External Validity and Evidence Accumulation

Quantitative and Computational Methods for the Social Sciences

DOI: 10.xxxx/xxxxxxxx (do not change)
First published online: MMM dd YYYY (do not change)

Tara Slough
New York University

Scott A. Tyson
Emory University

Abstract: TBD

Keywords: External validity, evidence accumulation, meta-analysis, replication

JEL classifications: C18, C81, C82, C90

ISBNs: xxxxxxxxxxxxx(PB) xxxxxxxxxxxxx(OC)
ISSNs: xxxx-xxxx (online) xxxx-xxxx (print)

Contents

1 The Accumulation of Evidence

The social sciences (e.g., political science, economics, sociology, etc.) have advanced through a process of conceptual development, quantification, and empirical inquiry (Hacking, 1990, 2006; Kitchin, 2014; Porter, 1996). When these three ingredients work together, they demonstrate the power that a scientific approach can wield. [1] Although social scientists draw heavily from techniques and concepts originally formulated in the natural sciences, there are important differences—especially when it comes to the generality of social phenomena—that demands attention.

Much thinking in the social sciences are conceptually organized around the idea or belief that general substantive forces drive human behavior through mechanisms. In particular, a mechanism—if characterized precisely enough—will yield the same influence in whatever context it arises. For example, collective action problems, where social groups, whether organized or not, have more difficulty accomplishing goals as the size of the group gets larger, has been articulated in every branch of social science (Olson, 1965). The intuition of collective action problems is thought to arise regardless of what goals a social group pursues, what individuals comprise it, or where it is located (geographically or temporally). Put simply, a collective action problem refers to a set of mechanisms that are general enough that they can arise in most settings or contexts. Indeed, when behavior departs from such an intuition, it becomes an avenue of research (Flache & Macy, 2013).

The generality of collective action problems is not a special feature of collective action problems: it is something that many substantive mechanisms are thought to exhibit. The existence of these mechanisms are an important feature of any scientific endeavor whose claim to relevance rests on the existence of discoverable phenomena that extend beyond the idiosyncratic circumstances in which they are observed. Moreover, generality of this sort is a necessary ingredient of any explanation, as "it seems altogether reasonable to maintain that any adequate explanation of a particular fact must be, in principle, generalizable into an explanation of a suitable sort of regularity." (Salmon, 1984, pg. 276). This kind of generality is about whether a mechanism has *external validity*, and we devote a fair amount of attention below to explicating this concept more precisely.

Isolating and understanding the mechanisms that exhibit external validity is a crucial ingredient for developing knowledge that transcends the idiosyncrasies inherent in observation, and in a way that can be applied across time, place, and circumstance. This type of general knowledge is necessary for the

[1] See Hamming (1980) for similar argument applied to the sciences broadly.

many experimentalists, in particular, who seek to use their findings to inform policy. Leading organizations that fund, organize, and promote experimental research in the social sciences like Evidence in Governance and Politics (EGAP) and the Abdul Latif Jameel Poverty Action Lab (J-PAL) center the use of evidence in their mission statements. EGAP "promotes rigorous knowledge accumulation, innovation, and evidence-based policy in various governance domains, including accountability, political participation, mitigation of societal conflict, and reducing inequality" (Evidence in Governance and Poltics, 2022). J-PAL's mission is "to reduce poverty by ensuring that policy is informed by scientific evidence. We do this through research, policy outreach, and training" (Abdul Latif Jameel Poverty Action Lab, 2022). These examples highlight the importance of *measuring* the influence of a substantive mechanism beyond that in a single study. Learning how to do this, and evaluate results, occurs through the *accumulation of evidence*, taking from studies of a common mechanism that are conducted in different places, at different times, and in different contexts.

Much of the recent development in empirical science has focused instead on *internal validity*, with great emphasis on ensuring that, for instance, an empirical finding from a single study can be given a causal interpretation. This perspective focuses on measuring the "effects of causes" (Holland, 1986), and has transformed the way that social scientists approach empirical research. Empiricists increasingly seek to measure some form of aggregate causal effect (e.g., the average treatment effect [ATE]), and they have increasingly adopted research designs—often called design-based approaches to causal inference—which reduce substantially the use of model-based approaches that were previously dominant in empirical social science.[2]

External validity poses severe challenges for both the accumulation and the application of evidence. Researchers typically seek to make inferences about broader substantive phenomena, which are not tied to the context and sample used to measure treatment effects. But, for example, when does evidence about voter preferences in Colombia provide information about voter preferences in South Africa? Similarly, Bond et al. (2012) showed that Facebook banners with get-out-the-vote messages have a mobilizing effect in 2010, but would this effect have emerged in later elections? This question asks whether there is temporal validity of Facebook as a mobilizing tool.[3] Questions like these, while difficult to answer, are central to understand causal mechanisms and general substantive phenomena, but ultimately have nothing to do with the quality of the studies of voter preferences in Colombia or Facebook mobilization in 2010.

[2] Much graduate coursework in statistical methods and econometrics increasingly emphasize design-based methods for causal inference (Cunningham, 2021).

[3] See Munger (2021) for a focused discussion of temporal validity.

The most prominent critique of studies that draw so heavily on techniques associated with the credibility revolution (e.g., experiments or natural experiments, etc.), and less on model-based techniques, is that these research designs place too much emphasis on internal validity—perhaps at the expense of—external validity (Deaton, 2010; Deaton & Cartwright, 2018; Gailmard, 2021; Huber, 2017). In other words, even if a research design permits credible estimation of a causal treatment effect in some context or sample, it ultimately says nothing about a treatment effect elsewhere. That a study's internal validity is somehow related to its external validity, either by enhancing it or detracting from it, rests on a core confusion. Specifically, it is *impossible to know* whether a similar effect would obtain in different contexts or samples, taking only information at the level of a single study—regardless of the quality of that single study. Put differently, asking whether a finding from a single study has external validity, based solely on the properties of that single study, is a category error. One cannot assess external validity—its presence or scope—without accumulating evidence from multiple places, at different times, and contexts.

To illustrate what we mean by external validity, consider an example. Suppose that we are interested in the effect of an undergraduate student advising program on student grades at New York University (NYU). To measure this effect, we select a simple random sample of 300 undergraduates at NYU, of whom we randomly assign 150 to control and 150 to the new advising program. Given the random sampling into the experimental sample, the results of our study, tell us something about the population of NYU undergraduates. To this point, nothing about external validity has been invoked, just sampling, estimation, and inference. In other words, external validity is not about transporting a treatment effect from a random sample to the population from which the sample is drawn, i.e., the support of the sampling distribution.

External validiity is instead about moving an inference from the population the sample's drawn from to a distinct population.[4] Unfortunately, the results obtained from our study using NYU undergraduates tells us nothing about Emory University undergraduates, at least, not without knowledge that the mechanism the advising program activates has external validity. This last point illustrates that external validity cannot be faithfully characterized as an entirely empirical concept, i.e., it always reflects some theoretical commitments. For instance, if we conducted our study on Emory undergraduates and found statistically indistinguishable findings to those from NYU, this ultimately provides no empirical information about the program's effects on undergraduates at the University of Rochester. However, our belief that the program would have a

[4] That distinct population could well contain the population of our experiment (all NYU undergraduates), but it would also contain some other units like Emory University undergraduates.

similar affect at Rochester, having done two studies instead of one, is ultimately a theoretical claim.

Given the importance of issues related to evidence accumulation, in recent years, scholars have made contributions to conceptualize external validity and developed various methodological "fixes" for dealing with the accumulation of evidence in the shadow of external validity. This *Element* provides a framework to help to synthesize, organize, and select among these approaches. One core idea of this book is that these ideas and tools vary in their fidelity to the guiding principles of the credibility revolution (see below). But in order to understand how these approaches relate to causal mechanisms within a single study, it is important to state precisely the guiding principles of the credibility revolution.

1.1 The Credibility Revolution

The credibility revolution refers to the response to the methodological critique of empirical work in economics by Leamer (1983), which is reviewed by Angrist and Pischke (2010) for economics and Samii (2016) for political science, among others. It outlines a set of goals as well as an intellectual framework that is used to define and evaluate various empirical problems. In this section we briefly discuss these ingredients which can be characterized by three guiding principles:

1. **A model of causality**: Causality is defined within the potential outcomes framework (or Neyman-Rubin Causal Model).
2. **Design-based assumptions**: A minimal set of assumptions, sometimes justified by a study's design, is invoked for identification or estimation.
3. **Unbiasedness**: Unbiased estimators should be prioritized over biased estimators.

