External Validity and Meta-Studies*

Tara Slough† Scott A. Tyson‡

August 27, 2021

Abstract

Meta-studies compare or combine estimates from multiple studies to test for differences in estimates between studies, or more commonly, to generate aggregate estimates of the findings across studies. Social scientists increasingly turn to these designs in order to aggregate evidence and learn about the generality of substantive phenomena and mechanisms. We develop a framework to examine the conceptual foundations of meta-studies, with emphasis on clarifying the role of external validity in meta-studies. We identify the conditions under which multiple studies identify *comparable* quantities of interest, highlighting the importance of several forms of design harmonization between the studies that comprise a meta-study. We apply our results to common formulations of meta-analysis that are increasingly employed in social science—fixed and random effects models—providing design-based identification results for each model. Our results reveal limits to agnostic approaches to the combination of causal evidence from multiple studies.

*We thank Peter Aronow, Neal Beck, Alex Coppock, Chris Fariss, Dorothy Kronick, Winston Lin, Kevin Munger, Mark Ratkovic, Cyrus Samii, Jessica Sun, panel participants at PolMeth XXXVIII and the Virtual Formal Theory Seminar, and workshop participants at EGAP.

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

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

1 Introduction

The identification revolution has transformed the way social scientists approach empirical research by advancing the use of research designs that credibly identify the effects of causal mechanisms (Angrist and Pischke, 2010; Samii, 2016). While it is natural to inquire how widespread the effects of a substantive mechanism may be, identification-driven empirical research is often not meant to establish whether causal effects generalize across different settings or time periods. An important critique of identification-driven studies stresses researchers' inability to know whether similar findings apply beyond the scope of an individual study or setting (Heckman, 2000; Deaton, 2010).

A common response to these critiques advocates for *meta-studies*, which is a research design that compares or combines the results from multiple studies conducted on different samples, in different contexts, and potentially at different times. Meta-studies are often posited as an "agnostic" or "barefoot" approach to probing the generalizability of empirical findings without invoking strong assumptions or relying on implausible statistical models (Rosenthal, 1986; Dunning et al., 2019; Blair and McClendon, 2021). However, because the studies that make up a meta-study may not be identical across contexts, there are a number of conceptual and statistical challenges that arise when trying to combine or compare different estimands. Because some aspects of study design—the treatments and outcome measures—are (to some degree) within researchers' control, it is important to understand how these design elements contribute to the potential viability of a meta-study.

In this article, we develop a framework to elucidate some concepts and properties of meta-studies in the social sciences. 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, including temporal context (Lovett and Munger, 2019). The set of settings outline the *scope conditions* of an argument or theory, i.e., those 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). It is natural to think of a contrast as representing a treatment and control comparison. Third, a study has a *measurement strategy*, which specifies the outcomes of interest and how they will be measured (Adcock and Collier, 2001; Fariss, 2014). Measurement strategies are connected to an underlying mechanism because they measure its observed effect.

We formalize, and highlight the importance of, two types of measurement validity when comparing or combining effects across studies: convergent and divergent validity (Campbell and Fiske, 1959; Shadish, Cook, and Campbell, 2002). First, a measurement strategy has *convergent validity* if it measures "no effect" when a mechanism is indeed absent, thus linking a measurement strategy, the estimand it produces, and the underlying causal mechanism. Second, and critical for meta-analysis, two measurement strategies have *divergent validity* when they can be distinguished, i.e. when they produce different treatment effects in the same setting and at the same contrast (in a single study). Moving to meta-studies, divergent validity in constituent studies ensures that differences between their effects is the result of different substantive mechanisms and not artifacts of differences in their measurement strategies.[1]

Our framework organizes different formulations of a mechanism's generalizability, which occupies a central role in meta-studies. We say 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 it produces systematically different effects across different settings when all other aspects of the study are identical (including the mechanism).

We focus on the *comparability* of constituent studies, which is when the estimands of constituent studies refer to a common substantive (theoretical) quantity, and are thus focused on

[1]We note that two measurement strategies that produce exactly the same treatment effect are equivalent, and not distinguished within our framework.

measuring the same underlying mechanism.[2] This is required to ensure that whatever conclusions are drawn from a meta-study comprised of those constituent studies are meaningful and interpretable. Our results stress the importance of design *harmonization*—of contrasts and measurement strategies—-between studies. Our first result shows that the estimands from constituent studies are comparable if and only if the studies are measurement harmonized, which is when the outcome of interest is the same and it is measured in the same way. Our second result shows that the estimands for two constituent studies are comparable if and only if the studies are contrast harmonized, which is where the substantive comparison across studies is the same.[3]

Our two main results, taken together, show that combining constituent studies which have not been both measurement and contrast harmonized will not necessarily find consistent evidence regarding an externally valid substantive mechanism *even when that substantive mechanism is, in fact, present in all the settings comprising the meta-analysis*. Our results imply that a meta-analysis comprised of nonharmonized studies may also find consistent evidence of a substantive mechanism—even when that mechanism is in fact absent. Contrast and measurement harmonization are important because they are typically the most amenable to researcher control, either through design of prospective studies or through inclusion criteria in retrospective studies.

We conclude by applying our results to the most common manifestations of meta-analyses, fixed- and random-effects meta-analysis estimators, and discuss how they constitute a structural approach to external validity (Koopmans and Reiersol, 1950; Goldberger, 1972). Specifically, in each of these models, the quantity of interest is a structural parameter that is common across all settings, and therefore the underlying mechanism is externally valid by construction. The structural approach has a number of desirable properties, namely, by making more explicit the theoretical structure between studies, further development of structural models that combine evidence from

[2]Izzo, Dewan, and Wolton (2020) use a similar notion of comparability in the context of field experiments on political accountability.

[3]To be precise this result holds *almost everywhere*, meaning on a set of full measure (e.g., with probability 1).

multiple studies can unpack various conceptual issues when the stringent design conditions that we highlight for the comparability of studies are not necessarily met. The cost of this approach is that reduces many of the conceptual issues we highlight to *estimation* problems (by assumption).

The framework and results we present are intentionally designed to articulate the conceptual foundations of meta-studies in the social sciences. We treat external validity as a theoretical property, associated with a mechanism (c.f., Gailmard, 2021), and thus it must be invoked or substantively justified in order to compare or combine treatment effects across different study settings. While statistical considerations can be nested into our framework, our goal is to articulate and stress other aspects of external validity and their implications for meta-studies. 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), the design of field experiments (Chassang, Miquel, and Snowberg, 2012; Izzo, Dewan, and Wolton, 2020), and the decision-theoretic foundations of experiments (Banerjee, Chassang, and Snowberg, 2017; Banerjee et al., 2020).

We illustrate the applicability of our framework by considering three applications. In the first, because of its use as a pedagogical example (e.g., Gerber and Green, 2012), we discuss the early experiment reported in James Lind's (1753) *A Treatise on the Scurvy*. In 1747, Lind (working as a naval surgeon) tested six proposed treatments against scurvy on twelve ailing ("scurvied") seamen. One of the treatments—lemons and oranges—led to rapid improvements in subjects' condition. While Lind (1753) did not identify the exact mechanism leading to this effect (Bartholomew, 2002), the study is famous for the (ex post) clarity of the Vitamin C mechanism. The second application draws from recent survey experiments that examine the effects of information on belief formation (or updating). In these experiments, participants are randomly assigned to consume information on a policy issue (treatment) or not (control). Afterward, these surveys elicit respondents' beliefs or opinions about the policy. Multiple mechanisms underlying the process of belief formation have been proposed (i.e., Lord, Ross, and Lepper, 1979; Nyhan and Reifler, 2010; Ger-

ber and Green, 1999; Coppock, 2021), and we focus on motivated reasoning. The third application considers the economic and political effects of conditional cash transfers (CCTs), which pay subsidies to families (often mothers) conditional on their childrens' school attendance and/or compliance with healthcare requirements (vaccines, primary care visits, etc.). Scholars have focused on the economic and political effects of these programs (Ana, 2012; Imai, King, and Rivera, 2020; Zucco and Power, 2006).[4]

2 External Validity

The most common approaches to external validity (see below) come from how scholars think about experimentation, and

> "Both critics and apologists of experimentation in the social sciences often focus on the statistical representativeness of the experimental population as if it were the most relevant problem to be solved. It is, in fact, just one aspect of the external validity issue—which is in reality much more complicated than that." Guala (2005: p. 145) .

The ways people think about external validity can, broadly, be distinguished between an *inductive view*, where external validity is a property of a single study, and a *deductive view*, where external validity is a property of a substantive mechanism.

