Sign-Congruence, External Validity, and Replication
Supporting Materials

Tara Slough[*] Scott A. Tyson[†]

Contents

[*] Assistant Professor, New York University. taraslough@nyu.edu

[†] Associate Professor, University of Rochester. styson2@ur.rochester.edu

A Proofs

Proof of Remark 1. For the first part, from Definition 4, the target discrepancy is

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

which, after applying the definition of exact external validity, implies that $\Delta_{\mathcal{D}}(\theta_i, \theta_j) = 0$ almost everywhere, establishing necessity and sufficiency.

For the second part, from Definition 7, the artifactual discrepancy is

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

which is 0 if and only if i and j are harmonized, i.e., when $\mathcal{D}_i = \mathcal{D}_j$. □

Proof of Theorem 1. Sufficiency is straightforward from the definitions of sign-congruent external validity and harmonization. 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 target-congruence holds almost everywhere 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 almost everywhere.

The treatment effect function is a smooth function (almost everywhere) that maps from the set of research designs and settings to its image, the set of empirical targets: $\tau_m(\omega', \omega'' \mid \theta) : M \times \mathcal{C} \times \Theta \to \mathbb{R}$. Its composition with the function $sign : \mathbb{R} \to \{-1, 0, 1\}$, allows us to partition the set of empirical targets, i.e., the image of τ, into three sets. Sign-congruent external validity implies that these sets do not depend on θ, which we drop for parsimony. Now, define 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

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

Since $sign(\tau_m(\omega', \omega'' \mid \theta)) = -sign(\tau_m(\omega'', \omega' \mid \theta))$, these sets are nonempty. Moreover, $E_m^+ \cup E_m^0$ and $E_m^- \cup E_m^0$ are each manifolds with boundary, and their common boundary is E_m^0.

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

For two studies, $\mathcal{E}_i$ and $\mathcal{E}_j$, define the set $H(\mathcal{E}_i, \mathcal{E}_j) = co(\tau_m^{-1}(E_{m_i}^0) \cup \tau_m^{-1}(E_{m_j}^0))$ as the convex hull of $\tau_m^{-1}(E_{m_i}^0) \cup \tau_m^{-1}(E_{m_j}^0)$. Note that the elements of $H(\mathcal{E}_i, \mathcal{E}_j)$ are precisely those that have a different sign in study i than in study j, implying that on this set target-congruence does not hold. Since measurement strategies are distinguishable almost everywhere, i.e., τ's derivative in m has full rank almost everywhere, the set $H(\mathcal{E}_i, \mathcal{E}_j)$ has a nonempty interior, and thus, positive Lebesgue measure, contradicting that target-congruence holds almost everywhere. An identical argument applies to harmonization of contrasts. □

Proof of Theorem 2. This result follows from the following straightforward lemma:

Lemma A.1. *Let $X, Y, Z \subset \mathbb{R}$ and define the convex hull $W = co(X \cup Y)$, then*

$$co(W \cup Z) = co(X \cup Y \cup Z).$$

Proof. By the definition of convex hull, for any $t \in co(W \cup Z)$, there exists some $\alpha \in [0, 1]$ such that $t = \alpha w + (1 - \alpha)z$, for some $w \in W$ and $z \in Z$. Since $W = co(X \cup Y)$, there exists a

$\gamma \in [0,1]$, an $x \in X$ and $y \in Y$, such that $w = \gamma x + (1-\gamma)y$. Thus,

$$t = \alpha w + (1-\alpha)z = \alpha(\gamma x + (1-\gamma)y) + (1-\alpha)z$$
$$= \alpha\gamma x + \alpha(1-\gamma)y + (1-\alpha)z.$$

Denoting $\beta_1 = \alpha\gamma$, $\beta_2 = \alpha(1-\gamma)$, and $\beta_3 = (1-\alpha)$, and noting that $\alpha\gamma + \alpha(1-\gamma) + (1-\alpha) = 1$, implies that any element of $co(W \cup Z)$ can be written as $\beta_1 x + \beta_2 y + \beta_3 z$, for some $x \in X$, $y \in Y$, and $z \in Z$, and where $\beta_1 + \beta_2 + \beta_3 = 1$. Thus, t is an element of $co(X \cup Y \cup Z)$. For the reverse direction, note that $X \cup Y \cup Z \subset W \cup Z$, hence $co(X \cup Y \cup Z) \subset co(W \cup Z)$. □

Suppose that one considers a set of studies $\{\mathcal{E}_i = (m_i, (\omega_i^{'}, \omega_i^{''}, \theta_i)\}_{i=1}^{N}$, where contrasts are harmonized, so that $(\omega_i^{'}, \omega_i^{''})$ are identical across i. Using Lemma A.1, observe that the set where target-congruence does not hold, as a function of the number of studies N, can be defined recursively as follows. Define $H(\{\mathcal{E}_i\}_{i=1}^{2}) = H(\mathcal{E}_1, \mathcal{E}_2)$ as in the proof of Theorem 1. For any $1 < n \leq N$, define the set

$$H(\{\mathcal{E}_i\}_{i=1}^{n}) = co(H(\{\mathcal{E}_i\}_{i=1}^{n-1}) \cup \tau_m^{-1}(E_{m_n}^0)).$$

That $H(\{\mathcal{E}_i\}_{i=1}^{n-1}) \subset H(\{\mathcal{E}_i\}_{i=1}^{n})$ is immediate. The argument for contrasts is similar. □

B Target-Equivalence

The relationship between harmonization, exact external validity, and target-equivalence is developed at length in Slough and Tyson (2023), applied to the case of meta-analysis; for completeness we present their results and proof here.

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

Proof. Sufficiency follows by noting that Remark 1 guarantees that exact external validity ensures that all target discrepancies are zero. Moreover, Remark 1 also shows that harmonization ensures that artifactual discrepancies are also zero. These observations show how exact external validity and harmonization are jointly sufficient for target-equivalence.

