Sign-Congruence, External Validity, and Replication*

Tara Slough† Scott A. Tyson‡

September 9, 2022

Abstract

We develop a framework for the accumulation of evidence across multiple 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 the empirical targets across studies are the same, and target-congruence, where the sign is the same across studies. We develop results that show how each of these assessment criteria are related to distinct formulations of external validity. We propose a new, less-demanding, formulation, *sign-congruent external validity*, which obtains when the empirical target's sign is the same across settings. Our results 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 conclude with practical guidance for designing studies with an eye toward accumulating empirical evidence in pursuit of general substantive knowledge.

*We thank Gleason Judd, Walter Mebane, John Patty, Dan Posner, Pia Raffler, and participants at Polmeth XXXIX for helpful comments.

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

‡Associate Professor, 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 (Gailmard, 2021). Without such knowledge, careful and credible empirical work may seem highly particular and developing methods for accumulating causal evidence should be a goal for any research community (Esterling, Brady, and Schwitzgebel, 2021; 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. In this article, we develop a framework to highlight key concepts to help understand the accumulation of empirical evidence using replication.

In our framework, an empirical study is how the influence of a mechanism (or set of mechanisms) is measured, i.e., 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 its measurement (Adcock and Collier, 2001). 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 (Findley, Kikuta, and Denly, 2021). These three ingredients combine to define an *empirical target*, or treatment effect, which corresponds to a study's primary estimand.[1]

Comparing estimates from two studies of the same phenomenon—as in a replication study—is challenging because there are multiple reasons that the estimates in these studies might differ

[1]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).

(Gailmard, 2014). 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 to these, 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 in a replication. First, following Slough and Tyson (2022), two studies are *target-equivalent* when they measure 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).[2] 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. We use these concepts to present a conceptual classification of different kinds of replication, nesting common classifications (Collins, 1992; Schmidt, 2009; Nosek and Errington, 2017).

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

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

has external validity then the target discrepancies across studies is zero. We then show that sign-congruent external validity constrains what kind of target discrepancy can arise.

Novel to our framework is the observation that different constituent studies often have different research designs, e.g., different treatments or different outcome measures. Such differences in research designs produce *artifactual discrepancies* between empirical targets because they make different comparisons or measure outcomes differently across studies. Artifactual discrepancies also reflect the inability to measure the influence of a mechanism under the same conditions. 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 contrasts and measurement strategy, we say they are *harmonized*, and show that by harmonizing two studies, researchers can eliminate artifactual discrepancies.

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-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. Second, 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 harmonized. These results, when taken together, highlight how different ways of assessing empirical targets correspond to different notions of external validity. Our analysis additionally clarifies the theoretical foundations for using qualitative comparisons of related studies and interpreting their findings (e.g., in comprehensive literature reviews). For example, statements of the form: "author A's study finds that X increases Y, whereas we find no evidence that X increases Y," is a comparison between studies of a common phenomenon, and implicitly invokes an expectation that it will yield similar things if probed empirically.

Using our results, 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). In particular, we assess the properties of two common statistical tests that are used in replication studies. The first, the *sign-comparison test*, probes target-congruence by examining the signs of estimates from different studies. The second, the *estimate-comparison test*, examines the difference in point estimates from constituent studies, thus probing target-equivalence. 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.

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 (Banerjee and Duflo, 2009; Gerber and Green, 2012; Dunning, 2016).

Motivating Example and Related Literature

Motivated by the poor health outcomes for children in rural Uganda, Björkman and Svensson (2009) present a 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 work-

ers 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.[3] The replication experiment was conducted a decade after the original experiment was fielded and was more heavily-resourced. Specifically, 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

[3]The small number of clusters in the original experiment also motivated the replication.

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, that was observed in Uganda 2004-2005, could be 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 for the discrepancies observed between Björkman and Svensson (2009) and Raffler, Posner, and Parkerson (2020). In particular, 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, whereas Björkman and Svensson (2009) measured outcomes at

[4]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).

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; Pearl and Bareinboim, 2014, 2011), or onto a grand population (e.g., Egami and Hartman, 2022; Findley, Kikuta, and Denly, 2021). Another formulation is parallelism, which is where a finding measured in an experimental settings transports to more natural settings (Guala, 2005).

In this article we formally define external validity, which allows us to conceptually distinguish two distinct versions that are invoked in replication studies. In our framework, external validity characterizes the relationship between multiple studies (or estimates) without reference to some external quantity, sample, or setting. 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] We argue that our formulations of external validity are appropriate for considerations related to replication. In particular, by doing a replication study, authors invest time and often substantial resources in trying to measure an effect in a new sample or setting. This is quite different than using information from a single study to *estimate* or *impute* the effect from one sample or setting 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 one of the estimators surveyed by Egami and Hartman (2022) to the Björkman and Svensson (2009) data.

