External Validity and Meta-Analysis*

Tara Slough† Scott A. Tyson‡

Abstract

Meta-analysis is a method that combines estimates from studies conducted on different samples, in different contexts, or at different times. Social scientists increasingly use meta-analyses to aggregate evidence and learn about general substantive phenomena. We develop a framework to examine the theoretical foundations of meta-analysis, with emphasis on clarifying the role of external validity. We identify the conditions under which multiple studies are target-equivalent, meaning they identify the same empirical target. Our main result shows that external validity and harmonization, in comparisons made and how outcomes are measured, are necessary and sufficient for target-equivalence. We examine common formulations of meta-analysis—fixed- and random-effects models—developing the theoretical assumptions that underpin them and providing design-based identification results for these models. We then provide practical guidance based on our framework and results. Our results reveal limits to agnostic approaches to the combination of causal evidence from multiple studies.

Keywords: Meta-analysis; External Validity; Measurement

*We thank Peter Aronow, Neal Beck, Alex Coppock, Chris Fariss, Don Green, Guy Grossman, Federica Izzo, Dorothy Kronick, Winston Lin, John Marshall, Kevin Munger, Pablo Querubín, Mark Ratkovic, Cyrus Samii, Jessica Sun, Stephane Wolton, panel participants at PolMeth XXXVIII, the summer 2021 Virtual Formal Theory Seminar, workshop participants at EGAP, and participants at the 17th Northeast Workshop in Empirical Political Science.

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

‡Assistant Professor, University of Rochester, styson2@ur.rochester.edu

The identification revolution transformed the way social scientists approach empirical research by advancing research designs that ascribe a causal interpretation to observed effects (Angrist and Pischke, 2010; Samii, 2016). The most prominent critique of designs intended to identify causal effects stresses a lack of external validity, or researchers' inability to know whether and how findings apply beyond the scope of an individual study (Deaton, 2010; Esterling, Brady, and Schwitzgebel, 2021). In response to this critique, proponents of causal identification-driven research often advocate the use of *meta-analysis* to combine results from multiple studies conducted on different samples, in different contexts, or at different times (Banerjee and Duflo, 2009; Imbens, 2010; Gerber and Green, 2012).

When used to combine treatment effects, a meta-analysis aims to reveal a quantitative measure of a general (causal) effect, thus helping to explain social phenomena and potentially advance policy recommendations. This exercise critically relies on two important theoretical properties: (i) constituent studies are unified by a common mechanism; and (ii) constituent studies are aiming at the same empirical target. Most expositions of meta-analysis take these conceptual concerns for granted, focusing instead on estimation issues (e.g., Field and Gillett, 2010). Whether a common mechanism exists, and whether constituent studies aim at the same empirical target, are questions about the *theoretical* relationship between constituent studies—which are distinct from the specific mechanism under investigation. In this article, we develop a framework to study the theoretical foundations of meta-analysis, and in doing so, elucidate concepts and establish identification results that are important for existing, as well as future, meta-analyses.

In our framework, a study consists of three components. First, a study is conducted in a *setting*, which includes features of the population, like subject attributes or behavioral types (Wilke and Humphreys, 2020), and features of the environment relevant to the behavior under investigation. The set of settings outlines the *scope conditions* of an argument or theory, i.e., when and where we may reasonably expect the mechanism to manifest. Second, a study focuses on a *contrast*, explicitly or implicitly, which defines the comparison of substantive and empirical interest

(Bueno de Mesquita and Tyson, 2020). Most commonly, a contrast is given by a treatment/control comparison. Third, a study has a *measurement strategy*, which specifies the outcomes of interest and how they are measured (Adcock and Collier, 2001).

A critical concept for meta-analysis is that constituent studies are *target-equivalent*, meaning they refer to a common substantive (theoretical) quantity, and are thus focused on capturing the same thing.[1] Target-equivalence ensures that constituent studies aim at the same empirical target, and is necessary to ensure that conclusions drawn from a meta-analysis are meaningful and interpretable. Our results stress the importance of design *harmonization* between studies—with respect to contrasts and measurement strategies. Constituent studies are contrast harmonized when the substantive comparison across studies is the same. Studies are measurement harmonized when the outcome of interest is the same and it is measured in the same way. Harmonization ensures that the same underlying construct is represented across studies, and thus what matters is construct validity across measurement strategies and contrasts, rather than equality in a literal sense. Contrast and measurement harmonization are important because they are typically the most amenable to researcher control, either through design or inclusion criteria.

Our results highlight the importance of two kinds of validity when comparing or combining effects across studies (Shadish, Cook, and Campbell, 2002). First, two measurement strategies satisfy *divergent validity* when they can be distinguished, i.e., they produce different treatment effects in the same setting and at the same contrast (in a single study). Divergent validity ensures that differences between the effects in constituent studies, when they occur at the theoretical level, are the result of different substantive mechanisms and not artifacts of differences in how outcomes are measured. Second, a mechanism has *external validity* when it produces identical treatment effects across settings (up to statistical noise), when evaluated at exactly the same contrast and with the same measurement strategy. Put another way, a mechanism *lacks* external validity if the

[1] Examining field experiments on political accountability, Izzo, Dewan, and Wolton (2020) use a similar notion, *comparability*, which is analogous to target-equivalence.

same mechanism produces systematically different effects across different settings when all other aspects of studies are identical (including the mechanism).

Our first result shows that estimands from contrast-harmonized constituent studies are target-equivalent if and only if the studies are measurement harmonized, and our second result shows that estimands for two measurement-harmonized constituent studies are target-equivalent if and only if the studies are contrast harmonized. Our main result shows that external validity and harmonization (in contrasts and measurement strategies) are necessary and sufficient for target-equivalence. This result implies that without harmonization, a meta-analysis may not find consistent evidence of an externally valid mechanism *even when that mechanism is, in fact, present in all the settings comprising the meta-analysis*. Also, our results imply that a meta-analysis comprised of nonharmonized studies may seemingly find consistent evidence of a substantive mechanism—even when that mechanism is, in fact, absent.

We conclude with some practical guidance by applying our results to a common meta-analysis model—random-effects—carefully isolating the theoretical from statistical dimensions. Many existing meta-analyses (implicitly or explicitly) invoke *design invariance*, a theoretical assumption that delivers identification of a common treatment effect in meta-analyses. Although design invariance implies target-equivalence, it is considerably stronger because it refers to when a treatment effect is not sensitive to how it's measured or what comparisons are being made. We show instead that external validity and harmonization provide identification in the random-effects model under far less stringent assumptions. We further provide guidance for prospective meta-analysts when harmonization or design invariance are unlikely to hold and where external validity may be more localized.

The framework and results we present are intentionally designed to articulate the theoretical foundations of meta-analysis in the social sciences given recent interest in this research design (see Table 1 and Appendix S2.1). We conceptualize external validity *of a mechanism* or causal pathway, and our framework requires a precise formulation of external validity to facilitate a formal

analysis.[2] We contribute to the theoretical understanding of external validity in two ways. First, we formulate external validity formally and develop theoretical results regarding the connection between external validity, harmonization, and empirical identification in meta-analysis. Second, our framework reconciles the relationship between a mechanism and a reduced-form effect, namely, a mechanism produces a reduced-form effect at a particular measurement strategy and contrast (which is different at different contrasts and measurement strategies). While statistical issues can be nested into our framework, our goal is to articulate and stress conceptual or theoretical issues related to meta-analysis of causal effects. This article thus contributes to an emerging literature on the *theoretical implications of empirical models*, where scholars have focused on the connection between theory and identification strategies (Bueno de Mesquita and Tyson, 2020; Abramson, Koçak, and Magazinnik, 2019; Tomasi, 2020; Slough, 2022), the design of field experiments (Chassang, Padró i Miquel, and Snowberg, 2012), and the decision-theoretic foundations of experiments (Banerjee et al., 2020).