For necessity, first notice that target-equivalence under harmonization is equivalent to exact external validity. Now suppose that studies $\mathcal{E}_1$ and $\mathcal{E}_2$ are target-equivalent, but not harmonized. Then, for $\mathcal{D}_1$ and $\mathcal{D}_2$:

$$\tau_{m_1}(\omega_1^{'}, \omega_1^{''} \mid \theta_1) = \tau_{m_2}(\omega_2^{'}, \omega_2^{''} \mid \theta_2). \tag{B.1}$$

Applying exact external validity at $\mathcal{D}_2$, it must be that for arbitrary θ_1 and θ_2:

$$\tau_{m_2}(\omega_2^{'}, \omega_2^{''} \mid \theta_1) = \tau_{m_2}(\omega_2^{'}, \omega_2^{''} \mid \theta_2). \tag{B.2}$$

Combining (B.1) and (B.2),

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

which, since the setting and contrasts were arbitrary, implies that the treatment effect function must be the same at m_1 and m_2 in any setting. Thus, since θ_1 and θ_2 were arbitrary, exact external validity allows us to suppress the dependence of the treatment effect function on θ.

Recalling that M is a manifold, define

$$\kappa \equiv \tau_{m_1}(\omega_1^{'}, \omega_1^{''} \mid \theta),$$

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

$$\dim \tau^{-1}(\kappa) = \dim M \times \mathcal{C} - \dim \mathbb{R} = 3 - 1 = 2.$$

Thus, dim $\tau^{-1}(\kappa) <$ dim $M \times \mathcal{C}$, implying that $\tau^{-1}(\kappa)$ is a Lebesgue measure zero subset of $M \times \mathcal{C}$, completing the argument. □

The key intuition for Theorem B.1 is illustrated in Figure B1. In panel (a), the treatment effect functions exhibit exact external validity, but a lack of harmonization induces artifactual discrepancies from using different levels of treatment, ω_1'' and ω_2''. These artifactual discrepancies undermine target-equivalence (except at exactly two points), illustrated in the grey regions. In contrast, Panel (b) shows that harmonization is insufficient to achieve target-equivalence when exact external validity is absent (even with sign-congruent external validity). The grey zones in each panel correspond to the set of treatment levels where target-equivalence fails due to a lack of exact external validity. These examples, depicted in Figure B1(a)-(b), are not unusual, and Theorem B.1 establishes that the sets where target-equivalence fail, due either to a lack of harmonization or a lack of exact external validity, have positive measure in general.

C Conceptual Illustrative Examples

In this section we present two models to illustrate our main points in idealized settings. We present these models to focus on conceptual features of our framework in a concrete manner without the additional complications that are introduced in practice. Appendix E presents a more applied

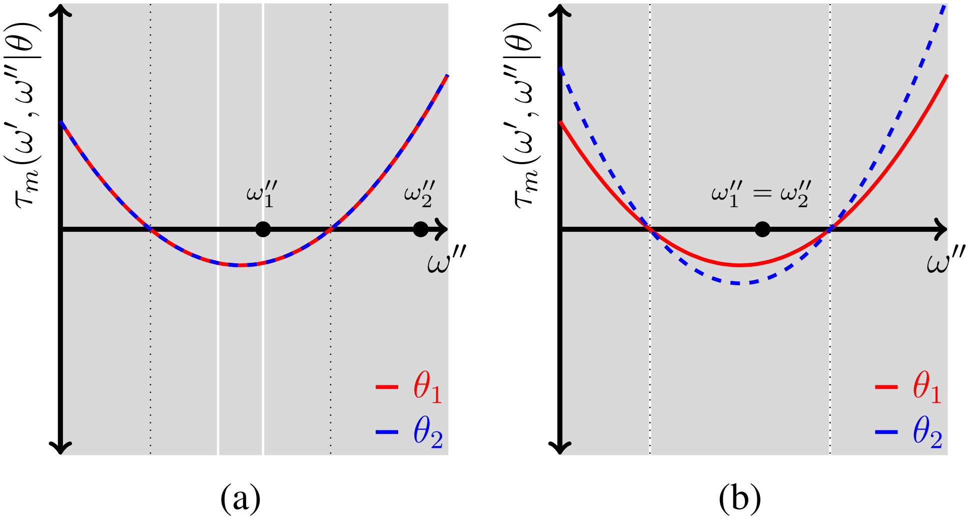

Figure B1: Illustration of Theorem B.1. The grey regions in panel (a) depict the regions where target-equivalence fails when ω''s are not harmonized. The grey regions in panel (b) depict the regions where target-equivalence fails due to a lack of exact external validity.

discussion of actual experiments and observational studies.

C.1 Potential Outcomes and Harmonization

We provide an illustrative example of the concepts that we introduce in a conventional formulation of the potential outcomes model. Building from the presentation of the framework in the main text, we introduce a few other terms. First, in setting θ there is a set of units $u \in I_\theta$. Second, there is a set of instruments, Ω, where $\omega \in \Omega$ corresponds to a particular experimental manipulation. In the main text, the set of contrasts corresponds to two values of an instrument, and so $\mathcal{C} = \Omega \times \Omega$ (see Bueno de Mesquita and Tyson (2020) for the same construction). Finally, potential outcomes are captured by the mapping $Y^u_m(\omega \mid \theta) : I_\theta \times M \times \Omega \times \Theta \to \mathbb{R}$. The empirical target corresponding to the average treatment effect is then given by

$$\tau_m(\omega', \omega'' \mid \theta) = \mathbb{E}_u[Y^u_m(\omega'' \mid \theta)] - \mathbb{E}_u[Y^u_m(\omega' \mid \theta)].$$

This shows how the empirical target, that is key to our analysis, follows from a single study, whose measured effects can be represented within the potential outcomes model. Questions about external validity correspond to specific ways that the empirical target, and hence potential outcomes, depend on the setting, θ.