Finally, this paper contributes to an emerging literature on the "theoretical implications of em-

[6]In Appendix B, 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.

pirical 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) and Banerjee et al. (2020).

Framework

We expand the framework originally presented by Slough and Tyson (2022) and develop new concepts that are important for replication. Suppose that 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 and is a smooth manifold. 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**, $(\omega', \omega'') \in \Omega \subset \mathbb{R}^2$, which defines the comparison of interest between two instrument values (e.g., Imbens and Angrist, 1994). The two instrument values are taken from Ω, 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

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

study is conducted.

An empirical exercise measures the presence and influence of a mechanism by looking at its effect, i.e., by focusing on "the effects of causes" (Holland, 1986). The effect in a particular study comprises its **empirical target**, and we formalize the empirical target 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 and whose derivative has full rank in measurement strategies and contrasts.*

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] The last part of Definition 1—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, M-STOUT etc., which are all special cases of our framework. In particular, UTOS, PICO, and M-STOUT follow from our framework by imposing design invariance, which is wwhen the effect of interest is independent of research design features.[10]

Finally, 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.

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

[10]PICO is common in medical meta-studies, UTOS comes from Shadish, Cook, and Campbell (2002), M-STOUT is elaborated by Findley, Kikuta, and Denly (2021). Unlike these frameworks, the MIDA framework described by Blair et al. (2019) does not necessarily assume 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.

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, $(m_j, (\omega_j^{'}, \omega_j^{''}))$, setting, θ_j, and random noise interrupting the direct measurement of that empirical target, $\varepsilon_j^{n_j}$. Notice that e_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 $n_i \to \infty$.

Classifying Replication

Expositions of replication generally describe three classes of replication designs: exact, direct, and conceptual (Collins, 1992; Schmidt, 2009; Nosek and Errington, 2017). We have described three features of studies that can differ between constituent studies in a replication: samples, setting, and design (contrasts and measurement strategies). These features map directly onto the more conventional typology for replication, as we show in Table 1.

The mapping between our framework and the standard classification generates several insights about replication studies. *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 (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 political science (and social science more broadly) change more than a study's sample, thereby conducting a *conceptual replication*. Our framework clarifies three sub-classes of conceptual replication designs. Like direct replications, constituent studies in conceptual replications use different samples, but this feature is not essential to conceptual replications. Instead, 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. Both sub-classes are different from conceptual replications that are conducted in multiple settings and may not be harmonized across settings, which we term non-harmonized, multi-setting conceptual replications.

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 consitutent studies, there are not established best practices for how these replications should be organized or assessed. Our categorization, and our results below, distinguish between different types of conceptual replication, and what can be learned from accumulating evidence across different replication designs.

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

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-congruent when the targets have the same sign and target-equivalent when their targets are the same. It is important to reiterate that estimates of these targets—the observed e_1 and e_2—are observed with random error. This means that even if two studies were target-equivalent, our estimates of the targets will be different (with probability 1) and they may even have different signs. Our focus is instead on the non-statistical reasons for difference in estimates across studies. Here, there are two possible discrepancies that emerge when we compare estimates between studies, which we term *target* and *artifactual* discrepancies, and which relate to important themes in the literature on research design.

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*

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

When considering the target discrepancy it is important to hold aspects of an empirical design fixed, i.e., holding the measurement strategy, m, and the contrast, (ω', ω'') to the same value across the two settings. This way $\Delta_{\mathcal{D}}(\theta, \theta')$ identifies the difference in empirical targets that is attributable to moving from setting θ to setting θ'. 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 as its defined in Definition 3. 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'.*

This result follows immediately by the definition of external validity, which highlights the link between external validity and target discrepancies. Remark 1 stresses that target discrepancies emerge *because* the mechanism lacks external validity between two settings θ and θ'. It is important to note that 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 can be a stringent condition, and it 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 weaker notion of external validity that is more closely-aligned with directional theories and hypotheses. To this end, we introduce a new concept: *sign-congruent external validity*, which is about the sign of an effect across settings.

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'')*

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

A mechanism has sign-congruent external validity 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 exactly the same (as in Definition 4). Indeed, sign-congruent external validity is a weaker condition (in a logical sense)

