Sign-Congruence, External Validity, and Replication*

Tara Slough† Scott A. Tyson‡

October 23, 2022

Abstract

We develop a framework for accumulating evidence across studies and apply it to understand the theoretical foundations of replication. We focus on two ways of assessing empirical results across studies: target-equivalence, where empirical targets across studies are the same, and target-congruence, where the sign is the same across studies. Our results show how each of these assessment criteria are related to distinct formulations of external validity. We stress the importance of research design harmonization when accumulating evidence across studies, which holds aspects of a research design fixed across settings, and ensures that external validity questions can be addressed using replication.

*We thank Scott Abramson, Gleason Judd, Walter Mebane, John Patty, Dan Posner, Pia Raffler, Cyrus Samii, Anna Wilke, and participants at Polmeth XXXIX, the Knowledge Accumulation and External Validity: Implications for Design and Analysis workshop at the University of Montreal, and the 2022 American Political Science Association annual conference for invaluable comments.

†Assistant Professor, New York University, tara.slough@nyu.edu

‡Associate Professor, Department of Quantitative Theory and Methods, Emory University, s.tyson@emory.edu

Accumulating empirical evidence about a phenomenon that manifests in multiple places, at different times, and is measured by different scholars is a critical step toward the production of substantive knowledge. Without such knowledge, careful and credible empirical work may seem highly particular, and because of this, developing methods for accumulating causal evidence should be a goal for any research community (Deaton, 2010; Deaton and Cartwright, 2018). An important tool toward this goal is *replication*, where the same substantive question is addressed by comparing the results of different studies (Banerjee and Duflo, 2009; Gerber and Green, 2012; Dunning, 2016). However, determining what features and considerations make the comparison of empirical evidence across studies productive is unclear since there is no general understanding—or best practices—to guide such efforts.[1] In this article, we develop a framework to highlight key concepts to help understand replication's role in the accumulation of empirical evidence.

In our framework, an empirical study measures the influence of a mechanism (or set of mechanisms) by assessing the "effects of causes" (Holland, 1986). A study consists of three key ingredients. First, each study includes a *contrast*, which defines the comparison of interest and consists of two values of an instrument, such as treatment/control (Imbens and Angrist, 1994). Second, conducting a study involves a *measurement strategy*, which encapsulates all considerations that go into measuring the effect of a contrast, such as the choice of an outcome and the various techniques involved in its measurement. Third, the *setting* gives the contextual features that are relevant to the empirical assessment of a mechanism, such as the time/place/population a study was conducted. These three ingredients combine to define an *empirical target*, or treatment effect, which corresponds to a study's primary estimand.[2]

Comparing estimates from two studies of the same phenomenon—as is the goal in a replication study—is challenging because there are multiple reasons that the estimates in these studies might differ. First, and as is well known, statistical noise stemming from random samples or chance imbalances in treatment assignment ensure that any two (realized) estimates will be different, leading to *statistical discrepancies*. In addition, we derive two *non-statistical* reasons that estimates may differ, which emerge from study design features and a mechanism's external validity. These are fundamentally *theoretical* concerns which are important because they determine whether different studies are "aiming at the same target" and thus speak to the same substantive question.

We develop two concepts to describe the theoretical relationship between constituent studies. First, following Slough and Tyson (2022), two studies are *target-equivalent* when they measure

[1] This may be part of the reason replication is undervalued by experimental economists (Rubinstein, 2001: pg. 625-626).

[2] When the estimand from a research design corresponds to the empirical target, they are *commensurate* (Bueno de Mesquita and Tyson, 2020; Ashworth, Berry, and Bueno de Mesquita, 2021).

the same empirical target, here a treatment effect. Second, and novel to this article, two studies are *target-congruent* when their empirical targets (treatment effects) have the same sign (positive or negative). We also develop two formal definitions of external validity (of a mechanism). First, a mechanism has *external validity* if it produces the same empirical target in different settings under an otherwise identical experiment (i.e., same contrast and measurement strategy).[3] Second, a mechanism has *sign-congruent external validity* when it produces an empirical target with the same sign in different settings. External validity is a stronger condition as it implies sign-congruent external validity, whereas a mechanism with sign-congruent external validity need not be externally valid.

When a replication study is conducted in different settings, at different times, and on different samples, it may not address the same empirical target as the original study. The *target discrepancy* between two studies measures the extent to which a mechanism produces a different effect in different settings (holding fixed other aspects of the research design). It reflects the degree to which external validity holds (or fails) between two settings, and we show that when a mechanism has external validity then the target discrepancy across studies is zero. Sign-congruent external validity allows for target discrepancies but constrains which types kind can arise.

Novel to our framework is the observation that different constituent studies often have different research designs, e.g., different treatments or different measurement strategies. Such differences in research designs produce *artifactual discrepancies* between empirical targets because they make different comparisons or measure things differently. The artifactual discrepancy between studies measures the extent to which empirical targets between two studies are not the same but for reasons that are distinct—and orthogonal—to issues of external validity. For example, if the contrasts in two studies are different, then studies implicitly make different comparisons, which leads to differences in observed treatment effects. When two studies employ the same contrast and measurement strategy, we say they are *harmonized*, and show that by harmonizing two studies, researchers can eliminate artifactual discrepancies. Artifactual discrepancies may also reflect constraints, e.g., measuring the influence of a mechanism under the same conditions may be impossible in some cases.

We show that evaluating a mechanism's external validity (or sign-congruent external validity) is a more demanding endeavor than is typically acknowledged (albeit informally). Our main results connect different notions of external validity and harmonization to target-equivalence and target-congruence. First, we show that a collection of studies is target-equivalent (meaning they have the same empirical target) if and only if all of the studies are externally valid and harmo-

[3] See Slough and Tyson (2022: Definition 7) and their discussion of external validity.

nized. Second, we show that a collection of studies is target-congruent (meaning their empirical targets have the same sign) if and only if all of the studies satisfy sign-congruent external validity and all are harmonized. The latter clarifies the theoretical foundations for qualitative comparisons of the form: "author A's study finds that X increases Y, whereas we find no evidence that X increases Y." Such comparisons implicitly invoke an expectation that similar things will be observed if probed empirically, but take for granted how differences in design can undermine such conclusions. Our results, when taken together, highlight how different ways of assessing empirical targets correspond to different notions of external validity.

A large majority of the literature on replication and external validity focus almost exclusively on statistical issues that arise when combining evidence across studies, or worse, assume that the kinds of theoretical issues highlighted by target and artifactual discrepancies can be conceptualized as statistical issues. To stress how this approach can be misleading, we include statistical noise in our framework and show that there is a tradeoff when increasing the number of studies considered in a replication. Specifically, we show that although increasing the number of studies alleviates the influence of idiosyncratic—or random—error in observation, it also *magnifies* the influence of artifactual discrepancies that arise when research designs are not harmonized across studies. Moreover, because researchers cannot distinguish random error from artifactual discrepancies, this severely limits their ability to isolate and measure the true influence of a substantive mechanism in practice. These results suggest that the guidance to "do more studies" as a method of mitigating potential problems arising from combining studies may be underappreciating the downsides of such an approach absent additional guidance on the design of replication agendas (Banerjee and Duflo, 2009; Gerber and Green, 2012; Dunning, 2016).

We then assess the properties of two common statistical tests that are used in replication studies. The first, the *estimate-comparison test*, examines the difference in point estimates from constituent studies, thus probing target-equivalence. The second, the *sign-comparison test*, probes target-congruence by examining the signs of estimates from different studies. We show that these tests are only indicative of the relevant type of external validity when all studies are harmonized and the estimators used in each study are unbiased and consistent. Otherwise, artifactual discrepencies become conflated with external validity and the tests cannot distinguish them.

We conclude with guidance for a replication agenda that involves a sequential process which more carefully moves from replicating an experiment to replicating a phenomenon with an eye toward understanding what artifactual discrepancies may be present because of different design features. Most expositions of replication generally describe three classes of replication designs differing in how much of the original experiment they repeat (Collins, 1992; Schmidt, 2009; Nosek

and Errington, 2017). In particular, Guala (2005: pg. 14) distinguishes between repetition, which is essentially a replication of a research or experimental design, and genuine or *conceptual* replication, which modifies the research design in an effort to see if the same phenomenon is present in multiple places. Our results highlight that this distinction is incomplete. In particular, even comparison of sign between studies can be misleading when there are design differences across studies. Because our framework distinguishes between a study's sample, setting, and design (contrasts and measurement strategies), it allows us to expand on common expositions of replication by distinguishing different kinds of conceptual replications. We describe a *design-based approach to conceptual replication*, which provides a more natural connection between research design and causal effects, and allows causal effects to be different for different designs. It therefore provides a way of giving a causal interpretation to effects that arise in multiple places and at different times.

1 Motivating Example and Related Literature

Motivated by the poor health outcomes for children in rural Uganda, Björkman and Svensson (2009) present an important study on community monitoring of health care workers from an experiment that was conducted in Uganda in 2004. The authors ask whether greater oversight of health care workers could improve service provision and thus health outcomes. The primary focus of their study is unofficial community oversight, and not oversight by the Ugandan government. To study this question, Björkman and Svensson measure the effects of an intervention that consisted of three things: (i) dissemination of a health report card containing information about local dispensaries in community meetings; (ii) health facility meetings; and (iii) a series of joint meetings between community members and health workers. This bundled treatment was randomly assigned to 25 communities with another 25 communities as control, i.e., who did not receive any part of the bundled treatment.

Björkman and Svensson (2009) show that their bundled treatment increased healthcare utilization by community members as well as increasing child health outcomes, including reductions in childhood mortality. Notably, the treatment effects in the study were large. In particular, many (standardized) treatment effects were more than a standard deviation in magnitude. Björkman and Svensson suggest that civilian pressure—monitoring and the threat of collective action—was the mechanism that best explains the dramatic improvement in health outcomes associated with their treatment.

Prompted by the large policy impact of Björkman and Svensson (2009), Raffler, Posner, and Parkerson (2020) conducted a carefully-designed, pre-registered replication experiment in rural Ugandan communities from 2014-2016. The replication experiment was conducted a decade after

the original experiment was fielded and was more heavily-resourced. Specifically, Raffler, Posner, and Parkerson (2020), who were partly motivated by the small number of clusters in the original experiment, suggest that it may have been underpowered. The replication experiment included 92 clusters in treatment and 95 clusters in control.