While these principals have not been articulated in this format before, they are all discussed at length (e.g., Angrist & Pischke, 2008, 2010; A. V. Banerjee & Duflo, 2009; Imbens, 2010; Imbens & Wooldridge, 2009; Samii, 2016). Our goal in elaborating these principals in a book about external validity is because we develop key assumptions that underlie the accumulation of evidence and we organize and assess them relative to the guiding principle of the credibility revolution.

A causal mechanism is a foundational concept in many sciences, and it responds to the core idea that "*if you perform such and such action, you will have such and such experiences*," or more generally, "*if anyone performs such and such actions, then such and such publicly observable events will take place*" (Putnam, 1981, pg. 180-182). The potential outcomes framework formalizes a causal mechanism and provides a particular definition of "causality" through causal effects.[5]

The potential outcomes framework has three ingredients: (i) a population of interest; (ii) an outcome measure; (iii) an intervention. The outcome measure, evaluated at different levels of the intervention (e.g., control and treatment) provide *potential outcomes*. These potential outcomes are generally assumed to be *fixed* or non-random.[6] Interventions are sometimes intentional, such as when they are designed and implemented by policymakers or researchers. However, an intervention can also be a naturally occurring thing (such as a historical event) that has nothing to do with human action (Woodward, 2005).

Within the potential outcomes model, causality is defined by looking at the effect of the intervention—the causal effect is measured by comparing (observed) potential outcomes. In particular, it is the answer to the question: what's the difference between the outcome, evaluated at treatment, and the outcome, evaluated at control, for a given unit? Since, in most cases, both values of potential outcomes cannot be observed for a single unit, comparisons can only be conducted at some level of aggregation, typically the average outcome between those exposed to treatment and those exposed to control.[7]

In the potential outcome model, the potential outcomes are *primitives of the model*, which is why it does not require explicit modeling of how potential outcomes arise from some underlying process, the details of which depend on the specific structural model. Proponents of the credibility revolution view this as an advantage of the potential outcomes framework over more traditional structural or econometric views of causality that require explicit modeling the source of treatment effects, and in particular, strong assumptions about how "the [external] world works." Imbens and Wooldridge (2009) write that: "One additional advantage of the potential outcome set up is that the parameters of interest can be defined, and the assumptions stated, without reference to particular statistical models" (p. 7).

Aronow and Miller (2019) describe the credibility revolution's reduced reliance on statistical models as stemming from "a growing acknowledgment that the evidence that researchers adduce for their claims is often predicated on unsustainable assumptions" (p. xv). Instead, proponents advise being explicit about the assumptions that support both *identification* and *estimation* of treatment effects, and when possible, adopt research designs that render these assumptions plausible. The embrace of more agnostic methods for identification and estimation stems from the concern that if assumptions do not obtain in

[5] The model is originally developed in Neyman (1923) and Rubin (1974).

[6] Permitting potential outcomes to be stochastic is not of consequence for most of our discussion. We will make a note when admitting stochastic potential outcomes changes our interpretation.

[7] That one is constrained to only see one potential outcome per unit is known colloquially as the *fundamental problem of causal inference* (Holland, 1986).

the world or the data, inferences could be wildly misleading. Moreover, when research is used to inform policy, recommendations taken from fallacious models *could be harmful*.

The most common assumptions at the core to the identification of causal estimands are: (i) exclusion (or excludability) and (ii) stable unit treatment value assumption (SUTVA). Exclusion is about whether the "only relevant causal agent is the receipt of treatment" (A. S. Gerber & Green, 2012, p. 39). SUTVA ensures that the mapping from intervention to outcome measure is one-to-one, i.e., there is a unique potential outcome for every level of the intervention. It also guarantees that there are "no spillovers" where potential outcomes for a unit depend on the intervention status of other units. A third assumption, ignorability of intervention assignment, facilitates unbiased estimation of treatment assignment. Ignorability holds that potential outcomes are independent of the assignment of intervention across units. The latter is typically guaranteed by randomizing an intervention and is necessarily to facilitate estimation of the average treatment effect through a difference in means.

The credibility revolution advocates for *design-based* strategies to ensure that the assumptions required to give a causal interpretation to one's estimand are satisfied and plausible. It is important to emphasize that designed-based strategies do not limit the need for *any* assumptions, nor do they necessarily render *every* assumption reasonable, but that they rest on assumptions that link to a practical design rather than assumptions about the external world.

Because of the heavy reliance of design-based methods for causal inference, which may be constraining, studies may favor settings or samples for convenience reasons. For example, close elections regression discontinuity designs focusing on "tied" elections may say little about places where electoral results have a wider margin of victory. In each of these examples, the design-based perspective leads researchers to focus on a part of the potential sample that is not representative, and hence, questions of external validity may be particularly salient. However, moving to designs with larger or more representative samples may not provide additional insight into external validity, because it is the mechanism that transports and not findings. Therefore, learning about the external validity (of treatment effects) requires *building on*—rather than supplanting—the advances of the credibility revolution.

Finally, another important and distinct part of any empirical study involves statistical estimation. When assessing an estimator of a causal effect, proponents of the credibility revolution have generally stressed the importance of statistical unbiasedness. Specifically, Imbens (2010, p. 417) writes: "by internal validity I mean the credibility of the estimator as an estimator of the causal effect of interest." Unbiasedness is just one property of an estimator. Blair, Cooper, Coppock, and Humphreys (2019) provide a litany of other "diagnosands" or properties on which estimators can be compared. In principle, one might trade off modest increases in bias for sufficient reductions in the variance (or increases in the power). Our point is simply that the approaches developed in the credibility revolution put substantial (if not exclusive) weight on unbiasedness as the objective when selecting an estimator. Unbiasedness of an estimator is about what the statistical measure conducted in a study "aims at," i.e., the estimation target.

1.2 Empirical Targets

What are the objects that an empirical exercise aims to measure? One view is that empirical studies are about uncovering descriptive features of a sample, and on this view, substantive mechanisms play no role—they are irrelevant. Is collecting and describing a sample the "aim" of empirical social science? The other view is that the goal of empiricism is to make inferences about substantive phenomena, by extracting the signal from the noise in what is observed.[8] This view gives meaning to the influence of a mechanism since it is predicated on the presence of some underlying relationship which mechanically links together various objects in the external world.

A research design (credible or not) produces an *empirical target*, which is what needs to transport or generalize across different contexts. Specifically, when looking across studies it is important to ensure that the targets across studies are related in the same way, either because research designs are "aiming at the same target," or because their targets are systematically related. Otherwise, the output of a meta-study is necessarily ambiguous. Empirical targets serve as the quantitative object that unites different studies of the same mechanism. For this reason, they comprise the object of evidence accumulation. Having a quantitative object uniting different studies is important because "What is desirable [about quantitative knowledge] is the strength and *severity of the argument* that is afforded by a special kind of experimental knowledge. As such, it makes sense to call all cases that admit of a specifiably severe or reliable argument "quantitative," so long as this special meaning is understood." (Mayo, 1996, pg. 44). The empirical target is the only quantitative object that can plausibly constitute a mechanism's influence or effect. Therefore, identifying empirical targets is a necessary step toward substantive explanation, which tells us what will happen when a particular policy intervention is implemented.

[8] This view was originally explicated by Venn (1888).

Our approach to external validity, where empirical targets are a core ingredient, is organized around the *effects produced by causal mechanisms*, and connects to the guiding principles of the credibility revolution. Our mechanistic view of external validity, as applied to experiments, answers critiques related to the external validity of findings from randomized-controlled trials. For instance, Deaton (2010, p. 448) argues that "for an RCT to produce 'useful knowledge' beyond its local context, it must illustrate some general tendency, some effect that is the result of mechanism that is likely to apply more broadly." This point is correct and the logic is straightforward. In the absence of common mechanisms that might produce similar treatment effects in multiple places, it is unclear why an internally valid estimate in setting 1 should provide any information about the analogous treatment effect in setting 2. But this has little to do with the knowledge coming from an RCT. Phrasing external validity in terms of empirical targets clarifies what kind of relationship internal and external validity can have. In particular, there is not an obvious tradeoff between internal and external validity, but rather, internal validity is a necessary condition for external validity.[9] Findings from a single study are a single empirical target and they simply do not project to other empirical targets without a theory of how they transport.