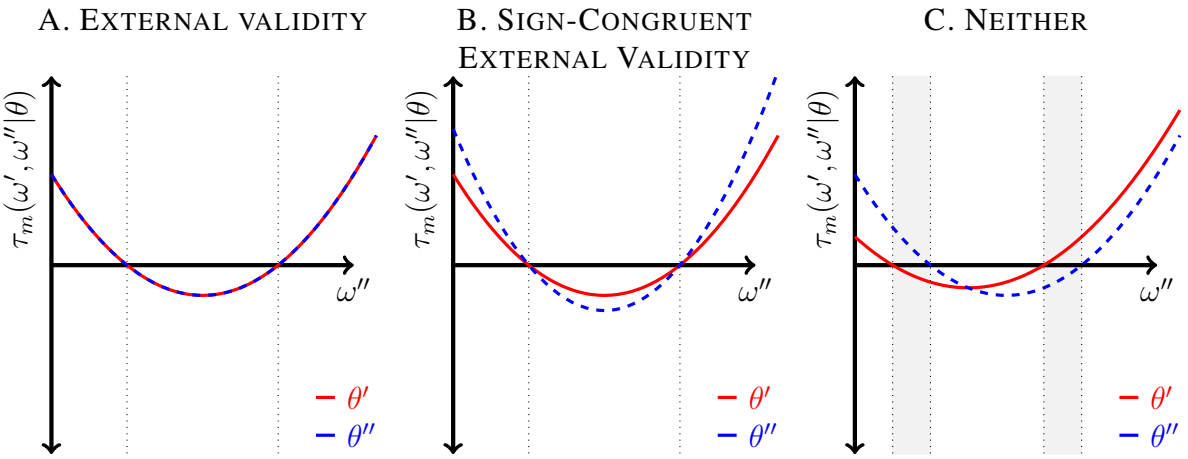

Figure 1: Illustration of external validity and sign-congruent external validity in harmonized experiments in two sites, θ' and θ''. We assume a fixed ω' and m in order to depict these concepts in two dimensions.

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.[12] A case where sign-congruent external validity might be well-suited to two related studies is when a mechanism is not activated for all members of a treatment group in two experiments with the same design. If a non-zero treatment effect is produced but the proportion of such members is different in the different studies, we should not expect resultant average treatment effects to be the same. Yet, we may expect that the average treatment effects to be scaled by the proportion of subjects for which the mechanism is activated. Here, we have good reason to expect sign-congruent external validity but not external validity.

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' (e.g., the control arm is harmonized across studies). We plot treatment effects, $\tau_m(\omega', \omega'' \mid \theta)$, in two settings, θ' and θ'', as a function of the other instrument ω''. In the left panel, we show that external validity implies that treatment effects are identical in both

[12] It is straightforward to construct examples.

settings.[13] Importantly, the plot shows that external validity makes no requirement of functional form, only that the treatment effect function is the same in both settings. The center 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).[14] The right panel depicts a mechanism that lacks sign-congruent external validity, which can be seen because the x-intercepts are different in the two sites. Indeed, in the shaded regions, the two treatment effect functions have opposite signs. Even though the shape of the treatment effect functions is quite similar, the mechanism does not exhibit sign-congruent external validity.

One might ask whether external validity or sign-congruent external validity can be considered as a property that is more "local," i.e., whether there is a subset of settings where it holds. Indeed it can be, however, one would need to know *exactly* the boundaries of this subset. We argue that it is very difficult to identify whether we are in a region of the parameter space (the set of possible research designs) where a given form of external validity holds in any given application. As such, both forms of external validity should be regarded as global and generic properties in practical applications.

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: research designs. In practice, and

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

[14] 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.

outside the special case of direct replications, 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. This leads to our next concept.

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*

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

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*. For example, In a drug trial we generally expect to observe different treatment effects if the dosage of the drug were doubled, even if it were administered to the same population and in the same setting. Similarly, 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 lead 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, $\mathcal{A}_{ij}(\theta) = 0$, almost everywhere if and only if i and j are harmonized.*

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 goes unobserved.

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." Since many features of the technology of intervention apply across a wide variety of settings, e.g., the provision of incentives using money (Guala, 2005: Ch. 11), it is important to understand how populations respond to a variety of interventions.

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, 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 replication research designs.

Theorem 1 (Target-equivalence)**.** *Target-equivalence holds almost everywhere if and only if a collection of studies* $\{\mathcal{E}_i = (m_i, (\omega_i^{'}, \omega_i^{''}, \theta_i)\}_{i=1}^{N}$ *all 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 and is based upon Theorem 3 in Slough and Tyson (2022).

We now consider the weaker concept of target-congruence. Theorem 2 establishes the relationship between sign-congruent external validity, harmonization of study designs, and target-congruence.

Theorem 2 (Target-congruence)**.** *For any 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.*

The proof of this result is relegated to the Appendix. A key component of the proof of Theorem 2 is the set where target-congruence doesn't hold, and the proof of Theorem 2 establishes that this set has positive measure. The importance of this "sign-flip set" distinguishes target-congruence, where this set is critical, from target-equivalence, where it is not important.[15] More concretely, the

[15]This is a major difference between this study and Slough and Tyson (2022).

sign-flip set is the set of contrasts where the sign is different between two different measurement strategies, which is important because it is where the the sign of an effect is different depending only on changing the research design—not because the mechanism's effect varies over settings. 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, and produce misleading results.