In contrast to the original study, Raffler, Posner, and Parkerson (2020) generally find greatly attenuated or null treatment effects on utilization and health outcomes when compared to those in Björkman and Svensson (2009). Why do Raffler, Posner, and Parkerson (2020) find qualitatively different results than Björkman and Svensson (2009)? In their article, they cite two classes of explanations. First, the presence of statistical noise, i.e., random error, could lead to differences between each study's results. Specifically, one may be concerned—as were Raffler, Posner, and Parkerson (2020)—that the small number of clusters in Björkman and Svensson (2009) invites noisier estimates of treatment effects, and as a consequence, the promising findings of the original study were the result of a statistical fluke. Second, Raffler, Posner, and Parkerson (2020) postulate that increases in the overall *level* of healthcare over the intervening decade between the studies made the intervention less effective. Other explanations include, for example, that the high number of experiments conducted in Uganda over the course of the decade could have changed how community members and healthcare workers respond to external interventions. Either of these explanations suggests that the original effect regarding community monitoring interventions (observed in Uganda 2004-2005), could have been a real effect, but one that lacks external validity.[4] Consequently, we should not necessarily expect similar findings in Uganda in 2014-2016.

There is another potential explanation. Since it was difficult for Raffler, Posner, and Parkerson (2020) to conduct *exactly* the same experiment as Björkman and Svensson (2009), there are a number of differences between their respective research designs.[5] If the interventions or outcome measures were sufficiently different between studies, such differences could be partly responsible for the differences between the effect observed in each study. For example, while Raffler, Posner, and Parkerson (2020) worked with implementing partners with no prior experience in treatment communities, Björkman and Svensson (2009) worked through 18 community-based organizations, some of which had previous experience working in treatment communities. Additionally, Raffler, Posner, and Parkerson (2020) measured outcomes at 8 month and 20 months post-treatment,

[4]Specifically, it would lack temporal validity (Munger, 2021). We provide more precise definitions of external validity below.

[5]Importantly, among other community-monitoring interventions in the field of healthcare, Raffler, Posner, and Parkerson (2020) remain most faithful to the treatments and outcome measures in the original experiment. See Raffler, Posner, and Parkerson (2020) for a discussion of other conceptual replications of Björkman and Svensson (2009).

whereas Björkman and Svensson (2009) measured outcomes at 12 months post-treatment. Ultimately, distinguishing between between these three possibilities—statistical noise, lack of external validity, and variation in study design—is of central importance to the productive use of replication.[6]

We contribute to the literature on external validity, which is best thought of as an umbrella term that encapsulates a number of related but distinct concepts—unified by their concern with target discrepancies. Many formulations of external validity is about *projecting* an empirical estimand onto a destination, which can include another study site (e.g., Shadish, Cook, and Campbell, 2002), or a grand population (e.g., Egami and Hartman, 2022; Findley, Kikuta, and Denly, 2021). Pearl and Bareinboim (2011, 2014) develop an imputation method, the "transport formula," which takes observational covariates collected in two settings and uses differences between them to reweight the observed causal effect from one setting to another. Other applications use both unit- and setting-level covariates for extrapolation (e.g., Bisbee et al., 2017; Dehejia, Pop-Eleches, and Samii, 2021). Another formulation consistent with projectivism, parallelism, refers to the projection of findings from experimental setting transports onto more "natural" settings (Smith, 1982; Guala, 2005). Fariss and Jones (2018) connect projective external validity to the predictive power of a study.

Recent elaboration of hierarchical models similarly take a projectivist view of external validity in which a "common effect" projects onto each site or constituent study. Meager (2019) develops a Bayesian hierarchical model to study the distributional effects of experimental microcredit expansion interventions. Additionally, Gechter and Meager (2021) adopt a hierarchical model to combine observational studies—where the key concern is selection into treatment—with experimental studies—where the key concern is site selection bias (Allcott, 2015). Both classes of hierarchical models feature a projection from a true underlying effect to study-specific estimates.

In this article we formally define external validity cross-sectionally. These definitions characterize the relationship between multiple studies (or estimates) without reference to some external destination, and are thus naturally suited to replication studies. Consequently, our formulations of external validity are a property of a cross-section of studies and not something that "projects" from one study to another.[7] By doing a replication study, authors invest time and often substantial resources in trying to measure an effect in a new sample or setting, which is quite different than using information from a single study to *estimate* or *impute* the effect from one sample or setting

[6] In Appendix S1, we show that our framework also applies to observational replication studies, by discussing recent dialogue on the effects college football game outcomes on pro-incumbent voting (Healy, Malhotra, and Mo, 2015; Graham et al., 2021, 2022; Fowler and Montagnes, 2015, 2022*a*,*b*).

[7] See Slough and Tyson (2022) for a classification of different formulations of external validity.

to another. Indeed, Raffler, Posner, and Parkerson (2020) laudably raised hundreds of thousands of dollars to replicate Björkman and Svensson (2009), instead of simply applying some estimator, e,g., Pearl and Bareinboim (2011), to the Björkman and Svensson (2009) data.

Finally, this paper contributes to an emerging literature on the "theoretical implications of empirical models" that focuses on the theoretical properties of commonly-used empirical research designs (Bueno de Mesquita and Tyson, 2020; Abramson, Kocak, and Magazinnik, 2022; Slough, 2022). We join Slough and Tyson (2022), Izzo, Dewan, and Wolton (2020), and Wilke and Samii (2022) in considering the properties of research designs aimed at the accumulation of empirical knowledge. Our characterization of meta-study design extends guidance on the design of individual experiments from Chassang, Padró i Miquel, and Snowberg (2012), who uses a principal-agent model to show how field experiments can exploit subject demand effects, and Banerjee et al. (2020), who show how randomization can act as a robust response to an adversarial audience.

2 Framework

We expand the framework originally presented by Slough and Tyson (2022) and develop new concepts that are important for replication. Suppose there is a collection of $J \geq 2$ studies on a common phenomenon which are indexed by j and can include experiments, quasi-experiments, or observational studies. What matters is that these studies are unified by the presence of a common (set of) mechanism(s), which motivates comparison of study estimates as an exercise in *knowledge accumulation*.

Each study is comprised of three key ingredients. Unless stated otherwise, all sets are measure spaces with strictly positive Lebesgue measure and are smooth manifolds.[8] First is a **measurement strategy**, denoted by $m \in M \subset \mathbb{R}$, where M represents the set of potential measurement strategies. A measurement strategy captures the choices a researcher makes when choosing an outcome of interest and devising a measure of that outcome. Second, every study involves a **contrast**, $(\omega', \omega'') \in \mathcal{C} \subset \mathbb{R}^2$, which defines the comparison of interest between two instrument values (e.g., Imbens and Angrist, 1994). The two instrument values are taken from the set of all potential comparisons, and are most commonly referred to as "treatment" and "control." We say that two studies are **harmonized** if they have the same measurement strategy and the same contrast. Third, every study takes place in a setting, $\theta \in \Theta \subset \mathbb{R}$. Settings capture attributes of individual units (i.e., subjects) as well as features of the environment in which the study is conducted.

An empirical exercise measures the presence and influence of a mechanism by looking at its

[8]These are not particularly restrictive as any set of probability distributions over a finite set satisfies these assumptions.

effect, and the effect in a particular study is its **empirical target**, which we formalize as follows.

Definition 1. *For a measurement strategy $m \in M$, a contrast $(\omega', \omega'') \in \Omega$, and setting $\theta \in \Theta$, the **treatment effect function** is a function, $\tau_m(\omega', \omega'' \mid \theta) : M \times \Omega \times \Theta \to \mathbb{R}$, that is smooth almost everywhere, whose derivative has full rank in measurement strategies and contrasts, and which is symmetric in contrasts, i.e., $\tau_m(\omega', \omega'' \mid \theta) = -\tau_m(\omega'', \omega' \mid \theta)$.*

The empirical target is the measured effect of a study as it relates to how things are measured, what comparison is made, and features of the setting (time, location, etc.) in which the study is conducted.[9] That the derivative of the treatment effect function has full rank in measurement strategies and contrasts captures that the observed effect of a particular design varies with that design. Our framework emphasizes the relationship between research design and empirical targets. This stresses an important feature that distinguishes our framework from others, e.g., UTOS, PICO, etc., which are all special cases of our framework.[10] Symmetry simply reflects that the order of the instrument value does not affect the magnitude of the treatment effect, only its sign. This symmetry holds for treatment effects defined in terms of differences in potential outcomes.

Empirical measurement is also concerned with *estimation*, which encapsulates the set of concerns that invariably arise because of "random noise" that interrupts the analyst's ability to precisely measure the empirical target. Such random noise typically stems from the sampling of units, chance imbalances in the assignment of instruments, and/or non-systematic measurement error. To capture the potential for estimation concerns in our framework, there is a collection of random variables $\varepsilon_j^{n_j}$, where n_j represents the sample size of study j. The observed, or *measured effect* in study j, conducted in site θ_j, is written as

$$e_j = \tau_{m_j}(\omega_j^{'}, \omega_j^{''} \mid \theta_j) + \varepsilon_j^{n_j}, \tag{1}$$

which is the empirical target in study j, as a consequence of the design, $\mathcal{D}_j \equiv (m_j, (\omega_j^{'}, \omega_j^{''}))$, setting, θ_j, and random noise interrupting the direct measurement of that empirical target, $\varepsilon_j^{n_j}$. Introducing distributions over this observation error induces a Blackwell experiment (Blackwell,

[9] That τ is smooth almost everywhere is not particularly restrictive, unless one expects it to be a nonmeasurable function or fractal.

[10] In particular, UTOS and PICO follow from our framework by imposing design invariance, which is when the effect of interest is independent of research design features (Slough and Tyson, 2022). PICO is common in medical meta-studies, UTOS comes from Shadish, Cook, and Campbell (2002), and is extended to M-STOUT in Findley, Kikuta, and Denly (2021) by including mechanism and time.

1953). An estimator of the target $\tau_{m_j}(\omega_j^{'}, \omega_j^{''} \mid \theta_j)$ is unbiased when $\mathbb{E}[\varepsilon_j^{n_j}] = 0$ and consistent when the variance of $\mathbb{E}(\varepsilon_i^{n_i} - \mathbb{E}[\varepsilon_j^{n_j}])^2 \to 0$ (in measure) as $n_i \to \infty$.

3 Concepts

When comparing two or more studies, there may be systematic differences that are not statistical, because they arise from differences between the design of constituent studies, the settings at hand, or the mechanisms producing the effects. As a result, these differences cannot be reduced to "error," and should not be treated as random. In this section we develop concepts that help organize some of the nonstatistical issues that can arise when accumulating evidence across settings.

We characterize the relationship between the empirical targets—the treatment effect functions—of two studies. Recall that these targets do not include statistical noise.