Artifactual and Target Discrepancies. To focus on how discrepancies arise consider another setting, $\theta' \in \Theta$, where the experiment uses the contrast (ω', ω'''), where $\omega''' \in \Omega$. In this case, we can take the difference between empirical targets, i.e.,

$$\begin{aligned}
&\tau_m(\omega', \omega'' \mid \theta) - \tau_m(\omega', \omega''' \mid \theta') \\
&\qquad = \mathbb{E}_u[Y_m^u(\omega'' \mid \theta)] - \mathbb{E}_u[Y_m^u(\omega' \mid \theta)] - (\mathbb{E}_u[Y_m^u(\omega''' \mid \theta')] - \mathbb{E}_u[Y_m^u(\omega' \mid \theta')]) \\
&\qquad = \mathbb{E}_u[Y_m^u(\omega'' \mid \theta)] - \mathbb{E}_u[Y_m^u(\omega''' \mid \theta)] + \qquad \text{(C.1)} \\
&\qquad\quad \mathbb{E}_u[Y_m^u(\omega''' \mid \theta)] - \mathbb{E}_u[Y_m^u(\omega''' \mid \theta')] + \mathbb{E}_u[Y_m^u(\omega' \mid \theta')] - \mathbb{E}_u[Y_m^u(\omega' \mid \theta)]. \qquad \text{(C.2)}
\end{aligned}$$

This expression reflects all the discrepancies between empirical targets. First, (C.1) is the artifactual discrepancy because the potential outcome at ω'' is different than at ω''' at setting θ. Second, (C.2) is the target discrepancy that arises because the potential outcome at the same value of the instrument (ω''' or ω') may be different in different settings. Note that the target discrepancy reflects the difference at two values of the instrument.

Harmonization. To examine harmonization, in the context of the potential outcomes model, we now consider two different measurement strategies. Specifically, suppose there are two measurement strategies, m and m', where for almost all ω and almost all θ:

$$Y_m^u(\omega \mid \theta) = Y_{h(m')}^u(\omega \mid \theta) + \varepsilon_u,$$

for some invertible function, h, that does not depend on ω or θ, and where ε_u is a random variable, independently and identically distributed across u, with mean 0 and some variance. Then, notice that since

$$\begin{aligned}\tau_{h^{-1}(m')}(\omega', \omega'' \mid \theta) &= \mathbb{E}_u[Y^u_{h^{-1}(m')}(\omega'' \mid \theta)] - \mathbb{E}_u[Y^u_{h^{-1}(m')}(\omega' \mid \theta)] \\ &= \mathbb{E}_u[Y^u_m(\omega'' \mid \theta)] - \mathbb{E}_u[Y^u_m(\omega' \mid \theta)] = \tau_m(\omega', \omega'' \mid \theta),\end{aligned}$$

harmonization holds between m and $h^{-1}(m)$.

In our model, a single measurement strategy, $m \in M$, should be thought of as an equivalence class of measurement strategies (relative to τ), and this is important for at least two reasons. First, the measurement strategy, m, is not a label of the different things that go into a single experiment. Consequently, literal differences between two measurement strategies may or may not reflect *substantive* differences, and only the latter would matter in applying our model. Second, measurement strategies and contrasts reflect theoretical considerations as well as practical concerns. For instance, assessing the mapping h in the example above can involve both theoretical considerations, e.g., knowing how two measures are related theoretically, and practical considerations that suggest that two measures might be interchangeable.

C.2 Political Accountability

We provide an illustrative example of the concepts that we introduce with a very simple formal model of moral hazard of politicians within the framework of electoral accountability. While the substantive problem that we describe is widely discussed, we consider an abstract experimental intervention and its effect on retention of an incumbent.

A large literature on electoral accountability suggests that voter welfare is improved when politicians "work" or exert effort on their behalf. Yet, voters observe politician effort imperfectly, generating a moral hazard problem. This limited ability to directly observe effort may lead the politician to "shirk," or provide less effort than the voter desires. Whereas a variety of experiments manipulate voter information about politician effort (e.g., Adida et al., 2019), we focus on a hypothetical intervention that aims to *reduce the politician's cost of effort* by providing training or information.[1] Our goal in departing from standard interventions in the literature is to focus on the concepts that we introduce.

Baseline model: To straightforwardly capture moral hazard, we consider a canonical model where a politician chooses whether to exert effort, e, with $e = 1$ corresponding to the choice to exert effort, at cost $c > 0$, and $e = 0$ to the choice to not exert effort. The expected utility-maximizing politician who values office and seeks re-election, gains a payoff normalized to 1 if they win re-election and 0 otherwise.

In the status quo, the voter observes a signal of effort, $x \in \{0, 1\}$, where:

$$P(x = 1 \mid e = 1) = P(x = 0 \mid e = 0) = p \in ({}^1\!/\!{}_2, 1].$$

This means that with probability p, the voter correctly observes the effort of the politician, and with probability $1 - p$, the voter observes an incorrect signal of the politician's effort.

To simplify the exposition, we will assume that the voter reelects the politician if and only if they observe $x = 1$.[2] Looking down the tree to the voter's decision, when should the politician choose to exert effort? Exerting effort is incentive compatible for the politician if and only if:

$$\underbrace{P(x = 1 \mid e = 1) - c}_{\text{Expected payoff if } e=1} \geq \underbrace{P(x = 1 \mid e = 0)}_{\text{Expected payoff if } e=0}.$$

This expression simplifies to $p - c \geq 1 - p$ and is equivalent to:

$$c \leq c_p^* \equiv 2p - 1. \tag{C.3}$$

[1] For one example of an intervention in this vein, see Raffler (2022), who trains politicians to better monitor bureaucrats. We abstract from the specific intervention and context in order to keep this example focused on concepts.

