Political Analysis (2025), **00**, 1–16
doi:10.1017/pan.2024.26

ARTICLE

Sign-Congruence, External Validity, and Replication

Tara Slough[1] and Scott A. Tyson[2]

[1] Assistant Professor, Department of Poltics, New York University, New York, NY, USA; [2] Associate Professor, Department of Political Science, University of Rochester, Rochester, NY, USA and Research Associate, W. Allen Wallis Institute of Political Economy, University of Rochester, Rochester, NY, USA

Corresponding author: Scott A. Tyson Email: styson2@ur.rochester.edu

(Received 6 November 2023; revised 7 March 2024; accepted 8 March 2024)

Abstract

We develop a formal framework for accumulating evidence across studies and apply it to develop theoretical foundations for replication. Our primary contribution is to characterize the relationship between replication and distinct formulations of external validity. Whereas conventional wisdom holds that replication facilitates learning about external validity, we show that this is not, in general, the case. Our results show how comparisons of the magnitude or sign of empirical findings link to distinct concepts of external validity. However, without careful attention to the research design of constituent studies, replication can mislead efforts to assess external validity. We show that two studies must have essentially the same research designs, i.e., be harmonized, in order for their estimates to provide information about any kind of external validity. This result shows that even minor differences in research design between a study and its replication can introduce a discrepancy that is typically overlooked, a problem that becomes more pronounced as the number of studies increases. We conclude by outlining a design-driven approach to replication, which responds to the issues our framework identifies and details how a research agenda can manage them productively.

Keywords: Replication; external validity; sign-congruence; harmonization

Edited by: Jeff Gill

When contextualizing empirical findings, researchers often make comparisons of the form: "study A finds that X increases Y, whereas we find no evidence that X increases Y" or "like study A, we find that X increases Y." Such comparisons are widespread in individual articles and literature reviews. They implicitly invoke an expectation that similar findings would be observed in different contexts if probed empirically. But they take for granted how differences in research design can undermine such conclusions. Measuring and assessing differences between empirical studies serves as an explicit goal in *replication*, which seeks to address the same substantive question by comparing results from distinct, yet similar, empirical studies.

Replication is frequently advanced as a means of accumulating evidence (Banerjee and Duflo 2009; Dunning 2016). But replication is invoked by different practitioners for different purposes. Some seek to establish the external validity (or generality) of a mechanism, while others want to evaluate the robustness of an empirical finding to statistical concerns (e.g., lack of power), and still others aim to evaluate research integrity (e.g., fraud). Under what conditions can replication be used for each of these purposes? Can a single replication study provide information about external validity, statistical robustness, and research integrity?

In this paper, we develop a formal framework for evidence accumulation, building on Slough and Tyson (2023, 2024a). We use this framework to examine the theoretical foundations of replication, and, by extension, less formal efforts to compare results across studies. In particular, we seek to understand when the comparison of empirical results provide evidence about the generality—or external validity—of a mechanism. In particular, we analyze two forms of external validity that are relevant for the comparison of empirical findings through replication.[1]

Within our framework, there are two theoretical—and non-statistical—reasons why a replication study can produce results that are different from an original study. First, empirical findings may differ because external validity fails, i.e., the phenomenon of interest does not transcend time, place, or circumstance. Second, differences between two empirical findings may result from differences in the research designs between the two studies. For example, if the treatment conditions in two studies are different, then studies implicitly make different comparisons, which leads to differences in observed treatment effects. We call differences in empirical results due to failures of external validity *target discrepancies* and differences due to variation in research designs *artifactual discrepancies*. Standard presentations of replication presume that these two discrepancies can be conceptualized as purely statistical—not theoretical—issues, and worse, presume that these discrepancies are drawn from some convenient probability distribution.

Our first set of results clarify when the comparisons invoked in replication are useful for the accumulation of empirical evidence. Our primary contribution is to characterize the relationship between replication and formulations of external validity. Conventional wisdom holds that replication facilitates learning about external validity (Banerjee and Duflo 2009). However, we show that this is not, in general, the case. Our results show that comparisons of the magnitude or the sign of empirical findings each link to distinct concepts of external validity, revealing that additional research design considerations are necessary to learn about external validity using replication. Thus, 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.

Comparing estimates from two studies of the same phenomenon (which is ostensibly the goal of replication) is challenging because constituent studies need to "aim at the same thing"—or have the same *empirical target*—to speak to the same substantive question. A mechanism has *exact external validity* if it produces the same empirical target in different settings under an otherwise identical experiment (Slough and Tyson 2023), and has *sign-congruent external validity* when it produces an empirical target with the same sign in different settings. Exact external validity is a stronger condition (in a logical sense) as it implies sign-congruent external validity, whereas a mechanism with sign-congruent external validity need not exhibit exact external validity. The analysis of sign-congruent external validity distinguishes us from Slough and Tyson (2023) who analyze only exact external validity. Consequently, a key contribution of this article is to articulate the concept of sign-congruent external validity, which accommodates directional theoretical implications (e.g., "an increase in X causes an increase in Y"), which are dominant in the social sciences.