Some of the larger replication studies in psychology, like Klein et al. (2014) and Open Science Collaboration (2015), pool several distinct replications into the same analysis, but leveraging data from several distinct replications. Although pooling more conceptual replications could facilitate learning about any statistical discrepancies between studies, the information the analyst gains is substantially complicated by 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 independently and identically distributed from a known distribution across different replication studies. 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 different from artifactual discrepancies 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, and follows from the observation that adding additional studies involves adding more artifactual discrepancies. Formally, the set where the sign is different between two studies

due to artifactual discrepancies is the convex hull over the null sets for different designs. Because adding studies means taking a convex hull over more points, the set of sign flips gets larger.[16]

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 treatment 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 $\varepsilon_j^{n_j}$, when these additional studies lack harmonization, they invariably introduce more artifactual discrepancies, 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. In particular, statistical noise most commonly associated with estimation, target discrepancies induced by mechanisms that lack external validity (however articulated), and novel to our framework, artifactual discrepancies that are induced by research designs that are insufficiently harmonized.

Practical Guidance: Comparing Study Estimates

When researchers seek to compare estimates across different studies, they typically adopt at least one of two approaches, which we outline formally.[17] The first approach involves comparison of the *sign* of estimates across studies. While we characterize this approach formally, it is important to

[16]This follows because any hull operator is nondecreasing in set containment, and the Krein-Milman Theorem guarantees that any convex set can be written as the convex hull of its extreme points. Adding more studies adds more extreme points.

[17]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. We return to these designs below.

note that this approach is frequently invoked informally when researchers describe the relationship between their study and related work. The second 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.

We state the results in this section in terms of two studies (or a study and its replication). However the logic extends to replication agendas with $N > 2$ studies. In these cases, researchers may test a joint null hypothesis that all estimates share the same sign or estimate. We focus on two studies throughout much of the verbal exposition of the framework, but our results—unless explicitly noted—hold generally for replication agendas with $N > 2$ studies. The first 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.

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

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

and tests the null hypothesis $H_0^z : \mathcal{Z} > 0$ against the alternative $H_a^z : \mathcal{Z} \leq 0$.

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. □

This result presents practical implications of Theorem 2 The requirement of 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 1 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 $e_1 \cdot e_2 > 0$. This event corresponds to the setting in which 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.

It is important to stress that the sign-congruence test is often conducted informally, and thus, an actual statistical test along the lines of Proposition 1 is not formally conducted. While the general intuition is similar, 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).

What do we learn from a sign-comparison test when studies are *not* necessarily harmonized? Theorem 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 there is also artifactual discrepancies, whose magnitude and direction are unknown. As such, we cannot construct the "reverse" test for harmonization with the sign-comparison test.

The second common approach to accumulating evidence compares the estimates directly (instead of relying only on their sign). Here, researchers seek to measure 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 of studies.

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

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

and test the null hypothesis $H_0^w : \mathcal{W} = 0$ against the alternative $H_+^w : \mathcal{W} > 0$ or $H_-^w : \mathcal{W} < 0$.

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

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 the 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. 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.[18]

Proposition 2 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.[19] Second, by assuming external validity, the estimate-comparison approach permits a test for harmonization—provided the analyst knows (or assumes) that the mechanism under study is externally valid.

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.

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

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

Our results indicate that this omission is consequential since a lack of harmonization can lead to Type-I or Type-II errors in our inferences about external validity in either the sign- or estimate-comparison tests.

Statistical Discrepancies

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 thereby underpowered) study. In other cases—largely in psychology—replication is used to diagnose researcher error, malfeasance, or publication bias (e.g., Open Science Collaboration, 2015; Klein et al., 2014). Our presentation so far has black-boxed the statistical issues that may arise in practical replications. We did this to focus on theoretical properties that are important issues in conducting replications 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.

Consider, for example, the estimate-comparison test between two studies, 1 and 2, which computes:

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

and can be written:

$$e_1 - e_2 = \overbrace{\varepsilon_1^{n_1} - \varepsilon_2^{n_2}}^{\text{statistical discrepancy}} + \underbrace{\Delta_{m_1}(\theta_1, \theta_2 \mid \omega_1^{'}, \omega_1^{''})}_{\text{target discrepancy}} - \overbrace{\mathcal{A}_{12}(\theta_2).}^{\text{artifactual discrepancy}} \tag{2}$$

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.[20]

To formulate valid tests for these statistical concerns, an analyst routinely 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 $\mathbb{E}[\varepsilon_i^{n_i} - \varepsilon_j^{n_j}] = 0$. These assumptions facilitate inference with either the estimate- or sign-comparison tests. While inference is straightforward in the estimate-comparison test, researchers often use the sign-comparison test heuristically without formally testing the null hypothesis that $e_1 \times e_2 > 0$. It is, however, straightforward to derive a statistical test of this hypothesis when the distribution of the error terms ($\epsilon_i^{n_i}$) are known.[21]

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:

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

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. 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,

[20] The same point applies to the sign-comparison test, which uses $e_1 \times e_2$, where a similar, albeit less elegant, substitution as in (2) can be made.