The Inductive Viewpoint. Some scholars view external validity as a property of a single study, which fall into (at least) four categories. The first focuses on the *extrapolation* of empirical results over variation in settings, units (sample), outcomes, or treatments (Shadish, Cook, and Campbell, 2002). The most practical conceptualization 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. Second, *transportability* (Pearl and Bareinboim, 2011, 2014) aims to generalize a study's findings from one context to another, by a "transport formula" which uses all the observed differences between two

[4]One of the first national CCT programs, PROGRESA, was subject to a randomized impact evaluation in Mexico during its initial rollout between 1997-2000 (Skoufias, 2001).

settings to reweight experimental findings and impute causal effects (which is essentially "selection on observables").

The third posits a *grand population* where the underlying phenomenon is present and manifests as a parameter. Because of sampling, physical, or temporal constraints, the grand population becomes subdivided into various subpopulations, and the analyst endeavors to estimate a population-level estimand from the observed sample or subpopulation (Gerber and Green, 2012). Critically, this reduces external validity to an *estimation* problem, and the question becomes whether the estimated treatment effect on the sample was an artifact of sample idiosyncrasies or the estimator employed. Various estimators have been developed to move from sample estimands to grand population parameters (i.e., Cole and Stuart, 2010; Kern et al., 2016), and recent expositions of this view describe forms of "external validity bias," which is the difference between sample estimates and population parameters (Egami and Hartman, 2020; Findley, Kikuta, and Denly, 2021).

The fourth, *parallelism*, which is not entirely inductive, focuses on whether empirical findings which are measured in an artificial setting (like a laboratory) extend to natural settings (Smith, 1982). Latour (1993) advocates for "extreme localism" arguing that manipulating natural phenomena are ineluctably non-generalizable (Guala, 2003). 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 in an effort to bridge the gap between experiments and observational studies in the case of microfinance.

The Deductive Viewpoint. We formalize external validity as a relational property between studies, and thus, it is a characteristic of a cross-section of studies, which is either present in that cross-section or not. Thus, external validity is fundamentally a theoretical property, a point that has been argued by Lucas (2003) and Gailmard (2021), albeit without formulating definitions. Our relational view of external validity accords neatly with meta-studies as a specific kind of research design where researchers conduct multiple studies to measure treatment effects on different samples, typically in different settings, as opposed to relying on extrapolation from a single study. It is

important to emphasize that our framework is conceptual—not statistical—and we have intentionally omitted statistical concerns (e.g., sampling) to better focus our analysis.

3 Meta-Analysis in Social Science

In current practice, social scientists gravitate toward two meta-study designs: meta-analyses and replication studies. Meta-analyses *combine* or synthesize the estimates from constituent studies, whereas replication studies *compare* the findings from multiple studies in a specific way. In this paper, we apply our framework to meta-analyses, and in a companion paper (Slough and Tyson, 2021), extend our framework to study conceptual replications, the modal form of replication study in the social sciences (Collins, 1992; Schmidt, 2009), where the issues (and results) are distinct.

While meta-analyses originated with Glass (1976) and Rosenthal and Rubin (1982), their current popularity in political science is more recent, and is related to the rise of treatment-harmonized randomized controlled trials (RCTs). Meta-analysis of these treatment-harmonized RCTs is advocated as a design that promotes accumulation of evidence across studies while (arguably) invoking a small number of assumptions (Dunning et al., 2019). Such meta-studies are advanced by Banerjee et al. (2015) and the Metaketa initiative championed by Evidence in Governance and Politics.

In Table 1 we identify and classify the most recent meta-analyses in political science, including the four complete Metaketa projects, and published meta-analyses in three political science journals (*American Journal of Political Science*, *American Political Science Review*, and *The Journal of Politics*) as well as meta-analyses on political subjects in general science journals. In the panels of the table, we distinguish between *prospective* and *retrospective* meta-analyses. The treatment-harmonized RCTs constitute prospective meta-analyses since the constituent studies (or sites) were designed with an eye to formal synthesis across sites through a meta-analysis. In retrospective meta-analysis, researchers collect and synthesize estimates from a variety of existing studies. We identify one study, Kalla and Broockman (2018), which uses both approaches to synthesize existing experiments on persuasion while incorporating a number of new experiments. Finally, all

of the meta-analyses that we identify use random effects or fixed effects meta-analysis estimators, which we relate to our framework below. A rare exception to fixed-and random-effects approaches is Meager (2019), who develops a Bayesian hierarchical model to study the distributional effects of experimental microcredit expansion interventions.

We are not the first scholars to consider the conceptual problems that emerge when comparing different studies. Ferraro and Agrawal (2021) consider the challenges posed by variation in treatments across studies while Barrett (2021) asks how non-compliance with treatment assignment and limits to measurement harmonization may affect the interpretation of meta-analytic findings. Izzo, Dewan, and Wolton (2020) argue that when the underlying mechanism varies across constituent studies, they lack *comparability* and may not estimate a common quantity of interest; we define comparability similarly below.

4 Framework

A meta-study compares or combines related studies that were performed in different places, and potentially, at different times. In our analysis, we emphasize novel conceptual issues that arise when thinking about external validity, with particular emphasis on the context of meta-studies. To focus on considerations that are unique to the meta-studies context, we abstract from two common sets of concerns in empirical studies: (1) internal validity and identification, and (2) sampling and estimation. Both are important and have received fairly 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.

4.1 Mechanisms

A *causal mechanism* is a way of representing a collection of causal factors that, when jointly activated, lead to a measurable effect (Rosenbaum, 1984). Separate causal factors can be thought of as comprising different *molecular mechanisms*, each of which comprises a necessary but insufficient condition for an effect (Mackie, 1965). For a single effect, these different molecular mechanisms,

Study	Stated Meta-Analysis Motivation			Component study design		Estimator		
	Generalizability	Precision	Lit. Synthesis	Experimental	Observational	RE	FE	Other
A: 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: Meta-Analysis based on Secondary Analysis of Existing Studies								
Blair, Christensen, and Rudkin (2021)			✓		✓	✓	✓	
Blair, Coppock, and Moor (2020)			✓		✓†	✓		
Incerti (2020)			✓	✓		✓	✓	
Kertzer (2021)	✓		✓		✓‡	✓		
Schwarz and Coppock (2020)	✓		✓	✓		✓		
C: Meta-Analysis based on Secondary Analysis and Original Studies								
Kalla and Broockman (2018)	(✓)		✓	✓		✓		

Table 1: A compilation of recent meta-analyses in political science.
†: Blair, Coppock, and Moor (2020) conduct meta-analyses on the difference in prevalence of a sensitive behavior between list experiments and direct questions. The latter quantity is observational so we classify the difference as observational.
‡ Kertzer (2021) examine differences in treatment effects between mass and elite respondents. Because mass and elite status are not randomly (or as-if randomly) assigned, this difference is observational.

however they are combined (Cartwright, 1983: Ch. 3), make up the *molar mechanism* that is linked to the effect of empirical interest, and is what scholars typically refer to when they talk about identifying a mechanism (Shadish, Cook, and Campbell, 2002: pg. 10). A single study measures the effect of a molar mechanism (where it may be present) and a meta-study combines the findings from different studies that pertain to the same mechanism at play in multiple contexts.

The status of causal mechanisms and the distinction between molecular and molar causation, is ultimately a philosophical statement about how the world works. Our view of mechanisms builds upon longstanding arguments that "mechanical" connections determine how the observable world unfolds, and, as a consequence, these connections help to explain observed variation in the world (Peirce, 1892). The link between causation and a mechanist philosophy is perhaps best expressed by Hume: "the same cause always produces the same effect, and the same effect never arises but from the same cause," and "the difference in the effects of two resembling objects must proceed from that particular, in which they differ." (Hume, 1739-40 (2003: Section XV, (4) and (6)).

For our analysis of meta-studies we need only represent the presence of a molar substantive mechanism. We denote the presence of a molar mechanism by the indicator $\gamma \in \{0, 1\}$, where $\gamma = 1$ when the molar mechanism is present, and $\gamma = 0$ when the molar mechanism is absent.[5] For the remainder, we will use the term mechanism to refer to a molar mechanism relevant to the studies under consideration unless otherwise stated.

4.2 Building Blocks: Individual 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. 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 that the mechanism indicated by γ could potentially

[5]The mechanism indicator γ can be broken up into molecular components in a straighforward way. Specifically, a molar mechanism indicator can be written as $\gamma = \Pi_i g_i$, where g_i is an indicator that takes the value 1 whenever the molecular mechanism i is present and 0 otherwise.

arise, and thus be subject to empirical scrutiny.