We say that two studies are *target-congruent* when their empirical targets (e.g., treatment effects) have the same sign (positive or negative). When two studies make the same comparisons (e.g., same treatment/control), and measure things the same way (including all considerations that go into measuring the effect of a contrast), then we say that they are *harmonized*. We show that only by harmonizing two studies can researchers eliminate artifactual discrepancies.[2] Our main results connect sign-congruent external validity and harmonization to target-congruence. Specifically, a collection of harmonized studies are target-congruent (meaning their empirical targets have the same sign) if and only if sign-congruent external validity holds across all studies. Moreover, if sign-congruent external validity holds, then a collection of studies are target-congruent if and only if all the studies are harmonized. Our results

[1] Our framework also encompasses other concepts of external validity, which are less directly relevant to replication. For a full treatment, see Slough and Tyson (2024a).

[2] Artifactual discrepancies may also reflect the constraints researchers face, e.g., measuring the influence of a mechanism under the same conditions may be impossible in some cases.

https://doi.org/10.1017/pan.2024.26 Published online by Cambridge University Press

show how evaluating a mechanism's sign-congruent external validity is a more demanding endeavor than is typically acknowledged (albeit informally).

To stress how heuristic approaches to evidence accumulation that rely on conducting more studies can be misleading, we identify a novel tradeoff that arises when increasing the number of studies. 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. Whereas adding more studies is *helpful* for addressing statistical concerns, it is potentially *harmful* in light of the theoretical concerns we present. These results suggest that the guidance to "do more studies" to assess a mechanism's external validity under-appreciates the downsides of this approach absent additional guidance on the structure of replication agendas.

Our second set of results assess properties of two common statistical tests that are used in replication. The first, the *estimate-comparison test*, examines the difference in point estimates from constituent studies, thus probing target-equivalence (i.e., that two studies have the same empirical target). The second, the *sign-comparison test*, probes target-congruence by comparing 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 discrepancies become conflated with external validity, and conventional tests cannot distinguish the source of differences in measured effects.

We conclude by outlining a *design-driven approach to replication*, providing guidance for a replication agenda that involves a sequential process that more carefully moves from replicating an experiment to replicating a phenomenon. Our approach keeps an eye toward understanding what artifactual discrepancies may be present because of different research design features and how to measure such discrepancies. Existing expositions of replication are qualitative and classify different replications as exact, direct, or conceptual, which differ according to how much of the original experiment they hold constant (Collins 1992; Guala 2005; Nosek and Errington 2017; Schmidt 2009). Our results highlight that this distinction is insufficiently precise. Since our framework distinguishes between a study's sample, setting, and research design, it allows us to expand on common expositions of replication. A design-driven approach to replication gives a more natural connection between research design and causal effects, and provides a way of bestowing a causal interpretation to effects that arise in multiple places and at different times. This is the key advantage of our framework: it is the only approach to evidence accumulation that remains consistent with the experimental approach to empirical studies (Slough and Tyson 2024a). By the experimental approach, we mean that empirical results seek to measure counterfactual comparisons, regardless of the method employed (Rosenbaum, 2017).

Existing presentations of external validity offer model-based accounts of the cross-study environment (e.g., Egami and Hartman 2022; Findley, Kikuta, and Denly 2021; Shadish, Cook, and Campbell 2002) that are more elaborate than what is typically invoked by replication practitioners. By stressing the importance of research design, we provide a design-driven approach to replication, thereby formally linking the literatures on replication and external validity. Finally, this paper contributes to an emerging literature on the "theoretical implications of empirical models" that examines theoretical properties of common empirical research designs (Abramson, Koçak, and Magazinnik 2022; Banerjee *et al.* 2020; Bueno de Mesquita and Tyson 2020; Slough 2023). We join Izzo, Dewan, and Wolton (2020) and Slough and Tyson (2023) in modeling the cross-study environment to study learning about external validity or generalizability of empirical findings.

1. Framework: Studies

We expand the framework originally presented by Slough and Tyson (2023) 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 or observational studies. What matters is that these studies are unified by the presence of a common (set of) mechanism(s), which motivates

https://doi.org/10.1017/pan.2024.26 Published online by Cambridge University Press

comparison of study estimates as an exercise in *knowledge accumulation*. Unless stated otherwise, all sets are measure spaces with strictly positive Lebesgue measure and are smooth manifolds.

A **measurement strategy**, denoted by $m \in M \subset \mathbb{R}$, captures the choices a researcher makes when choosing an outcome of interest and devising a measure of that outcome, where M represents the set of potential measurement strategies. Every study involves a **contrast**, $(\omega',\omega'') \in \mathcal{C} \subset \mathbb{R}^2$, where $\mathcal{C}$ is compact, which defines the comparison of interest between two instrument values. The two instrument values are taken from the set of all potential comparisons, and are most commonly referred to as "treatment" and "control." The **setting**, $\theta \in \Theta \subset \mathbb{R}$ captures attributes of individual units (i.e., subjects) as well as features of the environment where the study is conducted.[3]

Definition 1. A **study**, $\mathcal{E} = \{m,(\omega',\omega''),\theta\}$, is a research design, comprised of a measurement strategy, m, a contrast, (ω',ω''), and is conducted in a setting, θ.