[2] Note that by positing a simple re-election rule, we have assumed that the voter is not strategic. If the voter were strategic, the exact formulation of the re-election rule would need modification but this is straightforward and not relevant for our point here; hence, it is omitted.

The re-election rate for an incumbent politician is:

$$P(x=1) = \begin{cases} p & \text{if } c \leq c_p^* \\ 1-p & \text{if } c > c_p^*. \end{cases}$$

Note that the re-election rate is p when the politician chooses to exert effort, which occurs whenever the cost of effort is $c \leq c_p^*$. This is simply the probability with which the voter receives an accurate signal that the politician exerted effort. The re-election rate is $1-p$ when the politician does not exert effort, which occurs when $c > c_p^*$. This is the probability with which the voter sees an inaccurate signal that the politician exerted effort when in fact they did not.

The experimental intervention (contrast): Now consider an experimental intervention that reduces the cost of providing effort for politicians. Specifically, suppose that there is an initial cost of effort, denoted by c_0, and that a treatment reduces that cost to $c_\mu = c_0 - \mu$, where $\mu > 0$, thus ensuring that $c_\mu < c_0$. An experimental comparison, then, is one between politicians receiving the treatment and politicians not receiving the treatment. Specifically, the contrast is $(\omega', \omega'') = (c_0, c_\mu)$ and the treatment effect is

$$p - (1-p) = 2p - 1, \tag{C.4}$$

whenever $c_\mu \leq c_p^* < c_0$ and it is zero otherwise.

We can now consider efforts at replicating the above study. Since the model is so sparse, there are essentially two ways that two studies can differ. First, study's can differ in terms of the treatment administered, i.e., they can differ in μ. Second, studies can differ in terms of the accuracy of information about the incumbent, i.e., their settings can differ in p. We will consider each separately and show how such differences in this toy model reflect the discrepancies highlighted in the main manuscript.

Artifactual Discrepancies. To isolate when there are artifactual discrepancies we consider two different values of the experimental manipulation that differ in terms of the strength of their activation. Suppose there are two values $\mu > \mu'$ and suppose that at μ, we have that $c_\mu \leq c_p^* < c_0$, ensuring that the treatment effect in the study that uses μ is $2p - 1$ (from (C.4)). Now consider the case with μ'. Suppose first that $c_{\mu'} \leq c^* < c_0$, in which case as in (C.4), the treatment effect is $2p - 1$ and the difference in empirical targets is

$$\tau_m(c_0, c_\mu \mid \theta) - \tau_m(c_0, c_{\mu'} \mid \theta) = 2p - 1 - (2p - 1) = 0.$$

Suppose instead, that μ' is sufficiently smaller than μ, so that $c_p^* < c_{\mu'} < c_0$. Then the treatment effect at μ' is

$$1 - p - (1 - p) = 0,$$

and the difference in empirical targets becomes

$$\tau_m(c_0, c_\mu \mid \theta) - \tau_m(c_0, c_{\mu'} \mid \theta) = 2p - 1 - (0) = 2p - 1.$$

In this case the value $2p - 1$ is the treatment effect in one study as well as the artifactual discrepancy between studies (because the treatment effect in one is zero).

Because of the empirical targets in this model, harmonization is relatively straightforward. Specifically, consider the value of the experimental manipulation, $\hat{\mu}$, such that $c_{\hat{\mu}} = c_p^*$. It is the weakest value that flips the politician to providing effort. In this context, any value of $\mu > \hat{\mu}$ is harmonized with $\hat{\mu}$. Similarly, all values of μ such that $\mu < \hat{\mu}$ are harmonized.

Target Discrepancies. Now we consider differences between settings. Suppose that the manipulation, μ, is such that $c_\mu \leq c_p^* < c_0$, ensuring that the treatment effect is $2p - 1$ from (C.4). Now consider two settings, one in which the signal precision follows from p and another where the signal precision is $p' > p$. In this case, the two settings reflect different cutoffs, i.e., $c_p^* < c_{p'}^*$. There are two ways target discrepancies arise. First, similar to above, if $c_{p'}^*$ increases enough, then

$c^*_{p'} > c_0 > c_\mu$ and the treatment effect at p' is

$$p' - p' = 0.$$

The target discrepancy is then

$$\tau_m(c_0, c_\mu \mid \theta_p) - \tau_m(c_0, c_\mu \mid \theta_{p'}) = 2p - 1 - (0) = 2p - 1.$$

Second, supposing that $c_\mu < c^*_p < c^*_{p'} < c_0$, then the target discrepancy is

$$\tau_m(c_0, c_\mu \mid \theta_p) - \tau_m(c_0, c_\mu \mid \theta_{p'}) = 2p - 1 - (2p' - 1) = 2(p - p').$$

Notice that the change in setting does not only change whether the treatment effect is activated, but also changes the magnitude of the treatment effect, implying that in our simple example, exact external validity does not hold. It is important to note that there is a negative target discrepancy for all $p' > p$, implying that sign-congruent external validity holds.

The toy model presented in this section illustrates our main points without the added complexities that typically arise in practice. We do this to focus on the concepts that we develop in the main text in an idealized case to give them more (conceptual) concreteness. Notice that in the toy model, the derivative of the treatment effect function does not have full rank in contrasts (as is assumed in the main text). This was intentional because, although the full rank assumption was used in proving our main results, this example illustrates that the same issues highlighted by our main results arise even when those assumptions are not satisfied.

D Construction of p-values for Sign-Comparison Test

The sign-comparison test evaluates the null hypothesis that two measured effects share the same sign. In this section we give the construction of p-values for the sign-comparison test applied to N studies. In the main text, the p-values presented follow by setting $N = 2$.