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 influence of a manipulation (or treatment) on that outcome (experimentally, quasi-experimentally, etc). When applied to causal questions, this approach is consistent with the common view that causality is assessed by focusing on *measuring the effects of causes* (see Holland, 1986: pg. 945). To capture the essential features of this approach in our framework, we denote a (finite) set of **measurement strategies**, M, where different elements $m \in M$ can correspond to different outcome measures, different measurement scales, etc.[6] For example, one could measure the effect of a particular health policy by looking at how it affects hospitalizations, prevalence of preventable illness, mortality, etc., each of which would correspond to a different measurement strategy, i.e. a different m in the set 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 (after a scale-location transformation).

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

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 classifica-

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

tions 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 most formulations of the potential outcome framework handle treatment status with a binary 0-1 indicator, it is important to emphasize that this is a normalization, and does not generalize across different studies. Put differently, differences in treatment effects across studies are typically not the result of naming conventions.

We can now define what constitutes an individual study in our framework:

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 $\mathcal{M}(\mathcal{I}) = \{\mathcal{E}_i\}_{i \in \mathcal{I}}$.*

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

For a meta-study, each constituent study contributes one or more inputs. In experimental (or quasi-experimental) studies, the outputs of constituent studies are generally some form of causal estimand or treatment effect.

Definition 3. *For a study $\mathcal{E} = \{m, (\omega', \omega''), \theta\}$, a **treatment effect** is a smooth mapping (almost everywhere)*

$$T_m^{\gamma}(\omega', \omega'' \mid \theta) : \{0, 1\} \times M \times \mathcal{C} \times \Theta \to \mathbb{R}.$$

Moreover, the derivative of T has full rank for almost all contrasts.

Although we use the term treatment effect, the function $T_m^{\gamma}(\omega', \omega'' \mid \theta)$ can represent many different estimands depending on the application. To illustrate, consider a special case where the

treatment effect function takes the form:

$$T_m^{\gamma}(\omega'', \omega' \mid \theta) = f(Y_m(\omega'')|\mathcal{D}, \theta) - f(Y_m(\omega')|\mathcal{D}, \theta), \tag{1}$$

where $Y_m(\omega)$ are potential outcomes, and f is some operator (most commonly 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) conditional average treatment effects 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, depending on sample composition.

4.3 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."[7] Finally, note that assuming that T is smooth in contrasts almost everywhere is not particularly restrictive(Stein and Zygmund, 1964); unless the set of treatment effects is known to be a fractal.

Our framework assumes that all studies in a meta-study are internally valid, abstracting from issues of experimental design, identification, and analysis that are relevant to attaining credibility within individual studies (experiments, quasi-experiments, etc.). We make this assumption to focus

[7]We thus take for granted concerns of estimand identification and interpretation that depend on potential outcomes (see, e.g., Slough, 2020; Izzo, Dewan, and Wolton, 2020).

our analysis on issues that are unique to meta-studies. Indeed, many of the studies in Table 1 are quite explicit and careful to design or include only studies with plausible claims to internal validity. Relaxing internal validity in some studies is a separate conceptual problem and adds substantial complexity without alleviating problems of external validity or providing insights relevant to what we study.

We are certainly not the first to represent an empirical study as a theoretical object. Indeed, Shadish, Cook, and Campbell (2002: pg. 19), building on Cronbach (1982), think of a study as being comprised of four components: units, treatments, observations, and settings (UTOS). Similarly, Blair et al. (2019), building from King, Keohane, and Verba (1994), focus on four different components: a model, an inquiry, a data strategy, and an answer strategy (MIDA). Our notion of a study—which is comprised of a setting, a contrast, and a measurement strategy—overlaps with and complements both of these approaches. In our framework, contrasts essentially correspond to treatments in UTOS, and one piece of the model in MIDA; measurement strategies correspond to observations in UTOS and data strategies in MIDA; mechanisms constitute the other piece of models in MIDA but remain implicit in UTOS; and settings appear in UTOS but are implicit in MIDA. 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 omit—or abstract from—other things which are not critical.

4.4 Applications

Citrus and scurvy: The setting of Lind's 1753 experiment was the ship where the experiment was conducted in 1747 and the mechanism was subject-level Vitamin C absorption. The contrast in the (simplified) experiment is a treatment vs. control comparison where treatment consisted of lemon and orange consumption and control consisted of no citrus consumption, and his measurement strategy consisted of an indicator for the incidence of scurvy symptoms.

Because Lind's experimental population was conditioned upon presentation of scurvy symptoms, we expect $\mathcal{D}$ in (1) to include all experimental subjects. As such, $T_m^{\gamma}(\omega'', \omega' \mid \theta)$ becomes the average treatment effect (ATE). If, in a hypothetical second setting (a different ship), the experiment did not condition the sample on presentation of scurvy, $\mathcal{D}$ in (1) might similarly be defined as the subset of subjects presenting with scurvy symptoms at baseline. In this case, $T_m^{\gamma}(\omega'', \omega' \mid \theta)$ would be the conditional ATE (CATE) on that subgroup.

Information and belief formation: For concreteness, we will discuss one experiment by Guess and Coppock (2020); Coppock (2021), focusing on one informational treatment: informational videos about gun control, though similarly-oriented treatments in survey experiments are widespread. The setting of the experiment was a nationally-representative sample of 2,112 US residents in June 2016, about 10 days after the Pulse nightclub mass shooting in Orlando. They seek to test the motivated reasoning mechanism advanced by Lord, Ross, and Lepper (1979).[8] The experment included treatment conditions: an anti-gun control message, a placebo video message, and a pro-gun control message, and thus three different contrasts (pro-gun control vs placebo, anti-gun control vs. placebo, and pro- vs. anti-gun control), since treatment effects are defined as differences between two treatment arms. Coppock (2021)'s measurement strategy includes a binary indicator of support for gun control as the outcome variable.

Motivated reasoning suggests that information should have different effects on the opinions of ex ante supporters and opponents. Within our discussion of molar causation, motivated reasoning is comprised by two molar mechanisms, which differ in one molecular mechanism, ex ante supportive or opposed. For this reason, most analysts estimate treatment effects among ex ante supporters and ex ante opponents separately.

Conditional Cash Transfers: We focus on the first (and arguably) most famous CCT experiment: Mexico's PROGRESA (Skoufias, 2001). The setting of the experiment was rural Mexican

[8]Note that Guess and Coppock (2020) do not recover evidence consistent with motivated reasoning. Because ex ante consideration of the mechanism is critical in our framework, we proceed with the theory they sought to test.

localities (villages) in seven Mexican states between 1997 and 2000. Treatment was assigned at the locality level and consisted of the CCT program being delivered to means-tested eligible households contingent on school attendance and health visits. Given the absence of pre-existing social programs, similarly eligible households in control localities did not receive any contingent government transfers. The contrast is thus the operation of PROGRESA within a locality versus no operation of PROGRESA in the locality.

Discussions often focus on household consumption maximization as the mechanism which operates at the household level, but another, albeit under-studied, mechanism is local bureaucratic capacity, operating at the locality-level. A common measurement strategy for the former is school enrollment because parents are expected to enroll students in school in lieu of childrens' participation in the labor market when cash transfers exceed childrens' potential wages. Thus, this mechanism operates on schooling decisions for families whose children are otherwise not enrolled in school. As such, the set $\mathcal{D}$ in (1) includes households with children who were not enrolled in school at baseline, and $T_m^{\gamma}(\omega'', \omega' \mid \theta)$ is the CATE for this subset of households. With respect to the bureaucratic capacity mechanism, if we expect that changes in bureaucratic capacity increase citizen expectations of politicians, these effects might be measured using survey responses. In practice, we might expect that these changes are observed to all citizens in treatment clusters. If so, $\mathcal{D}$ in (1) includes all residents of experimental communities and $T_m^{\gamma}(\omega'', \omega' \mid \theta)$ is the average causal effect.

5 Concepts

Any empirical study is comprised of at least two phases—a design phase and an analysis phase (Morton and Williams, 2010). When designing and conducting a meta-study, $\mathcal{M}(\mathcal{I})$, achieving comparability of constituent studies generally represents a challenge for both the design and analysis phases. In the design phase, researchers aim to make constituent studies comparable, because this facilitates a more straightforward interpretation in the analysis phase, where various statisti-

cal techniques are used to combine data or estimates from multiple studies. In this section, we focus on the important concepts that are relevant for design decisions that researchers make when conducting meta-studies.