An empirical exercise measures the presence and influence of a mechanism by looking at its effect, and the effect in a particular study is its **empirical target**, which is mapped to from a study, and we formalize it as follows.[4]

Definition 2. For a measurement strategy $m \in M$, a contrast $(\omega',\omega'') \in \mathcal{C}$, and setting $\theta \in \Theta$, the **treatment effect function** is a function, $\tau_m(\omega',\omega'' \mid \theta) : M \times \mathcal{C} \times \Theta \to \mathbb{R}$, that is smooth almost everywhere, whose derivative has full rank in measurement strategies and contrasts, and for which $sign(\tau_m(\omega',\omega'' \mid \theta)) = -sign(\tau_m(\omega'',\omega' \mid \theta))$.

The empirical target is the measured effect of a study as it relates to how things are measured, which comparison is made, and features of the setting where the study is conducted (time, location, etc.). Our framework accommodates all standard causal estimands, including variations on the marginal treatment effect of Heckman and Vytlacil (2005). 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 locally with the research design. Our framework emphasizes the relationship between research design and empirical targets, distinguishing it from others (e.g., UTOS, PICO, etc.), which are special cases of our framework.[5] The final condition holds that reversing the order of the instrument value changes the sign of the empirical target, which holds for treatment effects defined in terms of differences in potential outcomes.[6]

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 random 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, $\varepsilon_j^{n_j}$. The index j is over different studies. Introducing distributions over this observation error induces a Blackwell experiment

[3] Munger (2023) identifies the importance of time as a feature of settings.

[4] Specifically, the empirical target is a point in the image of the treatment effect function.

[5] In particular, UTOS of Shadish *et al.* (2002), or PICO, which is common in medical meta-studies, follow from our framework by imposing that the effect of interest is independent of comparisons that are made (contrasts) or how outcomes are measured (measurement strategies).

[6] Appendix C.1 develops the connection between our framework and the potential outcomes model.

https://doi.org/10.1017/pan.2024.26 Published online by Cambridge University Press

(Blackwell 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 $\mathbb{E}(\varepsilon_i^{n_i} - \mathbb{E}[\varepsilon_j^{n_i}])^2 \to 0$ (in measure) as $n_i \to \infty$.

2. Concepts: Comparing Studies

When comparing two or more studies, there may be systematic differences that are not statistical, arisising from differences between the design of constituent studies, the settings at hand, or the mechanism(s) producing the treatment effects. As a result, these differences cannot be reduced to "statistical 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.

Definition 3. 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:

1. **target-equivalent** if their empirical targets are equal, i.e.,

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

2. **target-congruent** if their empirical targets share the same sign, i.e.,

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

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 generically different and may even have different signs. Our focus is instead on the non-statistical reasons for differences in estimates across studies, because such differences cannot be solved using statistical techniques.

2.1. 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.

Definition 4. For research design $\mathcal{D} = \{m,(\omega',\omega'')\}$, comprised of measurement strategy, $m \in M$ and contrast, $(\omega',\omega'') \in \mathcal{C}$, the **target discrepancy** from settings θ_i to θ_j is

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

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 two settings. As such, $\Delta_{\mathcal{D}}(\theta_i,\theta_j)$ identifies the difference in empirical targets that is attributable to moving from setting θ_i to θ_j, holding fixed the 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 discrepancies which typically falls under the label of "external validity."

Definition 5 (Slough and Tyson (2023)). A mechanism has **exact external validity** from settings θ_i to θ_j if for almost every measurement strategy $m \in M$ and almost every contrast (ω',ω'')

$$\tau_m(\omega',\omega'' \mid \theta_i) = \tau_m(\omega',\omega'' \mid \theta_j).$$

A mechanism is externally valid if it has exact external validity for almost all settings $\theta \in \Theta$.

Exact external validity may be more than one needs. A researcher may be interested in assessing the *sign*, rather than the precise *magnitude* of treatment effects across different settings. Moreover, if a mechanism is only activated for a subset of units—e.g., a drug therapy works only on women—differences in sample composition will differentially dilute treatment effects. In either case, 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 6. A mechanism has **sign-congruent external validity** from settings θ_i to θ_j if for almost every measurement strategy $m \in M$ and almost every contrast (ω', ω'')

$$sign(\tau_m(\omega', \omega'' \mid \theta_i)) = sign(\tau_m(\omega', \omega'' \mid \theta_j)).$$

A mechanism is sign-congruent externally valid if it has sign-congruent external validity for almost all settings $\theta \in \Theta$.

Sign-congruent external validity is similar to exact external validity in that each expresses a theoretical property of empirical targets across settings. Definition 6, however, only requires that the empirical targets across studies share the same sign, rather than having to be the same magnitude (as in Definition 5). Indeed, sign-congruent external validity is logically weaker in that any mechanism that has exact external validity has sign-congruent external validity, i.e., exact external validity implies sign-congruent external validity, but that a mechanism that has sign-congruent external validity need not have exact external validity.