External Validity

Meta-analysis necessarily takes a perspective on external validity (implicitly or explicitly). We argue that external validity is better understood as a cluster of related concepts rather than a singular notion. Formulations of external validity can be organized into two classes: **projectivism**, where external validity is a property of a single study, and **cross-sectionalism**, where external validity is a property of a collection (or cross-section) of studies. A meta-analysis necessarily relies upon the latter concept.

Projectivism conceptualizes external validity as assessing whether an empirical finding from a single study *projects onto a destination*, and can be categorized further depending on that destination. Shadish, Cook, and Campbell (2002) focus on projecting empirical results or estimands from a single study onto another setting, which can vary in units (sample), outcomes, or treatments. Es-

[2]Gailmard (2021) suggests informally that external validity is a theoretical property associated with a mechanism.

Study	Stated Meta-Analysis Motivation			Component study design		Estimator		
	Generalizability	Precision	Lit. Synthesis	Experimental	Observational	RE	FE	Other
A: Prospective Meta-Analysis of Harmonized, Original Studies								
Dunning et al. (2019)	✓			✓			✓	✓
de la O et al. (2021)	✓			✓		✓		
Slough et al. (2021)	✓	✓		✓		✓		
Blair et al. (2021)	✓			✓		✓		
Coppock, Hill, and Vavreck (2020)	(✓)	✓		✓		✓		
B: Retrospective Meta-Analysis based on Secondary Analysis of Existing Studies								
Blair, Christensen, and Rudkin (2021)			✓		✓	✓	✓	
Blair, Coppock, and Moor (2020)			✓		✓†	✓		
Eshima and Smith (2022)			✓	✓		✓		
Incerti (2020)			✓	✓		✓	✓	
Godefroidt (2021)	✓		✓		✓*	✓		
Kertzer (2020)	✓		✓		✓‡	✓		
Schwarz and Coppock (2022)	✓		✓	✓		✓		
C: Meta-Analysis based on Secondary Analysis and Original Studies								
Kalla and Broockman (2018)	(✓)		✓	✓		✓		

Table 1: Meta-analyses in three political science and general science journals. For study selection and classification information see Appendix S2.1.

†: Blair, Coppock, and Moor (2020) estimate the difference in prevalence of a sensitive behavior between list experiments and direct questions. This difference is observational.

*: Godefroidt (2021) pools experimental and observational studies in a meta-analysis.

‡ Kertzer (2020) examine differences in treatment effects between mass and elite respondents. This comparison is observational.

(✓) indicates that motivation is suggested but not explicitly stated in study.

terling, Brady, and Schwitzgebel (2021) define external validity as following from the truth status of an empirical claim projecting onto an abstract statement. Similarly, *transportability* of Pearl and Bareinboim (2011, 2014) aims to generalize a study's finding from the context where a study took place to a different setting by imputation, a "transport formula" which imputes the causal effect in a setting from the causal effect identified elsewhere. This is accomplished by reweighting experimental findings using observational data collected in both settings (a variant of "selection on observables"). Egami and Hartman (2020) develop a related idea of the "contextual exclusion restriction" which holds that unit-level treatment effects do not change with unobserved contextual factors at the study level. These formulations are poorly adapted for meta-analysis as they emphasize only transportation of study-level causal effects, rather than combining findings from actual studies.

Another form of projectivism follows when the destination is a *grand population*, where an underlying mechanism manifests, but because of sampling, physical, or temporal constraints, is divided into subpopulations. A "population-level estimand" can be estimated from a single study through statistical techniques applied to the observed sample (Gerber and Green, 2012). Critically, this approach assumes the existence of a common parameter uniting all samples obtainable from the grand population, thus reducing the conceptual issues we highlight to *estimation problems*. Various estimators have been developed to move from sample estimands to grand population parameters (i.e., Cole and Stuart, 2010; Kern et al., 2016; Gechter and Meager, 2021), and recent expositions of this view describe forms of "external validity bias," which is the difference between sample and population estimands (Egami and Hartman, 2020; Findley, Kikuta, and Denly, 2021).

Last, *parallelism* focuses on whether empirical findings measured in an artificial setting (like a laboratory) extend to natural settings (Smith, 1982). Latour (1993) argues that manipulating natural phenomena are ineluctably non-generalizable. Guala (2005) interprets parallelism as a particular type of robustness, of a study design or mechanism, which can be assessed from study to study. Pritchett and Sandefur (2015) build upon this view to bridge the gap between experiments and

observational studies in the case of microfinance. Mutz (2011) conveys a related notion focusing on the similarity between experimental treatments (rather than empirical findings) and the "real world."

Perhaps the most practical characterization of projectivism is Fariss and Jones (2018), who argue that the extrapolation of a single study's result should be viewed in terms of its predictive scope. But projectivist perspectives on external validity are not naturally suited for meta-analyses, which essentially treats constituent studies symmetrically. Complementing projectivism, we formulate a *cross-sectionalist* view of external validity, treating it as a symmetric characteristic of studies. Cross-sectionalism has no destination to which a study's findings need project, nor does it necessarily assume the existence of a theoretical grand population from which samples are drawn. Because a meta-analysis combines empirical findings across studies, it necessarily requires a cross-sectionalist formulation of external validity. The recent rise in prospective multi-site meta-analyses, including Evidence in Governance and Politics' (EGAP's) Metaketa initiative is a clear, if implicit, endorsement of cross-sectionalism. Specifically, by allocating resources over multiple studies instead of a better-powered individual study, meta-analysis practitioners invest in the idea that external validity represents more than simply extrapolation.

An Illustrative Example

Before proceeding to our main analysis, we briefly present a simple example of a hypothetical meta-analysis comprised of two studies, 1 and 2, to illustrate some of our main points. Both constituent studies, 1 and 2, study the influence of a single common mechanism, which uniquely generates an observed effect. Further, suppose that the observed treatment effects in studies 1 and 2, μ_1 and μ_2, are measured absent statistical noise or sampling variability (e.g., because the sample size in each study is infinite and there is no measurement error).

Now suppose that the effects in each study can be written as

$$\mu_1 = \eta + \delta_1 \quad \text{and} \quad \mu_2 = \eta + \delta_2,$$

where η is the common effect, and δ_1 and δ_2 are constants. If, for simplicity, one posits enough structural assumptions, the meta-analysis estimand is a weighted average of μ_1 and μ_2, with respective weights α and $1 - \alpha$.[3] In this case the meta-analysis estimand is $\alpha\mu_1 + (1 - \alpha)\mu_2$, and can be written as

$$\eta + \underbrace{\alpha\delta_1 + (1 - \alpha)\delta_2}_{b}.$$

If $b \neq 0$, this is because $\mu_1 \neq \mu_2$. Since there are no statistical or sampling reasons that the observed effects between studies 1 and 2 differ, by assumption, the term b is not statistical noise originating from sampling differences, treatment imbalance, or differential measurement error between 1 and 2. Moreover, the mechanism across 1 and 2 is exactly the same, so b is not the result of an additional mechanism (i.e., mediator) present in one study but not in the other. Why is it important, theoretically, for $b = 0$? When $b \neq 0$, *studies* 1 *and* 2 *are not aiming at the same empirical target*. Thus, b is an artifact of *non-random* discrepancies in measurement or comparison. Work-horse statistical models for meta-analysis effectively assume b away by assuming that it is random, mean-zero noise. Although b is similar in spirit to "bias" it is important to emphasize that it represents a separate theoretical—not statistical—issue: the studies' targets are not the same.

To make things more concrete, we discuss a (hypothetical) meta-analysis of get-out-the-vote interventions in off-cycle local elections in the United States. In these contests, turnout is generally very low (Hajnal and Lewis, 2003), and one possible explanation for low turnout is that voters do not know when and where elections are held.[4] In each of the two constituent experiments, researchers randomly assign households to receive a strictly informational pamphlet about where

[3] After normalization, this is without loss of generality.