Definition 2. *Two studies* $\mathcal{E}_1 = \{m_1, (\omega_1^{'}, \omega_1^{''}), \theta_1\}$ *and* $\mathcal{E}_2 = \{m_2, (\omega_2^{'}, \omega_2^{''}), \theta_2\}$ *are* ***target-equivalent*** *if*

$$\tau_{m_1}(\omega_1^{'}, \omega_1^{''} \mid \theta_1) = \tau_{m_2}(\omega_2^{'}, \omega_2^{''} \mid \theta_2),$$

and ***target-congruent*** *if*

$$sign(\tau_{m_1}(\omega_1^{'}, \omega_1^{''} \mid \theta_1)) = sign(\tau_{m_2}(\omega_2^{'}, \omega_2^{''} \mid \theta_2)).$$

In short, two studies are target-equivalent when their targets are the same and target-congruent when the targets have the same sign. It is important to reiterate that the estimates of these targets—the observed e_1 and e_2—include idiosyncratic random error. This means that if two studies are target-equivalent, estimates of the targets will be different (with probability 1) and may even have different signs. Our focus is instead on the non-statistical reasons for difference in estimates across studies, because they cannot necessarily be solved using statistical techniques.

3.1 Target Discrepancy and External Validity

We begin with differences between empirical targets that are the result of a mechanism's influence, which can potentially manifest differently across settings. We call such differences *target discrepancies* and note that they constitute an *all-else-equal* difference in observed effects resulting from differences in setting.

Definition 3. *For research design* $\mathcal{D} = \{m, (\omega', \omega'')\}$*, comprised of measurement strategy,* $m \in M$ *and contrast,* $(\omega^{'}, \omega^{''}) \in \Omega$*, the* ***target discrepancy*** *from setting* θ *to* θ' *is*

$$\Delta_{\mathcal{D}}(\theta, \theta') = \tau_m(\omega^{'}, \omega^{''} \mid \theta) - \tau_m(\omega^{'}, \omega^{''} \mid \theta').$$

Our definition of target discrepancy holds aspects of a research design fixed, i.e., harmonizing the measurement strategy, m, and the contrast, (ω', ω''), across the two settings. As such, $\Delta_{\mathcal{D}}(\theta, \theta')$ identifies the difference in empirical targets that is attributable to moving from setting θ to setting θ', and not due to differences in research design. Although our terminology, and focus on empirical targets, is new, there is a great deal of scholarly attention given to issues revolving around target discrepancy which typically falls under the label of "external validity."

Definition 4 (Slough and Tyson (2022))**.** *A mechanism has **external validity** from setting θ to θ' if for almost every measurement strategy $m \in M$ and almost every contrast (ω', ω'')*

$$\tau_m(\omega^{'}, \omega^{''} \mid \theta) = \tau_m(\omega^{'}, \omega^{''} \mid \theta').$$

A mechanism is externally valid if it has external validity for almost all contrasts and almost all measurement strategies.

Our definition of external validity has a clear link to target discrepancy and to develop an intuition for their relationship, we present a straightforward remark.

Remark 1. *The target discrepancy between studies is zero, $\Delta_{\mathcal{D}}(\theta, \theta') = 0$ for almost all $\mathcal{D}$, if and only if the mechanism of interest has external validity between settings θ and θ'.*

Necessity and sufficiency in this result follow from the definition of external validity, highlighting the conceptual link between external validity and target discrepancies. Remark 1 stresses that target discrepancies emerge *because* the mechanism lacks external validity between two settings. The absence of external validity does not make any statement about the magnitude or sign of target discrepancies, only that they are non-zero.

External validity may be more than one needs. For example, Morton and Williams (2010) distinguish between "point" and "relationship" predictions of formal models in experimental social science, and similarly, a researcher may be interested in assessing the *sign*, rather than the precise *magnitude* of treatment effects across different settings. As such, it is useful when considering practical applications to introduce a notion of external validity that is more closely-aligned with directional theories and hypotheses.

Definition 5. *A mechanism has **sign-congruent external validity** from setting θ to θ' if for almost every measurement strategy $m \in M$ and almost every contrast (ω', ω'')*

$$sign(\tau_m(\omega^{'}, \omega^{''} \mid \theta)) = sign(\tau_m(\omega^{'}, \omega^{''} \mid \theta')).$$

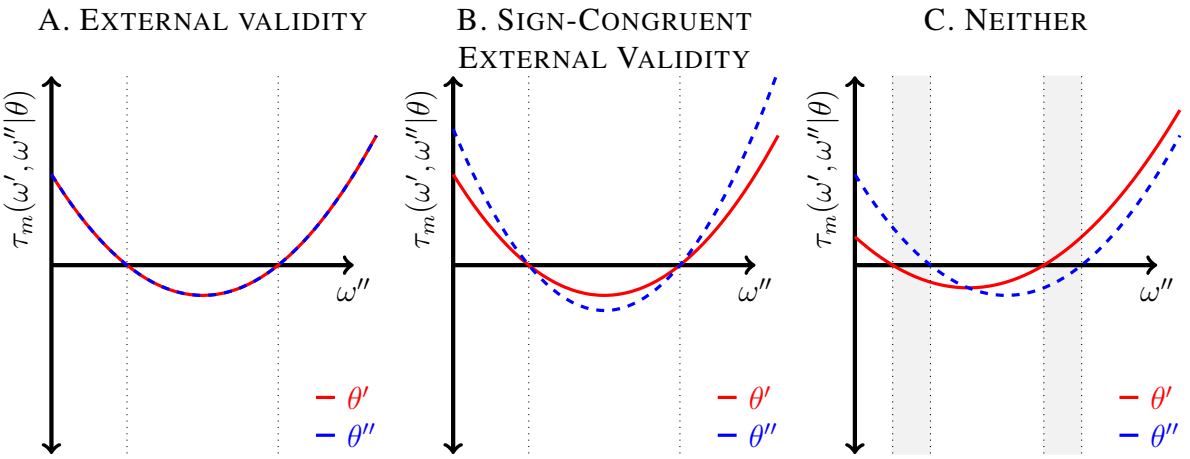


Figure 1: Illustration of external validity and sign-congruent external validity in harmonized experiments in two sites, θ' and θ''. We assume a fixed ω' and m in order to depict these concepts in two dimensions.

A mechanism is sign-congruent externally valid if it has sign-congruent external validity for almost all contrasts and almost all measurement strategies.

Sign-congruent external validity is similar to external validity in that each expresses a theoretical property of empirical targets across settings. Definition 5, however, only requires that the empirical targets across studies share the same sign, rather than having to be the same magnitude (as in Definition 4). Indeed, sign-congruent external validity is weaker in that any mechanism that has external validity has sign-congruent external validity, i.e., external validity implies sign-congruent external validity, but that a mechanism that is sign-congruent externally valid need not be externally valid. For example, when a mechanism is activated for different proportions of units across two experiments that use the same design, sign-congruent external validity is the relevant external validity concept. In this case, we should not expect resultant average treatment effects to be the same. Yet, we may expect one treatment effect is an attenuated version of the other.

Figure 1 illustrates external validity and sign-congruent external validity using graphical examples (to fix ideas). To plot these figures in two dimensions, we fix a measurement strategy, m, and one instrument, ω'. We plot treatment effects, $\tau_m(\omega', \omega'' \mid \theta)$, in two settings, θ' and θ'', as a function of the other instrument value ω'', representing the level of treatment. In the left-hand panel, we show that external validity implies that treatment effects are identical in both settings.[11] Importantly, the plot shows that external validity makes no requirement of functional form, only

[11]Recall that the treatment effect function is the empirical target. As such, it does not include statistical error.

that the treatment effect function is the same in both settings. The right-hand panel depicts a mechanism that has sign-congruent external validity but not external validity between settings θ' and θ''. This means that although the relationship between treatment, ω'', and the treatment effect, $\tau_m(\omega', \omega'' \mid \theta)$, can vary across settings, it can only do so in a particular way. Graphically, sign-congruent external validity requires that the treatment effect functions in the two settings must cross 0 in the same places (share all x-intercepts) and from the same direction (above or below 0).[12]

3.2 Artifactual Discrepancy and Harmonization

Almost all scholarly attention that is devoted to the accumulation of empirical evidence across studies is focused (informally) on issues related to target discrepancies. However, there is another feature that can frustrate efforts at accumulating evidence: variation in research designs. In practice, and outside the special case of replications that only vary samples, it can be very difficult to ensure that two studies are harmonized when conducted in different settings.

When two studies employ different measurement strategies, or make different comparisons (contrasts), their measured effects can vary for reasons unrelated to issues of estimation or external validity.

Definition 6. *For setting $\theta \in \Theta$, the **artifactual discrepancy** between studies $\mathcal{E}_i = \{m_i, (\omega_i^{'}, \omega_i^{''}), \theta\}$ and $\mathcal{E}_j = \{m_j, (\omega_j^{'}, \omega_j^{''}), \theta\}$ is*

$$\mathcal{A}(\mathcal{D}_i, \mathcal{D}_j \mid \theta) = \tau_{m_i}(\omega_i^{'}, \omega_i^{''} \mid \theta) - \tau_{m_j}(\omega_j^{'}, \omega_j^{''} \mid \theta).$$

Artifactual discrepancies are differences in empirical targets that emerge from using different contrasts or measurement strategies—they are discrepancies that come from *using different research designs*. In the Björkman and Svensson (2009) and Raffler, Posner, and Parkerson (2020) studies, measuring outcomes at different times relative to the rollout of the intervention may have led to different measured effects even if the underlying treatment effects (as a function of time) were the same.

Artifactual discrepancies highlight the importance of harmonization between different studies, which is illustrated by our next remark:

Remark 2. *The artifactual discrepancy is zero, $\mathcal{A}(\mathcal{D}_i, \mathcal{D}_j \mid \theta) = 0$, almost everywhere if and only if i and j are harmonized.*

[12] This illustrates an alternative characterization of sign-congruent external validity, namely, that $\tau_m(\omega', \omega'' \mid \theta)$ must have the same null (zero) set across θ and that the first-derivatives must have the same sign on that null set.