[21] Bootstrapping can be used for inference when the distribution of the error term is not known.

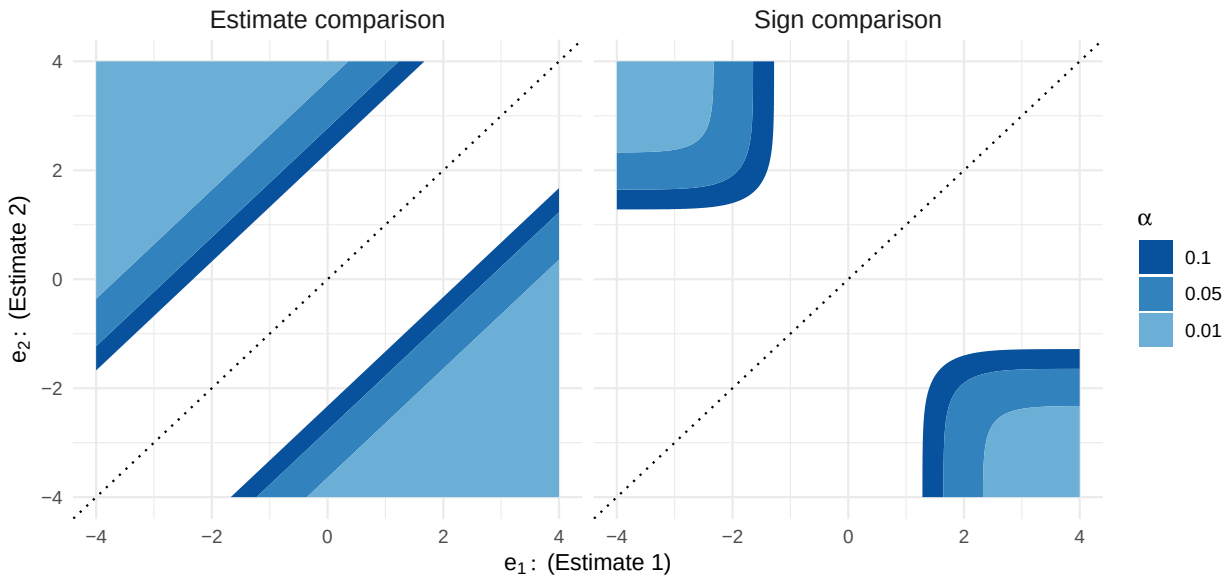

Figure 2: Rejection regions of the estimate- and sign-comparison approaches for Type-I error rates, $\alpha \in \{0.01, 0.05, 0.1\}$. Both plots fix $se_1 = se_2 = 1$ in order to visualize these regions in two dimensions.

the rejection regions for the sign-comparison test are strictly smaller than those of the estimate-comparison test.

Returning to our discussion of tests for statistical concerns, (2) reveals that, 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.[22]

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

Alternative Approaches to Replication

Conceptual replications can facilitate learning about external validity, and hence generate general knowledge about a substantive phenomenon. There are two approaches that can be used to accumulate knowledge across studies.

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.[23] In particular, by positing assumptions about the underlying structure linking together multiple studies and research designs, the analyst effectively constrains what kinds of target or artifactual discrepancies they permit to be present in the data. For example, researchers could make an assumption about how artifactual discrepancies vary in contrasts by specifying a functional form for $\mathcal{A}_{ij}(\theta)$ as a function of instruments. This assumption about artifactual discrepancies facilitates measurement of target discrepancies—and thus evaluation of external validity—in a non-harmonized, multi-setting replication.

For instance, in the context of multi-site meta-studies, Wilke and Samii (2022) advocate for a structural model of both target and artifactual discrepancies to study the efficiency of learning from meta-studies. Their model relies on three critical assumptions. First, there exists a universal treatment effect, which is independent of factors related to context or research design (Wilke and Samii, 2022: p. 17). Second, there is no external validity or sign-congruent external-validity (our Defintions 4 and 5). Third, there is no systematic influence of research design, but differences in context-specific components, and differences from research design, are independently "random" mean-zero error. These three features ensure that both artifactual and target discrepancies, $\mathcal{A}_{ij}(\theta)$ and $\Delta_{\mathcal{D}}(\theta, \theta')$ respectively, are independent of setting and design features and are normally distributed with mean zero. By reducing issues of external validity or design to random error, these

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

assumptions empower researchers to learn about the universal treatment effect, absent any concerns of harmonization.