[4] Gerber and Green (2019: Appendix B) present a meta-analysis similar to this example.

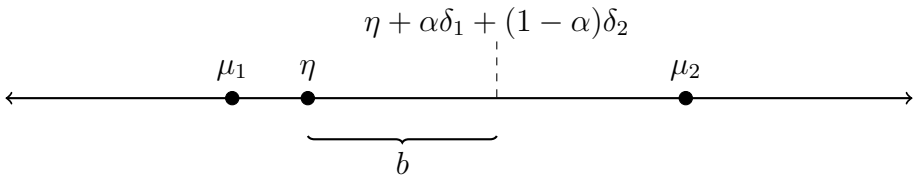

Figure 1: The meta-analytic estimate from studies 1 and 2 is $\eta + \alpha\delta_1 + (1-\alpha)\delta_2$. One cannot ascertain the sign or magnitude of η from the meta-analytic estimate because the sign and magnitude of b is unknown.

and when to vote in the local election. From these pamphlets, voters learn that they have the option to cast a ballot in an upcoming election. Researchers may be interested in synthesizing treatment effects on turnout. In this context, μ_1 and μ_2 measure the (true) average treatment effects of the informational pamphlets on household members' turnout decision in the off-cycle election. But, what if pamphlets were distributed at different times relative to the election between studies 1 and 2? or what if turnout is measured differently across 1 and 2?

Framework

We develop a theoretical framework to analyze conceptual issues that arise when combining estimates of causal effects in a meta-analysis. To focus on considerations that are unique to meta-analyses, we abstract from two concerns related to constituent empirical studies: (1) internal validity and within-study identification, and (2) sampling and estimation. Both are important and have received comprehensive textbook treatments (e.g., Angrist and Pischke, 2008). Our framework therefore should be viewed as complementary to those that focus on internal validity and estimation within single studies.

Building Blocks: Constituent Studies

The **setting** of a study is represented by θ, where the set of potential settings is a compact set, $\Theta \subset \mathbb{R}$, which has strictly positive Lebesgue measure.[5] Since all of our analysis is relative to a particular mechanism, the set of settings, Θ, characterizes the *scope conditions* of an argument or theory. As such, Θ contains only settings where the mechanism of interest could potentially arise, and thus be subject to empirical scrutiny.[6]

In general, an empirical study is a measurement exercise where researchers detect the presence of a mechanism by measuring its influence. This is accomplished by selecting a particular outcome and measuring the effect of a manipulation (or treatment) on that outcome. To capture the essential features of this approach in our framework, denote a (finite) set of **measurement strategies**, M, where different elements $m \in M$ correspond to different outcome measures, different measurement scales, etc.[7] For example, one could measure the effect of an informational pamphlet on turnout, or survey general awareness of upcoming elections, either of which would correspond to a different measurement strategy, i.e., a different $m \in M$. Note that this does not mean that two measurement strategies cannot be correlated, simply that our framework does not distinguish measurement strategies that are indistinguishable.

The second component of an empirical study is the selection of a comparison, which is chosen or designed to define the mechanism's effect. Formally, such a comparison is taken from a set of potential instruments (Imbens and Angrist, 1994), that we denote by the compact set $\Omega \subset \mathbb{R}$, which has strictly positive Lebesgue measure.

[5]Lebesgue measure, λ, formalizes volume and can be computed for a (measurable) set A with the formula

$$\lambda(A) = \int_A 1dx = \int \mathbb{1}_{\{A\}} dx.$$

For example, the Lebesgue measure of the interval $[a, b]$ is $\int_a^b dx = b - a$. For a complete construction see, e.g., Aliprantis and Border (2006: Ch. 10.6).

[6]We could include an indicator, where $\gamma = 1$ if the mechanism is present and 0 otherwise, but since the mechanism in our analysis is fixed, $\gamma = 1$ everywhere.

[7]That M is finite is not consequential for our results.

Definition 1. *A **contrast** is a pair, $(\omega', \omega'') \in \Omega^2$. The set of contrasts is*

$$\mathcal{C} = \{(\omega', \omega'') \mid \omega', \omega'' \in \Omega\}.$$

For a contrast, (ω', ω''), one can attach concrete labels like control (ω') and treatment (ω''), or more generally, ω' and ω'' could refer to different treatment conditions. Different contrasts, i.e., different elements in $\mathcal{C}$, capture different possible experimental manipulations, different classifications or dosages of treatment, and/or different underlying "untreated" states. For instance, comparing different control instruments (from different studies) to identical treatment instruments would imply different contrasts, and each would entail different substantive interpretations. Although an analyst typically normalizes their notation so that $(\omega'_1, \omega''_1) = (0, 1)$, the same normalization in another study, so that $(\omega'_2, \omega''_2) = (0, 1)$, can be misleading when $(\omega'_1, \omega''_1) \neq (\omega'_2, \omega''_2)$ since "0" and "1" have different meanings in the two studies. The set $\mathcal{C}$ is uncountably infinite and thus includes many potential treatment states.

We now define a constituent study:

Definition 2. *A **study** is a triple, $\mathcal{E} = \{m, (\omega', \omega''), \theta\} \in M \times \mathcal{C} \times \Theta$, comprised of a measurement strategy, m, a contrast, (ω', ω''), and a setting, θ. A **meta-study** is a collection, indexed by $i \in \mathcal{I}$, of constituent studies, $\mathcal{E}_i$, which we denote by $\mathcal{M}(\mathcal{I}) = \{\mathcal{E}_i\}_{i \in \mathcal{I}}$.*

Adding statistical noise to a constituent study, $\mathcal{E}$, would create a statistical model that determines the exact family of distributions over outcomes, constituting a Blackwell experiment (Blackwell, 1953). It is important to stress that two different studies, $\mathcal{E}_1$ and $\mathcal{E}_2$, investigating the same mechanism, can involve different measurement strategies, different contrasts, or different settings.

The output of a constituent study is some form of estimand or treatment effect.

Definition 3. *For a study $\mathcal{E} = \{m, (\omega', \omega''), \theta\}$, a **treatment effect** is a mapping $\tau_m(\omega', \omega'' \mid \theta) : M \times \mathcal{C} \times \Theta \to \mathbb{R}$, which is smooth almost everywhere and whose derivative has full rank for almost*

all contrasts.[8]

A mechanism manifests in our framework through its observable influence, which depends on both a measurement strategy and a contrast. We take this approach to be consistent with the common view that causality is assessed by focusing on *measuring the effects of causes* (Holland, 1986). This ensures that our approach remains agnostic since our results do not rely on an underlying theoretical structure that articulates a mechanism, but rather, on a mechanism's observable influence.[9] The function $\tau_m(\omega', \omega'' \mid \theta)$ represents the empirical target of a constituent study, answering a question like "what is the effect of A on B?"

The function $\tau_m(\omega', \omega'' \mid \theta)$ can represent many different estimands depending on the application and setting. To illustrate, consider

$$\underbrace{\tau_m(\omega', \omega'' \mid \theta)}_{\text{empirical target}} = \underbrace{f_{\mathcal{D}}(Y_m(\omega'') \mid \theta) - f_{\mathcal{D}}(Y_m(\omega') \mid \theta)}_{\text{estimand}}, \tag{1}$$

where $Y_m(\omega)$ are potential outcomes, f is some operator (generally the expectations operator or a quantile function), and $\mathcal{D}$ denotes the set of units over which the mechanism operates (according to the investigator). If the mechanism operates at the *study* level, $\mathcal{D}$ would include all subjects or units. If, however, we expected that a mechanism was only operative on women, $\mathcal{D}$ would include only subjects that identify as women. In this case, we would want to compare (for example) CATES on women instead of average treatment effects on the whole sample, since the treatment effect associated with the mechanism may be differentially diluted across sites, as a function of sample composition.