2.2. 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. 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 7. For setting $\theta \in \Theta$, the **artifactual discrepancy** between designs $\mathcal{D}_i = \{m_i, (\omega_i', \omega_i'')\}$ and $\mathcal{D}_j = \{m_j, (\omega_j', \omega_j'')\}$ 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 come from *using different research designs*. Design-induced discrepancies are "artifactual," but this does not imply that these discrepancies are "nuisance" parameters. To illustrate that artifactual discrepancies are fundamentally non-random, suppose that two studies observe the effect of a drug on patients but where each study administered different dosages. 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 in the same setting. Thus, failure to adjust for dosage differences would result in artifactual discrepancies.

Definition 8. 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** if they have the same measurement strategy, i.e., if $m_1 = m_2$, and the same contrast, i.e., if $(\omega_1', \omega_1'') = (\omega_2', \omega_2'')$.

Harmonization might be thought of as "design-equivalence" since it is about ensuring that the research designs between two studies are essentially the same, i.e., the same comparisons are being made and all quantities are measured in the same way. This does not imply that research designs are literally the same, but that they perform the same role in different settings. In Appendix C we use two conceptual examples to illustrate how theoretical and practical considerations help determine whether harmonization holds.

In contrast to arguments that a lack of harmonization is simply "another source of random error" in replication studies (Gilbert *et al.* 2016, 1037a), issues related to the harmonization between studies are fundamentally non-statistical concerns. They are instead issues of research design, and consequently, eliminating them is ultimately a theoretical and practical question.

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, and thus, may be of independent interest 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—can provide important information about the mechanism's effects or provide novel policy recommendations.[7] It also stresses that an intervention may interact with a mechanism or setting in ways that are not easy to disentangle.

3. Results

Our definitions of external validity and harmonization have clear links to target and artifactual discrepancies and to develop an intuition for these relationships we present a straightforward result.

Remark 1. For two studies, $\mathcal{E}_1 = \{\mathcal{D}_1 = (m_1, (\omega_1', \omega_1'')), \theta_1\}$ and $\mathcal{E}_2 = \{\mathcal{D}_2 = (m_2, (\omega_2', \omega_2'')), \theta_2\}$:

1. The target discrepancy between studies is zero, $\Delta_{\mathcal{D}}(\theta_1, \theta_2) = 0$ for almost all $\mathcal{D}$, if and only if the mechanism of interest has exact external validity between settings θ_1 and θ_2.
2. The artifactual discrepancy is zero, $\mathcal{A}(\mathcal{D}_1, \mathcal{D}_2 \mid \theta) = 0$, almost everywhere if and only if studies 1 and 2 are harmonized.

This follows from combining Definitions 4 and 5 and highlights the conceptual link between external validity and target discrepancies, and between harmonization and artifactual discrepancies. The first part of Remark 1 stresses that target discrepancies emerge *because* the mechanism lacks exact external validity between two settings. The absence of exact external validity does not make any statement about the magnitude or sign of target discrepancies, only that they are non-zero. The second part of Remark 1 shows how artifactual discrepancies highlight the importance of harmonization between different studies.

We now consider target-congruence and its relationship with harmonization of study designs and sign-congruent external validity.

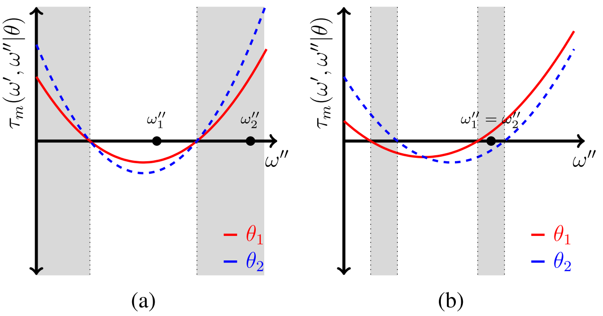
Theorem 1 (Target-congruence). *For any collection of studies,* $\{\mathcal{E}_i = (m_i, (\omega_i', \omega_i'', \theta_i)\}_{i=1}^N$,

(a) *if sign-congruent external validity holds across i then they are target-congruent if and only if every study is harmonized;*
(b) *if* $\mathcal{E}_i$ *is harmonized for all i, then they are target-congruent if and only if sign-congruent external validity holds across i.*

A key component of the proof of Theorem 1 is the "sign-flip" set, where target-congruence fails, and the details of its construction are in the appendix. This set is constructed for measurement strategies by focusing on the set of contrasts where the sign is different between two different measurement strategies.

[7] If a researcher were only interested in, e.g., integer values of an intervention (relative to the function τ), then this would involve a (plausibly) continuous distribution reflecting the analyst's uncertainty about the technology of intervention; this kind of uncertainty is outside of our model.

https://doi.org/10.1017/pan.2024.26 Published online by Cambridge University Press



(a)

(b)

Figure 1. Illustration of Theorem 1. The grey regions in panel (a) depict the sign-flip sets, or the regions where target-congruence fails when ω''s are not harmonized. The grey regions in panel (b) depict the regions where target-congruence fails due to a lack of sign-congruent external validity.