1.3 Accumulating and Transporting Evidence

The principles of the credibility revolution have motivated empirical researchers to be increasingly aware of their estimand, how its estimated, and the assumptions required to ascribe a causal interpretation to their estimand (e.g., Angrist & Pischke, 2008). When it comes to evidence accumulation, and necessarily external validity, most empirical responses take one of two paths. The first path involves *gathering*, advocating to "do more studies," or gather the findings across credible studies in multiple samples or contexts (A. V. Banerjee & Duflo, 2009; Dunning, 2016; A. S. Gerber & Green, 2012). For example, the Metaketa initiative by Evidence in Governance and Politics embraces this approach, completing at least four sets of coordinated field experiments on information and accountability (Dunning et al., 2019), formalization and tax compliance (de la O et al., 2021), community monitoring of common-pool resources (Slough et al., 2021), and community policing (Blair et al., 2021). The gathering path is also evident in some replication studies across the social sciences (e.g., Camerer et al., 2016, 2018; Open Science Collaboration, 2015).

A second approach to evidence accumulation and external validity focuses

[9] A feature noted by Gailmard (2021, pg. 96): "external validity cannot exist without internal validity."

on *extrapolation*, which advocates imputing a general causal effect that unites across findings across different samples or settings. Approaches focusing on generalization to other samples emphasize the reweighting of various findings to samples of units with different covariate profiles (Cole & Stuart, 2010; Egami & Hartman, 2020; Kern, Stuart, Hill, & Green, 2016). Current approaches focused on extrapolation to different contexts use (structural) models to “transport” estimates to different settings, an inferential strategy akin to selection on observables (Pearl & Bareinboim, 2011, 2014).

Both the accumulation and extrapolation approaches invoke distinct—and often implicit—definitions of external validity. In this book we precisely articulate and organize different concepts of external validity. In uncovering these various concepts of external validity, we identify approaches that are consistent with the principles that have guided the credibility revolution as they are articulated above. We show that more often than not, approaches to problems of external validity, causal generalization, and evidence accumulation, cannot ensure—and sometimes contradict—the principles of the credibility revolution. By identifying these tensions, we discuss the benefits—and the limitations—of the credibility revolution’s approach to causality and inference. We also outline an approach to external validity and evidence accumulation that are consistent with the principles of the credibility revolution. Our discussion is therefore largely theoretical, and focuses on the conceptual foundations of external validity and evidence accumulation, rather than statistical features.

1.4 A Motivating Example

Throughout the text, we focus on a motivating example to fix ideas and provide concrete intuition. We draw from the vast empirical and theoretical literature on political accountability (see Ashworth, 2012, for a review). In particular, the question of how voters, using the ballot box, can hold politicians accountable for their performance in office. Many studies, theoretical and empirical, have focused on various dimensions of political accountability, ranging from incentive/institutional issues, information problems, as well as psychological and cognitive deficiencies of voters.

1.5 Outline of this Element

[brief chapter summaries]

1.6 Appendix: The Potential Outcome Model

In this section we present a common version of the potential outcome model that will be helpful in later sections; we draw heavily from Imbens and Angrist (1994). We also present, in the context of the model, the three most common assumptions at the core of most studies of causal identification are: (i) exclusion; (ii) stable unit treatment value assumption (SUTVA); ignorability of intervention assignment. As is customary, we present everything in the context of a binary intervention, which can be interpreted as a treatment and control.

There is a *population of units*, indexed by $i = 1, ..., N$. There is a set of *instruments* Ω that correspond to different levels of an intervention. For a binary intervention, there are two values of the instrument, $\omega', \omega'' \in \Omega$. The value of the instrument, ω, can be thought of as representing the "dosage" of treatment for a subject.

From a theoretical perspective, it is natural to focus on the unit level, specifically, by defining *potential outcomes* as a mapping $Y^m : \Omega \to \mathbb{R}$. We index different measures of the outcome of interest by m and this will become important below. All units in the population have a potential outcome corresponding to each level or value of the instrument, i.e., $Y_i^m(\omega) \in \mathbb{R}$ (is defined) for each i. A *causal effect* is then defined as the difference between potential outcomes. In particular, for a single unit, i, the causal effect from an intervention changing the instrument from ω' to ω'' is

$$t_i = Y_i^m(\omega'') - Y_i^m(\omega'). \tag{11}$$

As mentioned above, the fundamental problem of causal inference means that one cannot observe both $Y_i^m(\omega'')$ and $Y_i^m(\omega')$ simultaneously, and hence, t_i is not observable. As a consequence, we need to introduce additional machinery and assumptions into the model.

Because of the fundamental problem of causal inference, we need to focus on which units receive what value of the intervention. Intervention *assignment* is a vector $\omega \in \Omega^N$ that gives the assignment status of each unit. Assignment determines the value of the intervention, ω' or ω'', for unit i determining which potential outcome is observable to the researcher for that unit. Thus, *observed potential outcomes* are determined by the function $Y^m(\omega) = (Y_1^m(\omega_1), Y_2^m(\omega_2), \ldots, Y_N^m(\omega_N)) : \Omega^N \to \mathbb{R}^N$.

The aggregate *treatment effect* is defined as

$$t_f = f(Y_i^m(\omega'') - Y_i^m(\omega')), \tag{12}$$

where $f(\cdot)$ represents some operator or function, typically the expectation operator or quantile function. Various forms of (aggregate) treatment effects can be constructed following (12) by changing the operator $f(\cdot)$ or changing the sample by conditioning on some pre-treatment characteristics of units.

We now present the assumptions common to studies of causal inference. These assumptions amount to important restrictions on the vectors ω and $Y^m(\omega)$. There is an assignment rule ρ, which is a distribution over the set of assignments. Since we are focusing on a binary intervention, where the only values of the instrument of relevance are ω' and ω'', the set of assignments is the set of vectors whose components are 1 or 0, formally, $\{0, 1\}^N$.

First, the *exclusion restriction* holds that the "only relevant causal agent is the receipt of treatment" (A. S. Gerber & Green, 2012, p. 39). One important implication of the exclusion restriction is that, when treatments are "bundled," we cannot attribute causal effects to any specific component of that bundle in isolation. Because the full bundle of treatment, captured by ω'' and/or ω' is, in principle, received simultaneously, potential outcomes $Y_i^m(\omega'')$ and $Y_i^m(\omega')$ are defined by the full bundle of characteristics of each treatment condition.

Second, SUTVA is that ω and $Y^m(\omega)$ have full rank in $\mathbb{R}^N$. The first part, that ω has full rank, captures that the assignment of one unit does not determine the assignment of another unit. The second part, that $Y^m(\omega)$ has full rank, implies that the potential outcome of one unit does not determine the potential outcome of another unit, and is sometimes referred to as a "no spillovers" or non-interference assumption.

Third, ignorability is that the assignment rule, ρ, has full support on $\{0, 1\}^N$ and is independent of the profile of potential outcomes, $(Y^m(\omega''), Y^m(\omega'))$. Ignorability ensures that the potential outcomes revealed in an experiment are equivalent in expectation to those that are unobserved. Formally, ignorability of treatment assignment ensures that $\mathbb{E}[Y_i^m(\omega'')|\omega = \omega''] = \mathbb{E}[Y_i^m(\omega'')|\omega = \omega']$ and $\mathbb{E}[Y_i^m(\omega)|\omega = \omega''] = \mathbb{E}[Y_i^m(\omega)|\omega = \omega']$. When the other identification assumptions are satisfied, ignorability facilitates our ability to estimate treatment effects like the average treatment effect through a difference-in-means estimator.