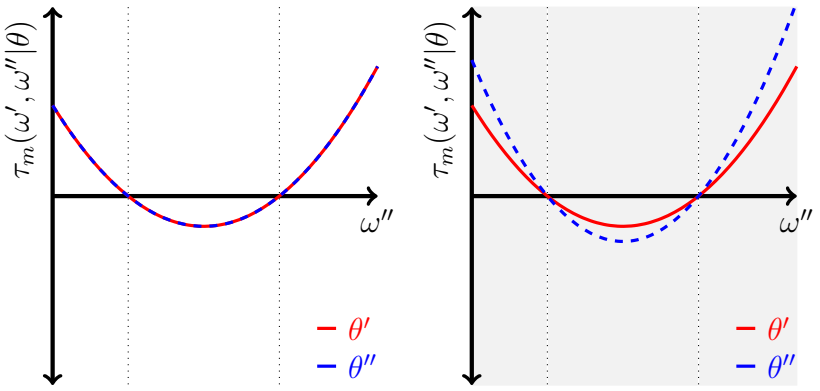
This remark follows immediately from the definition of harmonization and it says that when two studies are harmonized, the artifactual discrepancy is zero. It is important to note that design-induced discrepancies are "artifactual," but this does not imply that these discrepancies are "nuisance" parameters. Specifically, in contrast to arguments that a lack of harmonization is simply "another source of random error" in replication studies (Gilbert et al., 2016: p. 1037a), issues related to the harmonization between studies are fundamentally non-statistical concerns. Artifactual discrepancies are issues of research design, and consequently, eliminating them is ultimately a question of research design.

To illustrate that artifactual discrepancies are fundamentally non-random, suppose that two studies examine the effects of some mechanism such as nutritional intake on children's height. One study measures height in inches; the other measures height in centimeters. When the mechanism behind the treatment has external validity, the treatment effects across the studies will be different, but this difference is not random error—the measurements are deterministically related. Specifically, we expect the treatment effects in centimeters to be the treatment effects in inches scaled by a factor of 2.54. Researchers often purposefully select their contrasts and outcomes when designing a study. While some psychologists like Monin and Oppenheimer (2014) have advocated randomly varying the content of contrasts in conceptual replications, this practice remains far outside mainstream practice. As such, artifactual error should be understood as a form of *non-random* error in replication studies that nevertheless goes unobserved. As another example, In a drug trial we generally expect to observe different treatment effects if the dosage of a drug were doubled, even if it were administered to the same population and in the same setting. Failure to adjust for dosage differences would result in artifactual discrepancies.

Remark 2 stresses that there are two sources of artifactual discrepancy in our framework: (i) differences in measurement strategies and (ii) differences in contrasts. It is important to emphasize that artifactual discrepancies affect the connection between empirical targets that are unified by their study of a unique substantive phenomenon. However, they may be of independent interest in and of themselves since they provide information about the "technology of intervention." Learning how treatment effects vary in features of distinct interventions—like varying dosages of a treatment—provides important information. It also stresses that an intervention may interact with a mechanism or setting in ways that are not easy to disentangle.

4 Empirical Targets and External Validity

We now turn to some results that consider how external validity and harmonization relate to target-equivalence and target-congruence. The relationship between harmonization, external validity,



(a) Treatment effects are externally valid. (b) Treatment effects are not externally valid.

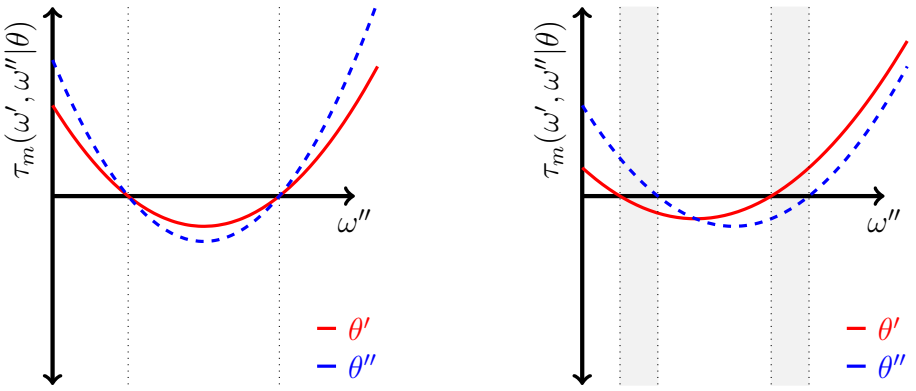
Figure 2: Illustration of the logic of Theorem 1. In panel (b), external validity does not hold. The grey regions in (b) depict the regions of the design space in which target equivalence fails.

and target-equivalence is developed at length in Slough and Tyson (2022), applied to the case of meta-analysis. However, they did not consider the role and importance of artifactual and target discrepancies, which are central to the comparison of treatment effects, and thus replication projects.

Theorem 1 (Target-equivalence)**.** *For a collection of studies* $\{\mathcal{E}_i = (m_i, (\omega_i^{'}, \omega_i^{''}, \theta_i)\}_{i=1}^{N}$*, target-equivalence holds across* i *almost everywhere if and only if all studies satisfy external validity and are harmonized almost everywhere.*

Recall that Remark 1 guarantees that external validity ensures that all target discrepancies are zero. Moreover, Remark 2 shows that harmonization ensures that artifactual discrepancies are also zero. These observations show how external validity and harmonization are jointly sufficient for target-equivalence. The argument for necessity is more involved, but follows from this simple analysis and is in the appendix. The key intuition for Theorem 1 is illustrated in Figure 2. Panel (a) shows treatment effect functions with external validity, and how when moving from θ' to θ'', there is no target discrepancy. Panel (a) also illustrates that even when external validity holds, if two studies use different ω'', then they target different treatment effects. Hence, target-equivalence is not maintained without harmonization. In contrast, panel (b) illustrates that when external validity does not obtain (although sign-congruent external validity does), there are target discrepancies except where the treatment effect functions intersect. The grey regions in panel (b) show the set of research designs (across ω'') where target equivalence fails even when both studies are harmonized (by fixing ω''). Theorem 1 shows that these sets have positive measure, and that the example in Figure 2 is not unusual.

We now consider target-congruence and its relationship with harmonization of study designs



(a) Treatment effects are sign-congruent externally valid. (b) Treatment effects are not sign-congruent externally valid.

Figure 3: Illustration of the logic of Theorem 2. The grey regions in (b) depict the "sign flip" set which has positive measure when sign-congruence does not hold.

and sign-congruent external validity.

Theorem 2 (Target-congruence)**.** *For a collection of studies* $\{\mathcal{E}_i = (m_i, (\omega_i^{'}, \omega_i^{''}, \theta_i)\}_{i=1}^{N}$*, they are target-congruent if and only if they all satisfy sign-congruent external validity and are all harmonized almost everywhere.*

A key component of the proof of Theorem 2 is the "sign-flip" set, where target-congruence doesn't hold, and the details of its construction are in the appendix. This is set is constructed for measurement strategies by focusing on the set of contrasts where the sign is different between two different measurement strategies. This is important because it is where the the sign of an effect is different depending only on changing the measurement strategy—not because the mechanism's effect varies over settings. The proof of Theorem 2 establishes that this set has positive measure. This is a problem because it implies that any empirical distribution over effects incorrectly identifies when a mechanism has the same effect in different places, and moreover, that the probability this happens can be taken arbitrarily close to 1. Another way of interpreting Theorem 2 is to observe that it also implies that a mechanism that lacks sign-congruent external validity, and hence produces effects with different signs in different settings, can produce the same sign in empirical studies because of artifactual discrepancies, thereby producing misleading results for the analyst.

Figure 3 illustrates Theorem 2, where panel (a) is the same as panel (b) in Figure 2. Panel (a), where sign-congruent external validity holds, illustrates how if experiments use different ω'', thereby violating design harmonization, the resultant empirical targets will not necessarily have the same sign. In panel (b), sign-congruent external validity does not hold, and even if researchers harmonize the design across studies by choosing the same ω'', the sign of the empirical targets may

differ. In particular, the grey regions correspond to the "sign-flip" sets from above (over contrasts here, not measurement strategies). In these regions target congruence does not hold, and Theorem 2, establishes that this set has positive measure whenever sign-congruent external validity does not hold. Moreover, the size of these sets can be arbitrarily large depending on how $\tau_m(\omega', \omega''|\theta)$ varies in setting, θ.

A natural question is to what extent external validity, or sign-congruent external validity, needs to hold globally, i.e., for almost all research designs and settings. What if, instead, external validity holds on a strict subset of Θ? In such a case target-equivalence, or target-congruence respectively, would fail unless the analyst is able to identify precisely where external validity, or sign-congruent external validity, holds. Moreover, if external validity held on some strict subset of research designs and settings, and sign-congruent external validity held on all, then only sign-congruent external validity can be taken to be satisfied, and target-congruence is the most the analyst can assess.

4.1 Increasing the Number of Studies

Some large replication studies conduct $N \geq 2$ independent replications of a single study (e.g., Klein et al., 2014). Although pooling more replications could facilitate learning about any statistical discrepancies between studies, the information the analyst gains is substantially complicated by the potential for the inclusion of studies with *different* target or artifactual discrepancies. Importantly, target and artifactual discrepancies are not, generally, random, and thus, cannot be treated as being drawn from a known distribution across different replication studies—this effectively sweeps the problem under the rug.

To illustrate the difference, we now briefly apply Theorem 2 to show that artifactual discrepancies are not solvable using standard statistical techniques, i.e., these problems cannot be mitigated by pooling multiple distinct replications without specific consideration of research design. In particular, we consider what happens to the sign-flip set discussed above when more studies are added to a replication.

Theorem 3. *Take a collection of studies, $\{\mathcal{E}_i = (m_i, (\omega_i^{'}, \omega_i^{''}, \theta_i)\}_{i=1}^{N}$, the set where the sign of empirical targets is (artifactually) different is nondecreasing (in the set inclusion order) in the number of studies N.*

This result establishes that increasing the number of studies does not make it "easier" to achieve target-congruence, but more difficult. This follows from the observation that adding additional studies involves expanding the sign-flip set discussed above, which is made up of artifactual discrepancies. Theorem 3 suggests that there is a trade-off when considering how many studies to include in a replication. While accumulating more studies to obtain more estimates of the treat-

ment effect certainly aids in addressing *statistical* concerns, it potentially exacerbates issues that arise from design issues. Specifically, although it is generally beneficial to observe more draws of the random variables $\varepsilon_j^{n_j}$, doing so involves adding studies that lack harmonization, they invariably introduce more artifactual discrepancies, $\mathcal{A}(\mathcal{D}_i, \mathcal{D}_j \mid \theta)$, which can complicate efforts to make inferences about both target-congruence *and* statistical properties of the random variables $\varepsilon_j^{n_j}$. Only when studies are harmonized does this trade off not arise.

While replication is an important tool for probing the breadth and robustness of observed treatment effects, it is not necessarily an "agnostic" empirical approach to accumulating empirical evidence. We identify three reasons why a replication study can produce results that are different from an original study: (i) statistical noise most commonly associated with estimation; (ii) target discrepancies induced by mechanisms that lack external validity (however articulated); and novel to our framework, (iii) artifactual discrepancies that are induced by research designs that are insufficiently harmonized.