Let there be a set of studies, indexed by j, each with measured estimates e_j with respective standard errors se_j. Denote the vector of measured effects by $\mathbf{e} = (e_1, e_2, \ldots, e_N)$. For each study, j, construct T-statistics, $T_j = \frac{e_j}{se_j}$, and compute the following:

1. Consider the joint event that all measured effects are negative, $\bigcap_j \{e_j < 0\}$. To test the null hypothesis

   $$H_0 : \left\{ \mathbf{e} \in \bigcap_j \{e_j < 0\} \right\},$$

   calculate the one-sided (lower) p-values for each T_j, denoted by $\underline{p}_j$. Implement a Bonferroni correction, denoted by $B(\cdot)$, and select the minimum Bonferroni-corrected p-value,

   $$\underline{p} = \min\{B(\underline{p}_1), B(\underline{p}_2), ..., B(\underline{p}_n)\};$$

2. Consider the joint event that all measured effects are positive, $\bigcap_j \{e_j > 0\}$. To test the null hypothesis

   $$H_0 : \left\{ \mathbf{e} \in \bigcap_j \{e_j > 0\} \right\},$$

   calculate the one-sided (upper) p-values for each T_j, denoted $\overline{p}_j$. As in Step #1, implement a Bonferroni correction and select the minimum Bonferroni-corrected p-value,

   $$\overline{p} = \min\{B(\overline{p}_1), B(\overline{p}_2), ..., B(\overline{p}_n)\};$$

3. The sign-comparison test tests the null hypothesis that the vector **e** is an element of one of the two sets described in steps #1 and #2. Following the intersection method of Berger (1982), as applied to p-values in Brinch, Mogstad and Wiswall (2017: Appendix B), the p-value for this test is given by
$$p = \max\{\underline{p}, \overline{p}\}.$$

The intuition of the sign-comparison test is that it evaluates a collection of one-sided tests, i.e., $\{e_j < 0\}$ for each j, because each event where study j yields a measured estimate of a different sign is evidence against sign-congruent external validity.

E Applications

E.1 Application I: Citizen Oversight of Healthcare Providers

E.1.1 Overview

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

Björkman and Svensson (2009) show that their bundled treatment increased healthcare utilization by community members as well as increasing child health outcomes, including reductions in childhood mortality. Notably, the treatment effects in the study were large. In particular, several measured (standardized) treatment effects were more than a standard deviation in magnitude. Prompted by the large policy impact of Björkman and Svensson (2009), Raffler, Posner and Parkerson (2022) conducted a carefully-designed, pre-registered replication experiment in rural Ugandan communities from 2014-2016. The replication experiment was conducted a decade after the original experiment was fielded and included 92 clusters in treatment and 95 clusters in control.

In contrast to the original study, Raffler, Posner and Parkerson (2022) 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 (2022) find qualitatively different results from Björkman and Svensson (2009)? In their article, they cite two explanations. First, the presence of statistical noise, i.e., random error, could lead to differences between each study's results. Specifically, one may be concerned—as were Raffler, Posner and Parkerson (2022)—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 (2022) 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 suggest that the original effect of community monitoring interventions (observed in Uganda 2004-2005), could have been a real effect, but one that lacks external validity because Uganda had changed sufficiently between 2004 and 2005. This kind of failure of external validity reflects a lack of *temporal validity* (Munger, 2023), because it is the passage of time that distinguishes different settings, i.e., the same place at two different times. Consequently, we should not necessarily expect similar findings in Uganda in 2014-2016.

There is another potential explanation. Since it was difficult for Raffler, Posner and Parkerson (2022) to conduct *exactly* the same experiment as Björkman and Svensson (2009), there are a number of differences between their respective research designs.[3] 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 (2022) 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 (2022) measured outcomes at 8 month and 20 months post-treatment, whereas Björkman and Svensson (2009) measured outcomes at 12 months post-treatment. We emphasize that these differences in design should not be viewed as a weakness of Raffler, Posner and Parkerson (2022), or any replication effort. Indeed, if the treatments, outcomes, and measurement strategies in an original study are flawed, replicators should not blindly repeat them. But these tweaks to the research design make isolating the source of differences in measured effects more challenging.

E.1.2 Statistical Tests and Interpretation

In this section, we apply both the estimate- and sign-comparison tests to three outcomes—under-5 mortality, under-3 vaccination rate, and under-18 month weight-for-age z scores—that are common to both Björkman and Svensson (2009) and Raffler, Posner and Parkerson (2022). We use this application to review the assumptions underpinning each test and the resultant interpretation of findings. In Table E1, we report intent-to-treat (ITT) estimates for three outcomes that are common to both studies. While both studies have substantially more outcomes, many are measured in

[3] Importantly, among other community-monitoring interventions in the field of healthcare, Raffler, Posner and Parkerson (2022) remain most faithful to the treatments and outcome measures in the original experiment.

Outcome measure	BS (2009)	RPP (2022)
Under-5 mortality per 1,000 live births	-49.9	-11
	(26.9)	(8)
Vaccination rate for children under 36 months (standardized)	2.01	0.054
	(0.67)	(0.035)
Weight-for-age z-scores, children under 18 months	0.14	0.00
	(0.07)	(0.048)

Table E1: Intent-to-treat (ITT) estimates of "Power to the People" treatments on three common outcomes. Standard errors are in parentheses. "BS (2009)" estimates come from Tables 5 and 6 of Björkman and Svensson (2009) and "RPP (2022)" estimates come from Tables H6, H13, and H14 of the supplemental information of Raffler, Posner and Parkerson (2022). Note that the Raffler (2022) mortality is scaled by a factor of 1,000 for comparability.

	p-value from comparison of	
Outcome measure	Estimates	Signs
Under-5 mortality rate per 1,000 live births	0.961	1
Vaccination rates for children under 36 months (standardized)	< 0.001	1
Weight-for-age z-scores, children under 18 months	< 0.001	1

Table E2: p-values from the estimate- and sign-comparison tests applied to the estimates from Table E1.

different ways.

