Karlan, Dean, Osei, Robert, Parienté, William, Shapiro, Jeremy, Thuysbaert, Bram, & Udry, Christopher. 2015. A multifaceted program causes lasting progress for the very poor: Evidence from six countries. *Science*, **348**(6236), 1260799–1.

Banerjee, Abhijit V., & Duflo, Esther. 2009. The Experimental Approach to Development Economics. *Annual Review of Economics*, **1**(1), 151–178.

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

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

Beach, Derek. 2017. Process-Tracing Methods in Social Science. *Oxford Research Encyclopedia of Politics*, 1.

Beach, Derek, & Kaas, Jonas Gejl. 2020. The Great Divides: Incommensurability, the Impossibility of Mixed-Methodology, and What to Do about It. *International Studies Review*, **22**(2), 214–235.

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

Borsboom, Denny, Mellenbergh, Gideon J., & van Heerden, Jaap. 2004. The Concept of Validity. *Psychological Review*, **111**(4), 1061–1071.

Bowler, Shaun, Esterling, Kevin, & Holmes, Dallas. 2014. GOTJ: Get Out the Juror. *Political Behavior*, **36**, 515–533.

Card, David. 2022. Design-Based Research in Empirical Microeconomics. *American Economic Review*, **112**(6), 1773–81.

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

Cartwright, Nancy, & Hardie, Jeremy. 2012. *Evidence-Based Policy: A Practical Guide to Doing It Better*. New York, N.Y.: Oxford University Press.

Collier, David. 2011. Understanding Process Tracing. *PS: Political Science & Politics*, **44**(4), 823–830.

Collier, David, Brady, Henry E., & Seawright, Jason. 2010. Outdated Views of Qualitative Methods: Time to Move On. *Political Analysis*, **18**(4), 506–513.

Cook, Thomas. 2012. Causal Generalization: How Campbell and Cronbach Influenced My Theoretical Thinking on This Topic, Including in Shadish, Cook, and Campbell. *Pages 89–112 of:* Alkin, Marvin C. (ed), *Evaluation Roots*. SAGE Publications, Inc.

Cook, Thomas D. 2014. 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, Thomas D., Tang, Yang, & Seidman Diamond, Shari. 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, Lee J. 1982. *Designing Evaluations of Educational and Social Programs*. San Francisco: Jossey-Bass Publishers.

Cronbach, Lee J., & Meehl, Paul E. 1955. Construct Validity in Psychological Tests. *Psychological Bulletin*, **52**(4), 281–302.

Dafoe, Allan, Zhang, Baobao, & Caughey, Devin. 2018. Information Equivalence in Survey Experiments. *Political Analysis*, **26**(4), 399–416.

Deaton, Angus. 2019. Randomization in the Tropics Revisited: A Theme and Eleven Variations. *In:* Bedecarrats, Flortent, Guerin, Isabelle, & Rouboud, Francois (eds), *Randomized Control Trials in the Field of Development*. New York, N.Y.: Oxford University Press.

Deaton, Angus, & Cartwright, Nancy. 2018. Understanding and Misunderstanding Randomized Control Trials. *Social Science and Medicine*, **210**, 2–21.

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

Dunning, Thad, Grossman, Guy, Humphreys, Macartan, Hyde, Susan D., McIntosh, Craig, & Nellis, Gareth. 2019. *Information, Accountability and Cumulative Learning*. New York, N.Y.: Cambridge University Press.

Edwards, Jessie K., Cole, Stephen R., & Westreich, Daniel. 2015. 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.

Egami, Naoki, & Hartman, Erin. 2022. Elements of External Validity: Framework, Design, and Analysis. *American Political Science Review*, **Forthcoming**.

Falleti, Tulia G., & Lynch, Julia F. 2009. Context and causal mechanisms in political analysis. *Comparative Political Studies*, **42**(9), 1143–1166.

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

Fenno, Richard F. 1978. *Homestyle: House Members in Their Districts*. Boston, Mass.: Little, Brown and Co.

Findley, Michael G., Kikuta, Kyosuke, & Denly, Michale. 2021. External Validity. *Annual Review of Political Science*, **in press**, 1–51.

Fong, Christian, & Grimmer, Justin. 2019. Causal Inference with Latent Treatments. *Working Paper, Stanford University*.