The key strength of the structural approach is that it allows a researcher to make strong empirical conclusions from data, potentially eliminating concerns about target or artifactual discrepancies. It is important to stress 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 wold, i.e., what structural assumptions are appropriate in what cases, and whether such things are faithfully represented as "nuisance" parameters. Many structural approaches assume external validity and, similar to above, that measured treatment effects do not vary in the design of the studies.[24] By prohibiting the external world from presenting target or artifactual discrepancies (other than as idiosyncratic error), researchers dodge the problems resulting from artifacts of research design or external validity we highlight. Yet, resolving questions of target or artifactual discrepancy by assuming them away undermines the causal interpretation some researchers may wish to impart to results from replication (or meta-analysis). Further exploration of structural approaches to replication should work to transparently clarify the assumptions and what is gained by downplaying the potential problems that might arise when combining evidence from multiple places.

The Design-based Approach

We conclude our discussion of practical concerns in replication by advocating for an approach that focuses more seriously on the importance of design features, and is thus more tightly connected with credibility approaches to internal validity (Banerjee and Duflo, 2009; Gerber and Green, 2012; Dunning, 2016; Samii, 2016). As a result, our approach provides a more natural connection between research design and causal effects, and allows causal effects to be different for different

[24] Slough and Tyson (2022) term this assumption design invariance. Alternatively, T- and Y-validity of Egami and Hartman (2022) jointly constitute an even more stringent version of design invariance.

designs. It therefore provides a way of giving a causal interpretation to effects that arise in multiple places and at different times.

Our 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. Moreover, learning about artifactual discrepancies may be of independent interest because it can be important for learning about intervention technologies. In particular, 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 showing how treatment effects vary in contrasts.

To learn about target or artifactual discrepancies, we propose the following *design-based approach to conceptual replication*, which takes a *sequential* method that proceeds by admitting one discrepancy at a time; see Table 2 for a summary.

1. Conduct harmonized (conceptual) replications in settings where the mechanism may be operative. Measure target discrepancies to evaluate the 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 the set of 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

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: Proposal for design-based replication agendas. Note that steps 1 and 2 can be pursued in reverse order.

whether artifactual discrepancies vary in settings. If artifactual discrepancies do not appear to vary in settings, the mechanism exhibits external validity.

This sequential approach to replication draws upon Collins' (1992) notion of sequential replications. If replications are pursued in sequence, as we suggest, it is important to consider whether treatment effects are stable over time (Lovett and Munger, 2019; Munger, 2021). 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. However, if treatment effects change over time—a manifestation a lack of external validity—single-setting replications cannot reliably measure artifactual discrepancies because time would introduce target discrepancies.

Conclusion

The accumulation of empirical evidence collected in multiple places, at different times, and measured by different scholars must confront a number of challenges. One of the most important being whether a mechanism has external validity. 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 testing for external validity—and the weaker sign-congruent external validity—is possible with replication studies, but not without cost. In particular, researchers need to articulate their cross-study environment, formally or otherwise. We focus on the *non-statistical* reasons that estimates in different studies may differ, which emerge from study design features and a mechanism's external validity.

We develop two sets of results about empirical targets and apply them to two statistical tests—the sign-comparison and estimate-comparison tests. We show that sign-congruent external validity and harmonization of studies is necessary and sufficient to establish target-congruence. 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).

Finally, we present 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 approaches the different components of the cross-study environment sequentially, and measures each of them in isolation. Of course, a desire for novelty arguably hamper 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.

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.

Adcock, Robert, and David Collier. 2001. "Measurement validity: A shared standard for qualitative and quantitative research." *American political science review* 95 (3): 529–546.

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.

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.

Blair, Graeme, Jasper Cooper, Alexander Coppock, and Macartan Humphreys. 2019. "Declaring and diagnosing research designs." *American Political Science Review* 113 (3): 838–859.

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.

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.

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.

Esterling, Kevin, David Brady, and Eric Schwitzgebel. 2021. "The Necessity of Construct and External Validity for Generalized Causal Claims." *Mimeo* .
URL: *https://osf.io/2s8w5*

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`.

Gailmard, Sean. 2014. *Statistical modeling and inference for social science*. Cambridge University Press.

Gailmard, Sean. 2021. "Theory, History, and Political Economy." *Journal of Historical Political Economy* 1 (1): 69–104.

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

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. 2010. "Irrelevant events affect voters'evaluations ofgovernment performance." *Proceedings of the National Academy of Sciences* 107 (29): 12804–12809.

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/.

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.

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.

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

A Proofs

Proof of Theorem 1. See Slough and Tyson (2022: Theorem 3). □

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 experiments 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.