2 A Framework for Empirical Research

Accumulating evidence across multiple contexts requires a more comprehensive view of the empirical enterprise than is usually required in single studies. In particular, the accumulation of evidence requires an understanding of what ties different studies together, how important context is to research design, as well as where these things lie between theoretical and measurement tools. In this chapter we provide a novel framework for thinking through empirical research. Crucially, we articulate the relationship between empirical research, theories, and "reality." This framework will help demarcate distinct ways that external validity is used conceptually, and thereby, clarify its role in meta-study approaches such as replication and meta-analysis. Our framework draws on prominent themes from philosophy of science, building on some conceptual distinctions in Putnam (1981), Bogen and Woodward (1988), Giere (2010), Bueno de Mesquita and Tyson (2020), and Ashworth, Berry, and de Mesquita (2021). Our contribution is to synthesize these themes in a novel framework for empirical research design that can be applied to better understand efforts to accumulate evidence.

Our conceptual framework consists of three elements and the relationship between them. First, the **external world** is the collection of objects that exist outside of social scientists' descriptions and characterizations. Second, the **conceptual world** is the collection of concepts, systems, and relationships that are used to define, classify, and organize the collective understanding of the external world. An important distinguishing feature of conceptual worlds, when compared with the external world, is that they are the product of human classification and reasoning, and are not typically unique. Third, and a key feature of our framework, is a bridge that connects the external world with a conceptual world. This bridge comprises all of the decisions and techniques that are employed in connecting the external world and the various concepts in the conceptual world that are meant to (or expected) to manifest in the external world as mechanisms. The choices and tools used to connect that external and conceptual worlds are best described as methodological, which is why we refer to this process as building a **methodological bridge**. The methodological bridge is a critical part of our framework, and will be important for thinking about how to accumulate evidence, which potentially involves comparison of different methodological bridges. Consequently, we need to be more precise about what goes into the methodological bridge, and connect this with some common approaches to empirical assessment of external world phenomena. Figure 1 illustrates the three components of our framework and shows how they fit together.

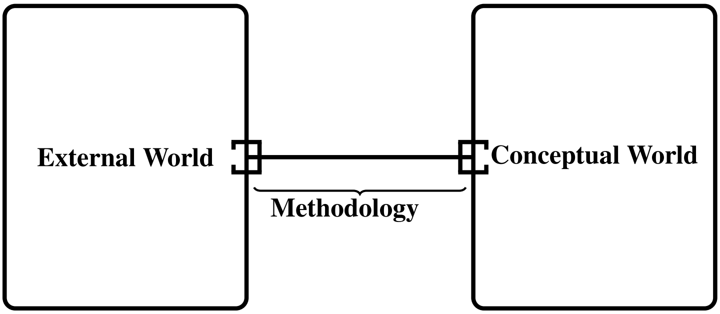

Figure 1 Conceptual framework for understanding empirical research.

When thinking about the processes that measure the external world it is important to keep in mind that "data are typically the result of complex interactions among a large number of disparate causal factors which are idiosyncratic to a particular experimental situation." (Bogen & Woodward, 1988, p. 319). Consequently, attributes of an empirical study such as sampling, and the (observed) realization of a treatment, ineluctably introduce some randomness into the observed data. Consequently, measurement error—whether systematic or non-systematic—is always present in the data to some degree or another. None of these features—sampling into an experiment, random error, or data processing—are inherent to the phenomena of interest that we aim to measure (by construction). As such, one set of data provides one snapshot of the external world, and a different set of data provide a potentially very different snapshot.

In this chapter, we provide a more thorough exposition of the the external and conceptual worlds, and then we describe how different methodological approaches/choices bridge these worlds in different ways. We place particular emphasis on reduced-form and structural approaches. Our framework is consistent with both structural and reduced-form approaches because theory plays the same role in both approaches to causality: it serves as the anchor to a select conceptual world. We argue against common skeptical assertions that any interpretation of causal estimands demands adopting structural or econometric models (implicitly or explicitly) (Deaton, 2010; J. J. Heckman, 2000). However, we also argue against the relatively common contention among proponents of the credibility revolution that reduced-form approaches circumvent problems of bridging data with concepts (Imbens, 2010). A recent literature on the theoretical implications of empirical models has made progress in incorporating theory and measurement in reduced-form research designs (e.g., Bueno de Mesquita & Tyson, 2020; Slough, 2022). Developing a precise distinction between structural and reduced-form approaches is important because they constitute different formulations of external validity, and hence, address different questions.

2.1 The External and Conceptual Worlds

Ultimately, the goal of empirical social science is to learn about *social phenomena*, or "what happens" in parts of the world that exist outside of the conceptual formulation of phenomena provided by social scientists. This extends from material objects (e.g., a tree) to nonmaterial phenomena, such as the responses/responsiveness of real people. Both manifest externally to the scientific enterprise. Objects and phenomena of interest comprise the external world that empirical research is focused on observing and measuring. We use the term external world to refer to "the real world" to be precise about our concept without having to detail what "reality" means. By distinguishing the external world from its description, we allow causal processes and regularities in human behavior to "have stable, repeatable characteristics" (Bogen & Woodward, 1988, p. 317), without requiring that they do in any given description.

Recall from Chapter 1, terms like "free-rider problem," "group size," and "selective incentives" are all concepts related to the phenomenon of "collective action problems." All of these are *concepts*, taken largely from Schelling (1960) and Olson (1965), that are used to describe, classify, and ultimately understand specific kinds of phenomena. However, they are all part of human classification, and not necessarily the mechanisms in the external world we, as social scientists, seek to understand.[10] While collective action has been widely theorized and studied, it is possible that there are completely distinct (and yet unknown) concepts for understanding this type of collective behavior. Here, we do not mean simply a semantic relabeling of Olson's (1965) concepts, as is too common in the social sciences. Instead, we mean distinct concepts or explanations for these phenomena. To this end, we allow for multiple conceptual worlds that contain distinct sets of concepts. However, readers need not view conceptual worlds as distinct to understand our framework.

The goal of formulating concepts is to describe, classify, and give collective meaning to phenomena in the external world. A conceptual world is thus the collections of concepts, systems, and relationships that are used to describe, classify, and organize observed phenomena in the external world. Another feature of a conceptual world is that it is the result of collective effort and communication, and thus, is "public property" among a community of users and populates its discourse. Consequently, concepts, including their formalization and characterization, provide the language used to to discuss and interpret phenomena in the external world. To stay focused on our contribution, we do

[10] We suppose that mechanisms are part of the external world, although admittedly, mechanisms are also concepts, and may themselves not be part of the external world. For conceptual clarity we sideline this concern.

not unpack the relationship between concepts and language used to express them any further, and their relationship to truth, referring to classic work on this topic such as Frege (1948), Kripke (1972), Putnam (1974), Goodman (1978), and Putnam (1981).

For example, many formulations of political accountability suggest that agency problems, including moral hazard, are characteristic of the relationship between the politician and voters. In particular, because voters cannot directly observe the effort a politician provides toward crafting policy, they must rely on imperfect signals of politician effort, such as the state of the economy, crime rates, or public goods provision. While these indicators are easily observed, they are only imperfectly related to the effort provided by the politician. Such moral hazard problems undermine political accountability regardless of whether we, as social scientists, recognize the concept or understand how it works. Symmetrically, if moral hazard does not arise in the context of political accountability, our collective belief that it does cannot change this feature. The politician's effort and voters' experiences with crime or or economic fundamentals are elements of the external world. Moral hazard, instead, is a concept that has emerged as part of the development of social science, and is thus part of the conceptual world.

2.2 Methodology as a bridge

The external world is not directly observed, but instead, it *becomes* observed when it is measured. Observation is thus something that is engaged in, with activity and care, rather than something which is passively conducted, a distinction which is often overlooked. Put differently, Ryle (1949, pg. 222) write: "We use the verb 'to observe' in two ways. In one use, to say that someone is observing something is to say that he is trying, with or without success, to find out something about it by doing at least some looking, listening, savouring, smelling or feeling. In another use, a person is said to have observed something, when his exploration has been successful, i.e. that he has found something out by some such methods." It is precisely to keep this distinction clear that, in our presentation of the potential outcome model in section 1.6, we introduced an index to keep track of different outcome measures (denoted by m). By observing that observation is an active process, we can see the importance of being attentive to the process of measuring the external world across contexts.