5.1 Convergent and Divergent Validity

We now define two important concepts—convergent and divergent validity—which we take from the literature on research design (Shadish, Cook, and Campbell, 2002) and measurement theory (Adcock and Collier, 2001; Fariss, 2014). Each kind of validity encapsulates an important aspect of the conceptual "match" between a measurement and the concept it is meant to capture. Consequently, both convergent and divergent validity are properties of measurement strategies.

Since a measurement strategy is key toidentifying the effect of a mechanism, a necessary feature of any valid measurement strategy is that it should not measure a mechanism's effect if the mechanism is, in fact, not present.

Definition 4. *A measurement strategy, m, has **convergent validity** if for almost every contrast (ω', ω'') and almost every setting θ,*

$$T_m^0(\omega', \omega'' \mid \theta) = 0.$$

Convergent validity is a property of a single measurement strategy, requiring that it detect a mechanism's presence when, and only when, the mechanism is present, and importantly, it applies in single studies. Convergent validity essentially ensures that there are no systematic type-I or type II-errors, although such errors could still arise from statistical issues.

Since we are concerned with meta-studies, we also need to consider the relationship between distinct measurement strategies.

Definition 5. ***Divergent validity** holds between measurement strategies $m \in M$ and $m' \in M$, if*

$$T_m^1(\omega', \omega'' \mid \theta) \neq T_{m'}^1(\omega', \omega'' \mid \theta), \tag{2}$$

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

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. 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 two measurement strategies. Otherwise, two measurement strategies become conflated in an unknown and unpredictable way on a set of positive measure. Note that a similar property holds (locally) for contrasts because the treatment effect mapping is smooth and its derivatives have full rank on $\mathcal{C}$.[9]

To illustrate, Figure 1 provides a visualization of treatment response functions for three different measures. The two instruments in this example are ω' and ω'', on the x-axis, and the y-axis gives the expected potential outcomes, denoted by $E[Y(\omega) \mid \theta]$, at different measurement strategies, m_1, m_2, and m_3, illustrating how the measurement strategy shapes the relationship between contrasts and treatment effects. As can be seen, measurement strategies m_1 and m_2 lack divergent validity, while m_3 exhibits divergent validity relative to both m_1 and m_2.

We now present our first result, which links convergent validity and the mechanism indicator, γ, and provides a characterization of the treatment effect function.

Lemma 1. *A measurement strategy, m, has convergent validity if and only if there exists a smooth function, $\tau_m(\omega', \omega'' \mid \theta) : \mathcal{C} \times \Theta \to \mathbb{R}$, such that*

$$T_m^{\gamma}(\omega', \omega'' \mid \theta) = \gamma \cdot \tau_m(\omega', \omega'' \mid \theta),$$

i.e., T is linear in the mechanism indicator, γ, for almost every setting θ, and almost every contrast

[9] Specifically, divergent validity is essentially the same as full rank of T's derivative.

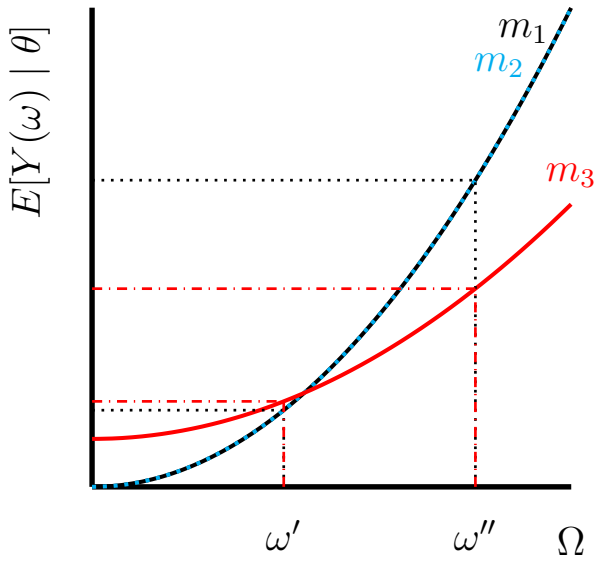

Figure 1: Each line represents the treatment response function of instrument ω for a different measurement strategy, m_1, m_2, or m_3. m_1 and m_2 do not exhibit divergent validity, however m_1 and m_3 do, as shown by the difference in treatment effects.

(ω', ω'').

The proof is in the Appendix. The molar mechanism corresponding to γ is present or it is not, and hence, the mechanism indicator only takes two values. This fact, in combination with convergent validity, implies that whenever $\gamma = 0$, the treatment effect, $T_m^{\gamma}(\omega', \omega'' \mid \theta)$, must also be zero for almost every contrast and setting. Then, the treatment effect function for a measurement strategy with convergent validity is flat in contrasts and settings when $\gamma = 0$, and critically, not when $\gamma = 1$. This implies that the treatment effect function can be written as a linear function of the mechanism indicator

A useful consequence of Lemma 1 is that it structures how the mechanism enters the remainder of the analysis. Specifically, we focus on the treatment effect function, τ, which is induced by a measurement strategy that has convergent validity. Since a failure of convergent or divergent validity would fundamentally undermine any meta-analysis, we proceed by considering only measurement strategies that have both convergent and divergent validity.

5.2 External Validity and Comparability

External validity is a term common in social science discourse, and it typically refers to whether a particular mechanism has the same effect in different settings. In this section we decouple two distinct concepts that are often conflated in discussions of external validity, but as we show, are conceptually distinct.

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. When two studies are conducted to address and measure the effect of the same mechanism, the core issue is whether the mechanism will yield the same evidence in different studies.

Definition 6. *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 **comparable** 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 **constituent comparability** if all constituent studies i in $\mathcal{M}(\mathcal{I})$ are comparable.*

Constituent comparability is the core concept underlying whether the treatment effects from individual constituent studies can be combined, since when it is satisfied, all empirical issues reduce to statistical issues, i.e. issues regarding estimation of the true effect. It is important to emphasize that when positing a (grand) population parameter to be estimated from constituent studies, which differ only by statistical noise, implicitly assumes constituent comparability.

Constituent comparability, and meta-study comparability, combines theoretical and empirical concepts. Consequently, it mixes together things that may be subject to empirical scrutiny from things that rely on substantive and theoretical arguments. To separate these different things, we distinguish external validity, which has to do with the theoretical generalizability of a mechanism across different settings, from comparability.

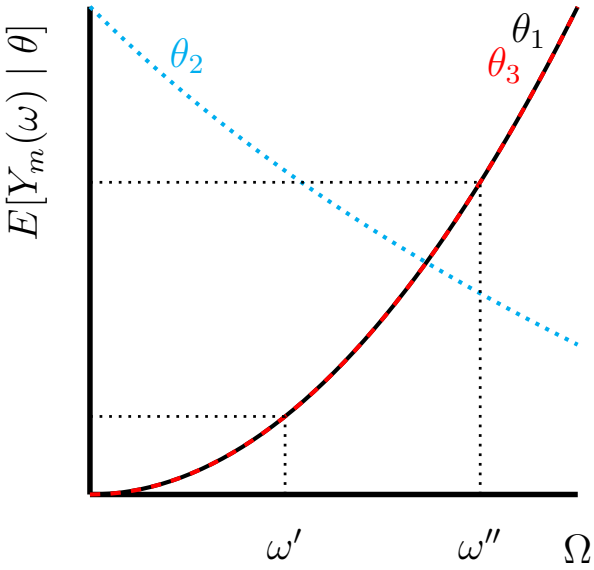

Figure 2: Each line gives the treatment response function in a different setting, θ_i. In this plot, the mechanism has external validity from setting θ_1 to θ_3; it does not have external validity from θ_1 to θ_2.

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 $\theta \in \Theta$.*

Definition 7 states precisely how external validity manifests in our framework. Figure 2 illustrates external validity in three cases. Each line gives the treatment response function for a different setting θ_i. The mechanism has external validity in settings θ_1 and θ_3 since the same instruments produce the same treatment effect in both settings. But the mechanism does not exhibit external validity between θ_2 and the other contexts, as the mechanism generates a qualitatively different relationship between contrasts and settings.

In our analysis we assume that convergent, divergent, and external validity all hold for all measurement strategies in M and for the mechanism of interest, indicated by γ. We do this because

we are interested in addressing *when and how an externally valid mechanism can be analyzed and incorporated into a meta-study*.

5.3 Harmonization

We consider two forms of *harmonization*, or design congruence between studies. There are a number of dimensions on which different studies can be harmonized, and for conceptual clarity, we present each separately.