This set is important because it is where the the sign of an empirical target is different depending only on changing the measurement strategy—not because the sign of the mechanism's effect varies over settings. The proof of Theorem 1 establishes that this sign-flip set has strictly positive measure, and this is a problem because it implies that any distribution over effects incorrectly identifies when a mechanism's effect has the same sign in different places.[8] The main intuition for Theorem 1 is that despite sign-congruent external validity being less demanding that exact external validity, for target-congruence to hold, exact external validity must hold when the empirical target is 0. Another way of interpreting Theorem 1 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.

Figure 1 illustrates Theorem 1. Panel (a) shows that even when sign-congruent external validity holds, a lack of harmonization—as indicated by the different ω''s—creates the sign-flip sets indicated by the grey regions. In Panel (b), sign-congruent external validity does not hold, and even if researchers harmonize treatment levels across studies (choosing the same ω''), the signs of the empirical targets differ in the grey regions, which is where target-congruence does not hold. Theorem 1 establishes that these sets have positive measure whenever harmonization or sign-congruent external validity do not hold. Moreover, the size of these sets can be arbitrarily large depending on how $\tau_m(\omega',\omega''|\theta)$ varies in setting, θ.[9]

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 statistical discrepancies between studies, the information the analyst gains is substantially complicated when the inclusion of studies introduces target or artifactual discrepancies. Importantly, target and artifactual discrepancies are not 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 apply Theorem 1 to show that artifactual discrepancies are not solved 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 2. *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.*

[8] The probability this happens can be arbitrarily close to 1.

[9] What if exact external validity holds only a strict subset of Θ? In such a case, one needs to identify precisely which settings exhibit exact external validity or sign-congruent external validity.

This result establishes that increasing the number of studies does not make it "easier" to achieve target-congruence but instead 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 2 suggests that there is a dilemma when considering how many studies to include in a replication. While accumulating more studies to obtain more estimates of the treatment effect certainly aids in addressing statistical concerns, it potentially exacerbates problems that arise from research design issues. Only when studies are harmonized does this dilemma not arise. Specifically, although it is generally beneficial to observe more draws of the random variables $\varepsilon_j^{n_j}$, when doing so involves adding nonharmonized studies, it introduces 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 $\varepsilon_j^{n_j}$.

4. Testing External Validity

Replications are increasingly used to study the statistical properties of a study (or collection of studies). Our presentation so far has focused on theoretical issues which are distinct from sampling and estimation, and thus, are independent of statistical issues. Anyone conducting a replication will, in practice, also confront *statistical discrepancies*, and our framework straightforwardly extends to include these concerns.

The first approach to replication compares the measured effects of two studies directly, assessing 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 measured effects 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 by substitution 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 measured effects e_1 and e_2 contains more than just random error, i.e., statistical discrepancies. It also includes target discrepancies (when external validity fails) and artifactual discrepancies (when research 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 the likelihood of observing a given difference in estimates (or signs) under a 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.

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

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

and tests 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*

https://doi.org/10.1017/pan.2024.26 Published online by Cambridge University Press

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

The proof of this result follows from Theorem B.1, which is developed in Supplemental Appendix B. The requirement of unbiasedness and consistency reflects conventional statistical concerns and shows the importance of internal validity of *all* constituent studies. 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 exact external validity the analyst must be able to (credibly) fix one of these discrepancies to zero in order to assess the other. Proposition 1 establishes two findings that are relevant for replication. First, by assuming harmonization, the estimate-comparison test allows for a test of a mechanism's external validity. Second, by assuming exact external validity, the estimate-comparison test permits a test for harmonization—provided the analyst knows independently, or assumes, that the mechanism under study is exact externally valid.

In the presence of non-zero target or artifactual discrepancies, the estimate comparison test risks rejecting the null hypothesis that $\tau_{m_1}(\omega_1', \omega_1'' | \theta_1) = \tau_{m_2}(\omega_2', \omega_2'' | \theta_2)$ 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 exact external validity or harmonization. Direct replications, where the setting is held constant and the design is harmonized, eliminate target and artifactual discrepancies. This replication design allows researchers to learn about statistical discrepancies and is well-suited to questions about publication bias or researcher integrity.[10]

It is important to consider the relationship between Proposition 1 and approaches that leverage replications of multiple distinct studies (e.g., Camerer *et al.* 2016). 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. This analysis suggests that artifactual and target discrepancies can bias estimates of a literature's replicability, but where even the direction of this bias is unknown.