The goal of empirical inquiry is to bridge the external world with a conceptual world, and this process involves the combination of theoretical development as well as measurement, and is *methodological*. In our framework, the methodological bridge is what connects a conceptual world to the external world and is illustrated in Figure 1. The methodological bridge has two claws, hooks, or anchors by which it affixes to the conceptual and external worlds. The first hook connects the methodological bridge to the external world, and corresponds building a **measurement model**. The second hook connects the methodological bridge with a conceptual world and corresponds to the construction of a **theoretical model**. The necessity of the methodological bridge reflects that "...reality is a richer causal structure than the one assumed by the theory of causality." **?**, pg. 154

In our framework, a measurement model refers to all of the processes, techniques, and choices involved in creating a measure. It encapsulates a number of procedures and strategies that are used to produce data about the external world, all of which likely vary by subfield and substantive question.[11] As an example, when applied to causal studies, a measurement model, or strategy, refers to a number of components that are chosen with the aim of detecting the presence, and quantifying the effect, of a particular causal mechanism. These components include the selection of an outcome of interest, and choice about how to measure it. Moreover, the intervention corresponds to the selection of a comparison, or contrast, that is used to evaluate and quantify the effect of a particular mechanism. In relation to the potential outcome model presented in section 1.6, the measurement model reflects the choice, conceptualization, and definition of potential outcomes.

For example, while it is convention for quantitative researchers to represent a binary treatment numerically as a 1 for treatment and a 0 for control, this is simply a normalization, and one that has important implications when trying to accumulate evidence across studies, where 1 and 0 may take on different meanings across studies. An intervention is necessarily associated with an object in the external world, and causal mechanisms are objects in the external world that an intervention is designed to "turn on" or "turn off." As such, ensuring that two interventions are the same *in the external world* requires knowledge of the external world beyond what is normally covered in a single study.

A measurement model is also concerned with the estimation of a quantity of interest. Consequently, estimation concerns are important. These concerns occupy a disproportionate share of methodological discussions in the social sciences. Indeed, of the previous 14 Elements in this series, 12 focus primarily on issues of estimation. It is crucially important that we understand the properties of the estimators we use—and thereby select better estimators (Manski, 1995). Our framework stresses that estimation is just one piece of the broader methodological

[11] The concrete details involved in this process are generally complex and vary from case to case, so an accounting is well beyond the scope of this book. However, some important themes are reviewed in Jerven (2013) and Fariss (2014).

bridge.

Much like a measurement model anchors a research design onto the external world, on the other end, *theoretical models* form an anchor onto the conceptual world. An important part of a formulating a theoretical model is operationalizing concepts, which corresponds to detailing how a particular concept manifests, and interacts with other concepts, as part of a framework. A theoretical model thus *articulates* a particular concept, which is where the concept becomes part of the model. Returning to our example, moral hazard, is a general concept, but it takes on slightly different formulations in different models, i.e., in a career-concerns model of political accountability versus one that is static.

We make an important distinction between conceptual development—the central task of constructing a conceptual world—and articulating how concepts interact (or do not interact) to produce measured causal effects. Social scientists across disciplines have different notions of theory and criteria through which to evaluate theories. Our framework is agnostic as to what types of theory are more insightful, or how such a determination would be reached. Instead, our main evaluative criteria is the degree to which a theoretical model aligns with the concepts that it entails to represent.

Many conceptual frameworks for empirical research omit discussion (and often recognition) of the external world, instead equating it with data. For example, King, Keohane, and Verba (1994) and Blair et al. (2019) start with data, as though it is equivalent to the external world (passively observed), and is independent of perception—a perspective referred to as *naïve realism*. Instead, *perspectivism* refers to when the observation of phenomena necessarily occurs through the "lens" of how the phenomena is measured (e.g., Giere, 2010). Maintaining a perspectivist viewpoint is necessary for thinking about the accumulation of evidence because when looking across different contexts, distinguishing phenomena from the tools used to measure them is necessary. Take as an example, two experiments on participation in collective action where group size is manipulated. In one, an experimentalist increases group size (relative to some control size) and in another the experimentalist decreases group size (relative to some other control size). Although both interventions should tap into the same phenomenon, there is no reason that they should do so in exactly the same way. As a result, they should not produce the same measured effect.

Our framework provides a more detailed account of the measurement process behind empirical science than is common in the social sciences. In particular, Shadish, Cook, and Campbell (2002) focus on *construct validity*, which is when a measure reflects the concept it is designed to capture. Most discussions of construct validity (e.g., Adcock & Collier, 2001), treat an empirical measure as something that can directly connect to a concept, without stressing the need to theoretically articulate that concept. Our framework identifies that construct validity is instead the combination of several components of the methodological bridge and emphasizes that the reliance of a "construct" on how a theory is articulated and connects to concepts.

In sum, a core premise of our framework is that the external world cannot be directly observed and this is relevant for the accumulation of empirical evidence. Instead, when social scientists conduct studies, they become anchored to the external world through measurement. But critically, the way one anchors oneself to the external world affects what one observes, and different anchors will likely produce a different snapshot of a phenomenon. Because scientific experience is contrived, it can never be independent of methodology, and "More to the point, *problems* of all sorts (including empirical ones) *arise within a certain context of inquiry* and are partly defined by that context." Laudan (1977, pg. 15). Within our framework, the methodological bridge accounts for the connection between a conceptual world and the external world. We described each side of the methodological bridge, measurement and theoretical modeling, each of which produce a target, but we have not addressed how these targets "meet" or align. This question is itself and separate concern from those discussed so far.

2.3 Commensurability

While a measurement model anchors to the external world, and a theoretical model anchors to a conceptual world, there is no guarantee that the targets of these models necessarily align. In particular, measurement and theory do not tell us whether the the two "sides" of the methodological bridge connect, or for that matter, whether they should. Bueno de Mesquita and Tyson (2020) formulate the question by asking: "When do the estimates generated by actual research designs correspond to quantities of theoretical interest?" Outside of our bridge analogy, Bueno de Mesquita and Tyson (2020) and Ashworth et al. (2021) refer to the lack of alignment between the targets of the theory and measurement models as the **commensurability problem**. We depict the commensurability problem in Figure 2.

The commensurability question in our framework is fundamentally about a measurement model specifying an estimand, and a theoretical model specifying an empirical target, that together connect the two sides of the methodological bridge. Our framework links this problem together with other problems related to theoretical and measurement considerations, and thus, calls for a more thorough unpacking of what goes into the estimand of a measurement model in terms of empirical measures. Distinguishing the external world from the

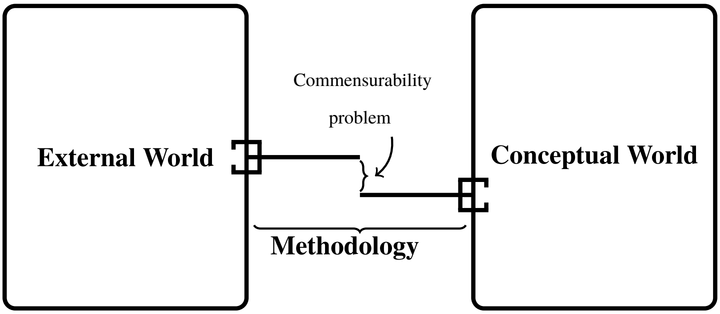

Figure 2 Commensurability problems emerge when the measurement and theory do not align within a research design

measurement model highlights how the selection of empirical measures and choice of an estimand are separate but related methodological decisions. For example, the average treatment effect (ATE) of a natural disaster on voters' *beliefs* about an incumbent is a different quantity from the ATE of a natural disaster on voters' pro-incumbent *votes*. So the targets of the measurement model, i.e., the estimands, are defined with respect to the empirical measures that comprise them, and different measures may change the estimand (or set of estimands) that is commensurable with a given theory.[12]

Since any pair of measurement model and theoretical model can have multiple targets, choosing between these targets reflects at least two decisions by the researcher. First, the subset of targets that is estimable is likely to be substantially smaller than the set of estimands that are commensurable with an empirical target from a theoretical model. Further concerns over properties of estimators may further shrink this set. In practice, researchers typically discriminate between estimands on the basis of estimations concerns, such as bias or efficiency, but neglect the importance of commensurability. Our framework stresses why ignoring the commensurability problem can lead to misleading inferences—the bridge between concepts and the external world is broken.[13]

Existing treatments of the commensurability problem focus exclusively on reduced-form approaches. Our characterization of theory as one side of the methodological bridge reflects the influence of criticisms of the reduced-form approach to causality that we describe in Chapter 1. Specifically, estimation and identification, in the absence of a theory, do not provide an anchor into any conceptual world, and without this anchor, whatever empirical findings are

[12] In general, there exist many (possibly infinite) commensurable estimands given a set of measures and a theory, and by a similar token, there exist infinite estimands that are not commensurable.