Definition 8. *Studies $\mathcal{E}_1$ and $\mathcal{E}_2$ are **contrast harmonized** if $(\omega_1', \omega_1'') = (\omega_2', \omega_2'')$ in almost every setting.*

A contrast represents the comparison that leads to the measured treatment effect, which is usually between two different treatments, or equivalently, between a treatment and a control. Consequently, measuring the levels of both values of the instruments constituting a contrast, $(\omega', \omega'') \in \omega$, 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. Discussion of treatment harmonization in existing work, including the Metaketa studies, is relatively common.

Definition 9. *Studies $\mathcal{E}_1$ and $\mathcal{E}_2$ are **measurement harmonized** if $m_1 = m_2$ at almost every setting.*

A measurement strategy in our framework encapsulates all the considerations that go into measuring the effect of a contrast on an outcome of interest. 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. To illustrate measurement harmonization, consider a well-known challenge from conflict studies, where sometimes the effect of an instrument is best measured with conflict incidence, while in other cases the instrument's effect is best measured as conflict duration. Clearly, with perhaps only a few exceptions, such studies are not measurement

harmonized. In these examples, considerations regarding identification and estimation, within an individual study, may lead to a lack of measurement harmonization between different studies.[10]

Definition 10. *Studies $\mathcal{E}_1$ and $\mathcal{E}_2$ are **harmonized** if they are both contrast harmonized and measurement harmonized. A meta-study $\mathcal{M}(\mathcal{I})$ is harmonized if every constituent study is harmonized, i.e. every $\mathcal{E}_i$ is harmonized across $i \in \mathcal{I}$.*

We reserve the term harmonization for situations where both contrast harmonization and measurement harmonization are satisfied, that is, studies that use the same outcome, measure that outcome the same way, and employ the same comparison. We do this to keep different conceptual issues distinct, which helps focus our discussion. The Metaketa studies highlight harmonization of a common treatment as a important feature of prospective study design. We contend that these harmonization efforts do not necessarily achieve contrast harmonization because they do not endeavor to harmonize (or measure) the control instrument across sites.

5.4 Illustration of Concepts

We illustrate convergent, divergent, and external validity with respect to the Vitamin C in Lind's study, which importantly, included additional treatment arms, in particular, one that instructed ailing participants to drink vinegar. Since vinegar does not contain Vitamin C, there was no improvement among treated seamen (relative to nontreated seamen). This is consistent with our concept of convergent validity: there was no apparent treatment effect in the absence of the Vitamin C mechanism. To illustrate divergent validity, which requires that different measurement strategies produce distinct treatment effects, note that if Lind had measured survival of the voyage, then the success of the citrus treatment would have been different: scurvy was only one driver of mortality on ships in the 18th century.

Our assumption of external validity corresponds to the (theoretical) expectation that if Lind were to conduct the survey on a different ship or different voyage, if the contrast between the

[10]For a substantive discussion of empirical and identification concerns regarding the use of conflict onset vs conflict duration see Miguel, Satyanath, and Sergenti (2004).

citrus and control treatment conditions and measurement strategies were identicial, the treatment effect on both voyages would produce identical treatment effects (up to issues of sampling and estimation). How might harmonization between Lind's actual ship and the hypothetical ship fail? Contrast harmonization would be violated if, for example, one ship replaced the lemon in the lemon/orange treatment with a lime/orange treatment, since limes have half the Vitamin C per weight ratio. Measurement harmonization would be violated if, instead of measuring outcomes using an indicator for any symptom (the union of all symptoms), the second ship measures only an indicator for bloody gums (one among multiple symptoms of scurvy).

6 External Validity, Comparability, and Harmonization

In this section, we explore the relationship between external validity, comparability, and different forms of harmonization. We consider a mechanism that is externally valid and measurement strategies that have convergent and divergent validity.[11] We maintain these theoretical assumptions to focus specifically on the importance of harmonization in establishing a mechanism's comparability between different studies.

Our definition of constituent comparability (Definition 6) captures a key assumption of most meta-studies. Specifically, in order to ensure comparability across the constituent studies comprising a meta-study, it must be that any differences in the observed effect of a mechanism reflect statistical or estimation issues.

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, indicated by γ, be externally valid, then $\mathcal{E}_1$ and $\mathcal{E}_2$ are comparable if and only if they are measurement harmonized.*

[11] It is straightforward to show that our results do not rely on convergent validity, but a lack of convergent validity would also give rise to different problems.

The proof of Theorem 1 is straightforward and is contained in the appendix. This result shows that when two studies are contrast harmonized, then measurement harmonization is both necessary and sufficient for two studies to be comparable. The intuition is that when two studies are comparable, then either their measurement strategies are exactly the same, or their measurement strategies cannot satisfy divergent validity.

Divergent validity is a key piece of the argument for measurement harmonization in Theorem 1, so it is important to emphasize its importance. 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. Moreover, 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 whether differences are due to inconsistencies in measurement, or instead due to substantive differences that she is trying to detect.

Going back to our example from conflict studies, Theorem 1 suggests that combining studies, some measuring conflict onset while others measure conflict duration, will not necessarily find consistent evidence of a substantive mechanism *even when that substantive mechanism is in fact present in all the settings and produces the exact same effect (i.e. it is externally valid)*. Moreover, 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]

Moving on, we next address what happens when we relax contrast harmonization, but following the same approach as before, focus on the case where the measurement strategies between studies are harmonized, so that we isolate the importance of contrast harmonization absent other concerns.

Theorem 2. *Let studies $\mathcal{E}_1$ and $\mathcal{E}_2$ be measurement harmonized, and the mechanism, indicated by γ, be externally valid, then $\mathcal{E}_1$ and $\mathcal{E}_2$ are comparable if and only if they are contrast harmonized*

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

almost everywhere.

The proof establishing Theorem 2 is in the appendix and proceeds over two steps. The first step is key, where we show that by combining external validity with comparability across studies, 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, θ. The second step, then, 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]

That the treatment effect function is responsive to changes in the contrast is key to establishing Theorem 2, and moreover, it reflects the substantive relationship between a contrast, i.e. the comparison of interest, and the treatment effect, the measure of the causal effect. To illustrate what this feature of the treatment effect function means, consider one way that it might fail. If potential outcomes were not responsive to different values of an instrument, i.e. the effect of treatment doesn't depend on the treatment administered, this would translate to the treatment effect being independent of the contrast. In an individual study, such a feature is typically ruled out, for example, by supposing that there is a first-stage relationship in an instrumental variables setup.

Recalling that in our framework we reserve the term harmonization for studies that are both measurement and contrast harmonized, our analysis culminates in the following:

Theorem 3. *A meta-study, $\mathcal{M}(\mathcal{I})$, is constituent comparable if and only if the mechanism, indicated by γ, is externally valid, every measurement strategy m_i, where $i \in \mathcal{I}$, satisfies divergent validity, and every study is harmonized, i.e. measurement and contrast harmonized.*

Proof. Follows by combining Theorems 1 and 2. □

[13] Specifically, dimension refers to the number coordinates in Euclidean space that are needed to identify an element of the set, Guillemin and Pollack (see 1974).

This result elucidates when comparing or combining the treatment effects from constituent studies leads to valid and interpretable conclusions—provided one has taken care of estimation or sampling concerns. It is important to stress that the external validity of the mechanism of interest is a theoretical property, and so it manifests in empirical studies as an assumption, which cannot, in general, be assessed empirically—it relies ultimately on a substantive argument. The only thing that can be assessed empirically is the comparability across and between studies, which Theorem 3 shows is only meaningful whenever a meta-study's constituent studies are measurement and contrast harmonized. This implies that meta-studies should approach comparability by measuring external validity as directly as possible, meaning that effort should be devoted to minimizing as much as possible differences between contrasts and measurement strategies, thus ensuring that all differences are only reflective of the setting, rather than how outcomes are measured or the comparisons being made.

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 a studies treatment arm across studies, but where less care has been devoted to harmonizing the control arm.

Corollary 1. *Consider two studies, $\mathcal{E}_1$ and $\mathcal{E}_2$, that are measurement harmonized and where $\omega_1'' = \omega_2'' = \omega''$, then $\mathcal{E}_1$ and $\mathcal{E}_2$ are comparable 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, has focused most of its attention on treatment harmonization, which corresponds to the kind of partial harmonization exhibited in Corollary 1. Our results suggest that just as much attention needs to be devoted to ensuring harmonization for the

control condition 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 conceptual foundations of meta-studies. We have intentionally kept our framework abstract, which has kept our analysis somewhat distant from concrete aspects that scholars who conduct a meta-study usually confront. Consequently, our approach has two important consequences. First, by abstracting from a more concrete framework, we ensure that our results have more to do with the conceptual foundations of external validity and meta-analysis, and will thus be true across several different meta-study 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 without invoking a larger set of theoretical assumptions than is currently articulated. An important implication of this point, which we expand on, is that treating some of the conceptual concerns we highlight with statistical methods essentially *assumes that those conceptual problems are estimation problems*.