Our second test focuses on the signs of the measured 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 of estimates heuristically without formally testing a null hypothesis. 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:*

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

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

https://doi.org/10.1017/pan.2024.26 Published online by Cambridge University Press

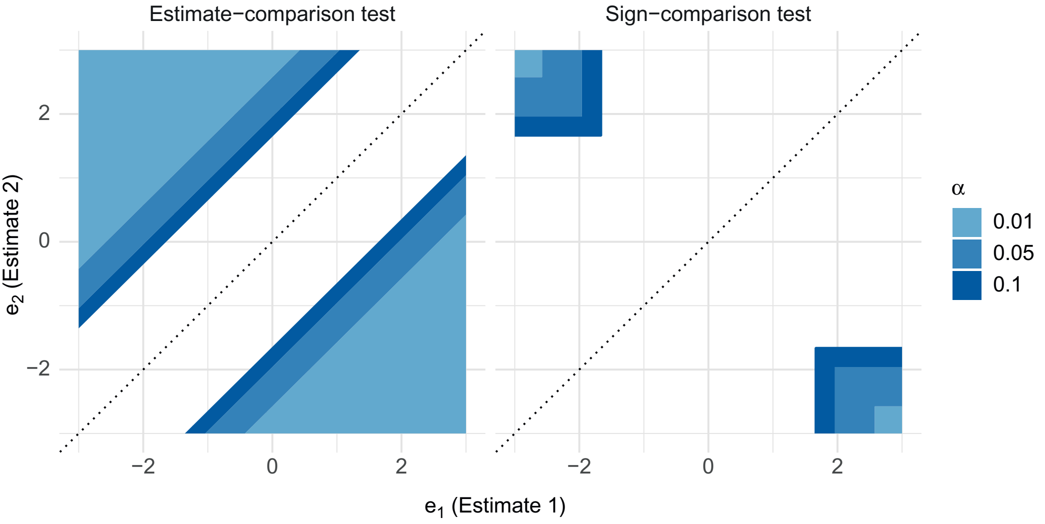


Figure 2. Rejection regions of the estimate- and sign-comparison tests for Type-I error rates, $\alpha \in \{0.01, 0.05, 0.1\}$. Both plots fix $se_1 = se_2 = 1$ in order to visualize these regions in two dimensions.

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

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 1. □

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*. The null hypothesis of the sign-comparison test corresponds to the event in which both empirical targets have the same sign. As such, rejection of this null hypothesis constitutes a rejection of target-congruence. When studies are harmonized, this is equivalently a test for sign-congruent external validity.

Figure 2 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. The details for constructing the p-values in the sign-comparison test are in Appendix D.

What do we learn from a sign-comparison test when studies are *not* necessarily harmonized? Remark 1 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, since 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 some form of external validity or harmonization in the case of the estimate-comparison approach. However, we show that any test for exact 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

https://doi.org/10.1017/pan.2024.26 Published online by Cambridge University Press

Table 1. Classification of replication studies.

		Studies differ in&			
Class	Sub-class	Samples	Settings	Design	Example(s)
Exact		–	–	–	–
Direct		✓	–	–	Camerer *et al.* (2016)
Conceptual	Harmonized	✓	✓	–	Heinrich *et al.* (2006)
Conceptual	Single-setting	✓	–	✓	Boas, Hidalgo, and Melo (2019)
Conceptual	Non-harmonized, multi-setting	✓	✓	✓	Fowler and Montagnes (2015)

(unbiasedness and consistency) as well as cross-study properties (harmonization and 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.

Before moving on, we note that one can, in principle, apply the estimate- or sign-comparison tests to sets of $N > 2$ studies (beyond individual pairwise comparisons), and we provide details on construction of p-values for the sign-comparison test with $N > 2$ studies in the appendix. However, Theorem 2 cautions that if not all studies are harmonized, sign-congruent external validity becomes harder—not easier—to assess through replication.

5. The Design-Driven Approach to Replication

We have established how replication can facilitate learning about different formulations of external validity, and hence generate knowledge about substantive phenomena. We now outline the *design-driven approach to replication*, and use our framework to provide a concept-driven classification of replication studies.

We have described three features that can differ between constituent studies in a replication: samples, setting, and research design (contrasts and measurement strategies). These features map directly onto a replication classification, shown in Table 1, that expands on common expositions of replication, including exact, direct, and conceptual replication (Collins 1992; Guala 2005; Nosek and Errington 2017; Schmidt 2009). Our categorization distinguishes between different types of conceptual replication, and our results stress what can be learned from accumulating evidence through replication. All three sub-classes of conceptual replication are utilized in the social sciences at present—though sometimes not classified as replications—and we provide examples of each in the right column.

Exact replication implies that all aspects of two studies' research design are identical, including the sample, which is typically impossible in the social sciences.[11] 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 (Ou and Tyson 2022; Schmidt 2009). 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). Large-scale efforts to replicate laboratory experiments using identical treatments and

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

https://doi.org/10.1017/pan.2024.26 Published online by Cambridge University Press

outcomes in similar laboratory environments make a credible claim to be direct replications (Camerer *et al.* 2016).[12]

Most replications in social science change more than a study's sample, thereby conducting a *conceptual replication*. 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. Conceptual replications use different samples (like direct replications), but also differ in either the setting a study is conducted or in aspects of research design. Our framework identifies three sub-classes of conceptual replication. In harmonized conceptual replications, researchers implement the same design (i.e., contrasts and measurement strategy) on samples from different settings (and thus different populations). For example Heinrich *et al.* (2006) use simple lab-in-the-field games with common treatments and outcomes to measure how 15 distinct populations engage in costly punishment. While they do not use the term "replication" to describe the multi-setting design, they compare the resultant treatment effects to assess the generality of the phenomenon.