[13] See Slough (2022) for an applied example where the commensurability problem is always present for some of a research design's estimands because these estimands are undefined.

collected cannot speak to, or ascribe meaning to, the phenomena of interest. This broad point echoes critics of "atheoretical" reduced-form approaches to estimation and causal inference, which emphasize concerns of interpretation of the resultant estimands (Deaton, 2010; J. J. a. Heckman, 2010; Keane, 2010; Rust, 2010).

We stress that commensurability is only about the connection in the methodological bridge, and for this reason, it is a necessary, but insufficient, condition for a "good" methodological bridge. Returning the the accumulation of evidence, commensurability is an essential ingredient, as we will show in the coming chapters. What links together different studies—a uniting principle—is often conceptual. Commensurability problems break that link and undermine any uniting principle. This is the subject of the next chapter.

2.4 Structural and reduced-form approaches

Questions of commensurability are applicable to—and essential for—two common empirical approaches which differ in their methodological bridge. Any discussion of existing approaches to evidence accumulation must allow for both reduced-form and structural research designs. Whereas the commensurability problem asks *if* the estimand aligns a measurement model with a theoretical model, thereby closing the bridge, it can be solved in different ways by using different types of estimands and empirical targets.

In reduced-form research, the presence of a mechanism is measured with an estimand resulting from a research design that admits a causal interpretation. In structural approaches, the presence of a mechanism is measured by estimating a model parameter associated with a mechanism (within the context of the model). In most reduced-form approaches, the estimands are some type of aggregate causal effect, such as the average treatment effect (ATE), the average treatment effect on the treated (ATT), or a quantile treatment effect. By contrast, in structural work, the quantities of interest are parameters that are associated with some structural model (Goldberger, 1972). The distinction between structural and reduced-form research designs is of a question of *where* theory meets measurement on the methodological bridge. Consequently, each approaches the commensurability problem differently, as we depict in Figure 3.

Structural and reduced-form approaches can be distinguished in our framework by comparing where the two sides of the methodological bridge meet. We illustrate each in Figure 3, emphasizing where each deals with the commensurability problem, detailing the reliance of each on a theoretical model and a measurement model. A structural approach relies heavily on a theoretical model that anchors the bridge to a conceptual world. In this sense, structural approaches

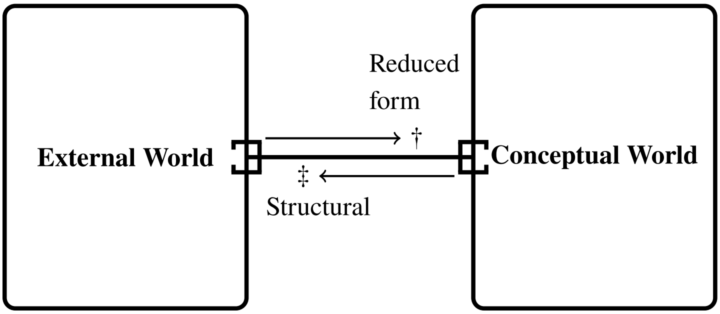

Figure 3 Depiction of structural and reduced form designs

cover the majority of the bridge between the external world and a conceptual world with a theoretical model. In Figure 3, with the structural research design, denoted ‡, much more of the methodological bridge is comprised by theory. This occurs because a theoretical model is doing a majority of the methodological work. Critically, structural approaches cannot directly "penetrate" the external world, and thus, still require a measurement model to connect with the external world.

A structural model ultimately contains a number of assumptions, both explicit and implicit, about the external world. These assumptions serve to constrain how the external world presents data through the measurement model. These constraints on the data that the external world can present are precisely what the heavy reliance on a structural theory in the methodological bridge entails. In particular, the structural model presents a version of the external world, and then under the presumption that the model is correct, proceeds to estimate parameters of that model. The structural approach is a descriptive exercise about the theoretical model, not the external world. Ultimately, the external world need not be subject to the constraints of the structural model.

Reduced-form approaches, in contrast, cover a majority of the methodological bridge with a measurement model. Here, theory does not constrain the data that the external world can supply to the same degree. Yet, this does not mean that reduced-form approaches are "atheoretical" or "model-free." Quite the contrary, as our framework emphasizes, a measurement model cannot directly connect to the conceptual world, because an estimand only gains meaning through some kind of interpretation. The reduced-form approach is a descriptive exercise about the measurement model, not a conceptual world, which is not subject to the constraints of the measurement model.

The commensurability problem is fundamental to connecting concepts with the external world, and its presence emphasizes that an empirical study of a substantive question can never fully dispense with a theoretical model or a measurement model. The examples of the structural and reduced-form approaches starkly illustrate this feature of empirical research. The commensurability problem shows that *the methodological bridge can never be completely comprised of a measurement model or a theoretical model.*

2.5 Our Framework and Evidence Accumulation

Our framework—highlighting the external world, conceptual worlds, and the methodological bridge that connects them—thus spans multiple approaches to empirical research in the social sciences. To this point, for the purposes of exposition of our framework, we have focused on a single study, with a single conceptual world and one methodological bridge. Consider the case when there are multiple studies, that might be combined into a meta-study. This makes comparison across the studies a central consideration. Our model of empirical research provides a framework to discuss issues of evidence accumulation. We now use our framework to better understand meta-studies.

The goal of accumulating evidence is to learn about the generality of phenomena in the external world. The first consideration is whether measurement models permit measurement of the same phenomena in different settings or studies. Figure 4 illustrates this with two distinct bridges between the external and a conceptual world. When two distinct measurement models affix to the external world in different ways, and thus, may pick up on the same phenomena in the external world differently.

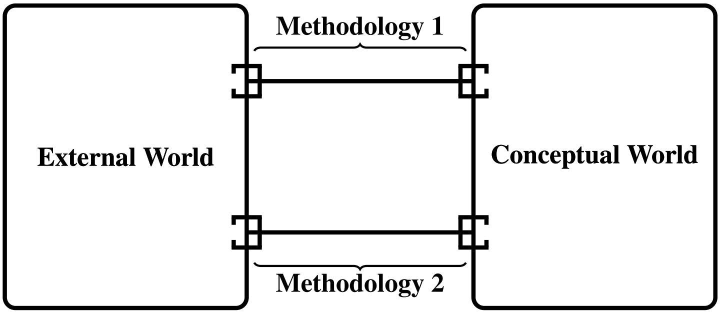

Figure 4 Two constituent studies of a meta-analysis do not target the same quantities.

A second consideration is whether there reason to believe that a mechanism is present, or active, across two studies. Put differently, is there a theoretical reason suggesting that a common concept unites different studies? This question is ultimately about whether a single set of concepts can be organized in the same way across studies. When studies are not united by a common set of concepts, there is no reason to expect that mechanisms operate in the same way across studies. Consequently, the decision to compare or combine studies relies on some conceptual world uniting phenomena in two or more settings or studies. Most efforts to accumulate evidence take this conceptual world as given, and this gives rise to a uniting principle in meta-studies.

3 Uniting Assumptions and Accumulation

The central goal of evidence accumulation is to build an evidential case about a general social phenomenon: whether it exists, how general it is, and to learn about its properties. Part of this process is about concept formation, i.e., is a mechanism conceptualized generally enough for evidence accumulation to be well-posed. But a large part of this process relies on a theoretical understanding of the cross-study environment, i.e., what is the meta-context that holds together all the individual contexts from individual studies? In this chapter, we show that any effort at evidence accumulation relies on a **uniting principle**, which details how the different studies under consideration in a meta-study are brought together. Ultimately, a uniting principle is a theoretical property and its credibility relies on substantive, rather than statistical, arguments.