6.1 Applications

Citrus and scurvy: Had Lind (1753) sought to conduct his experiment on multiple ships, and then combine what he found, would his findings have been comparable? It depends. Lind didn't administer Vitamin C directly, but instead gave out lemons and oranges. If Lind had administered limes instead of lemons, then, because of differences in Vitamin C content between lemons and limes, Theorem 2 suggests that the effect on scurvy would be different. In fact, the argument of Theorem 8 in this example essentially looks for the likelihood that the oranges used on the lime-ship happen to exactly offset the Vitamin C deficiency of the limes, and finds that this event is excessively unlikely. Harmonization of treatment would require the standardization of the amount of citrus consumed by seamen in the treatment group on different ships, and examining the effect on

scurvied seaman through experimental sample selection (like Lind), or by examining conditional effects on the scurvied subgroup, to ensure a comparable "control" condition. By ensuring that we measure the treatment effect on subjects presenting with scurvy—either by conditioning the experimental sample or by estimating the CATE on that subgroup—helps to ensure harmonization of the control instrument. Theorem 1, suggests that had Lind measured other symptoms (like bloody gums), we should not necessarily expect to detect the same effect on individual symptoms as with an indicator for any symptom of scurvy. In this simple example, the importance of harmonization is straightforward: we only expect to see the effects of Vitamin C to be comparable if we look at the same contrast or if we examine the same measures.

Information and attitude formation: Our second example suggests that while the conditions identified by Theorem 1 are stringent, they arguably can be met through careful design in some social science experiments. While Guess and Coppock (2020) was not a meta-study, its design is easily transportable to other settings (populations or samples). Specifically, the contrasts in this experiment between any two of the three treatments (pro- or anti-gun control information or placebo videos) do not rely on a "status quo" pure control condition. This reduces concerns about the separability of the setting from the control instrument (a concern in many field-based meta-studies). As such, harmonized versions of these treatments could be assigned to different populations or samples. Similarly, the survey outcome measure—support for stricter gun control policies—could be implemented more broadly. This suggests that this experiment—and similar survey experiments testing the effects of information on attitude formation—may achieve comparability along with other (to date, hypothetical) studies.

Conditional Cash Transfers: Theorem 1 points to the challenge of achieving comparability in many field experimental meta-studies. First, when we study actual policies (the cash transfer programs) across multiple settings, there are limits to the harmonization of treatment instruments, here conditional cash transfer programs, as reported by Banerjee et al. (2017). In theory, one could design better harmonized programs, for example a CCT that makes transfers equivalent to

10% of household consumption or of an equivalent value of $75/month. However, more tedious harmonization of program attributes does not imply contrast harmonization when control conditions like household income vary, as shown by Gechter et al. (2019). When either instrument—the experimental program or the control condition—are not harmonized, the same externally valid mechanism will not necessarily produce comparable treatment effects across settings. Similarly, on school enrollment, administrative records may not define enrollment in the same way across contexts—even ones that adopt the same binary scale—may not be harmonized, undermining the comparability of site estimates. While researchers may have more discretion over the design of questions about expectations of government, ensuring that the questions measure a harmonized construct remains a challenge.

7 Concrete Manifestations of Meta-Analysis

Inspired by existing meta-analyses, we now apply our framework to some common ways meta-studies are conducted. For the purposes of this discussion, we will continue to assume that all constituent studies in a meta-study are internally valid, since such well-studied concerns are beyond our substantive scope.

In most meta-analyses, the (estimated) treatment effects from individual constituent studies are viewed as *reduced-form* estimates which are stochastically related to a population-level treatment effect. There are a few statistical approaches designed to measure the population-level treatment effect from the sample-level treatment effects from constituent studies, the two most common are fixed-effects and random-effects estimators. In both approaches, it is important to emphasize that external validity is assumed and a structural model for differences across settings is posited. This means that the dominant meta-analysis estimators are *structural* estimators.

7.1 Fixed-Effects Meta-Analysis

Suppose that each study, indexed by i (as in our theoretical framework), produces one estimate of an average treatment effect, relative to a given outcome measure, m. We denote this treatment

effect as $\widehat{t_i}$.[14] When these average treatment effects are combined in a meta-analysis, utilizing a fixed-effects model, two crucial theoretical assumptions are made. First, the existence of a true average treatment effect, denoted by μ, is assumed constant across studies, which corresponds to assuming external validity (Definition 7). Second, differences between the observed effect across studies can be modeled as idiosyncratic error that may be generated by sampling variability, chance imbalance in the random assignment of the instruments, or measurement error, denoted by ε_i. Combining these pieces, the structural relationship between the population-level treatment effect and the constituent-level treatment effect for study i is

$$\widehat{t_i} = \mu + \varepsilon_i. \tag{3}$$

The goal of the fixed effects approach is to estimate the structural parameter μ, which implicitly depends on the particular comparison made in constituent studies, i.e. the contrast (ω', ω''), and the outcome being assessed, i.e. the measurement strategy, m. Critically, identifying μ also requires care and attention about cross-study properties, such as divergent validity, as well as contrast and measurement harmonization across studies. Assuming the structural model summarized by (3) is correct:

Remark 1. *The structural parameter μ is identified in a fixed-effects meta-analysis if each measurement strategy in a constituent study, m_i, satisfies divergent validity, and every study in $\mathcal{M}(\mathcal{I})$ is both contrast and measurement harmonized.*

Remark 1 follows by observing that $\mu = \tau_m(\omega', \omega'' \mid \theta)$, and combining (3) with Theorem 3. An implication of Remark 1 is that the internal validity of constituent studies is a necessary, but not sufficient, condition for identification of a treatment effect, $\tau_m(\omega', \omega''|\theta)$, from an externally valid mechanism. Since positing the existence of the structural parameter μ implicitly assumes that the mechanism of interest is externally valid, and proceeds to estimate the measured effect of

[14] We suppress the dependence on contrasts, measures, and setting for notational convenience.

that externally valid mechanism, one cannot establish whether a mechanism has external validity, since the same techniques can be implemented regardless.

7.2 Random-Effects Meta-Analysis

Some scholars view the fixed-effects model as too restrictive for meta-analysis applications and also note weaknesses of the estimator such has high Type-I error rates and poor coverage (Slough et al., 2021; Blair, Coppock, and Humphreys, 2022), moving instead to the random-effects meta-analysis estimator (see Table 1). The primary difference between the fixed-effects and random-effects models, as they apply to meta-analyses, arises from the specification of the mechanism's underlying theoretical structure.

In contrast to fixed-effects meta-analysis, in a random-effects meta-analysis, the analysts assumes that study-specific parameters, denoted by λ_i, are drawn from a common normal distribution with mean μ (the structural parameter from above), and variance υ^2.[15] The decomposition of the variance essentially allows for differences in the means between studies. The observed estimates taken from constituent studies are again denoted by $\widehat{t_i}$, and can be expressed as $\widehat{t_i} = \lambda_i + \varepsilon_i$, where, as above, ε_i represents study-specific estimation error, and it is assumed to come from a normal distribution with mean 0 and variance σ_i^2.[16] The standard random-effects model can be summarized by:

$$\begin{aligned} \widehat{t_i} &= \lambda_i + \varepsilon_i \\ \lambda_i &= \mu + u_i \\ \epsilon_i &\sim \mathcal{N}(0, \sigma_i) \\ u_i &\sim \mathcal{N}(0, \upsilon) \end{aligned} \tag{4}$$

Assuming the structural model summarized by (4) is correct:

[15] Note that when $\upsilon = 0$, the random-effects model reduces to the fixed-effects model.

[16] Our notation for this model may be non-standard given our usage in the conceptual framework. In this framework, λ (study means) is usually denoted θ and the between-site variance is usually denoted τ, not υ.

Remark 2. *The structural parameter μ is identified in a random-effects meta-analysis if each measurement strategy in a constituent study, m_i, satisfies divergent validity, and every study in $\mathcal{M}(\mathcal{I})$ is both contrast and measurement harmonized.*

In the random-effects model, the true site-level treatment effects are different, but are related to the "grand" mean effect μ in that they equal to the structural parameter of interest, μ with additive noise that is drawn from a normal distribution with nondegenerate variance υ^2. Since the main difference between the random effects model and the fixed effects model is to add another dimension of estimation challenges, Remark 2 follows by noting that, as above, $\mu = \tau_m(\omega', \omega''|\theta)$, taking the combined $u_i + \varepsilon_i$ and applying Remark 1. The similarities in the arguments reflect the similarities between the random-effects and fixed-effects estimators.