In single-setting conceptual replications, researchers implement a different design (perhaps on a different sample) in the same setting. For example, Boas *et al.* (2019) seek to measure voter responses to corruption revelation using two research designs, a survey and a field experiment, among the same population of voters. These designs probe a common mechanism—voter learning—but use different treatments and different survey outcome measures. These authors similarly do not classify the two experiments as constituent studies in a replication, but they do compare treatment effects to measure how treatment effects vary across technologies of intervention (the two experimental designs). Finally, most conceptual replications in the social sciences should be considered non-harmonized, multi-setting replications. In the appendix we apply our framework to experimental and observational replication studies, by discussing the Raffler, Posner, and Parkerson (2022) replication of Björkman and Svensson (2009)'s study of citizen oversight of healthcare providers and a recent dialogue on the effect of college football game outcomes on pro-incumbent voting (Fowler and Pablo Montagnes 2022; Graham *et al.* 2022).

Motivated by the distinctions highlighted in our framework, we propose a *design-driven approach to replication*, which stresses the importance of a *replication agenda* and how such agendas should be structured. This approach proceeds by admitting one potential discrepancy at a time and is more tightly connected with credibility approaches to internal validity:

1. Conduct **harmonized (conceptual) replications** in settings where the mechanism may be operative. Measure target discrepancies to evaluate 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 a 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 exact external validity.

[12] In campus-based experimental economics laboratories like those in Camerer *et al.* (2016), sampling strategies may be hard to precisely characterize. If one were skeptical of our characterization of a common sampling strategy, these studies could instead be classified as harmonized conceptual replications. Camerer *et al.* (2016, 1433) describes these as "direct replications."

https://doi.org/10.1017/pan.2024.26 Published online by Cambridge University Press

Our categorization of replication, and how each kind of replication is used, as follows:

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

This highlights how each kind of replication addresses a different kind of discrepancy between empirical targets.

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 often costly to implement.

6. Conclusion

The accumulation of empirical evidence collected in multiple places, at different times, and by different scholars presents numerous challenges. Perhaps most important is whether a mechanism is externally valid. Replication (direct and conceptual) is 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 sign-congruent external validity and harmonization of studies are required to guarantee target-congruence between studies. We then develop two sets of results about empirical targets and apply them to two statistical tests—the estimate-comparison and sign-comparison tests. These results have 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 closely resembles the way scholars informally discuss related studies (even outside the context of replication).

Our theoretical results stress the importance of design harmonization, where the measurement strategy and contrast across studies are the same. However, achieving harmonization in some settings may be extremely difficult, or even impossible. Future research should consider the theoretical implications of imperfect harmonization, where, for instance, two treatments which are "sufficiently close" should lead to closeness of empirical targets (i.e., continuity). Another natural extension of our framework involves the role of describing settings using covariates. In particular, if there exists some "reduction set" between the set of settings and the θ argument of τ. This is potentially valuable because two concrete settings may not differ in a meaningful way relative to τ, in which case both settings would map to the same value in the reduction set.

Finally, we introduce a design-driven approach to 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. We conclude by highlighting two important issues that arise in replication agendas. First, a desire for novelty arguably hampers any replication-based research agenda. These concerns are ultimately about professional incentives rather than the accumulation of knowledge. However, a benefit of a sequential replication research agenda 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 guard against researcher malfeasance, and as a result, independence of research teams conducting replications is an important concern. Our notion of harmonization does not in any way preclude independent replication, however, more transparent characterization and reporting of measurement strategies and comparisons will likely be necessary to facilitate independent productive replication.

Acknowledgments. We thank Scott Abramson, Michaël Aklin, Agustín Casas, Ray Duch, Chris Fariss, Naoki Egami, Gleason Judd, Walter Mebane, John Patty, Dan Posner, Pia Raffler, Cyrus Samii, Fredrik Sävje, Ian Turner, Ed Vytlacil, Anna Wilke, Hye Young You and workshop/seminar participants at Columbia, Yale, Polmeth XXXIX, the Knowledge Accumulation and External Validity: Implications for Design and Analysis workshop at the University of Montreal, the Methodological Challenges of Meta-analysis in the Social Sciences Workshop in Lisbon, the Petralia Applied Economics Workshop, and the 2022 American Political Science Association annual conference for invaluable comments.

Funding Statement. We received no funding for this project.

Competing Interests. None.

Data Availability Statement. Replication code for this article has been published in Code Ocean, a computational reproducibility platform that enables users to run the code, and can be viewed interactively at https://doi.org/10.24433/CO.0867235.v2. A preservation copy of the same code and data can also be accessed via Dataverse at https://doi.org/10.7910/DVN/BDCJBZ (*Slough and Tyson 2024b*).