[8] A property holds *almost everywhere* when it holds with the exception of sets that have zero measure, e.g., events with probability zero.

[9] Thus, our framework is not tied to a particular formulation of a causal principle, making it more agnostic than the model utility criteria of Findley, Kikuta, and Denly (2021: pg. 377).

Discussion

Equation (1) shows that our framework accommodates many estimands including, but not limited to, the average treatment effect, the treatment effect on the treated, and the local average treatment effect, all of which are related to the marginal treatment effect of Heckman and Vytlacil (2005). Differences in these estimands are inconsequential for our results. Because of this, and in the interest of clarity, we abstract from specific estimands for the remainder of our framework, referring only to "treatment effects." Assuming that $\tau_m(\omega', \omega''|\theta)$ is smooth in contrasts almost everywhere is not particularly restrictive, unless for instance, the set of treatment effects is known to be a fractal. Finally, our framework can be extended to incorporate a sample, n, into our treatment effect function. For instance, taking two samples n and n', a direct replication would assess whether $\tau_m(\omega', \omega''; n \mid \theta)$ is statistically different from $\tau_m(\omega', \omega''; n' \mid \theta)$.

We are not the first to represent an empirical study as a theoretical object. Shadish, Cook, and Campbell (2002: pg. 19), building on Cronbach and Shapiro (1982), think of a study as being comprised of four components: units, treatments, observations, and settings (UTOS), which is expanded in Findley, Kikuta, and Denly (2021) to include mechanisms and time under the acronym M-STOUT.[10] Similarly, Blair et al. (2019), focus on four different components: a model, an inquiry, a data strategy, and an answer strategy (MIDA). PICO—population, intervention, comparison, and outcome—is popular in medicine. We map the aspects of our characterization of studies onto each of these frameworks in Table 2. It is important to emphasize that no framework is entirely comprehensive—and this is perhaps the greatest attribute of conceptual frameworks. In our characterization of a study, we include only the pieces that we require for our results and the points those results highlight, and we intentionally abstract from other features which are not critical.

Application. We apply our framework to studies on informational pamphlets about off-cycle elections, where the settings consist of local elections. These settings could be described by both

[10]See Munger (2021) on the importance of temporal validity.

Our framework	Analogous feature of... UTOS	M-STOUT	MIDA	PICO
Setting	Setting, Units	Setting, Units, Time	–	Population
Contrast	Treatments	Treatment	Model	Intervention, Comparison
Measurement strategy	Observations	Observations	Data strategy	Outcome
Mechanism	–	Mechanism	Model	–

Table 2: Relation between elements of our framework and other conceptual frameworks.

the locality (the constituency) and the specific election. The mechanism of interest is informational: pamphlets inform (some) registered voters that they can vote for local offices. The contrast we focus on is treatment-vs-control comparison where households in treatment were mailed the pamphlets and households in control were not. The baseline measurement strategy that we consider is an indicator for turnout from the state voter file.

We might not expect the posited informational mechanism to be present for all residents of experimental households. If the pamphlets simply teach voters that they have the opportunity to vote, it should not produce a treatment effect on subjects who already know the election is coming up. One natural concern, therefore, is that in different settings (local elections), different shares of the electorate might be affected by the mechanism we propose. To account for these differences, researchers might condition on past voting behavior. For example, the set $\mathcal{D}$ from (1) may be specified to include only voters who have never voted in past off-cycle elections. Researchers would then meta-analyze the conditional ATES (CATES) on this group of past non-voters for whom this mechanism may be most plausible.

Concepts

Any empirical study is comprised of at least two phases—a design phase and an analysis phase (Morton and Williams, 2010). In this section, we focus on concepts that are relevant for design

decisions that researchers make when conducting meta-analyses.

Target-equivalence and Harmonization

Study $\mathcal{E}_1 = \{m_1, (\omega_1', \omega_1''), \theta_1\}$ is conducted in setting θ_1, where outcomes are assessed with measurement strategy m_1, and the substantive comparison of interest is given by the contrast (ω_1', ω_1''). In study $\mathcal{E}_2$, where the same mechanism is at play, the setting is θ_2, outcomes are measured with measurement strategy m_2, and the contrast (ω_2', ω_2'') defines the comparison of interest.

Definition 4. *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*

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

A meta-study has ***target-equivalence*** *if all constituent studies* i *in* $\mathcal{M}(\mathcal{I})$ *are target-equivalent.*

When treatment effects in constitutent studies are aiming at the same empirical target, they can be usefully combined. As such, when target-equivalence holds, all empirical issues reduce to statistical issues, i.e., issues regarding estimation of the common effect. Note that $\mathcal{D}$ from (1) need not be the same across constituent studies to ensure target-equivalence. It is important to emphasize that when an analyst posits a grand population parameter to be estimated from constituent study estimates, they implicitly invoke target-equivalence. Specifically, with a fixed-or random-effects estimator, as well as the Bayesian approach of Gechter and Meager (2021), there is an underlying population parameter—which is the same across studies—and the goal of the estimation approach is to identify this parameter. Assuming the existence of such a parameter in these empirical models *invokes target-equivalence as an underlying assumption*.

Studies can be harmonized along different dimensions, and for conceptual clarity, we present each separately.

Definition 5. *Studies* $\mathcal{E}_1$ *and* $\mathcal{E}_2$ *are*

1. ***measurement harmonized*** *if* $m_1 = m_2$ *at almost every setting;*

2. ***contrast harmonized*** *if* $(\omega_1', \omega_1'') = (\omega_2', \omega_2'')$ *in almost every setting;*

3. ***harmonized*** *if they are both contrast and measurement harmonized.*

A meta-study $\mathcal{M}(\mathcal{I})$ *is harmonized if every constituent study,* $\{\mathcal{E}_i\}_{i \in \mathcal{I}}$*, is harmonized.*

When two studies, $\mathcal{E}_1$ and $\mathcal{E}_2$, employ the same outcome—measured in the same way—to assess the effect of a mechanism, then we say that those two studies are measurement harmonized. A contrast represents the comparison that leads to the measured treatment effect, which is usually between two different treatments, or equivalently, between treatment and control. Consequently, measuring the levels of both values of the instruments, $(\omega', \omega'') \in \Omega$, is a critical ingredient for generalizing across studies (formally or otherwise). When contrasts are harmonized between two studies, $\mathcal{E}_1$ and $\mathcal{E}_2$, then the instruments used to make comparisons are equivalent. In practice, this corresponds to ensuring that the control and treatment arms between two studies, $\mathcal{E}_1$ and $\mathcal{E}_2$, are the same.

We reserve the term harmonization for when studies are both contrast and measurement harmonized, that is, studies that use the same outcome, measure that outcome the same way, and employ the same comparison. The Metaketa studies highlight harmonization of a common treatment as an important feature of prospective study design. These harmonization efforts do not necessarily achieve contrast harmonization because they do not endeavor to harmonize (or measure) the control instrument across sites. In the Supplementary Information to this article, we discuss harmonization of the meta-analyses listed in Table 1.

Construct Validity

Our framework is theoretical, and thus, the concepts we employ are theoretical constructs that relate to empirical objects. Consequently, harmonization—both in terms of contrasts and measurement strategies—is evaluated in terms of *construct validity* in each study, i.e. the extent to which an empirical object corresponds to an underlying substantive concept (Shadish, Cook, and Campbell, 2002; Adcock and Collier, 2001). Harmonization means that measurement strategies and contrasts

are identical *in the model*, meaning they represent the same construct, but does not mean that they are the same in a literal sense (see also Ferraro and Agrawal, 2021).[11] To illustrate, consider the example from Gilbert et al. (2016), who critique a replication study that treated taking military leave for an Israeli as the same as an American going on a honeymoon, arguing that these two things do not represent the same construct. Military leave for an American similarly may not be the same as it is for an Israeli, and whether it is depends on a substantive argument that hinges on specialized contextual knowledge. Verifying harmonization relies on a positive argument by the analyst that follows from substantive and contextual factors, the details of which will vary from case to case.