Gailmard, Sean. 2021. Theory, History and Political Economy. *Journal of Historical Political Economy*, **1**(1).

Gelman, Andrew, & Imbens, Guido. 2013 (11). *Why ask Why? Forward Causal Inference and Reverse Causal Questions*. Tech. rept. 19614. National Bureau of Economic Research.

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

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

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

Hainmueller, Jens, & Hopkins, Daniel J. 2014. Public Attitudes Toward Immigration. *Annual Review of Political Science*, **17**(1), 225–249.

Heckman, James J. 2005. The Scientific Model of Causality. *Sociological Methodology*, **35**(1), 1–97.

Henrich, Joseph, Heine, Steven J., & Norenzayan, Ara. 2010. The weirdest people in the world? *Behavioral and Brain Sciences*, **33**(2-3), 61–83.

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

Humphreys, Macartan, & Scacco, Alexandra. 2020. The Aggregation Challenge. *World Development*, **127**, 104806.

Imai, Kosuke, King, Gary, & Stuart, Elizabeth A. 2008. Misunderstandings among Experimentalists and Observationalists about Causal Inference. *Journal of the Royal Statistical Society, Series A (Statistics in Society)*, **171**(April), 481–502.

Imbens, Guido W., & Rubin, Donald B. 2015. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction.* New York, N.Y.: Cambridge University Press.

Jiménez-Buedo, María. 2011. 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, Luke. 2015. The Statistics of Causal Inference: A View from Political Methodology. *Political Analysis*, **23**(3), 313–335.

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

Kelly, Truman Lee. 1927. *Interpretation of Educational Measurements.* Yonkers-on-Hudson, N.Y.: World Book.

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

Kocher, Matthew A., & Monteiro, Nuno P. 2016. Lines of Demarcation: Causation, Design-Based Inference, and Historical Research. *Perspectives on Politics*, **14**(4), 952–975.

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

Lewis, David. 1973. Causation. *Journal of Philosophy*, **70**, 556–567.

Lundberg, Ian, Johnson, Rebecca, & Stewart, Brandon M. 2021. What is Your Estimand? Defining the Target Quantity Connects Statistical Evidence to Theory. *American Sociological Review*, **in press**.

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

Manski, Charles. 1995. *Identification Problems in the Social Sciences.* Cambridge, MA: Harvard University Press.

Morgan, Stephen L., & Winship, Christopher. 2015. *Counterfactuals and Causal Inference: Methods and Principles for Social Research.* 2nd edn. New York, N.Y.: Cambridge University Press.

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

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

Nosek, Brian A., & Errington, Timothy M. 2020. What is replication? *PLOS Biology*, **18**(3), e3000691.

Paul, L.A., & Hall, Ned. 2013. *Causation: A User's Guide*. Oxford: Oxford University Press.

Pearl, Judea. 2000. *Causality: Models, Reasoning and Inference*. 2 edn. New York, N.Y.: Cambridge University Press.

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

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

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

Petersen, Maya L., & van der Laan, Mark J. 2014. Causal Models and Learning from Data: Integrating Causal Modeling and Statistical Estimation. *Epidemiology*, **25**(3), 418–426.

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

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

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

Rothman, KJ. 1976. Causes. *American Journal of Epidemiology*, **104**, 587–592.

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

Samii, Cyrus. 2016. Causal Empiricism in Quantitative Research. *The Journal of Politics*, **78**(3), 941–955.

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

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

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

Slough, Tara. 2022. Phantom Counterfactuals. *American Journal of Political Science*, **Forthcoming**.

Slough, Tara, & Tyson, Scott A. 2022. External Validity and Meta-Analysis. *American Journal of Political Science*, **Forthcoming**.

Stokes, Susan. 2014. A Defense of Observational Research. *In:* Teele, Dawn Langan (ed), *Field Experiments and Their Critics*. New Haven, Conn.: Yale University Press.

Stuart, Elizabeth A., Cole, Stephen R., Bradshaw, Catherine P., & Leaf, Philip J. 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, Jacqueline A. 2009. The multiplicity of experimental protocols: A challenge to reductionist and non-reductionist models of the unity of neuroscience. *Synthese*, **167**(3), 511–539.