To see the importance of a credible uniting principle, consider two hypothetical studies. Study #1 examines the contact hypothesis forwarded by Allport (1954) by estimating the effects of intergroup contact on attitudes toward outgroups. To study this, researchers create a new sports league and randomly assign participants to ingroup-only versus mixed teams (e.g., Asimovic, Ditlmann, & Samii, 2022; Lowe, 2021; Mousa, 2020). After the season, they construct an index measure that summarizing players' (participants') attitudes toward outgroups. Study #2 examines whether voters punish corrupt politicians when they learn of corruption. To study this, researchers randomly assign voters or communities to receive information on the results of government audits of local governments (e.g., Arias, Larreguy, Marshall, & Querubin, 2022; Boas, Hidalgo, & Melo, 2019; Ferraz & Finan, 2008). They then measure vote choice for the incumbent. Under standard assumptions, the ATE (and the intent-to-treat effect) is identified. Further, researchers can select an unbiased estimator of these effects.

Do we learn anything about a phenomenon from comparing or combining the ATEs of intergroup contact on outgroup attitudes and of corruption information on pro-incumbent voting? We argue that we do not.[14] On one hand, it is very simple to compare or combine these estimates using off-the-shelf hypothesis tests or estimators. Both experiments should produce real-valued estimates of the ATE. As such, these tests or estimators cannot tell us about a general substantive phenomenon because there is not a theoretical mechanism that unites these studies. Instead, researchers must rely on a convincing theoretical argument to unite these two studies and provide a rationale for the tests and

[14] Of course, if one gathers enough estimates, we can ask some methodological or meta-scientific questions. For example, we can look for evidence of publication bias by examining the distribution of T-statistics across unrelated published studies (Brodeur, Cook, & Heyes, 2020; A. Gerber & Malhotra, 2008). But this research does not speak to evidence accumulation.

estimators that are used. Thus, the united principle is about uniting conceptual worlds.

In this chapter, we articulate the necessity of *uniting principles* for the accumulation of evidence. These are assumptions about the conceptual world that justify comparing or combining the results of multiple studies. We then formulate the uniting principles in three common approaches to evidence accumulation. We show that these unifying principles differ across different approaches to accumulation.

3.1 Uniting Principles

Our examples of an experiment on the contact hypothesis, and one on electoral accountability, shows that although it is possible to combine evidence collected in two places, it is not always sensible to do so, as there is no theoretical motivation to unite these studies. This observation is quite obvious, but only because we picked examples that were very clearly unrelated.[15] Not all cases of evidence accumulation are so clear cut.

As another example, take two different studies on information and electoral accountability in different countries—e.g., in Mexico and Brazil—which may or may not tap into some common concepts or theoretical link into the conceptual world. In the latter case, Dunning et al. (2019) and Izzo, Dewan, and Wolton (2020) disagree about the conditions under which estimates from multiple accountability experiments can be compared or combined. Dunning et al. (2019) conduct a meta-study premised (implicitly) on the idea that common mechanisms related to voter updating are apt to be present in constituent contexts.[16] Izzo et al. (2020) argue that when circumstances (e.g., the state of the economy) vary, there is no reason to believe that the same voter updating mechanisms are activated across contexts. This means that we should not expect the same treatment effects to manifest across contexts.

The contrast between the arguments of Dunning et al. (2019) and Izzo et al. (2020) suggests that researchers' theoretical model about the relationship between constituent studies, or between a study and other hypothetical studies, are crucial to the exercise of evidence accumulation. When the goal is to learn about some phenomenon that is common across a number of contexts, it is critical that these contexts can be united through a common conceptual world, and even better if their theoretical models can be united. Combining, comparing, or extrapolating empirical findings invokes a set of uniting principles, whether they are explicitly formulated or not, and what is learned about the generality of

[15] It is straightforward to come up with even sillier examples.

[16] **?** conduct a meta-analysis of six information and accountability experiments.

phenomena from meta-studies depends centrally on what theoretical assumptions are invoked in the uniting principles.

When conducting a exercise in evidence accumulation, uniting principles specify two properties of the empirical targets across studies:

1. **Common Conceptual World**: Targets in different studies connect to a common concept or set of concepts.
2. **Common Methodological Bridge**: Targets in different studies are related theoretically in some known or assumed way.

The first property of uniting principles is straightforward—it requires there be a common conceptual world that unites across different studies. If there is no conceptual link between studies—as in the contact hypothesis and electoral accountability example—there is no reason to believe that comparing or combining targets can teach us anything about the external world. The second property is more subtle but equally important. It requires not only that different studies make reference to the same concepts, but that they are articulated in a common way—which is about the theoretical model in the methodological bridge. When this second property is lacking, then the theoretical relationship about how empirical targets across studies relate to each other is left unspecified, and we cannot know whether different studies are speaking to the same operationalization. There is no doubt that treatments of information experiments by Dunning et al. (2019) and Izzo et al. (2020) emphasize similar sets of concepts, satisfying property #1, but they disagree on whether the empirical targets in different studies relate to each other, thereby departing on property #2.

Uniting principles are about the relationship across and between studies (realized or hypothetical), and present an orthogonal set of concerns apart from those about within-study issues like identification, estimation, or commensurability. However, we cannot accumulate evidence from multiple studies without knowledge of how these studies are related to each other, and thus, the exercise of accumulating evidence is necessarily less “agnostic” than single study settings. While every example of evidence accumulation (e.g., replication, meta-analysis, etc.) necessarily invokes a set of uniting principles, they are rarely articulated. This creates problems because it leaves the scope and interpretation of such studies’ finding unclear. In the remainder of this chapter, we outline three common approaches to evidence accumulation. We then unearth the relevant uniting principles for each approach, and importantly, show how these uniting principles vary across different studies.

Approach	Research design	Meta-study?	Applied Examples
Comparing	Replication (direct or conceptual)	✓	Camerer et al. (2018); Open Science Collaboration (2015); Raffler, Posner, and Parkerson (2020)
Combining	Meta-analysis	✓	A. Banerjee et al. (2015); Blair et al. (2021); Dunning et al. (2019); Slough et al. (2021)
Extrapolating	(No standardized name)	–	Dehejia, Pop-Eleches, and Samii (2021)

Table 1 Three approaches to accumulation.

3.2 Three Approaches to Evidence Accumulation

There are multiple ways that one can accumulate evidence, and we focus on the three approaches which are most common: comparing, combining, and extrapolating. The most straightforward application of *comparing* is a replication study. However, the same logic that goes into formally comparing studies can be found more broadly in empirical studies. For example, when researchers in a single-setting study contextualize their estimates to (arguably) related findings in the literature, they engage in a similar comparing effort. Accumulation through comparing requires at least two studies that measure an effect of interest. Second, meta-analyses, in all of its various forms, is a case of *combining*. Specifically, meta-analysis takes as an input data or estimates from a collection of multiple studies that are united by a common structure. Last, when someone uses treatment effect estimates from a single study to make inferences abouta treatment effect in a grand population, or another setting, they are *extrapolating* Extrapolating differs from comparing or combining because it typically requires only a single study as an input.[17] Consequently, while comparing and combining involve a meta-study, extrapolating does not. Extrapolation-based approaches have been the subject of more recent *statistical* developments relative to comparison- or combination-based approaches. However, these methods are not yet widely utilized in applied work. We summarize these three approaches to accumulation of evidence in Table 1.

[17] Note that these methods are typically validated using existing meta-studies for which there are estimates from multiple settings.

3.2.1 Comparing

The simplest approach to accumulating evidence is to take two or more studies, i.e., a meta-study, and look to see if the same effect arises in multiple settings. This outlines the general logic of a replication, which is well known among experimentalists, but the comparing approach extends well beyond formal replications in experimental settings. Specifically, the logic of comparing often finds its way into literature reviews , at least informally, as a way of assessing the evidence accumulated over a particular research topic over time, albeit without a formal statistical test. Heuristic comparisons of the form: "author A's study finds that X increases Y, whereas we find no evidence that X increases Y" invoke the comparing logic (see Slough & Tyson, 2021, p. 3). Another common formulation of comparing checks for *sign-congruence* by looking across studies to see if their measured effects yield the same sign.

Comparing facilitates learning about general substantive phenomena whenever there is reason to believe that effects in different studies are produced by the same mechanism. As such, discussions of concepts or theory should focus on those mechanisms.[18] Put differently, what are the uniting principles that underlie a comparing approach, and are they different depending on the severity of the comparison (cardinal vs sign)?