Returning to our get-out-the-vote example, consider two possible failures of harmonization. First, suppose that the pamphlets in experiment 1 were distributed one month prior to the off-cycle election but that the pamphlets in experiment 2 were distributed two days before the election. If sequence relative to the election changes the "strength" of the informational effect of these pamphlets (Kalla and Broockman, 2018), then contrast harmonization could be violated by the differences in treatment-control comparisons. Measurement harmonization would be violated if one site used voter file (behavioral) turnout measures while another site relied on self-reported turnout in a post-election survey, given well-known concerns about overstatement of turnout in surveys (Burden, 2000).

Divergent and External Validity

We next consider the relationship between distinct measurement strategies.

Definition 6. ***Divergent validity*** *holds between measurement strategies $m \in M$ and $m' \in M$ if*

$$\tau_m(\omega', \omega'' \mid \theta) \neq \tau_{m'}(\omega', \omega'' \mid \theta), \tag{2}$$

when $m \neq m'$ at almost every setting, $\theta \in \Theta$, and almost every contrast, $(\omega', \omega'') \in \mathcal{C}$.

[11] See also Barrett (2021) for a discussion of measurement.

This implies that two distinct measurement strategies, m and m', do not produce the same treatment effect for a fixed contrast and setting (almost everywhere). The importance of divergent validity can be illustrated by considering two distinct measurement strategies that *do not* satisfy divergent validity, and thus produce indistinguishable treatment effects (substantively, not statistically) for some nontrivial set of settings or contrasts (i.e. on a set of positive measure). In such cases, the analyst would have to know precisely when the two measurement strategies are distinguishable and when they are not, i.e. when setting-contrast pairs are distinguishable relative to different measurement strategies. Otherwise, measurement strategies become conflated in an unknown and unpredictable way with strictly positive probability. A similar property holds (locally) for contrasts because the treatment effect mapping's derivative has full rank for almost all contrasts.

Definition 7. *A mechanism has **external validity** from setting θ to setting θ' if for 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 across almost all settings $\theta \in \Theta$.*

Importantly, our agnostic approach means mechanisms manifest through their observable effects, and this perspective is reflected in our formalization of external validity. In our analysis, we assume that divergent validity holds for all measurement strategies in M. We do this because we are interested in addressing *when and how an externally valid mechanism can be analyzed and incorporated into a meta-analysis*.

Returning to our running example, divergent validity requires that different measurement strategies produce distinct treatment effects. The informational mechanism we propose may affect multiple outcomes. For example, the treatment may affect our primary outcome, turnout, but it should also increase voter awareness that the election is being held, which could be measured using a survey. Here we would expect the treatment effects to be different. In principle, a voter must

know about an election in order to vote, however, many individuals abstain from voting even if they know about the election. This suggests that the treatment should have different measured effects on these two outcomes, satisfying divergent validity. In contrast, external validity of our informational mechanism corresponds to the theoretical expectation that if we were to implement the same experiment by harmonizing both the contrast and the outcome measure in a different election, we expect identical treatment effects (here, CATEs) in both sites, up to issues of sampling and estimation.

Results on Target-equivalence

Our definition of target-equivalence (Definition 4) captures a key assumption of most meta-analyses. Specifically, in order to ensure target-equivalence across constituent studies, it must be that any differences in the observed effect of a mechanism reflect statistical issues of sampling or estimation. We focus on measurement strategies that have divergent validity to focus specifically on the importance of harmonization in establishing target-equivalence between studies.

To isolate the importance of measurement harmonization, absent other potential issues, we begin our analysis by focusing on measurement strategies, i.e., different elements of M, while holding fixed a particular contrast, (ω', ω''), thus assuming that studies are contrast harmonized.

Theorem 1. *Let studies $\mathcal{E}_1$ and $\mathcal{E}_2$ be contrast harmonized, and the mechanism be externally valid, then $\mathcal{E}_1$ and $\mathcal{E}_2$ are target-equivalent if and only if they are measurement harmonized.*

This result shows that, under external validity, when two studies are contrast harmonized, measurement harmonization is both necessary and sufficient for two studies to be target-equivalent. The proof of Theorem 1 is in the appendix, and shows that when two studies are target-equivalent, either their measurement strategies cannot satisfy divergent validity, or their measurement strategies are exactly the same. Since we restrict attention to measurement strategies that satisfy divergent validity, target-equivalence of two studies necessarily requires measurement harmonization.

Divergent validity is a key piece of the argument for measurement harmonization in Theorem 1. If two measurement strategies do not have divergent validity, then there exists a set of contrast-setting pairs (with positive measure) where those two measurement strategies produce the same treatment effect, and are thus indistinguishable. There also exists a set of contrast-setting pairs (also with positive measure) where the measurement strategies produce different treatment effects. Unless the analyst knows the boundaries of these sets exactly, then she can never know (or even bound) whether differences are due to measurement inconsistencies, or due to substantive differences that she might be trying to detect.

From an empirical perspective, if two studies that are not measurement harmonized yield similar estimates of a substantive effect, i.e. one cannot distinguish the effect across two studies, then this evidence suggests either that the measurement strategies lack divergent validity or the mechanism is not externally valid.[12] Going back to our get-out-the-vote example, Theorem 1 suggests that combining non-harmonized studies, some measuring turnout while others measure awareness of the election through a survey, will not necessarily find consistent evidence of the informational role of pamphlets *even when that substantive mechanism is, in fact, present in all the settings and is externally valid*.

Following the same approach as before, to isolate the importance of contrast harmonization, we next focus on the case where measurement strategies between studies are harmonized.

Theorem 2. *Let studies $\mathcal{E}_1$ and $\mathcal{E}_2$ be measurement harmonized, and the mechanism be externally valid, then $\mathcal{E}_1$ and $\mathcal{E}_2$ are target-equivalent if and only if they are contrast harmonized almost everywhere.*

The proof establishing Theorem 2 is in the appendix and proceeds over two steps. First, we show that by combining external validity with target-equivalence, at two different contrasts (ω_1', ω_1'') and (ω_2', ω_2''), the treatment effect at either contrast must be the same in any fixed setting.

[12] This does not include the inability to *statistically* distinguish between them.

Second, the proof addresses how "large" the set of contrasts can be that produce the same treatment effect in a fixed setting, and shows that such a set is "small." To be more precise, we show that the set of contrasts that produce the same treatment effect in a single setting is small by showing that its dimension is smaller than the dimension of the set of contrasts, and hence, constitutes a measure zero subset of the set of contrasts, $\mathcal{C}$.[13]

Suppose that the respective treatments in two of the get-out-the-vote experiments vary along two dimensions: the timing of pamphlet distribution (relative to the election) and the attractiveness of the pamphlet. Here, we may expect that a pamphlet arriving immediately prior to the election may activate the informational mechanism more strongly. Moreover, an attractive pamphlet may garner more attention from recipients, increasing the share of treatment household members that learn about the upcoming election. Under either of these differences in the treatment instruments—and thus the contrast—Theorem 2 suggests that the effect on turnout would be different. The proof of Theorem 2 in this example essentially looks for the likelihood that the attractiveness of the pamphlet happens to exactly offset differences in the timing of pamplet distribution. Theorem 2 implies that this precise offsetting is very unlikely, specifically, has probability zero.[14]

That the treatment effect function is responsive to changes in the contrast is key to establishing Theorem 2. It reflects the substantive relationship between a contrast, i.e. the comparison of interest, and the treatment effect. One way that this property might fail is if potential outcomes were not responsive to different values of an instrument, meaning the treatment effect is independent of the contrast. Such a feature is typically ruled out, for example, by supposing a first-stage relationship.