The estimates in Table E1 serve as the inputs to the estimate- and sign-comparison tests. Table E2 reports the p-values from both tests for each of the three outcomes listed above. Inspection of the p-values suggests that the null hypothesis can be rejected only in the case of the estimate-comparison test for the vaccination and weight-for-age outcomes. Our theoretical results discipline interpretation of the relevant null hypotheses.

Treating the Raffler (2022) study as a replication of Björkman and Svensson (2009), our tests developed in the main manuscript yields the following results:

1. Under the assumption that both experimental designs are harmonized, *we reject the null hypothesis of exact external validity* for the vaccination rate and weight-for-age outcomes. Given a reasonable belief that these outcomes are affected by a similar mechanism, one might reasonably argue that these inferences provide evidence against external validity of the effect of

power-to-the-people interventions on health more generally.

2. Under the assumption that both experimental designs are harmonized, *we fail to reject the null hypothesis of sign-congruent external validity*.

3. Under the assumption that mechanism probed by both experiments is exact externally valid, *we reject the null hypothesis of harmonization*. Note that a rejection of harmonization for a given outcome may be local to that outcome. In other words, it may be that the contrast is harmonized and another outcome measurement strategy could be harmonized.

One limitation of these tests is that the Björkman and Svensson (2009) design is underpowered. This limits our ability to reject the relevant null hypotheses, even when a replication study is better- or well-powered. In the context of adversarial replication (or replication by different teams), a null hypothesis of exact external validity or sign-congruent external validity incentivizes the replicating team to design a better-powered study whenever feasible.

E.2 Application II: Sports Outcomes and Pro-Incumbent Voting

E.2.1 Overview

Our framework is also useful for understanding replication efforts that employ observational data, where determining what constitutes a replication is less clear. To illustrate how to use our framework in such contexts, we consider an ongoing debate about whether sporting game outcomes affect pro-incumbent voting. We use this example because it comes from a lengthy published back-and-forth about how replication should be conducted in observational research, and thus provides a unique opportunity where issues about how to analyze and interpret replications with observational data are discussed in print.

In an important contribution, Healy, Malhotra and Mo (2010) find that college football victories, which occur in the two weeks before general elections for president, governor, and senator, increase the incumbent party's vote share in the county where the university is located. The sample used in their study consists of presidential elections from 1960-2004, and gubernatorial and senate elections from 1967-2006. In terms of the mechanism, the authors posit 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 higher incumbent vote share. Healy, Malhotra and Mo (2015: p. 12804) further elaborate this mechanism, writing:

> Voters who are in a positive state of mind on Election Day are likely to use their mood as a signal for the incumbent party's success...and access positive memories about the incumbent party...and/or interpret past actions taken by the incumbent party more favorably...Additionally, positive emotions may cause voters to be more satisfied with the status quo...Those voters may then be more likely to choose the incumbent party in the election.

Since college football victories are thought to be outside the purview of presidents, governors, and senators, this finding—if it arises elsewhere—raises important questions about voter rationality and, as a result, the limits of democratic accountability.[4]

In response to the original Healy, Malhotra and Mo (2010) paper, and sparking the debate about replication that interests us, Fowler and Montagnes (2015) argue that the finding that college football victories increase pro-incumbent voting is likely a false positive. Their argument is built upon a number of analyses. First, they extend the panel from the original Healy, Malhotra and Mo (2010) study to presidential elections from 1960-2012 (i.e., adding 2004-2012), and gubernatorial and senate elections from 1960-2006 (i.e., adding 1960-1967). Using the extended sample, Fowler and Montagnes (2015) test a number of ancillary hypotheses that are consistent with Healy, Malhotra and Mo (2010)'s proposed mechanism, and also include an alternative set of specifications with county-year fixed effects. In addition, Fowler and Montagnes (2015) conduct what is best

[4] We do not contribute to the discussion of voter rationality, or what constitutes an "irrelevant event," for a discussion see Ashworth, Bueno de Mesquita and Friedenberg (2017, 2018).

described as a conceptual replication using NFL games, arguing that the mechanism proposed by Healy, Malhotra and Mo (2010) should operate on such victories as well, especially since NFL games enjoy higher viewership and a more loyal following by fans. The additional specifications, additional data, and conceptual replication analyzed by Fowler and Montagnes (2015) ultimately do not recover evidence that is consistent with the findings originally reported in Healy, Malhotra and Mo (2010). This lack of evidence led Fowler and Montagnes (2015) to conclude that there is no systematic evidence to support the argument that sporting outcome shocks, which may influence voter well-being, benefit incumbent politicians electorally.

In a direct response to Fowler and Montagnes (2015)'s critique, Healy, Malhotra and Mo (2015) argue that Fowler and Montagnes (2015) do not conduct a true replication. Specifically, they claim that Fowler and Montagnes (2015) do not consider the totality of the evidence presented because they do not consider the survey evidence on NCAA basketball games that was discussed in Healy, Malhotra and Mo (2015). Moreover, Healy, Malhotra and Mo (2015) claim that Fowler and Montagnes (2015) neglect their preferred specification, which accounts for a team's *ex-ante* probability of victory, thereby isolating the effect of unexpected victories.[5]

Following up with a different set of replications, Graham et al. (2021) conduct a pre-specified replication of several studies of voter competence/rationality, including Healy, Malhotra and Mo (2010). In addition to correcting several data errors in Healy, Malhotra and Mo (2010) (see the supplemental information of Graham et al. (2021)), they extend the time series slightly. Their preferred specification pools the (corrected) in-sample data with the new (previously) out-of-sample

[5] The authors' preferred operationalization of treatment measures 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.