Supplementary Material For supplementary material accompanying this paper, please visit https://doi.org/10.1017/pan.2024.26.

References

Abramson, S. F., K. Koçak, and A. Magazinnik. 2022. “What Do We Learn about Voter Preferences from Conjoint Experiments?” *American Journal of Political Science* 66 (4): 1008–1020.

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

Banerjee, A. V., S. Chassang, S. Montero, and E. Snowberg. 2020. “A Theory of Experimenters: Robustness, Randomization, and Balance.” *American Economic Review* 110 (4): 1206–1230.

Björkman, M., and J. 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, D. 1953. “Equivalent Comparisons of Experiments.” *The Annals of Mathematical Statistics* 24 (2): 265–272.

Boas, T. C., F. Daniel Hidalgo, and M. A. Melo. 2019. “Norms Versus Action: Why Voters Fail to Sanction Malfeasance in Brazil.” *American Journal of Political Science* 63 (2): 385–400.

Bueno de Mesquita, E., and S. 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, C. F., et al. 2016. “Evaluating Replicability of Laboratory Experiments in Economics.” *Science* 351 (6280): 1433–1436.

Collins, H. 1992. *Changing Order: Replication and Induction in Scientific Practice*. University of Chicago Press, Chicago, IL.

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

Egami, N., and E. Hartman. 2023. “Elements of External Validity: Framework, Design, and Analysis.” *American Political Science Review* 117(3): 1070–1088.

Findley, M. G., K. Kikuta, and M. Denly. 2021. “External Validity.” *Annual Review of Political Science* 24: 365–393.

Fowler, A., 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, A., and B. Pablo Montagnes. 2022. “Distinguishing Between False Positives and Genuine Results: The Case of Irrelevant Events and Elections.” *Journal of Politics* 85(1): 304–309.

Gilbert, D. T., G. King, S. Pettigrew, and T. D. Wilson. 2016. “Comment on “Estimating the Reproducibility of Psychological Science.” *Science* 351 (6277): 1037–1038.

Graham, M. H., G. A. Huber, N. Malhotra, and C. H. Mo. 2023. “How Should We Think about Replicating Observational Studies? A Reply to Fowler and Montagnes.” *Journal of Politics* 85(1): 310–313.

Guala, F. 2005. *The Methodology of Experimental Economics*. Cambridge University Press, New York, NY.

Heckman, J. J., and E. Vytlacil. 2005. “Structural Equations, Treatment Effects, and Econometric Policy Evaluation 1.” *Econometrica* 73 (3): 669–738.

Heinrich, J., et al. 2006. “Costly Punishment across Human Societies.” *Science* 312 (June): 1767–1770.

Izzo, F., T. Dewan, and S. Wolton. 2020. “Cumulative Knowledge in the Social Sciences: The Case of Improving Voters’ Information.” Available at SSRN 3239047.

Klein, R. A., et al. 2014. “Investigating Variation in Replicability: A “Many Labs” Replication Project.” *Social Psychology* 45 (3): 142.

Munger, K. 2023. “Temporal Validity as Meta-Science.” *Research & Politics* 10(3): 1–10.

Nosek, B., and T. M. Errington. 2017. “Making Sense of Replication.” *eLife* 6(e23383): 1–4.

Open Science Collaboration. 2015. “Estimating the Reproducibility of Psychological Science.” *Science* 349 (6251): 1–8.

Ou, K., and S. A. Tyson. 2022. “Elicitation in Voting Experiments.” Working Paper. https://drive.google.com/file/d/1QuH_HNGU8Txfqmgp0yaJUOlYpiuKn4EG/view.

Raffler, P., D. N. Posner, and D. Parkerson. 2022. “Can Citizen Pressure be Induced to Improve Public Service Provision?” Working Paper. http://danielnposner.com/wp-content/uploads/2022/04/RPP-ACT-Health-220323.pdf.

Rosenbaum, P. 2017. *Observation and Experiment: An Introduction to Causal Inference*. Cambridge: Harvard University Press.

Schmidt, S. 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, W., T. D Cook, and D. T. Campbell. 2002. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston: Houghton Mifflin.

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

Slough, T. 2023. “Phantom Counterfactuals.” *American Journal of Political Science* 61 (1): 137–153.

Slough, T., and S. A Tyson. 2023. “External Validity and Meta-analysis.” *American Journal of Political Science* 67 (2): 440–455.

Slough, T., and S. A Tyson. 2024a. *External Validity and Evidence Accumulation*. Cambridge University Press, New York, NY.

Slough, T., and S. A. Tyson. 2024b. “Sign-Congruence, External Validity, and Replication: Replication Package.” Working Paper. https://doi.org/10.24433/CO.0867235.v2.

Cite this article: Slough, T. and Tyson, S. A. (2025). Sign-Congruence, External Validity, and Replication. *Political Analysis*, 1–16. https://doi.org/10.1017/pan.2024.26