Recalling that in our framework we reserve the term harmonization for studies that are both measurement and contrast harmonized, our main result is:

[13]Specifically, a set's dimension refers to the number of coordinates in Euclidean space needed to identify an element of the set.

[14]Harmonization of treatment would require some standardization of the amount of learning from the informational pamphlets. In practice, this is quite difficult to do when multiple attributes of the treatment instrument vary.

Theorem 3. *A meta-study, $\mathcal{M}(\mathcal{I})$, is target-equivalent if and only if the mechanism is externally valid and every study, $i \in \mathcal{I}$, is harmonized, i.e. measurement and contrast harmonized.*

Proof. The first part follows by noticing that under harmonization, target-equivalence and external validity are equivalent. The rest follows by combining Theorems 1 and 2. □

This result elucidates the conditions under which combining treatment effects from constituent studies leads to valid and interpretable conclusions—because they aim at the same empirical target. Our results suggest that practitioners of meta-analysis who want to remain agnostic about structural features of the mechanism need to focus on harmonization, thereby minimizing differences between contrasts and measurement strategies, to ensure target-equivalence. In conjunction with external validity, which ultimately relies on a substantive argument, harmonization ensures that any observed differences in constituent study estimands can be attributed to sampling idiosyncracies, random error, or other estimation issues.

Before moving on, we briefly consider what might arise when contrasts are only partially harmonized, meaning that there are two contrasts, (ω_1', ω'') and (ω_2', ω''), where $\omega_1' \neq \omega_2'$. This is relevant in cases where care has been taken to harmonize treatment arms across studies, but where less care has been devoted to harmonizing control arms.

Corollary 1. *Consider two studies, $\mathcal{E}_1$ and $\mathcal{E}_2$, that are measurement harmonized, where the mechanism is externally valid, and $\omega_1'' = \omega_2'' = \omega''$, then $\mathcal{E}_1$ and $\mathcal{E}_2$ are target-equivalent if and only if $\omega_1' = \omega_2'$.*

This result follows directly from Theorem 2, and we provide additional details in the appendix. We present Corollary 1 separately for two reasons. First, to better stress the importance of contrast harmonization, this result establishes that harmonizing both parts of the contrast—control and treatment—is critical. Second, the Metaketa initiative has been an important proponent of harmonization between different constituent studies. However, these projects focus attention on

treatment harmonization, corresponding to the kind of partial harmonization exhibited in Corollary 1. We show that just as much attention needs to be devoted to ensuring harmonization for the control arm across constituent studies. In field-based studies, like the Metaketas, that compare treatment to a "pure control" condition, full contrast harmonization may not be possible, a theme we revisit below.

Thus far, we have focused our analysis on the theoretical foundations of meta-analysis and have intentionally kept our framework abstract. This abstraction yields two benefits. First, we ensure that our results have more to do with the conceptual foundations of external validity and meta-analysis, and will thus apply across several different meta-analysis designs. Second, as mentioned previously, by abstracting from statistical concerns, we show that our results do not follow from issues of estimation or sampling, and thus, cannot be solved solely with statistical techniques, at least not without invoking a larger set of theoretical assumptions than is currently articulated. An important implication of this point is that treating some of the conceptual concerns we highlight with statistical methods essentially *assumes that those theoretical problems are estimation problems*.

Practical Advice for Meta-Analyses

We now apply our framework to a common statistical model used for meta-analysis, the random-effects model, which posits an explicit model of the data and seeks to identify a common parameter (e.g., a population average treatment effect). Observation of this parameter is distorted by two sources of random error, one emerging from sampling constituent studies from a population, and the other from sampling populations from a grand population (Field and Gillett, 2010: pg. 672). For the purposes of this discussion, we continue to assume that all constituent studies are internally valid.

The random-effects model follows by supposing that each study, indexed by i, produces one estimate of a treatment effect, $t^i_{m_j}(\omega'_j, \omega''_j)$; which, to keep ideas straight, depends on an outcome

measure, $m_j \in M$, and contrast, $(\omega_j', \omega_j'') \in \mathcal{C}$. Now consider the following structural model of the environment where studies take place:

$$\begin{aligned} \beta^i_{m_j}(\omega_j', \omega_j'') =& \lambda_i + \varepsilon_i \\ \lambda_i =& t^i_{m_j}(\omega_j', \omega_j'') + u_i \end{aligned} \tag{3}$$

where $\beta^i_{m_j}(\omega_j', \omega_j'')$ is the estimated treatment effect in study i, λ_i are study-specific parameters, and ε_i is study-specific estimation error. The random-effects model follows after making two crucial assumptions.

Assumption 1 (Statistical Structure)**.** *The study-specific error, ε_i, is drawn across i from a normal distribution with mean 0 and variance σ_i^2, and mean-level random error, u_i, is drawn across i from a normal distribution with mean 0 and variance υ^2.*

Differences between the observed effect across studies is formally articulated in the random-effects model as random error that may be generated by sampling variability, chance imbalance in the assignment of the instruments, or random measurement error, denoted by ε_i. The u_i term allows for differences in the means between studies due to random error. Assumption 1 outlines the statistical components of the random-effects model, that are not present in our framework, which is focused on theoretical issues.

The second assumption that is made in the random-effects model is a theoretical assumption, which is necessary for identification.

Assumption 2 (Design invariance)**.** *There is an underlying structural parameter, $\mu \in \mathbb{R}$, that is constant across studies:*

$$\mu \equiv t^i_{m_j}(\omega_j', \omega_j'') \quad \textit{for all } i.$$

Combining Assumptions 1 and 2 makes (3) into a random-effects model that formulates the treatment effects from constituent studies as *reduced-form* estimates that are stochastically related

to a common treatment effect, μ. It outlines the structural relationship between the the common treatment effect, μ, and $\beta^i_{m_j}(\omega'_j, \omega''_j)$, the study-level treatment effects. The fixed- and random-effects models are distinguished only by statistical assumptions, i.e., Assumption 1, since the fixed-effects model also assumes design invariance.[15] Consequently, the following results also apply to the fixed-effects model.

In the random-effects model, the true site-level treatment effects are different, but are related to a common treatment effect, μ, and the goal is to estimate this structural parameter. Identification in the random-effects model thus depends on the existence of μ, which unites studies across i in (3). We now consider two ways this can come about.

Proposition 1. *Suppose that* (3) *is the correct model of constituent studies, and that Assumptions 1 and 2 hold, then the common treatment effect, μ, is identifiable.*

The proof follows from standard statistical arguments. Proposition 1 is implicitly invoked when using the random-effects model to conduct a meta-analysis aiming to estimate μ. Design invariance corresponds exactly to one of the key concepts developed in our framework above: target-equivalence (Definition 4). Importantly, positing the existence of μ implicitly assumes external validity by invoking Assumption 2. Consequently, using a fixed- or random-effects model to *estimate a common treatment effect does not establish whether a mechanism has external validity*, since both models assume external validity for identification of that effect.

Just how strong is design invariance? and what does it imply about the kinds of mechanisms being studied in random-effects meta-analysis? To explore this, we return to our framework, where Assumption 2 would imply that the treatment effect function, $\tau_m(\omega', \omega'' \mid \theta)$, is constant in measurement strategies, m, contrasts, (ω', ω''), and settings, θ. With design invariance, Theorem 3 is vacuously true (i.e., true by construction), because target-equivalence is assumed and not derived. This kind of restriction is extremely strong since it supposes that the common treatment effect,

[15] The fixed-effects model follows by assuming $E[u_i] = 0$, $Var[u_i] = 0$, and $E[\varepsilon_i] = 0$.