VanderWeele, Tyler J., & Hernán, Miguel A. 2006. From counterfactuals to sufficient component causes and vice versa. *European Journal of Epidemiology*, **21**(12), 855–858.

Vivalt, Eva. 2020. How Much Can We Generalize from Impact Evaluations? *Journal of the European Economic Association*, **in press**.

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

Weller, Nicholas, & Barnes, Jeb. 2014. *Finding Pathways: Mixed Method Research for Studying Causal Mechanisms*. New York, N.Y.: Cambridge University Press.

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

A Appendix A: Formal Statements of Construct and External Validity

Recall that we stated our general causal claim as the sentence: "α causes β in γ." Causal specification requires assumptions about each of these aspects of a causal process as well as their relationship. We formalize those assumptions in this appendix.

Causal Specification for Internal Validity. In an RCT, identification requires internal validity, that is, that the value of β under the counterfactual of α being either true or false is in fact unrelated to the realized value of α within an experiment, or

$$\text{“}[\beta_i(\alpha = 1), \beta_i(\alpha = 0)] \perp\!\!\!\perp \alpha_i,\text{”} \tag{8}$$

which is analogous to equation 1, except it is stated at the level of the causal relata, and we enclose it in quotes to highlight its status as a claim that might depart from the truth.

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

$$\text{“}\tau_{CVCW} = \mathbb{E}_{\mathcal{S}}\{B_i([(\alpha_i = 1) \wedge \theta_{\alpha_i}], C_i)\} - \mathbb{E}_{\mathcal{S}}\{B_i([(\alpha_i = 0) \wedge \theta_{\alpha_i}], C_i)\} \;\forall\; \theta_\alpha \tag{9a}$$

$$\tau_{CVCW} \neq 0,\text{”} \tag{9b}$$

where the $\forall$ symbol means "for each" – that is, the cases where θ_α is either true or false, ignoring cases in which θ_α is true on one side of the statement and false on the other. When the claim is about a direction of the causal effect, such as if GSL were to claim that the postcards increase turnout, the inequality in statement 9b should be directional using either $>$ or $<$, depending on the direction. Under this claim, the causal effect τ compares expected potential outcomes when α is present to when α is absent, both when θ_α is present and when it is not. The claim in 9 holds that the cause the researcher postulates to be the actual cause is in fact an actual cause. In other words, in order to support a deduced causal generalization, the cause α must be *specified*. If Claim 9 is false, 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. In particular, to enable valid generalized causal claims, the RCM would need to relax the requirement that potential outcomes are defined over measured variables only (Edwards *et al.* 2015). Claim 9 demonstrates the inadequacy of the exclusion restriction and SUTVA as a substitute for construct validity. 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 the intervention, without specifying α.

A claim of *strong construct validity of the cause* would add the following to claim 9:

$$\text{“}0 \approx \mathbb{E}_{\mathcal{S}}\{B_i([\alpha_i \wedge (\theta_{\alpha_i} = 1)], C_i)\} - \mathbb{E}_{\mathcal{S}}\{B_i([\alpha_i \wedge (\theta_{\alpha_i} = 0)], C_i)\} \;\forall\; \alpha.\text{”} \tag{10}$$

When α is present, the expectation of B is the same irrespective of whether θ_α is present

or absent, and likewise when α is absent. Under our background assumptions, unless this statement is true the claimed inert ingredient "θ_α" is not the real inert ingredient θ_α, and hence strong construct validity is lacking. 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.

Causal Specification for External Validity. Our definition of external validity relies on understanding conditions as all features of the setting, units, or design that are constant or balanced between values of α. Recall we define the conditions in the setting, $C \triangleq (\gamma \wedge \theta_\gamma)$. That is, C is true if both γ and θ_γ are true; γ are the active ingredients in the setting that enable or disable the cause α, and θ_γ are the inert ingredients that are also in the setting. Under our definition, the causal conditions γ must be correctly specified to make a valid causal generalization. The presence of θ_γ clarifies there are features that are constant or balanced, many of which are ignorable. A claim of *weak external validity* is

$$\text{“}\tau_{EV} = \mathbb{E}_{\mathcal{S}}\{B_i(A_i, [(\gamma_i = 1) \wedge \theta_{\gamma_i}])\} - \mathbb{E}_{\mathcal{S}}\{B_i(A_i, [(\gamma_i = 0) \wedge \theta_{\gamma_i}])\} \;\forall\; \theta_\gamma\text{,”} \tag{11a}$$