Next, we focus on the preimage of $sign$ in Ω. Since τ is smooth and regular on Ω, the sets $\tau_m^{-1}(E_m^+ \cup E_m^0) \subset \Omega$ and $\tau_m^{-1}(E_m^- \cup E_m^0) \subset \Omega$ are manifolds with common boundary $\tau_m^{-1}(E_m^0) \subset \Omega$. 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 to be the convex hull of $\tau_m^{-1}(E_{m_i}^0)$ and $\tau_m^{-1}(E_{m_j}^0)$, and note that the elements of H_{ij} 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. The establishes that the set H_{ij} has a nonempty interior, and thus, has positive measure, contradicting that sign-congruence holds almost everywhere. An identical argument applies to harmonization of contrasts. □

Proof of Theorem 3. Follows from footnote 16. □

B Application II: Sports Outcomes and Electoral Accountability

Overview: In the main text, we discuss Raffler, Posner, and Parkerson's (2020) experimental replication of Björkman and Svensson (2009). Our framework is also useful for understanding replication efforts with observational data. To this end, we consider an ongoing debate about whether sporting game outcomes affect pro-incumbent voting. Table A1 summarizes the published (or forthcoming) papers associated with this debate. We use the "keys" from the table to refer to these papers in our discussion.

Key	Citation	Summary
HMM2010	Healy, Malhotra, and Mo (2010)	Finds that college football victories in the two weeks before general elections for president, governor, and senator increase the incumbent party's vote share. The sample consists of presidential elections from 1960-2004, and gubernatorial and senate elections from 1967-2006. The mechanism of interest is that shocks to voter well-being (football victories) increase voter-satisfaction with the status-quo. Because the incumbent party represents the status-quo, this increased satisfaction translates into incumbent vote share.
FM2015	Fowler and Montagnes (2015)	Argues that HMM2010 is likely a false positve. Using an extended panel of presidential elections from 1960-2012 and gubernatorial and senate elections from 1960-2006, FM2015 test ancillary hypotheses consistent with the mechanism proposed by HMM2010 and include alternative set of specifications with county-year fixed effects. They also conduct a replication with a different treatment: NFL outcomes.
HMM2015	Healy, Malhotra, and Mo (2015)	Argues that FM2015 do not consider the totality of the evidence presented because they do not consider the survey evidence on NCAA basketball games or the preferred specfication that adjusts for the probability of victory.
GHMM2021	Graham et al. (2021)	Conduct a pre-specified replication of several studies of voter competence/rationality including HMM2010. To do so, they correct several data errors (see the supplemental information of Graham et al. (2021)) in HMM2010 and extend the time series slightly.† Their preferred specification pools the (corrected) in-sample data with the new (previously) out-of-sample data and show that estimates are attenuated, but in the same direction as the original finding.
FM2022a	Fowler and Montagnes (2022*a*)	Argue that Graham et al. (2021) overstate the strength of evidence consistent with Healy, Malhotra, and Mo (2010). They distinguish between in-sample and out-of-sample data and conduct a simulation to show that the evidence on the pooled sample and cannot reject (statistically) the possibility that the Healy, Malhotra, and Mo (2010) was a false positive.
GHMM2022	Graham et al. (2022)	Contest Fowler and Montagnes' 2022*a* equal treatment of multiple specifications, instead advocating for prioritization of main effects over heterogeneous treatment effects and advocating replication on an expanded sample that consists of both in-sample and out-of-sample observations.
FM2022b	Fowler and Montagnes (2022*b*)	Justifies the focus on multiple pre-specfied tests and argues for the merits of out-of-sample replication.

Table A1: Summary of replications and responses to Healy, Malhotra, and Mo (2010).
†: FM2022a note that GHMM2021 rely on a subset of the original data starting in 1985 rather than using the full (original) sample.

Concepts: It is important to the debate over whether sports outcomes affect voter assessment of incumbents has, to date, focused on *statistical* discrepancies. Specifically, FM2015, FM2022a, and FM2022b suggest that the original results in HMM2010 are likely false positives.

Our framework helps to elucide other possible *non-statistical* explanations for differences in findings.

- Settings: This study uses panel data. The inclusion criteria for counties (the cross-sectional unit) is: "counties home to a college football team that was a member of a Bowl Championship Series (BCS) major conference (plus Notre Dame), the selection system that determined the champion of college footballs top division from 1998 to 2013" (see pre-analysis plan of GHMM2021). The studies vary the elections they analyze. We will consider the following temporal samples (from the later replications/extensions) as two distinct settings:
  - Original setting ("in sample" observations): presidential, senate, and gubernatorial elections from 1985-2006
  - Replication setting ("out of sample" observations): presidential, senate, and gubernatorial elections from 2007-2016 (per FM2022a) or the union of presidential, senate, and gubernatorial elections from 1980-1984 and from 2007-2016 (per GHMM@022).
- Contrasts: The authors use different contrasts to measure (unexpected) football outcomes.
  - HMM2010 and GHHM2021 employ two distinct instruments. Their preferred operationalization, used on samples where they have data on betting odds (point spreads), they measure treatment—a surprise football victory as:

    $$W_{it} = \text{Win}_{it} - \Phi\left(\frac{-x}{13.89}\right)$$

    where Win_{it} is a binary indicator that takes a value of 1 when the county's team wins game t; x is the game's points spread; and Φ is the standard normal cdf. They define this at different points (two weeks before the election, one week before the election, and both games). Note that $W_{it} \in (-1, 1)$, where -1 is a completely unexpected loss and 1 is a completely unexpected victory.