5 Testing External Validity

Replications are increasingly used to learn about the statistical properties of a study or body of work. For instance, Raffler, Posner, and Parkerson (2020) sought to replicate Björkman and Svensson (2009), in part, because it was such a small (and perhaps underpowered) study. In other cases, replication is used to diagnose researcher error, malfeasance, or publication bias in a given subset of the literature (e.g., Camerer et al., 2016, 2018; Open Science Collaboration, 2015; Klein et al., 2014). Our presentation so far has black-boxed statistical issues that may arise in replications. We did this to focus on properties that are important theoretical issues but which are distinct from sampling and estimation. Anyone conducting a replication will, in practice, however, also confront *statistical discrepancies*, and our framework straightforwardly extends to include such things.

When researchers seek to compare estimates across different studies, they typically adopt at least one of two approaches, which we outline formally.[13] The first approach involves comparing the *point estimates* of effects in different studies to assess whether a particular intervention/treatment, assessed in different settings, produces the same effect. The second approach involves comparison of the *sign* of estimates across studies. While we characterize this approach formally, it is important to note that this approach is frequently invoked informally when researchers describe the relationship between their study and related work. We state the results in this section in terms of two studies (or a study and its replication). However, the logic and results extend to

[13]Other approaches in the published literature rely on the statistical properties of a set of unrelated replications in which each replication consists of two or more studies and researchers assess properties of the distribution of estimates across replications.

replication agendas with more than two studies. In these cases, researchers may test a joint null hypothesis that all estimates share the same sign or are equivalent.

The first approach to accumulating evidence compares the estimates directly, measuring whether a mechanism generates *the same effect* in multiple studies. This approach is used in some formal replications but is less common in informal descriptions. To compare the estimates of two studies, 1 and 2, compute

$$e_1 - e_2 = \tau_{m_1}(\omega_1^{'}, \omega_1^{''} \mid \theta_1) + \varepsilon_1^{n_1} - \tau_{m_2}(\omega_2^{'}, \omega_2^{''} \mid \theta_2) - \varepsilon_2^{n_2},$$

which can be written:

$$e_1 - e_2 = \overbrace{\varepsilon_1^{n_1} - \varepsilon_2^{n_2}}^{\text{statistical discrepancy}} + \underbrace{\Delta_{\mathcal{D}_1}(\theta_1, \theta_2)}_{\text{target discrepancy}} - \overbrace{\mathcal{A}(\mathcal{D}_1, \mathcal{D}_2 \mid \theta_2).}^{\text{artifactual discrepancy}} \tag{2}$$

This expression highlights that the difference between the observed effects in 1 and 2, $e_1 - e_2$, contains more than just random error, i.e., statistical discrepancies, but also includes target discrepancies, when external validity fails, and artifactual discrepancies, emerging when the designs in 1 and 2 are not harmonized. Empirical researchers will never observe the statistical noise terms $\varepsilon_1^{n_1}$ and $\varepsilon_2^{n_2}$ directly, but instead, rely on properties of their probability distributions to estimate how likely we are to observe a given difference in estimates (or signs) under a the relevant null hypothesis. By writing (2) in terms of target and artifactual discrepancies, it is straightforward to see that the interpretation of these tests changes in the presence of these non-random discrepancies. To formulate statistical tests that facilitate inference, an analyst makes some assumptions about the distribution of $\varepsilon_j^{n_j}$ across j, as well as sampling properties. For instance, an analyst typically assumes that $\varepsilon_j^{n_j}$ are independently and normally distributed with mean-zero, which ensures $\mathbb{E}[\varepsilon_i^{n_i} - \varepsilon_j^{n_j}] = 0$.

Proposition 1. *The **estimate-comparison test** computes:*

$$\mathcal{W} = e_1 - e_2$$

and test the null hypothesis $H_0^w : \tau_{m_1}(\omega_1', \omega_1''|\theta_1) = \tau_{m_2}(\omega_2', \omega_2''|\theta_2)$ against the alternative $H_a^w : \tau_{m_1}(\omega_1', \omega_1''|\theta_1) \neq \tau_{m_2}(\omega_2', \omega_2''|\theta_2)$.

Let two studies $\mathcal{E}_1 = (m_1, (\omega_1^{'}, \omega_1^{''}), \theta_1)$ and $\mathcal{E}_2 = (m_2, (\omega_2^{'}, \omega_2^{''}), \theta_2)$ have unbiased and consistent estimation errors, then

1. If studies 1 and 2 are harmonized, then the estimate-comparison test assesses a null hypothesis

that the mechanism is externally valid;

2. *If the mechanism has external validity, then the estimate-comparison test assesses a null hypothesis that studies* 1 *and* 2 *are harmonized.*

Proof. Follows from Theorem 1. □

The estimate-comparison test permits an analyst to explore both external validity and harmonization—but not simultaneously. Generally, the test addresses whether $\Delta_{\mathcal{D}_1}(\theta_1, \theta_2) - \mathcal{A}(\mathcal{D}_1, \mathcal{D}_2 \mid \theta_2)$ is statistically distinguishable from zero.

In other words to test either harmonization or external validity the analyst must be able to (credibly) fix one of these features in order to assess the other.[14] Proposition 1 establishes two findings that are relevant for replication. First, by assuming harmonization, the estimate-comparison approach allows for a test of a mechanism's external validity.[15] Second, by assuming external validity, the estimate-comparison approach permits a test for harmonization—provided the analyst knows independently (or assumes) that the mechanism under study is externally valid.

In the presence of non-zero target or artifactual discrepancies, the estimate comparison test risks rejecting the null hypothesis that $e_1 - e_2 = 0$ because of non-statistical discrepancies. In other words, we could mistakenly infer that an observed estimate was a statistical fluke, or worse, a result of researcher malfeasance, because of a lack of external validity or harmonization. Direct replications, where the design is harmonized, eliminate target and artifactual discrepancies, and consequently, allow researchers to learn about statistical discrepancies.[16]

It is important to consider the relationship between Proposition 1 and replication designs that leverage replications of multiple distinct studies (for examples in economics, see Camerer et al., 2016, 2018). These tests rely on properties of the distribution of the error terms ($\varepsilon_i^{n_i}$). For example, if there were no publication bias or selective reporting, it should be the case that $E[\varepsilon_i^{n_i}] = 0$ (for unbiased estimators used to analyze experiments). There are various tests used in these herculean replication studies (see also Open Science Collaboration, 2015), but all of these tests are premised on a similar null hypothesis to Proposition 1, which assumes that $\mathcal{A}(\mathcal{D}_i, \mathcal{D}_j \mid \theta) = 0$ and $\Delta_{\mathcal{D}}(\theta, \theta') = 0$, for each constituent replication. But $\mathcal{A}(\mathcal{D}_i, \mathcal{D}_j \mid \theta)$ and $\Delta_{\mathcal{D}}(\theta, \theta')$ are not necessarily random and do not follow a known distribution absent additional assumptions. This analysis

[14] This is in stark contrast to meta-analysis, where target-equivalence is generally assumed for identification of the empirical models, and hence, is a key ingredient of such approaches.

[15] This test, of course, relies further on an assumption that all of the constituent studies employ unbiased estimators of the treatment effect.

[16] Obviously, direct replication is more feasible in some contexts—like surveys—than others (i.e., large-scale field experiments).

suggests that artifactual and target discrepancies can bias estimates of a literature's replicability, but he direction of this bias, however, is unclear ex ante.

The second test focuses on the signs of the observed effects, e_j, across studies and is meant to probe information about the consistency of the sign of a mechanism's effect. It is important to stress that researchers often informally compare the sign across studies heuristically when comparing studies, without formally testing a null hypothesis. Such heuristic versions of the sign-comparison test that differentiate between, for example, a positive (and significant) estimate versus a "null" estimate are prone to exceptionally high rates of Type-I error (incorrect rejections of the null hypothesis of sign congruence) (Simonsohn, 2015).

Proposition 2. *The **sign-comparison test** computes:*

$$\mathcal{Z} = e_1 \cdot e_2$$

and tests the null hypothesis $H_0^z : sign(\tau_{m_1}(\omega_1', \omega_1''|\theta_1)) \cdot sign(\tau_{m_2}(\omega_2', \omega_2''|\theta_2)) > 0$ *against the alternative* $H_a^z : sign(\tau_{m_1}(\omega_1', \omega_1''|\theta_1)) \cdot sign(\tau_{m_2}(\omega_2', \omega_2''|\theta_2)) < 0$.

If two studies $\mathcal{E}_1 = (m_1, (\omega_1^{'}, \omega_1^{''}), \theta_1)$ *and* $\mathcal{E}_2 = (m_2, (\omega_2^{'}, \omega_2^{''}), \theta_2)$ *are harmonized, and estimation errors,* $\varepsilon_1^{n_1}$ *and* $\varepsilon_2^{n_2}$*, are unbiased and consistent, then the sign-comparison test assesses a null hypothesis of sign-congruent external validity.*

Proof. Follows from Theorem 2. □

The requirement of unbiasedness and consistency reflect conventional statistical concerns and shows the importance of internal validity of *all* constituent studies. The novel and important part of Proposition 2 is that it shows that the sign-comparison test can be used to test a null hypothesis that a set of studies exhibits sign-congruent external validity, but *only if the constituent studies are harmonized*. Recall that the null hypothesis of the sign-comparison test holds that $e_1 \cdot e_2 > 0$, an event corresponding to when both estimates have the same sign. As such, rejection of this null hypothesis constitutes a rejection of target-congruence. Given Theorem 2, combined with harmonization, this is equivalently a test for sign-congruent external validity.