$$\tau_{EV} \neq 0\text{.”} \tag{11b}$$

Note the close parallel with weak construct validity of the cause in claim 9. When the claim is about a direction of the causal effect, such as in GSL's claim that affluence reduces the effect of the enforcement message, the inequality in statement 11b should be directional using either $>$ or $<$, depending on the direction. The causal effect of A on B depends on whether γ is present or absent, regardless of whether θ_γ is present or absent. The RCM 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 the notion of "transportability" described in Bareinboim & Pearl (2016). However, as the appendix shows, we clarify that 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 11:

$$\text{“}0 \approx \mathbb{E}_{\mathcal{S}}\{B_i(A_i, [\gamma_i \wedge (\theta_{\gamma_i} = 1)])\} - \mathbb{E}_{\mathcal{S}}\{B_i(A_i, [\gamma_i \wedge (\theta_{\gamma_i} = 0)])\} \;\forall\; \gamma\text{.”} \tag{12}$$

The strong external validity claim adds that θ_γ is irrelevant to the expected outcome, provided that A is present and γ is constant. If the equality is false the claimed inert condition "θ_γ" is not a real inert condition θ_γ, and strong external validity is absent.

B Appendix B: DAG Representation

In this appendix we approximate the argument in our text using directed acyclic graphs (DAGs) (Pearl 2000). A DAG cannot represent the full argument for two reasons. First, as we show in definition 4 of the main text, we conceive of the measured variables, A, B and C, as bundles of active and inert causal ingredients. In this sense, the measured variables

are compositions. While the measured variables are bundles and hence not exactly the causes, neither are the measured variables the *effects* of the causes – the part does not necessarily cause the whole, nor does the whole necessarily cause the part. However, under the consistency rule (Pearl 2010, 872), only causal nodes are permissible within a DAG, and hence the compositions that are at the core of our definition of validity are not permitted. As a result, a DAG lacks this flexibility and cannot represent the part-whole relationship as a subset relationship. Instead, it has to represent part-whole as either the parts causing the whole or the whole causing the parts. Given this limitation, we can only approximate our framework in a DAG by also assuming the elements themselves – the active and inert ingredients – cause the measured variables (Borsboom *et al.* 2004), which is in the traditional framework of measurement theory but not fully consistent with our causal specification framework.

Second, it is well-known that DAGs cannot visually represent an effect modification (Pearl 2019), which is also central to our model of causality. Instead, the DAG can only reference a separate formal statement of the effect modification, such as the one we provide in claim 7 of the main text.