    For samples for which they are missing some point spreads, they use a binary indicator of Win_{it} as the treatment instruments.
  - FM2015 use a measure akin to the latter measure of HHM2010 and GHHM2021. For each game, a win takes a value of 1, a tie takes a value of 0.5, and a loss takes a value of a loss. When considering the two games before an election, they take the average of both indicators, such that the contrast takes five possible values $W_{it} \in \{0, 0.25, 0.5, 0.75, 1\}$.
- Measurement strategies: All authors use the incumbent's county-level vote share as the outcome of interest.[25]

[25]Note that FM2015 use Democratic vote share as an outcome but estimate the ATT using an

We note that FM2015 include a second study that measures the effect of NFL game outcomes on incumbent vote share. This design utilizes a different setting: the cross-section is counties with NFL teams and a different contrast: the instrument measures the results of NFL games, not NCAA football games.

We might first ask whether the effect of sports outcomes on pro-incumbent voting is **externally valid**. External validity across the different temporal settings would hold if the mechanism produces the same effect in both periods. When comparing NCAA to NFL games in Fowler and Montagnes (2015), a discussion fo external validity might also ask whether the mechanism produces the same effect the subset of counties with NCAA teams as in the subset of counties with NFL teams.

We might then ask whether the designs are **harmonized**? We have shown that the authors adopt different measurment of contrasts. In particular, HHM2010, HHM2015, and GHHM2021 favor a treatment measured in terms of deviation from expected game outcomes whereas FM2015 favor an indicator for a game outcome—not how surprising it is. In the context of the NCAA/NFL comparison, we may consider whether NFL and NCAA game results constitute the same contrasts.
Discussion: Neither set of authors discuss the possibility of non-statistical discrepancies. We discuss several potential target discrepancies and artifactual discrepancies across the various replications.

Target discrepancies: We discuss two possible target discrepancies. First, and relevant to the in-sample and out-of-sample discussion in FM2022a, GHMM2022, and FM2022b, it is possible that the mechanism lacks external validity and thus produces different effects in the earlier period ("in sample" data) but had dissipated by the later period ("out of sample" data). There are several plausible explanations for these target discrepancies. College football viewership has held steady over the past 20 years,[26] whereas the population—and hence the pool of eligible voters—has grown. We might, then, expect effects to attenuate toward zero if a smaller subset of the population is watching college football.

Second, FM2015 advocate the use of county-year fixed effects which leverage variation in pro-incumbent voting when incumbents of two different parties stand for re-election at the same time. This effectively redefines the sample in HMM2010. FM2015 find precisely zero effect in this subset of the data. They interpret this as evidence that the main effect in HHM2010 is a false positive. But it could also be that the mechanism does not operate in this subsample.

Artifactual discrepancies: There are several sources of artifactual discrepancies. First, and most prominently, the authors consider different contrasts. The treatment indicators are measured on distinct scales and one seeks to capture the element of surprise (or deviation from expectations). These outcomes are related deterministically (both rely on a measure of Win_{it}). In general, we would expect the ATTs to be *different* under these different treatment indicators.

Second, FM2015 creatively also examine whether voters respond to NFL games. Here, the authors change the contrast (from NCAA victories to NFL victories). However, it is unclear whether NCAA and NFL football victories represent the same treatment. For example, football games

interaction between a measure partisan incumbency and Win_{it}.

[26]See, for example, `https://www.sportsmediawatch.com/2021/01/national-championship-ratings-record-low-audience-alabama-ohio-state/`.

around early-November general elections fall later within the NCAA schedule than the NFL schedule. As such, the games may be more emotion-laden as the conference championships or playoff picture come into fuller view. If it is the case that NCAA and NFL games represent different treatments, we need not expect an externally valid mechanism to produce the same effect. The resultat differences are artifactual discrepancies.

It is important to note that neither set of authors conducts an estimate- or sign-comparison test. They test a null hypothesis of zero in different subsets of the data (with different specifications). Our approach advocates FM2022a and FM2022b's use of out-of-sample testing as this permits cleaner comparison across studies. However, we advocate the use of statistical tests that directly compare the in-sample and out-of-sample findings.