To calculate p-values and conduct inference using the sign-comparison test, let $\varepsilon_i^{n_i}$ be normally distributed with mean 0 and let the standard error of e_i be se_i, then the p-value of the null hypothesis of sign-congruence is:

$$\begin{aligned} p &= \Pr(e_1 > 0)\Pr(e_2 > 0) + \Pr(e_1 < 0)\Pr(e_2 < 0) \\ &= \Phi(e_1/se_1)\Phi(e_2/se_2) + (1 - \Phi(e_1/se_1))(1 - \Phi(e_2/se_2)), \end{aligned}$$

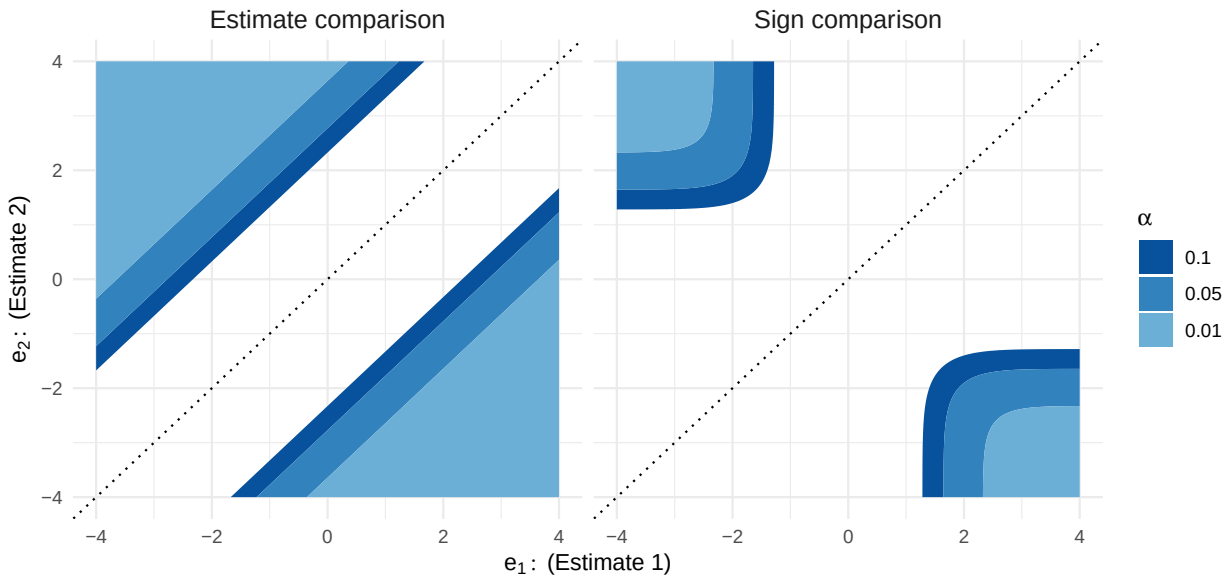


Figure 4: Rejection regions of the estimate- and sign-comparison approaches for Type-I error rates, $\alpha \in \{0.01, 0.05, 0.1\}$. Both plots fix $se_1 = se_2 = 1$ in order to visualize these regions in two dimensions.

where $\Phi(\cdot)$ is the cdf of the standard normal distribution. It is important to note that because the sign-comparison test assesses a weaker (less stringent) null hypothesis than the estimate-comparison test, it should be more difficult to reject the null with the sign-comparison test as compared to the estimate-comparison test. Figure 4 plots the regions in which one would reject the null hypothesis under both approaches, for varying Type-I error rates (α). Consistent with the intuition about the stringency of the null hypotheses, the rejection regions for the sign-comparison test are strictly smaller than those of the estimate-comparison test.

What do we learn from a sign-comparison test when studies are *not* necessarily harmonized? Theorem 2 shows that relaxing harmonization leads to the introduction of artifactual discrepancies. But because sign-congruent external validity does not pin down the target discrepancies we cannot ascertain the sign of treatment effects when artifactual discrepancies are also present, because their magnitude and direction are unknown. As such, we cannot construct the "reverse" test for harmonization with the sign-comparison test.

Propositions 1 and 2 show that tests that are commonly employed in replication studies can be used to assess external validity, depending on the approach, or harmonization in the case of the estimate-comparison approach. However, we show that any test for external validity, or sign-congruent external validity, makes further assumptions about the design of constituent studies than is typically acknowledged. In particular, a replication study makes assumptions about both the

statistical properties of constituent studies (e.g., unbiasedness, consistency) as well as cross-study properties (e.g., harmonization, external validity). Although the former is commonly discussed explicitly in practice, the latter is rarely considered or discussed explicitly in applied replications. Our results indicate that this omission is consequential since a lack of harmonization can lead to Type-I or Type-II errors in inferences about external validity in either the sign- or estimate-comparison tests.

6 Alternative Approaches to Replication

We have established how replication can facilitate learning about different formulations of external validity, and hence generate general knowledge about a substantive phenomenon. Before concluding we outline two approaches that can be used to accumulate knowledge across studies, a more common structural approach and the *design-based approach* that more heavily uses our framework and provides a concept-driven classification of replication studies.

6.1 The Structural Approach

We begin with the most common approach to combining evidence across multiple studies which relies on positing a structural model of cross-study properties.[17] Such approaches posit a model of the underlying structure that links together multiple studies, sometimes explicitly modeling aspects of a research design. The model and assumptions associated with the structrual approach effectively constrain what kinds of target and artifactual discrepancies are permitted to be present in the data. As an example, an analyst might suppose that the empirical target takes the following functional form:

$$\tau_m(\omega', \omega''|\theta) = f(\omega', \omega'', m) + g(\theta). \tag{3}$$

In this formulation, the function f specifies how treatment effects vary in contrasts and measurement strategies, which pins down artifactual discrepancies, and critically, does not allow artifactual discrepancies to depend on the setting θ. Instead, the function g specifies how empirical targets, or treatment effects, vary in setting (perhaps through contextual variables). Consequently, the function g pins down target discrepancies. Further assumptions about the functional form of f (like linearity) facilitate measurement of target discrepancies—and thus evaluation of external validity—in a non-harmonized, multi-setting replication. Specifically, in this case, it is straightforward to specify a null hypotheses analogous to those that in the estimate-comparison tests. For example, one

[17] For a conceptual overview of structural models see Koopmans and Reiersol (1950) or Goldberger (1972).

could evaluate a null hypothesis of the form:

$$e_1 = \lambda(e_2; m, \omega', \omega''), \tag{4}$$

where λ specifies the relationship between observed effects, e_1 and e_2, and how that relationship depends on contrasts and measurement strategies. The structural approach is most commonly used to *combine* rather than *compare* estimates across studies. Indeed, this formulation in the context of replication represents a natural extension of Pearl and Bareinboim (2011)'s approach to transportability and is commonly invoked—if unstated—in meta-analyses (Slough and Tyson, 2022). But, if one is willing to posit such a model, and the assumptions about how treatment effects can change across contexts, a similar approach can also be applied to replication studies.

The key strength of the structural approach is that it allows an analyst to make strong empirical conclusions from data, potentially eliminating concerns about target or artifactual discrepancies. It is important to stress, however, that these benefits result from modeling assumptions that constrain the kind of data the external world is permitted to supply. Moreover, there is little consensus on how to constrain the external world, i.e., what structural assumptions are appropriate in what cases, and whether such things are faithfully represented as "nuisance" parameters, especially when applied to evidence accumulation. Many structural approaches assume external validity and, similar to above, that measured treatment effects do not vary in the design of the studies.[18] By prohibiting the external world from presenting target or artifactual discrepancies (other than as idiosyncratic error), analysts dodge the problems resulting from artifacts of research design or external validity we highlight. Yet, assuming away target or artifactual discrepancies undermines the causal interpretation some analysts may wish to impart to results from replication (or meta-analysis). Further exploration of structural approaches to replication should stress transparently what assumptions are involved, and state precisely what is gained when downplaying the potential problems that might arise when combining evidence from multiple places.

6.2 The Design-based Approach

We have described three features that can differ between constituent studies in a replication: samples, setting, and design (contrasts and measurement strategies). These features map directly onto a replication typology, shown in Table 1, that expands on common expositions of replication, which include exact, direct, and conceptual replication (Collins, 1992; Schmidt, 2009; Nosek and Errington, 2017). Our categorization distinguishes between different types of conceptual replication, and our results stress what can be learned from accumulating evidence across different replications.

[18] Slough and Tyson (2022) term this assumption design invariance.

Class	Sub-class	Studies differ in... Samples	Settings	Design
Exact		–	–	–
Direct		✓	–	–
Conceptual	Harmonized	✓	✓	–
Conceptual	Single-setting	✓	–	✓
Conceptual	Non-harmonized, multi-setting	✓	✓	✓

Table 1: Mapping between conventional classification of replication studies and our framework. Note that the disaggregation of conceptual replications into sub-classes is non-standard in existing literature.

Exact replication implies that all aspects of two studies' research design are identical, including the sample, which is typically impossible in the social sciences.[19] The most faithful replications in the social sciences are *direct replications*, which hold fixed the setting and research design while varying the sample realizations across constituent studies (Schmidt, 2009; Ou and Tyson, 2022). Each sample is drawn from the same population (encompassed in settings in our framework) using the same sampling strategy. This design allows researchers to analyze differences in estimates that are generated by sampling (i.e., statistical noise).

Most replications in social science change more than a study's sample, thereby conducting a *conceptual replication*. The vast majority of replication studies in social science, including Raffler, Posner, and Parkerson (2020), are conceptual replications. While these conceptual replications vary different attributes of constituent studies, there are not established best practices for how these replications should be organized or assessed. Our framework clarifies three sub-classes of conceptual replication designs. Like direct replications, constituent studies in conceptual replications use different samples. Beyond different samples, conceptual replications differ in either the setting a study is conducted or in aspects of research design. In harmonized conceptual replications, researchers implement the same design (i.e., contrasts and measurement strategy) on samples from different settings (and thus different populations). In single-setting conceptual replications, researchers implement a different design on a different sample in the same setting.

Motivated by the distinctions highlighted by our framework, we propose a *design-based approach to conceptual replication*, which stresses the importance of a *replication agenda* and how such agendas should be structured. This approach takes a *sequential* method that proceeds by admitting one discrepancy at a time; summarized in Table 2. This approach, by focusing on research

[19]This is different from *reproduction* of results, which is what many journals do when computationally "replicating" the findings of accepted articles.

Step	Description	Learning	Caveats/limitations
1.	Harmonized (conceptual) replications	Evaluate external validity	No evidence about target discrepancies or external validity under different designs.
2.	Single-setting (conceptual) replications	Evaluate how τ changes in contrasts or measurement strategies	No guarantee artifactual discrepancies are equivalent across settings
3.	Non-harmonized multi-study (conceptual) replications, varying contrasts or measurement strategies.	With steps 1 and 2, evaluate whether artifactual discrepancies vary in settings.	

Table 2: Design-based replication agendas. Note that steps 1 and 2 can be pursued in reverse order.

design across settings, is more tightly connected with credibility approaches to internal validity (Banerjee and Duflo, 2009; Gerber and Green, 2012; Dunning, 2016; Samii, 2016).

1. Conduct harmonized (conceptual) replications in settings where the mechanism may be operative. Measure target discrepancies to evaluate the external validity, or sign-congruent external validity, of the mechanism. This allows for learning about the set of settings where the mechanism exhibits external validity under the harmonized design. This step does not provide evidence about target discrepancies or external validity under different designs.

2. Conduct single-setting (conceptual) replications in some setting by varying contrasts or measurement strategies. Measure artifactual discrepancies by evaluating how treatment effects change in contrasts or measurement strategies. This step does not guarantee that artifactual discrepancies are equivalent across settings.