Citation	Summary
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 mechanism is shocks to voter well-being (football victories) increase voter-satisfaction with the status-quo, translating to increased incumbent vote share.
Fowler and Montagnes (2015)	Argues that HMM2010 is likely a false positve. They re-analyze the HMM2010 data using alternative specifications, conduct the HMM2010 analysis on a longer panel, as well as seeing if the same result holds also for NFL game outcomes. They do not find evidence consistent with the posited mechanism (shocks to voter well-being).
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.
Graham et al. (2021)	Conduct a pre-specified replication of voter competence/rationality including HMM2010, extending the original time series.† Their preferred specification shows that estimates are attenuated, but in the same direction as the original finding.
Fowler and Montagnes (2022*a*)	Argue that Graham et al. (2021) overstate the strength of evidence consistent with Healy, Malhotra and Mo (2010), noting that they cannot reject (statistically) the possibility that the Healy, Malhotra and Mo (2010) was a false positive.
Graham et al. (2022)	Contest equal treatment of multiple specifications and advocate for replication on an expanded sample that consists of both in-sample and out-of-sample observations.
Fowler and Montagnes (2022*b*)	Justifies the focus on multiple pre-specfied tests and argues for the merits of out-of-sample replication.

Table E3: 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.

data and show that estimates are attenuated, but in the same direction as the original finding. In a response, Fowler and Montagnes (2022*a*) argue that Graham et al. (2021) overstate the strength of evidence consistent with Healy, Malhotra and Mo (2010)'s claims. In particular, they distinguish between in-sample and out-of-sample data, and conduct a simulation to show that the evidence on the pooled sample cannot reject (statistically) the possibility that the Healy, Malhotra and Mo (2010) was a false positive.[6] Table E3 summarizes the published (or forthcoming) papers associated with this debate.

Fowler and Montagnes (2015, 2022*a*,*b*) all suggest that the original results in Healy, Malhotra and Mo (2010), that sports outcomes affect voter assessment of incumbents, are likely false

[6] Continuing the back-and-forth, Graham et al. (2022) criticize Fowler and Montagnes (2022*a*)'s treatment of multiple specifications, which weigh the results from different specifications equally. They instead argue for prioritization of average effects over heterogeneous treatment effects.

positives (Type-I errors). The responses of Healy, Malhotra and Mo (2015), and Graham et al. (2021, 2022) claim that the results of Healy, Malhotra and Mo (2010), updated in Graham et al. (2021) reflect a genuine effect. The substance of the debate, as it pertains to replication, centers on statistical questions, mostly about what constitutes the appropriate sample or the right regression specification. While we do not weigh in on the substantive debate, it is worth pointing out that this debate treats replication conceptually. Our framework clarifies some core disagreements between the two teams of scholars when thinking about replication.

E.2.2 Two comparisons of note

We focus on two features of the debate that speak to issues of replication: (i) the presence (or lack thereof) of a similar effect with respect to NFL games; and (ii) the disagreement regarding the treatment of in- and out-of-sample data.

NFL versus NCAA victories: Does the effect of NFL game victories (expected or otherwise) on incumbent vote share constitute a replication of the Healy, Malhotra and Mo (2010) finding that college football victories improve incumbent vote share? Recalling that an empirical target is denoted by τ, the key theoretical claim by Fowler and Montagnes (2015) is that:

$$\tau_{NFL} > \tau_{NCAA},$$

i.e., NFL victories should have a larger (positive) effect on incumbent vote share than NCAA victories on incumbent vote share. They argue "we would expect NFL games to have a greater effect than college football games, because the NFL is significantly more popular-television ratings are $\sim$10 times greater and NFL teams receive strong regional support just like college teams" (p. 13802-3). This argument supposes that NFL victories are a *stronger treatment*, which suggests that the hypothesized claim about NFL games arises as an *artifactual* discrepancy. Specifically, the contrast using NFL games, relative to baseline, produces a stronger activation of voter mood, relative to NCAA games. As a measure of voter mood, the difference between NFL games and NCAA games in the empirical target is an artifactual discrepancy. Yet, they find no detectable effect of NFL victories on incumbent support. We consider two possible scenarios.

Scenario #1: The presence of artifactual discrepancies does not rule out the possibility of target discrepancies. In this case, a positive artifactual discrepancy could be present but attenuated by a (negative) target discrepancy. Consider one possible rationale for target discrepancies (a failure of exact external validity). It may be the case that the voter mood mechanism produces different effects on different cross sections of counties. NFL teams tend to be located in larger metropolitan areas, on average, than NCAA teams. It could be the case that the effect of the mechanism depends on metro-area population (due to, for example, the availability of non-football related activities). If it is true that NFL victories (relative to NCAA victories) produce a positive artifactual discrepancy, then the finding of no detectable effect of NFL victories could be evidence that the mechanism lacks external validity.

Scenario #2: It is possible that the hypothesis forwarded by Fowler and Montagnes (2022*b*) about artifactual discrepancies is not correct. Suppose that for non-obvious reasons, the artifactual discrepancies of NFL victories relative to NCAA victories was *negative* and of a similar magnitude to the positive effect of the mechanism identified by Healy, Malhotra and Mo (2010). If this were the case, the artifactual discrepancies could counteract the effect of an externally valid mechanism, misleading assessments of the external validity of the mechanism. Substantively this would mean that NFL victories are not “similar enough” to provide an alternative measure of the mechanism at play in the NCAA result. Consequently, one should not expect the same kind of effect for NFL victories.

In- versus out-of-sample: Graham et al. (2022) and Fowler and Montagnes (2022*b*) disagree on whether the original sample should be pooled with out-of-sample replication data. The primary argument of Fowler and Montagnes (2015) is that the result suggesting that college football victories improve incumbent vote share is a false positive. To show this, they conduct an analysis similar to Healy, Malhotra and Mo (2015) but on data that is outside the original sample. We note that they test whether treatment effects are different from zero but do not conduct a formal test comparing the original and replication estimates. Graham et al. (2021), instead, take the new data and *combine it with the original sample where the purported false positive is present*, to see if the result still maintains. They find that the original result is attenuated. Specifically, Graham et al. (2021) find a reduced influence of college football victories on incumbent vote share when combining additional data with the original sample. Graham et al. (2022) argue for pooling the original sample with new data, while Fowler and Montagnes (2022*b*) argue for the merits of out-of-sample replication and comparison of findings.