μ, is the same regardless of how it is measured or what comparisons are being made. It does not hold in most drug trials, for instance, if a drug's effect depends on dosage, or if assessing its effect depends on the outcome measure used. For example, the effect of a cholesterol drug is unlikely to be the same if measuring cholesterol via a blood test vs an indicator of heart disease. Absent a substantive argument for design invariance, we argue that Assumption 2 is too strong and rarely satisfied in practical applications. Design invariance relates to two key concepts developed by Egami and Hartman (2020: pg. 10-11): T-validity and Y-validity. These refer to when various treatments induce the same treatment effects and when treatment effects are the same for different outcome measures, respectively. Assuming that T-validity and Y-validity hold across studies is the same as invoking design invariance, which for the reasons above, we argue is overly restrictive.

Using our framework, we next show when the common treatment effect can be identified without invoking design invariance, which follows by focusing more on design features between studies, in particular, ensuring that studies are harmonized.

Proposition 2. *Suppose that* (3) *is the correct model of constituent studies, the mechanism of interest has external validity, and Assumption 1 holds. If every constituent study is both contrast and measurement harmonized, i.e.* $m_j = m$ *and* $(\omega'_j, \omega''_j) = (\omega', \omega'')$ *for all* j*, then the common treatment effect,* $t_m(\omega', \omega'')$*, which depends on* m *and* (ω', ω'')*, is identifiable.*

Proof. Observe that

$$t^i_{m_j}(\omega'_j, \omega''_j) = \tau_m(\omega', \omega'' \mid \theta),$$

and combine (3) with Theorem 3. □

Identification of a common treatment effect relies on suppressing differences across i for the effect of interest, and harmonization ensures that such differences do not arise at the theoretical level, meaning that study-level differences are the result of random error, and hence, can be handled with common statistical models.

An implication of Proposition 2 is that the internal validity of constituent studies is a necessary, but not sufficient, condition for identification of a treatment effect from an externally valid mechanism. Our framework also highlights that there are many ways a mechanism can manifest rather than a unique common treatment effect. Treatment effects depend on the measurement strategies and contrasts used to observe and measure them, which, we argue, is a desirable feature of our framework. This key feature is implicitly assumed away under design invariance, because treatment effects are independent of measurement strategies and the comparison being made. This is exceptionally strong and—as we show—unnecessarily narrow.

A Harmonization Test. The interpretation of the u_i terms in (3), whose variance is υ^2, can be thought of as capturing (statistical) departures from different kinds of harmonization (assuming the mechanism is externally valid). Thus, a correctly specified statistical test of the null hypothesis $\upsilon = 0$ can provide evidence of a lack of harmonization, provided the treatment effect is correctly specified. An important caveat is that this interpretation of the test is predicated on the structural assumption that measurement strategies and contrasts are normally distributed over their respective supports. Consequently, a rejection of the null hypothesis that $\upsilon = 0$ does not imply that the random-effects model is correctly specified.

What if Harmonization is not Achievable? A structural model of cross-study relationships, that specifies the functional relationship between non-harmonized design features, can facilitate learning from a meta-analysis absent contrast or measurement harmonization. For example, suppose that the the analyst knows that treatment effects were increasing *linearly* in the difference, $\omega_j^{''} - \omega_j^{'}$, so that

$$t^i_{m_j}(\omega_j', \omega_j'') = \gamma_0 + \gamma_1(\omega_j'' - \omega_j'),$$

for some constants, γ_0 and γ_1. If studies are not contrast harmonized, so that $\omega_j'' - \omega_j'$ varies in j, then, assuming contrasts are measured on a common scale, γ_0 and γ_1 could be estimated using

off-the-shelf meta-regression estimators that employ the measured difference, $\omega_j^{''} - \omega_j^{'}$, as a right-hand-side covariate. Similarly, if constituent studies lack measurement harmonization, the analyst would need to specify the functional relationship between measures, e.g., specify a function, h, such that $m' = h(m)$, for every $m' \neq m$. The functional form of h will determine how easily existing statistical models can be adapted to address departures from measurment harmonization.

What if External Validity is Local? Suppose a mechanism is externally valid in only a subset of the settings for which data is available. This implies that despite harmonization, there will be at least two distinct treatment effects. In this case, analysts should identify the subsets of studies (or estimates) where external validity holds within those sets, and meta-analyze only within those sets of estimates.[16] This is the approach taken by Dunning et al. (2019) when analyzing separately the (conditional) average treatment effects on voters that observed "good news" and voters that observed "bad news" about incumbents.

When researchers suspect both violations of the assumption of external validity and failures of harmonization, the approaches we advocate can be combined, given a sufficient number of studies and overlap across design features. Early efforts in this vein include Blair, Christensen, and Rudkin (2021) and Godefroidt (2021). We note that accounting for harmonization failures or variation in the active mechanisms across settings will result in a far less "standardized" (or off-the-shelf) set of meta-analytic research designs than is current practice (e.g., Table 1).

Conclusion

Critics of the identification (or credibility) revolution in empirical social science regularly cite limited external validity as a primary weakness of these research designs (Deaton, 2010; Deaton and Cartwright, 2018). In response, practitioners like Imbens (2010) advocate that such issues should be addressed with meta-studies, and increasingly, empirical scholars have turned to meta-analysis as a potential tool to address these kinds of concerns. We develop a framework to understand

[16] Alternatively, one could use meta-regression to estimate common treatment effects for subsets of estimates for which different mechanisms are believed to be externally valid.

external validity and elucidate the potential role of meta-analysis in generalizing empirical findings. We present a number of results that highlight external validity, design harmonization, and target-equivalence. Specifically, we show that external validity and harmonization are necessary and sufficient for target-equivalence. Our framework thus complements empirical frameworks that focus on internal validity and estimation within single studies.

Our results highlight the dangers of conflating conceptual differences across studies with statistical sources of variation (i.e., sampling variablility). Although such statistical concerns are important and need to be accounted for in any meta-analysis, there remain important conceptual issues that can arise when comparing or aggregating estimates which take information from different sources. Such conceptual issues, when not addressed or discussed, can lead to misleading inferences about the influence or generality of a causal mechanism. Our results suggest that prior to conducting a study, more effort should be devoted to ensuring that constituent studies are harmonized. Absent harmonization, explicitly modeling the structure of the relationships between study designs offers a promising path forward. While such an approach typically invokes many assumptions–both theoretical and empirical—we have shown that strong assumptions are often made in meta-analyses, without an explicit articulation or analysis.

While the embrace of "barefoot" or agnostic approaches to the study of causality has improved the credibility of empirical findings in the social sciences, the promise of such approaches in meta-analysis is less straightforward. Our results suggest that more attention into design-based strategies of harmonization—both of contrasts and measures—is critical for improving the credibility and interpretability of the evidence presented in meta-analyses. At the same time, the conditions we identify, and the additional assumptions that are implicitly invoked by meta-study approaches, suggest there are underappreciated limits to the kind of agnosticism advocated by the credibility revolution as we seek to cumulate evidence.

References

Abramson, Scott F, Korhan Koçak, and Asya Magazinnik. 2019. "What do we learn about voter preferences from conjoint experiments?" *Mimeo* .
URL: `https://www.korhankocak.com/publication/cp/CP.pdf`

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

Aliprantis, Charalambos, D., and Kim C. Border. 2006. *Infinite Dimensional Analysis: A Hitchhiker's Guide*. 3 ed. Springer-Verlag Berlin.

Angrist, Joshua D, and Jörn-Steffen Pischke. 2008. *Mostly harmless econometrics: An empiricist's companion*. Princeton university press.

Angrist, Joshua D., and J orn-Steffen Pischke. 2010. "The Credibility Revolution in Empirical Economics: How Better Research Design Is Taking the Con out of Econometrics." *Journal of Economic Perspectives* 24 (2): 3–30.