An interesting aspect of the random-effects model is the interpretation of the u_i terms, which can be thought of as capturing (statistically) departures from different kinds of harmonization (assuming the mechanism is externally valid). Thus, when constituent studies are believed to be fully harmonized, a correctly specified statistical test of the null hypothesis $\upsilon = 0$ can provide evidence of that assumption. An important caveat is that such an interpretation is predicated on the structural assumption that the space of potential harmonizing 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.

7.3 Alternate Structural Approaches

Our results focus on comparability between constituent studies and show the importance of harmonization in establishing such comparability. However, it is important to note that, consistent with "barefoot" (or agnostic) approaches to integrating evidence from multiple internally valid studies, our framework makes no assumptions about the structure of the set of contrasts, settings, or measurement strategies. We thus show that absent such structural assumptions, the necessary design conditions required are both demanding and stringent for valid identification of treatment effects

across studies; assumptions that may be implausible when studies are not intentionally harmonized (in contrasts or measurements). We did not develop such additional assumptions in our conceptual framework, since imposing structural assumptions was not the focus of our analysis. However, our results suggest researchers can approach problems of design that are related to harmonization by imposing stronger (structural) assumptions.

First, consider contrast harmonization, where studies may be partially (i.e., treatment) harmonized or completely unharmonized. In these cases, it is important that the instrument itself be measured (like the endogenous treatment in a first-stage regression in an instrumental variables setup). Second, when studies are not measurement harmonized, researchers must specify the relationship between unharmonized outcome measures. In common practice, researchers routinely employ scale-location transformations (i.e., Z-score transformations) to standardize the outcome scales across studies (Slough et al., 2021). This standardization implies an unstated assumption, namely, that measures across sites constitute an equivalence class of measures. Yet, there are meta-studies in which distinct outcome measures arguably do not form an equivalence class, implying that researchers need to be more explicit about the relationship between (i) instruments and measures in each study; and (ii) the relationship between measures across studies. Wider application (or integration) of latent-variable measurement models may provide the necessary structure on the relationship between measures in different studies (Fariss, Kenwick, and Reuning, 2020). These assumptions may require specifying the relationship between a mechanism and a particular measurement strategy, and will provide guidance for combining or synthesizing estimates on studies that lack measurement harmonization.

When conducting a meta-study, one's methodological approach and research goal ultimately determine the importance of modeling the relationships between contrasts, measurement strategies, and the substantive outcome of interest. While the structural approach calls for the invocation of substantially stronger assumptions when combining studies than the minimal assumptions made within constituent studies, we have shown that identification of the structural parameters estimated

in existing meta-analyses require much stronger design properties and assumptions than is generally acknowledged. As such, we see substantial room for the growth of structural estimation in meta-studies to allow for the synthesis of studies estimating non-comparable treatment effects. Where harmonization of studies is not possible, or where the assumption of external validity may be untenable, we view innovation in structural methods for meta-studies as an important and necessary complement to existing approaches.

8 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 (Heckman, 2000; Deaton, 2010; Deaton and Cartwright, 2018). In response, practitioners like Imbens (2010) advocate that such issues should be addressed with replication, and increasingly, empirical scholars have turned to meta-analysis as a potential tool to address these kinds of concerns. We develop a conceptual framework to understand external validity and elucidate the potential role of meta-studies in generalizing empirical findings. We present a number of results that highlight study comparability in the context of an externally valid mechanism. Our results stress the importance of harmonization, both in terms of what substantive comparison is being considered (contrast) as well as how outcomes are assessed and measured (measurement strategy). Our framework thus complements empirical frameworks that focus on internal validity and estimation within single studies.

Our conceptual framework points to the dangers of conflating conceptual differences across studies with statistical sources of variation (i.e., sampling) in treatment effects. Although such statistical concerns are important and need to be addressed in any meta-study, there remain important conceptual issues that can arise when comparing or aggregating estimates which take information from different studies. Such conceptual issues, when not addressed or discussed, can lead to incorrect or misleading inferences about the presence 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, and retrospectively, structural approaches should be applied to combine nonharmonized studies.

Finally, 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 to meta-studies is less straightforward. Our results suggest that more attention into design-based strategies of harmonization—of both contrasts and measures—is critical for improving the credibility and interpretability of the evidence presented in meta-studies. At the same time, the stringency of the conditions we identify, and the additional assumptions that are implicitly invoked by meta-study approaches, suggest there may be underappreciated limits to the kind of agnosticism advocated by the credibility revolution as we seek to cumulate evidence.

Appendices

A Proofs

Proof of Lemma 1. For a fixed measurement strategy, m, the treatment effect function is linear in the treatment indicator, γ, if there exists a smooth function, which we denote by $\tau_m(\omega', \omega'' \mid \theta) : \mathcal{C} \times \Theta \to \mathbb{R}$, such that

$$T_m^{\gamma}(\omega', \omega'' \mid \theta) = \gamma \cdot \tau_m(\omega', \omega'' \mid \theta),$$

for almost every θ and (ω', ω'').

Sufficiency follows immediately. For necessity, recall that a measurement strategy, m, has convergent validity if

$$T_m^0(\omega', \omega'' \mid \theta) = 0$$

for almost every contrast (ω', ω'') and almost every setting θ. Thus, convergent validity implies

$$\frac{T_m^1(\omega', \omega'' \mid \theta) - T_m^0(\omega', \omega'' \mid \theta)}{1 - 0} = T_m^1(\omega', \omega'' \mid \theta) \equiv \tau_m(\omega', \omega'' \mid \theta),$$

which is constant in γ, establishing that $T_m^{\gamma}(\omega', \omega'' \mid \theta)$ is affine in γ. This means that for some $a \in \mathbb{R}$ we can write

$$T_m^{\gamma}(\omega', \omega'' \mid \theta) = \gamma \cdot \tau_m(\omega', \omega'' \mid \theta) + a,$$

and again applying convergent validity,

$$T_m^0(\omega', \omega'' \mid \theta) + a = 0,$$

implying that $a = 0$, completing the argument. □

Proof of Theorem 1. Sufficiency is obvious. For necessity, suppose not. Since studies $\mathcal{E}_1$ and $\mathcal{E}_2$ are comparable, but not measurement harmonized, then for m_1 and m_2:

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

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

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

Combining (5) and (6),

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

contradicting divergent validity. □

Proof of Theorem 2. Sufficiency is obvious. For necessity, comparability implies that there are two contrasts, (ω_1', ω_1'') and (ω_2', ω_2''), where

$$\tau_m(\omega_1', \omega_1'' \mid \theta_1) = \tau_m(\omega_2', \omega_2'' \mid \theta_2), \tag{7}$$

and proceeding by contradiction, suppose that $(\omega_1', \omega_1'') \neq (\omega_2', \omega_2'')$. Applying external validity at m and (ω_1', ω_1''), we have that

$$\tau_m(\omega_1', \omega_1'' \mid \theta_1) = \tau_m(\omega_1', \omega_1'' \mid \theta_2). \tag{8}$$

Combining (7) and (8) yields

$$\tau_m(\omega_1', \omega_1'' \mid \theta_2) = \tau_m(\omega_2', \omega_2'' \mid \theta_2),$$

which, since the setting and contrasts were arbitrary, implies that the the treatment effect must be the same at (ω_1', ω_1'') and (ω_2', ω_2'') in any setting. Thus, external validity allows us to suppress the dependence of the treatment effect function on θ. Because $\mathcal{C}$ is a compact subset of $\mathbb{R}^2$, it is a two-dimensional manifold. Define

$$\kappa \equiv \tau_m(\omega_1', \omega_1'' \mid \theta),$$

which by external validity, is the same at almost any $\theta \in \Theta$. We are interested in the level set $\tau_m^{-1}(\kappa) \subset \mathcal{C}$. Since the derivative of $\tau_m(\omega', \omega'' \mid \cdot)$ has full rank for almost every contrast, $(\omega', \omega'') \in \mathcal{C}$, the set of regular points of τ_m is of full measure on $\mathcal{C}$. Thus, if κ is not a regular value, then $\tau_m^{-1}(\kappa)$ does not contain any regular points, and is thus of Lebesgue measure zero. Suppose, instead, that κ is a regular value, and thus, $\tau_m^{-1}(\kappa)$ is a set of regular points. By the Preimage Theorem (e.g., Guillemin and Pollack, 1974: pg. 21),[17] the set $\tau_m^{-1}(\kappa)$ is a submanifold of $\mathcal{C}$, and

[17]The Preimage Theorem is also called the Regular Level Set Theorem and is equivalent to the Constant Rank Theorem, see Tu (2011: Ch. 9-10).

moreover,

$$\dim \tau_m^{-1}(\kappa) = \dim \mathcal{C} - \dim \mathbb{R} = 2 - 1 = 1.$$

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

Proof of Corollary 1. The proof is the same as that of Theorem 2, with the smooth map $\tau_m(\omega_1', \omega'' \mid \theta)$, and replacing $\mathcal{C}$ with Ω, and noting that the Preimage Theorem then implies that

$$\dim \tau_m^{-1}(\kappa; \omega'') = \dim \Omega - \dim \mathbb{R} = 1 - 1 = 0,$$

thus completing the argument. □

References

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

Ana, de al O. 2012. "Do Conditional Cash Transfers Affect Electoral Behavior? Evidence from a Randomized Experiment in Mexico." *American Journal of Political Science* 57 (1): 1–14.