Why does the in- and out-of-sample definition matter when considering a replication? A replication is about a comparison between empirical targets (through estimates) that reflect different settings. Fowler and Montagnes (2015)'s analysis splits up the available data into two distinct samples that reflect different "settings" essentially making their exercise a replication assessing whether the empirical targets are similar, i.e., whether the voter mood mechanism has sign-congruent external validity. Graham et al. (2022)'s argument to pool all the data reflects a statistical concern, specifically, more data is better. However, their statistical exercise much more closely resembles a meta-analysis. Slough and Tyson (2023) show that standard meta-analyses *assume* target-equivalence through assertion of a common parameter across constituent studies. By analogy, this pooling exercise makes sense if one believes that the effect of the mechanism is the same in both samples/settings. But there are reasons that one may be skeptical of this claim. For instance, college football viewership has held steady over the past 20 years,[7] whereas the population—and hence the pool of eligible voters—has grown. This renders college football victories a *weaker* treatment for the electoral application. We might, then, expect the effect of college football on incumbent vote share to attenuate toward zero (artifactually) since a smaller subset of the voting population is

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

watching college football.

Before concluding, it is worth mentioning that statistical discrepancies are necessarily present. We know that the estimates using any of the above settings or measurement strategies will be measured with error. In the best case, Fowler and Montagnes (2015) test a null hypothesis of zero in different subsets of the data (with different specifications). Figure 2 in the main text of our paper shows how independent hypothesis tests of a null hypothesis of zero can lead to misleading inferences using heuristic versions of the sign-comparison test. Our estimate- or sign-comparison test both provide formal tests that can be applied in experimental and observational replications. In sum, our framework and approach to comparison of estimates in replication studies can be productively applied to observational studies.

F A Structural Approach to Replication

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

$$\tau_m(\omega', \omega''|\theta) = f(\omega', \omega'', m) + g(\theta). \tag{F.1}$$

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

$$\tau_1 = \lambda(\tau_2; m, \omega', \omega''), \tag{F.2}$$

where λ specifies the relationship between observed effects, e_1 and e_2, and how that relationship depends on contrasts and measurement strategies. This allows (2) to be written in terms of a single target:

$$e_1 - e_2 = \varepsilon_1^{n_1} - \varepsilon_2^{n_2} + \lambda(\tau_2; m, \omega', \omega'') - \tau_2,$$

where target and artifactual discrepancies can be written as properties of λ.

The structural approach is most commonly used to *combine* rather than *compare* estimates across studies. Indeed, this formulation in the context of replication represents a natural extension of Pearl and Bareinboim (2011)'s approach to transportability and is commonly invoked—if unstated—in meta-analyses (Slough and Tyson, 2023). But, if one is willing to posit such a model, and the assumptions about how treatment effects can change across contexts, a similar approach can also be applied to replication studies.

The key strength of the structural approach is that it allows an analyst to make strong empirical conclusions from data, potentially eliminating concerns about target or artifactual discrepancies. It is important to stress, however, that these benefits result from modeling assumptions that constrain the kind of data substantive phenomena are permitted to supply. Moreover, there is little consensus on how to constrain substantive phenomena, i.e., what structural assumptions are appropriate in what cases, and whether such things are faithfully represented as "nuisance" parameters, especially when applied to evidence accumulation. Many structural approaches assume exact external validity and that measured treatment effects do not vary in the design of the studies.[8] By prohibiting substantive phenomena from presenting target or artifactual discrepancies (other than as idiosyncratic error), analysts dodge the problems resulting from artifacts of research design, or lack of external validity, that we highlight, undermining the causal interpretation some analysts may wish to impart to results from replication. Further exploration of structural approaches to replication should stress transparently what assumptions are involved, and state precisely what is gained when downplaying the potential problems that might arise when combining evidence from multiple places.

[8] Slough and Tyson (2023) term this assumption “design invariance.”

References

Adida, Clair L., Jessica Gottlieb, Eric Kramon and Gwyneth McClendon. 2019. *Information and Political Accountability: A New Method for Cumulative Learning*. Cambridge University Press chapter Under what conditions does performance information influence voting behavior? Lessons from Benin, p. 81*117.

Ashworth, Scott, Ethan Bueno de Mesquita and Amanda Friedenberg. 2017. "Accountability and information in elections." *American Economic Journal: Microeconomics* 9(2):95–138.

Ashworth, Scott, Ethan Bueno de Mesquita and Amanda Friedenberg. 2018. "Learning about voter rationality." *American Journal of Political Science* 62(1):37–54.

Berger, Roger L. 1982. "Multiparameter Hypothesis Testing and Acceptance Sampling." *Technometrics* 24(4):295–300.

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.

Brinch, Christian N., Magne Mogstad and Matthew Wiswall. 2017. "Beyond LATE with a Discrete Instrument." *Journal of Political Economy* 125(4):985–1039.

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.

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`.

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.

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.

Munger, Kevin. 2023. "Temporal Validity as Meta-Science." *Research & Politics* Forthcoming.
URL: *https://osf.io/4utsk/*

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

Raffler, Pia. 2022. "Does Political Oversight of the Bureaucracy Increase Accountability? Field Experimental Evidence from a Dominant Party Regime." *American Political Science Review* 116(4):1443–1459.

Raffler, Pia, Daniel N. Posner and Doug Parkerson. 2022. "Can Citizen Pressure be Induced to Improve Public Service Provision?" Working paper, available at `http://danielnposner.com/wp-content/uploads/2022/04/RPP-ACT-Health-220323.pdf`.

Slough, Tara and Scott A Tyson. 2023. "External Validity and Meta-analysis." *American Journal of Political Science* 67(2):440–455.