3. Conduct non-harmonized multi-study (conceptual) replications in other settings by varying contrasts or measurement strategies in different settings. With steps 1 and 2, one can evaluate whether artifactual discrepancies vary in settings. If artifactual discrepancies do not appear to vary in settings, the mechanism exhibits external validity.

Our theoretical results show that the presence of non-zero artifactual discrepancies limit our ability to learn about target discrepancies—because artifactual discrepancies are not simply nuisance parameters. Consequently, a replication agenda must prioritize learning about artifactual discrepancies. In addition, estimating these discrepancies may be of independent interest. For example, by varying a study's design within a setting, we can understand how the treatment effect

function varies in contrasts or measurement strategies. Learning about artifactual discrepancies enables analysts to answer questions like "do treatment effects increase monotonically in the strength of treatment?" Because researchers can typically employ more than one measurement strategy in a given study, replication experiments can be particularly useful for learning how treatment effects vary in contrasts, which are generally more costly to implement. We note one limitation of this sequential replication agenda is apparent if treatment effects change over time—a manifestation a lack of external validity. If this were the case, single-setting replications cannot reliably measure artifactual discrepancies because time would introduce target discrepancies.[20] Within our framework, settings can be defined with respect to time in order to distinguish between a setting at times t and $t + 1$, as in the Björkman and Svensson (2009) and Raffler, Posner, and Parkerson (2020) examples.

7 Conclusion

The accumulation of empirical evidence collected in multiple places, at different times, and measured by different scholars presents a number of challenges. Perhaps the most important question related to evidence accumulation is whether a mechanism is externally valid (or sign-congruent externally validi). Replication (direct and conceptual) is advocated as a tool that informs researchers about the generalizability of their empirical findings. We develop a theoretical framework for the accumulation of evidence across multiple studies and apply it to understand the theoretical foundations of replication.

We show that external validity and harmonization of studies is necessary and sufficient to establish target-equivalence, whereas sign-congruent external validity and harmonization are necessary and sufficient to establish target-congruence. We then develop two sets of results about empirical targets and apply them to two statistical tests—the estimate-comparison and sign-comparison tests. This result has implications for the use of the sign-comparison test as a means to assess sign-congruent external validity. Specifically, this test is informative if and only if researchers examine harmonized studies. Consequently, our results provide a theoretical foundation for the most common statistical test in replication studies, which is also the way empiricists informally discuss related studies (even outside the context of replication).

We introduce a design-based approach to conceptual replication, which approaches learning about external validity through replication. We argue that researchers should invest more in conducting replications, but approach the different components of the cross-study environment sequentially, and measure each of them in isolation.

[20] See Lovett and Munger (2019); Munger (2021) on the importance of temporal validity.

We conclude by highlighting two important issues that arise in replication agendas. First, a desire for novelty arguably hampers any replication-based research agenda, and in this respect, our proposal is not unique (Koole and Lakens, 2012; Nosek, Spies, and Motyl, 2012; Galiani, Gertler, and Romero, 2017). These concerns are ultimately about professional incentives rather than the accumulation of knowledge. However, a benefit of our proposed sequential replication is that it more clearly articulates the contribution of each stage of the replication process. Second, in some communities replication is largely considered as a method to alleviate researcher malfeasance, and as a result, independence of research teams conducting replications is an important concern. Importantly, our notion of harmonization does not in any way preclude independent replication. However, more transparent characterization and reporting of treatments and all outcome measures will likely be necessary to facilitate independent replication within our proposed replication agenda.

A Proofs

Proof of Theorem 1. Sufficiency follows from he discussion in the text. For necessity, first notice that target-equivalence under harmonization is equivalent to external validity. Now suppose that studies $\mathcal{E}_1$ and $\mathcal{E}_2$ are target-equivalent, but not measurement harmonized. Then, for m_1 and m_2:

$$\tau_{m_1}(\omega', \omega'' \mid \theta_1) = \tau_{m_2}(\omega', \omega'' \mid \theta_2). \tag{5}$$

Applying external validity, at m_2 and (ω', ω''), it must be that for θ_1 and θ_2

$$\tau_{m_2}(\omega', \omega'' \mid \theta_1) = \tau_{m_2}(\omega', \omega'' \mid \theta_2). \tag{6}$$

Combining (5) and (6),

$$\tau_{m_1}(\omega', \omega'' \mid \theta_1) = \tau_{m_2}(\omega', \omega'' \mid \theta_1),$$

which, since the setting and contrasts were arbitrary, implies that the treatment effect must be the same at m_1 and m_2 in any setting. Thus, external validity allows us to suppress the dependence of the treatment effect function on θ.

Because $\mathcal{C}$ is a compact subset of $\mathbb{R}^2$, it is a two-dimensional manifold. Define

$$\kappa \equiv \tau_{m_1}(\omega', \omega'' \mid \theta),$$

which by external validity, is the same at almost any $\theta \in \Theta$. We are interested in the level set $\tau^{-1}(\kappa; \omega', \omega'') \subset M$. Since the derivative of $\tau_m(\omega', \omega'' \mid \cdot)$ has full rank for almost every measurement strategy, $m \in M$, the set of regular points of $\tau_m(\cdot)$ is of full measure on M. Thus, if κ is not a regular value, then $\tau^{-1}(\kappa; \omega', \omega'')$ does not contain any regular points, and is thus of Lebesgue measure zero. Suppose, instead, that κ is a regular value, and thus, $\tau^{-1}(\kappa; \omega', \omega'')$ is a set of regular points. By the Preimage Theorem (e.g., Guillemin and Pollack, 1974: pg. 21), the set $\tau^{-1}(\kappa; \omega', \omega'')$ is a submanifold of M, and moreover,

$$\text{dim } \tau^{-1}(\kappa; \omega', \omega'') = \text{dim } \mathcal{M} - \text{dim } \mathbb{R} = 1 - 1 = 0.$$

Thus, dim $\tau^{-1}(\kappa; \omega', \omega'') <$ dim M, implying that $\tau^{-1}(\kappa; \omega', \omega'')$ is a Lebesgue measure zero subset of M, completing the argument.[21] The argument for contrasts is similar and can be found

[21]The Preimage Theorem applies since all sets in our framework are in $\mathbb{R}$. Otherwise, similar arguments would follow applying the Regular Level Set Theorem, which is equivalent to the Constant Rank Theorem, see Tu (2011: Ch. 9-10).

in Slough and Tyson (2022: Theorem 2). □

Proof of Theorem 2. Sufficiency is obvious. For necessity, notice first that target-congruence, when combined with harmonization, is equivalent to sign-congruent external validity. To establish the necessity of harmonization over measurement strategies we suppose that sign-congruent external validity holds and proceed by contradiction. In particular, suppose that there exist two studies, $\mathcal{E}_i$ and $\mathcal{E}_j$, which are contrast harmonized but not measurement harmonized, but where target-congruence is satisfied.

The treatment effect function is a smooth function (almost everywhere) that maps from the set of designs and settings to the set of effects: $\tau_m(\omega', \omega'' \mid \theta) : M \times \Omega \times \Theta \to \mathbb{R}$. Its composition with the function $sign : \mathbb{R} \to \{-1, 0, 1\}$, allows us to partition the set of effects, i.e., the image of τ, into three sets. Sign-congruent external validity implies that these sets do not depend on θ, and so for parsimony we drop θ unless needed to avoid confusion. Explicitly, we have the following sets:

$$E_m^+ \equiv \{x \in \mathbb{R} \mid \tau_m(\omega', \omega'') = x > 0\},$$

and

$$E_m^0 \equiv \{x \in \mathbb{R} \mid \tau_m(\omega', \omega'') = x = 0\},$$

and

$$E_m^- \equiv \{x \in \mathbb{R} \mid \tau_m(\omega', \omega'') = x < 0\}.$$

Note that $E_m^+ \cup E_m^0$ and $E_m^- \cup E_m^0$ are each manifolds with boundary, and their common boundary is E_m^0. Moreover, since $E_m^+ \cap E_m^- = \emptyset$, the set E_m^0 is separating, and symmetry of τ ensures nonemptiness.

Next, we focus on the preimage of $sign$ in $\mathcal{C}$. Since τ is smooth and regular on $\mathcal{C}$, the sets $\tau_m^{-1}(E_m^+ \cup E_m^0) \subset \mathcal{C}$ and $\tau_m^{-1}(E_m^- \cup E_m^0) \subset \mathcal{C}$ are manifolds with common boundary $\tau_m^{-1}(E_m^0) \subset \mathcal{C}$. Moreover, the set $\tau_m^{-1}(E_m^0)$ is a boundaryless 1-dimensional manifold (see Guillemin and Pollack (1974: pg. 59)).

Define the set $H(\mathcal{E}_i, \mathcal{E}_j) = \tau_m^{-1}(E_{m_i}^0) \cup \tau_m^{-1}(E_{m_j}^0)$, and note that the elements of $H(\mathcal{E}_i, \mathcal{E}_j)$ are precisely those that have a different sign in study i than in study j. By assumption, the measurement strategies in i and j, m_i and m_j, are distinct. Now consider the boundary sets, $\tau_m^{-1}(E_{m_i}^0)$ and $\tau_m^{-1}(E_{m_j}^0)$, each of which have positive Lebesgue measure in $\mathbb{R}$. Since measurement strategies are distinguishable almost everywhere, i.e., τ's derivative in m has full rank almost everywhere, the set $\tau_m^{-1}(E_{m_i}^0) \cap \tau_m^{-1}(E_{m_j}^0)$ has dimension 0, and hence has Lebesgue measure 0 in $\mathbb{R}$. Hence, $\tau_m^{-1}(E_{m_i}^0)$ and $\tau_m^{-1}(E_{m_j}^0)$ are generically distinct. This establishes that the set $H(\mathcal{E}_i, \mathcal{E}_j)$ has a nonempty interior, and thus, positive Lebesgue measure, contradicting that sign-congruence holds

almost everywhere. An identical argument applies to harmonization of contrasts. □

Proof of Theorem 3. Follows by observing that the relevant set becomes $H(\mathcal{E}_i, \mathcal{E}_j, \mathcal{E}_k) = H(\mathcal{E}_i, \mathcal{E}_j) \cup \tau_m^{-1}(E^0_{m_k})$. □

References

Abramson, Scott F, Korhan Kocak, and Asya Magazinnik. 2022. "What Do We Learn About Voter Preferences From Conjoint Experiments?" *American Journal of Political Science* Forthcoming.

Allcott, Hunt. 2015. "Site selection bias in program evaluation." *The Quarterly Journal of Economics* 130 (3): 1117–1165.