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, Abhijit, Esther Duflo, Nathanael Goldberg, Dean Karlan, Robert Osei, William Parienté, Jeremy Shapiro, Bram Thuysbaert, and Christopher Udry. 2015. "A multifaceted program causes lasting progress for the very poor: Evidence from six countries." *Science* 348 (6236): 1260799.

Banerjee, Abhijit V., Rema Hanna, Gabriel E. Kreindler, and Benjamin A. Olken. 2017. "Debunking the Stereotype of the Lazy Welfare Recipient: Evidence from Cash Transfer Programs." *The World Bank Research Observer* 32 (2): 155–184.

Banerjee, Abhijit V, Sylvain Chassang, and Erik Snowberg. 2017. "Decision theoretic approaches to experiment design and external validity." In *Handbook of Economic Field Experiments*. Vol. 1 Elsevier pp. 141–174.

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.

Bartholomew, M. 2002. "James Lind's *Treatise of the Scurvy* (1753)." *Postgraduate Medical Journal* 78: 695–696.

Blackwell, David. 1953. "Equivalent comparisons of experiments." *The annals of mathematical statistics* pp. 265–272.

Blair, Graeme, Alexander Coppock, and Macartan Humphreys. 2022. *Research Design: Declaration, Diagnosis, Redesign*. Princeton, N: Princeton University Press.

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* First View: 1–8.

Blair, Graeme, and Gwyneth McClendon. 2021. "Conducting Experiments in Multiple Contexts." *Advances in Experimental Political Science* p. 411.

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 Weinstein, Fotini Christia, Eric Arias, Emile Badran, Robert A. Blair, Ali Cheema, Ahsan Farooqui, Thiemo Fetzer, Guy Grossman, Dotan A. Haim, Zulfiqar Hameed, Rebecca Hanson, Ali Hasanain, Dorothy Kronick, Benjamin S. Morse, Robert Muggah, Fatiq Nadeem, Lily Tsai, Matthew Nanes, Tara Slough, Nico Ravanilla, Jacob N. Shapiro, Barbara Silva, Pedro C. L. Souza, and Anna M. Wilke. 2021. "Does Community Politicing Build Trust in Police and Reduce Crime? Evidence from Six Coordinated Field Experiments in the Global South." Working paper.

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.

Campbell, Donald T, and Donald W Fiske. 1959. "Convergent and discriminant validation by the multitrait-multimethod matrix." *Psychological bulletin* 56 (2): 81.

Cartwright, Nancy. 1983. *How the Laws of Physics Lie*. Clarendon Paperbacks, Oxford: Oxford University Press.

Chassang, Sylvain, 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.

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

Coppock, Alexander. 2021. *Persuasion in Parallel*. Chicago Studies in American Politics Chicago: University of Chicago Press.

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. 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.

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." Working paper.

Fariss, Christopher J. 2014. "Respect for human rights has improved over time: Modeling the changing standard of accountability." *American Political Science Review* pp. 297–318.

Fariss, Christopher J., Michael R. Kenwick, and Kevin Reuning. 2020. *The SAGE Handbook of Reseasrch Methods in Political Science and International Relations*. Number 20 SAGE London chapter Measurement Models.

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.

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

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

Gechter, Michael, Cyrus Samii, Rajeev Dehejia, and Cristian Pop-Eleches. 2019. "Evaluating Ex Ante Counterfactual Predictions Using Ex Post Causal Inference." Working paper available at `https://arxiv.org/abs/1806.07016`.

Gerber, Alan S., and Donald P. Green. 1999. "Misperceptions about Perceptual Bias." *Annual Review of Political Science* 2: 189–210.

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

Glass, Gene V. 1976. "Primary, secondary, and meta-analysis of research." *Educational researcher* 5 (10): 3–8.

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

Guala, Francesco. 2003. "Experimental localism and external validity." *Philosophy of science* 70 (5): 1195–1205.

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

Guess, Andrew, and Alexander Coppock. 2020. "Does Counter-Attitudinal Information Cause Backlash? Results from Three Large Survey Experiments." *British Journal of Political Science* 50: 1497–1515.

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

Heckman, James J. 2000. "Causal parameters and policy analysis in economics: A twentieth century retrospective." *The Quarterly Journal of Economics* 115 (1): 45–97.

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.

Hume, David. 1739-40 (2003). *A Treatise of Human Nature*. Penguin Books.

Imai, Kosuke, Gary King, and Carlos Velasco Rivera. 2020. "Do Nonpartisan Programmatic Policies Have Partisan Electoral Effects? Evidence from Two Large-Scale Experiments." *Journal of Politics* 82 (2): 714–730.

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: Journal of the Econometric Society* pp. 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. 2021. "Re-Assessing Elite-Public Gaps in Political Behavior." *American Journal of Political Science* Forthcoming.

King, Gary, Robert O. Keohane, and Sidney Verba. 1994. *Designing Social Inquiry: Scientific Inference in Qualitative Research*. Princeton: Princeton University Press.

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

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

Lind, James. 1753. *A Treatise of the Scurvy in Three Parts*. London: Sands, Murray, and Cochran.

Lord, C.G., L. Ross, and M.R. Lepper. 1979. "Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence." *Journal of Personality and Social Psychology* 37 (11): 2098–2109.

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

Lucas, Jeffrey W. 2003. "Theory-testing, generalization, and the problem of external validity." *Sociological Theory* 21 (3): 236–253.

Mackie, John L. 1965. "Causes and conditions." *American philosophical quarterly* 2 (4): 245–264.

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

Miguel, Edward, Shanker Satyanath, and Ernest Sergenti. 2004. "Economic shocks and civil conflict: An instrumental variables approach." *Journal of political Economy* 112 (4): 725–753.

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

Nyhan, Brendan, and Jason Reifler. 2010. "When Corrections Fail: The Persistence of Political Misperceptions." *Political Behavior* 32: 303–330.

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.

Peirce, Charles S. 1892. "The doctrine of necessity examined." *The Monist* 2 (3): 321–337.

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

Rosenbaum, Paul R. 1984. "From association to causation in observational studies: The role of tests of strongly ignorable treatment assignment." *Journal of the American Statistical Association* 79 (385): 41–48.

Rosenthal, Robert. 1986. *Meta-Analytic Procedures for Social Science Research*. Sage publications Sage CA: Thousand Oaks, CA.

Rosenthal, Robert, and Donald B Rubin. 1982. "Comparing effect sizes of independent studies." *Psychological bulletin* 92 (2): 500.

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

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

Schwarz, Susanne, and Alexander Coppock. 2020. "What Have We Learned About Gender From Candidate Choice Experiments? A Meta-analysis of 67 Factorial Survey Experiments." *Journal of Politics* Forthcoming.

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

Skoufias, Emmanuel. 2001. "PROGRESA and its Impacts on the Human Capital and Welfare of Households in Rural Mexico: A Synthesis of the Results of an Evaluation by IFPRI." International Food Policy Research Institute.

Slough, Tara. 2020. "On Theory and Identification: When and Why We Need Theory for Causal Identification." *Mimeo—New York University* .

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.

Slough, Tara, and Scott A. Tyson. 2021. "Conceptual Replication under External Validity." Working paper, New York University.

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

Stein, E, and Antoni Zygmund. 1964. "On the differentiability of functions." *Studia Mathematica* 23: 247–283.

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

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.

Zucco, Cesar, and Timothy J. Power. 2006. "Bolsa Família and the Shift in Lula's Electoral Base 2002-2006: A Reply to Bohn." *Latin American Research Review* 48 (2): 3–24.