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.

Barrett, Christopher B. 2021. "On design-based empirical research and its interpretation and ethicas in sustainability science." *Proceedings of the National Academy of Sciences* 118 (29): e2023343118.

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

Blair, Graeme, Alexander Coppock, and Margaret Moor. 2020. "When to Worry about Sensitivity Bias: A Social Referent Theory and Evidence from 30 Years of List Experiments." *American Political Science Review* 114 (4): 1297–1315.

Blair, Graeme, Darin Christensen, and Aaron Rudkin. 2021. "Do Commodity Price Shocks Cause Armed Conflict? A Meta-Analysis of Natural Experiments." *American Political Science Review* 115 (2): 1–8.

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

Blair, Graeme, Jeremy M. Weinstein, Fotini Christia, Eric Arias, Emile Badran, Robert A. Blair, Ali Cheema, Ahsan Farooqui, Thiemo Fetzer, Guy Grossman, Dotan Haim, Zulfiqar Hameed, Rebecca Hanson, Ali Hasanain, Dorothy Kronick, Benjamin S. Morse, Robert Muggah, Fatiq

Nadeem, Lily L. Tsai, Matthew Nanes, Tara Slough, Nico Ravanilla, Jacob N. Shapiro, Barbara Silva, Pedro C. L. Souza, and Anna M. Wilke. 2021. "Community policing does not build citizen trust in police or reduce crime in the Global South." *Science* 374 (6571): eabd3446.

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.

Burden, Barry. 2000. "Voter Turnout and the National Election Studies." *Political Analysis* 8 (4): 389–398.

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.

Cole, Stephen R., and Elizabeth A. Stuart. 2010. "Generalizing evidence from randomized clinical trials to target populations: The ACTG 320 trial." *American Journal of Epidemiology* 172 (1): 107–15.

Coppock, Alexander, Seth J. Hill, and Lynn Vavreck. 2020. "The small effects of political advertising are small regardless of context, message, sender, or receiver: Evidence from 59 real-time randomized experiments." *Science Advances* 6 (eabc4046): 1–6.

Cronbach, L. J., and Karen Shapiro. 1982. *Designing evaluations of educational and social programs*. Jossey-Bass.

de la O, Ana, Donald P. Green, Peter John, Rafael Goldszmidt, Anna-Katharina Lenz, Martin Valdivia, Cesar Zucco, Darin Christensen, Francisco Garfiras, Pablo Balán, Augustin Bergeron, Gabriel Tourek, Jonathan Weigel, Jessica Gottlieb, Adrienne LeBas, Janica Magat, Nonso Obikili, Jake Bowers, Nuole Chen, Christopher Grady, Matthew Winters, Nikhar Gaikwad, Gareth Nellis, Anjali Thomas, and Susan Hyde. 2021. "Fiscal Contracts? A Six-country Randoized Experiment on Transaction Costs, Public Services, and Taxation in Developing Countries." Working paper.
URL: `https://nikhargaikwad.com/resources/De-La-O-et-al_2021.pdf`

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, Guy Grossman, Macartan Humphreys, Susan D. Hyde, Craig McIntosh, and Gareth Nellis, eds. 2019. *Information, Accountability, and Cumulative Learning: Lessons from Metaketa I*. New York: Cambridge University Press.

Egami, Naoki, and Erin Hartman. 2020. "Elements of external validity: Framework, design, and analysis." *Design, and Analysis (June 30, 2020)* .

Eshima, Shusei, and Daniel M. Smith. 2022. "Just a Number? Voter Evaluations of Age in Candidate Choice Experiments." *Journal of Politics* 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*

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

Ferraro, Paul J., and Arun Agrawal. 2021. "Synthesizing evidence in sustainability science through harmonized experiments: Community monitoring in common pool resources." *Proceedings of the National Academy of Sciences* 118 (29): e2106489118.

Field, Andy P, and Raphael Gillett. 2010. "How to do a meta-analysis." *British Journal of Mathematical and Statistical Psychology* 63 (3): 665–694.

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

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

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

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

Gerber, Alan S., and Donald P. Green. 2019. *Get Out the Vote: How to Increase Voter Turnout*. Fourth ed. Washington DC: Brookings Institution Press.

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.

Godefroidt, Amélie. 2021. "How Terrorism Does (and Does Not) Affect Citizens' Political Attitudes: A Meta-Analysis." *American Journal of Political Science* forthcoming.
URL: *https://onlinelibrary.wiley.com/doi/10.1111/ajps.12692*

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

Hajnal, Zoltan L., and Paul G. Lewis. 2003. "Municipal Institutions and Voter Turnout in Local Elections." *Urban Affairs Review* 38 (5): 645–668.

Heckman, James J, and Edward Vytlacil. 2005. "Structural equations, treatment effects, and econometric policy evaluation 1." *Econometrica* 73 (3): 669–738.

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

Imbens, Guido W. 2010. "Better LATE than nothing: Some comments on Deaton (2009) and Heckman and Urzua (2009)." *Journal of Economic literature* 48 (2): 399–423.

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

Incerti, Trevor. 2020. "Corruption Information and Vote Share: A Meta-Analysis and Lessons for Experimental Design." *American Political Science Review* 114 (3): 761–774.

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

Kalla, Joshua L., and David E. Broockman. 2018. "The Minimal Persuasive Effects of Campaign Contact in General Elections: Evidence from 49 Field Experiments." *American Political Science Review* 112 (1): 148–166.

Kern, Holger L., Elizabeth A. Stuart, Jennifer Hill, and Donald P. Green. 2016. "Assessing methods for generalizing experimental impact estimates to target populations." *Journal of Research on Educational Effectiveness* 9: 103–127.

Kertzer, Joshua D. 2020. "Re-Assessing Elite-Public Gaps in Political Behavior." *American Journal of Political Science* forthcoming.
URL: *https://onlinelibrary.wiley.com/doi/10.1111/ajps.12583*

Latour, Bruno. 1993. *The pasteurization of France*. Harvard University Press.

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/*

Mutz, Diana C. 2011. *Population-based survey experiments*. Princeton University Press.

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.

Pritchett, Lant, and Justin Sandefur. 2015. "Learning from experiments when context matters." *American Economic Review* 105 (5): 471–75.

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

Schwarz, Susanne, and Alexander Coppock. 2022. "What Have We Learned about Gender from Candidate Choice Experiments? A Meta-Analysis of Sixty-Seven Factorial Survey Experiments." *The Journal of Politics* 84 (2).

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

Slough, Tara. 2022. "Phantom Counterfactuals." *American Journal of Political Science* forthcoming.
URL: *http://taraslough.com/assets/pdf/phantom_counterfactuals.pdf*

Slough, Tara, Daniel Rubenson, Ro'ee Levy, Francisco Alpizar Rodriguez, María Bernedo del Carpio, Mark T. Buntaine, Darin Christensen, Alicia Cooperman, Sabrina Eisenbarth, Paul J. Ferraro, Louis Graham, Alexandra C. Hardman, Jacob Kopas, Sasha McLarty, Anouk S. Rigterink, Cyrus Samii, Brigitte Seim, Johannes Urpelainen, and Bing Zhang. 2021. "Adoption of Community Monitoring Improves Common Pool Resource Management Across Contexts." *Proceedings of the National Academy of Sciences* 10.1073: 1–10.

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

Tomasi, Arduino. 2020. "The Stakes and Informativeness Trade-Off: Electoral Incentives to Implement Programmatic Transfers." *Mimeo* .
URL: `https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3646289`

Wilke, Anna, and Macartan Humphreys. 2020. "Field experiments, theory, and external validity." In *SAGE Handbook of Research Methods in Political Science and International Relations*. SAGE London pp. 1007–35.