Ashworth, Scott, Christopher R Berry, and Ethan Bueno de Mesquita. 2021. *Theory and Credibility: Integrating Theoretical and Empirical Social Science*. Princeton University Press.

Banerjee, Abhigit V., and Esther Duflo. 2009. "The Experimental Approach to Development Economics." *Annual Review of Economics* 1: 151–178.

Banerjee, Abhijit V, Sylvain Chassang, Sergio Montero, and Erik Snowberg. 2020. "A theory of experimenters: Robustness, randomization, and balance." *American Economic Review* 110 (4): 1206–30.

Bisbee, James, Rajeev Dehejia, Cristian Pop-Eleches, and Cyrus Samii. 2017. "Local instruments, global extrapolation: External validity of the labor supply–fertility local average treatment effect." *Journal of Labor Economics* 35 (S1): S99–S147.

Björkman, Martina, and Jakob Svensson. 2009. "Power to the People: Evidence from a Randomized Field Experiment on Community-Based Monitoring in Uganda." *Quarterly Journal of Economics* 124 (2): 735–769.

Blackwell, David. 1953. "Equivalent comparisons of experiments." *The annals of mathematical statistics* 24 (2): 265–272.

Bueno de Mesquita, Ethan, and Scott A Tyson. 2020. "The commensurability problem: Conceptual difficulties in estimating the effect of behavior on behavior." *American Political Science Review* 114 (2): 375–391.

Camerer, Colin F., Anna Dreber, Eskil Forsell, Teck-Hua Ho, Jürgen Huber, Magnus Johannesson, Michael Kirchler, Johan Almenberg, Adam Altmejd, Taizan Chan, Emma Heikensten, Felix Holzmeister, Taisuke Imai, Siri Isaksson, Gideon Nave, Thomas Pfeiffer, Michael Razen, and Hang Wu. 2016. "Evaluating replicability of laboratory experiments in economics." *Science* 351 (6280): 1433–1436.

Camerer, Colin F., Anna Dreber, Felix Holzmeister, Teck-Hua Ho, Jürgen Huber, Magnus Johannesson, Michael Kirchler, Gideon Nave, Brian A. Nosek, Thomas Pfeiffer, Adam Altmejd, Nick Buttrick, Taizan Chan, Yiling Chen, Eskil Forsell, Anup Gampa, Emma Heikensten, Lily Hummer, Taisuke Imai, Siri Isaksson, Dylan Manfredi, Julia Rose, Eric-Jan Wagenmakers, and Hang Wu. 2018. "Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015." *Nature Human Behaviour* 2 (9): 637–644.
URL: *https://doi.org/10.1038/s41562-018-0399-z*

Chassang, Sylvain, Gerard Padró i Miquel, and Erik Snowberg. 2012. "Selective trials: A principal-agent approach to randomized controlled experiments." *American Economic Review* 102 (4): 1279–1309.

Collins, Harry. 1992. *Changing order: Replication and induction in scientific practice*. University of Chicago Press.

Deaton, Angus. 2010. "Instruments, randomization, and learning about development." *Journal of economic literature* 48 (2): 424–55.

Deaton, Angus, and Nancy Cartwright. 2018. "Understanding and misunderstanding randomized controlled trials." *Social Science & Medicine* 210: 2–21.

Dehejia, Rajeev, Cristian Pop-Eleches, and Cyrus Samii. 2021. "From Local to Global: External Validity in a Fertility Natural Experiment." *Journal of Business & Econonic Statistics* 39 (1): 217–243.

Dunning, Thad. 2016. "Transparency, Replication, and Cumulative Learning: What Experiments Alone Cannot Achieve." *Annual Review of Political Science* 19: S1–S23.

Egami, Naoki, and Erin Hartman. 2022. "Elements of external validity: Framework, design, and analysis." *American Political Science Review* Forthcoming.

Fariss, Christopher J, and Zachary M Jones. 2018. "Enhancing validity in observational settings when replication is not possible." *Political Science Research and Methods* 6 (2): 365–380.

Findley, Michael G, Kyosuke Kikuta, and Michael Denly. 2021. "External Validity." *Annual Review of Political Science* forthcoming: 1–51.

Fowler, Anthony, and B. Pablo Montagnes. 2015. "College football, elections, and false-positive results inobservational research." *Proceedings of the National Academy of Sciences* 112 (45): 13800–13804.

Fowler, Anthony, and B. Pablo Montagnes. 2022*a*. "Distinguishing between False Positives and Genuine Results: The Case of Irrelevant Events and Elections." *Journal of Politics* Forthcoming.

Fowler, Anthony, and B. Pablo Montagnes. 2022*b*. "On the Importance of Independent Evidence: A Reply to Graham et al." Working paper, available at `https://drive.google.com/file/d/16bV6Cyhau6spf6ahz4P1eO1k-O7lR2lt/view`.

Galiani, Sebastian, Paul Gertler, and Mauricio Romero. 2017. "Incentives for Replication in Economics." NBER Working Paper No. 23576.

Gechter, Michael, and Rachael Meager. 2021. "Combining Experimental and Observational Studies in Meta-Analysis: A Mutual Debiasing Approach." *Mimeo* .
URL: *https://www.personal.psu.edu/mdg5396/MGRM_Combining_Experimental_and_Observational_Studies.pdf*

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

Gilbert, Daniel T., Gary King, Stephen Pettigrew, and Timothy D. Wilson. 2016. "Comment on "Estimating the reproducibility of psychological science"." *Science* 351 (6277): 1037–1038.

Goldberger, Arthur S. 1972. "Structural equation methods in the social sciences." *Econometrica* pp. 979–1001.

Graham, Matthew H., Gregory A. Huber, Neil Malhotra, and Cecilia Hyunjung Mo. 2021. "Irrelevant Events and Voting Behavior:Replications Using Principles from Open Science." *Journal of Politics* Forthcoming.

Graham, Matthew H., Gregory A. Huber, Neil Malhotra, and Cecilia Hyunjung Mo. 2022. "How Should We Think About Replicating Observational Studies? A Reply to Fowler and Montagnes." *Journal of Politics* Forthcoming.

Guala, Francesco. 2005. *The methodology of experimental economics*. Cambridge University Press.

Guillemin, Victor, and Alan Pollack. 1974. *Differential topology*. AMS Chelsea Publishing.

Healy, Andrew J., Neil Malhotra, and Cecilia Hyunjung Mo. 2015. "Determining false-positives requires consideringthe totality of evidence." *Proceedings of the National Academy of Sciences* 112 (48): E6591.

Holland, Paul W. 1986. "Statistics and causal inference." *Journal of the American statistical Association* 81 (396): 945–960.

Imbens, Guido W, and Joshua D Angrist. 1994. "Identification and Estimation of Local Average Treatment Effects." *Econometrica* 62 (2): 467–475.

Izzo, Federica, Torun Dewan, and Stephane Wolton. 2020. "Cumulative knowledge in the social sciences: The case of improving voters' information." *Available at SSRN 3239047* .

Klein, Richard A, Kate A Ratliff, Michelangelo Vianello, Reginald B Adams Jr, Štěpán Bahník, Michael J Bernstein, Konrad Bocian, Mark J Brandt, Beach Brooks, Claudia Chloe Brumbaugh et al. 2014. "Investigating variation in replicability: A "many labs" replication project." *Social psychology* 45 (3): 142.

Koole, Sander L., and Daniël Lakens. 2012. "Rewarding Replications: A Sure and Simple Way to Improve Psychological Science." *Perspectives on Psychological Science* 7 (6): 608–614.

Koopmans, Tjalling C, and Olav Reiersol. 1950. "The identification of structural characteristics." *The Annals of Mathematical Statistics* 21 (2): 165–181.

Lovett, Adam, and Kevin Munger. 2019. "Temporal Validity, Prediction and the Problem of Replicability." Working paper, available at https://osf.io/yzghn/.

Meager, Rachael. 2019. "Understanding the average impact of microcredit expansions: A bayesian hierarchical analysis of seven randomized experiments." *American Economic Journal: Applied Economics* 11 (1): 57–91.

Monin, Benoît, and Daniel M. Oppenheimer. 2014. "The Limits of Direct Replications and the Virtues of Stimulus Sampling." *Social Psychology* 45 (4): 1–2.

Morton, Rebecca B, and Kenneth C Williams. 2010. *Experimental political science and the study of causality: From nature to the lab*. Cambridge University Press.

Munger, Kevin. 2021. "Temporal validity.".
URL: *https://osf.io/4utsk/*

Nosek, Brian A., Jeffrey R. Spies, and Matt Motyl. 2012. "Scientific Utopia II: II. Restructuring Incentives and Practices to Promote Truth Over Publishability." *Perspectives on Psychological Science* 7 (6): 615–631.

Nosek, Brian, and Timothy M. Errington. 2017. "Making Sense of Replication." *eLife* 6 (e23383).

Open Science Collaboration. 2015. "Estimating the reproducibility of psychological science." *Science* 349 (6251): 1–8.

Ou, Kai, and Scott A. Tyson. 2022. "Better Observation Leads to Better Inference.".

Pearl, Judea, and Elias Bareinboim. 2011. Transportability of causal and statistical relations: A formal approach. In *Twenty-fifth AAAI conference on artificial intelligence*.

Pearl, Judea, and Elias Bareinboim. 2014. "External validity: From do-calculus to transportability across populations." *Statistical Science* 29 (4): 579–595.

Raffler, Pia, Daniel N. Posner, and Doug Parkerson. 2020. "Can Citizen Pressure be Induced to Improve Public Service Provision?" Working paper, Harvard University.

Rubinstein, Ariel. 2001. "A theorist's view of experiments." *European Economic Review* 45 (4-6): 615–628.

Samii, Cyrus. 2016. "Causal empiricism in quantitative research." *The Journal of Politics* 78 (3): 941–955.

Schmidt, Stefan. 2009. "Shall We Really Do it Again? The Powerful Concept of Replication is Neglected in the Social Sciences." *Review of General Psychology* 13 (2): 90–100.

Shadish, William, Thomas D Cook, and Donald T. Campbell. 2002. *Experimental and quasi-experimental designs for generalized causal inference*. Houghton Mifflin Boston, MA.

Simonsohn, Uri. 2015. "Small Telescopes: Detectability and the Evaluation of Replication Results." *Psychological Science* 26 (5): 559–569.

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

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

Smith, Vernon L. 1982. "Microeconomic systems as an experimental science." *The American Economic Review* 72 (5): 923–955.

Tu, Loring W. 2011. *An Introduction to Manifolds*. Springer.

Wilke, Anna, and Cyrus Samii. 2022. "To Harmonize or Not? Research Design for Cross-Context Learning." Working paper, New York University.