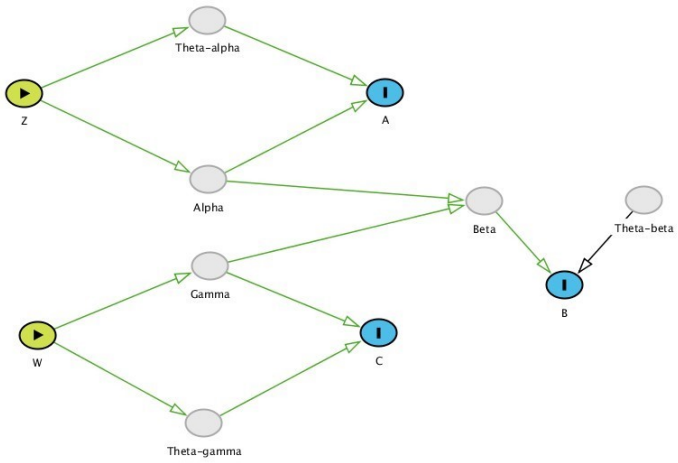


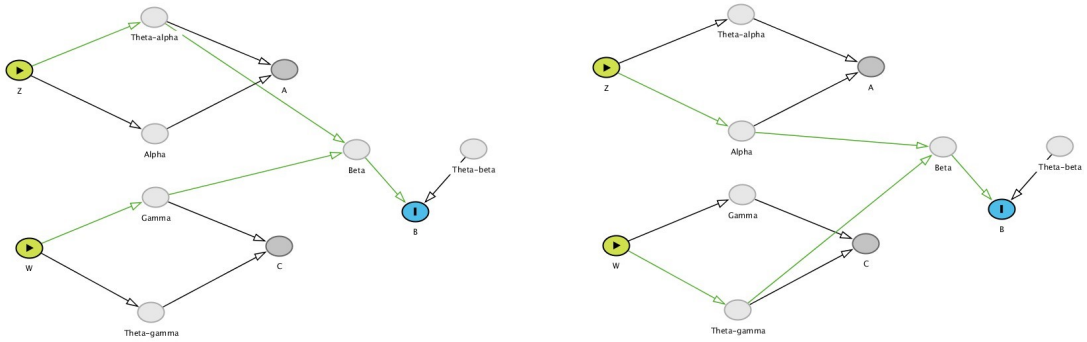
Figure 1: Preferred model for the claim "α causes β in γ" specified within the causal process 7 of the main text. Grey nodes are unobserved causes, effects and conditions of interest. Blue nodes with an "I" are measured variables. Yellow nodes with an arrow indicate *do* commands; Z is an assignment mechanism and W is a choice of conditions in the setting.

The DAG in figure 1 is a representation of the generalized causal claim "α causes β in γ" as defined in claim 7 of the main text; this representation is *valid* if it corresponds to nature. For completeness we introduce two assignment mechanisms that we leave implict in the main text: Z is the assignment to treatment and control and W is a choice over the conditions in the setting such as the characteristics of the units, the research design and the time and location; the $do()$ operator is indicated by an arrow inside of the yellow nodes. All of the other nodes are defined in the text. Grey nodes are unobserved or

latent and are represented by Greek letters. The blue nodes with an "I" represented by Latin letters are measured variables and hence are outcomes of a measurement process (Borsboom *et al.* 2004), and hence this figure is consistent with the measurement view of the observed variable bundles. Among the latent nodes, α and γ are "active" ingredients in that they cause the outcome of interest β. 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, but they can affect the measurement of the observed variable.

The diagram represents the causal process α causes β in γ that we represent in claim 7; that is, γ is an effect modifier that is necessary for the cause associated with α to occur. An ideal experiment would execute a $do(\alpha)$ procedure, in both the presence and absence of γ, but since α and γ are ontological referents (that is, events in nature that we ordinarily do not observe directly) such a procedure typically is not possible. Thus, the causal relationship between the relata (α and β) and the causal conditions (γ) can only be assumed, and the validity of those assumptions depends on their correspondence with the truth that resides in nature.

The DAGs are useful to demonstrate that a strongly valid generalized causal claim based on an observed statistical relationship between A and B requires all of the Greek letter nodes to be specified correctly. To support a weakly valid causal claim, the $\boldsymbol{\theta}$ vector does not need to be specified.



(a) No Construct Validity

(b) No External Validity

Figure 2: DAG representation of causal processes where the claim "α causes β in γ" lacks validity

Figure 2 shows DAG representations of (true) causal processes where the (semantic) claim "α causes β in γ" lacks validity. That is, in this figure, the DAGs are not a claim but instead are a representation of the ontological causal process. The left panel 2a shows when the claim "α causes β in γ" lacks construct validity of the cause. The right panel 2b shows when that claim lacks external validity. Note that in each case, the claim does not match the causal process found in nature, and hence is not valid. Clearly, internal validity that results from a $do()$ operation on Z or W is not sufficient to ensure a valid causal claim.

The DAG in figure 1 shows that a valid generalized causal statement is never with

respect to the measured variables since this would execute the $do()$ operator on an outcome of the measurement process (which is a collider variable) rather than on the cause of interest. For example, consider the consequence of erroneously taking the measured variable A to be the actual cause. In this case, placing the $do()$ operator on A demonstrates that the causal paths are not able to recover the causal effect of interest; since A is itself an outcome, the $do()$ operator in this case does not send any information along the causal path. Placing the $do()$ operator on a given node deletes the arrows that point toward the node. Since A is a collider, this results in A simply disconnecting from the graph. The analogous problem occurs when taking C as the necessary conditions instead of γ.

Note that in all of these figures we are using a $do()$ operator, and hence the assumption of internal validity holds in each case, but even then the causal effect cannot be recovered or understood deductively without causal specification.