The uniting principle invoked by a comparing approach is that the same mechanism arises in more than a single setting. In particular, that the mechanism of interest is something that can motivate phenomena in multiple places and at different times. This is the weakest uniting principle among these approaches because it says nothing about whether the mechanism has the same effects across settings, as this might naturally depend on how those effects are assessed (i.e., what is the research design). For clarity we call Θ the set of setting where a particular mechanism is active, where the boundary and properties of this set entail the *scope conditions* of a mechanism. Moreover, two mechanisms will be associated with different sets as will combinations of mechanisms.

The uniting principle of scope is (often implicitly) invoked in studies that compare estimated treatment effects in different settings. A study's setting includes features of the population and features of the environment that are relevant to measuring the effect of interest (see Slough and Tyson (2022)). The features of the population could include subject attributes or behavioral types that vary across the population (Wilke & Humphreys, 2020). The environment includes attributes of the context (e.g., institutional features), the temporal

[18] Note that while we refer to a single mechanism, our argument can also be applied to a bundle of mechanisms, so long as the contents of that bundle are not believed to vary across settings. In this case, the bundle of mechanisms can be interpreted as a compound mechanism.

context in which the effects are measured (Munger, 2021), and the realizations of stochastic features of the environment (termed “circumstances” by Izzo et al. (2020)). We will represent the setting of a given study as an element θ.

Given our emphasis on finding and measuring the effects of *general* mechanisms, several clarifications on the relationship between settings and mechanisms are necessary. Falleti and Lynch (2009), among others, argue that the same mechanism may produce different effects in different contexts. Within our framework, this concern conflates two distinct issues: mechanism activation and the relationship between multiple mechanisms. On one hand, mechanism—even those that are relatively generalizable—may not be activated within every setting. When a mechanism is inert (not activated), it should not produce a causal effect. Indeed, the mechanisms in our running electoral accountability example are, in principle, scoped to democracies where citizens can use the ballot box. We are interested in mechanisms that can be generalized even if they are not “universal” (within some time).

Second, there is debate about whether, conditional on activation, mechanisms produce different measured effects in different settings. In this sense, there is a question of whether setting conditions the effect that mechanisms produce (Falleti & Lynch, 2009). Sometimes this is is described as an “interaction” between setting and mechanism. Others contend that, conditional on activation, mechanisms deterministically produce the same effect on a given outcome [cites]. We adopt the latter view. We can reconcile this view with variation in observed effects across contexts by admitting multiple mechanisms. For example, if mechanism 1 presents in settings θ' and θ'' but mechanism 2 presents only in θ'', then we should not (necessarily) observe the same treatment effects in both settings. This does not mean that mechanism 1 “interacts” with setting. Nor does it preclude complementarity or substitutability between the mechanisms.

This discussion of mechanisms clarifies some properties—and possible limitations—of this uniting assumption. Specifically, the set Θ includes settings where the same unique mechanism or unique bundle of mechanisms is believed to be activated. It is worthwhile to consider the consequences of relaxing thes uniting assumption. Suppose, for example, that we had two estimates of a treatment effect from two settings, θ' and θ''. While mechanism 1 is believed to be activated in both settings, mechanism 2 is only believed to be activated in setting θ''. We then test the null hypothesis that the treatment effects are equivalent. If we were to reject the null hypothesis of equivalence, it could be because we were wrong that mechanism 1 is present in both contexts and our uniting assumption was wrong. But it could also be that our uniting assumption was correct and mechanism 2 yielded a different treatment effect in setting θ''. Ultimately, without additional assumptions that are stronger than our uniting assumption, we cannot distinguish between these possibilities. Moreover, if we were to fail to reject the null hypothesis, it could be the case that mechanism 2 was actually not activated in θ'' and our uniting assumption is correct. Or it could be because deviations from our uniting assumption offset sufficiently the effect of mechanism 2. Again, we cannot distinguish between these possibilities without additional assumptions.

While the uniting assumption that underpins accumulation through the comparison of measured effects may seem limiting, it is a *weaker* assumption than those that underpin other approaches to accumulation.

3.3 Combining

The combining approach is a kind of meta-study in which different effects, measured in multiple settings, are combined in an effort to construct a measure of the "aggregate" or "underlying" effect. The standard tool used in combining approaches is meta-analysis. The workhorse meta-analysis estimators—fixed- and random-effects models—although not typically associated with a structural model, when applied to meta-studies, they constitute a structural approach because of how they handle the environment across studies. Both the fixed- and random-effects models are derived from a common multi-level structural model, and we will center our analysis on that common model.

3.4 The Role of External Validity

Now that we have clarified the technologies for accumulation, we turn to the concepts of external validity invoked by approaches that compare, combine, and extrapolate.

4 Concepts of External Validity

5 Practical Applications

References

Abdul Latif Jameel Poverty Action Lab. (2022). Retrieved 23 September 2022, from `https://www.povertyactionlab.org/`

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

Allport, G. (1954). *The nature of prejudice*. Boston: Addison-Wesley.

Angrist, J. D., & Pischke, J.-S. (2008). *Mostly harmless econometrics: An empiricist's companion*. Princeton university press.

Angrist, J. D., & Pischke, J.-S. (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.

Arias, E., Larreguy, H., Marshall, J., & Querubin, P. (2022). Priors rule: When do malfeasance revelations help or hurt incumbent parties? *Journal of the European Economic Association*, *20*(4), 1433–1477.

Aronow, P. M., & Miller, B. T. (2019). *Foundations of agnostic statistics*. New York: Cambridge University Press.

Ashworth, S. (2012). Electoral accountability: Recent theoretical and empirical work. *Annual Review of Political Science*, *15*(1), 183–201.

Ashworth, S., Berry, C. R., & de Mesquita, E. B. (2021). *Theory and credibility: Integrating theoretical and empirical social science*. Princeton University Press.

Asimovic, N., Ditlmann, R., & Samii, C. (2022). *Estimating the effect of intergroup contact over years: Evidence from a youth program in israel.* (Available at `https://tinyurl.com/54njuby4`)

Banerjee, A., Duflo, E., Goldberg, N., Karlan, D., Osei, R., Parrienté, W., ... Udry, C. (2015). A multifaceted program causes lasting progress for the very poor: Evidence from six countries. *Science*, *348*(6236).

Banerjee, A. V., & Duflo, E. (2009). The experimental approach to development economics. *Annual Review of Economics*, *1*, 151-178.

Blair, G., Cooper, J., Coppock, A., & Humphreys, M. (2019). Declaring and diagnosing research designs. *American Political Science Review*, *113*(3), 838–859.

Blair, G., Weinstein, J. M., Christia, F., Arias, E., Badran, E., Blair, R. A., ... Wilke, A. M. (2021). Community policing does not build citizen trust in police or reduce crime in the global south. *Science*, *374*(6571), eabd3446.

Boas, T. C., Hidalgo, F., & Melo, M. (2019). Norms versus action: Why voters fail to sanction malfeasance in brazil. *American Journal of Political*

Science, *63*(2), 385-400.

Bogen, J., & Woodward, J. (1988). Saving the phenomena. *The philosophical review*, *97*(3), 303–352.

Bond, R. M., Fariss, C. J., Jones, J. J., Kramer, A. D. I., Marlow, C., Settle, J. E., & Fowler, J. H. (2012). A 61-million-person experiment in social influence and political mobilization. *Nature*, *489*, 295-298.

Brodeur, A., Cook, N., & Heyes, A. (2020). Methods matter: p-hacking and publication bias in causal analysis in economics. *American Economic Review*, *110*(11), 3634-3600.

Bueno de Mesquita, E., & Tyson, S. A. (2020). The commensurability problem: Conceptual difficulties in estimating the effect of behavior on behavior. *American Political Science Review*, *114*(2), 375–391.

Camerer, C. F., Dreber, A., Forsell, E., Ho, T.-H., Huber, J., Johannesson, M., . . . Wu, H. (2016). Evaluating replicability of laboratory experiments in economics. *Science*, *351*(6280), 1433-1436. doi: 10.1126/science.aaf0918

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T.-H., Huber, J., Johannesson, M., . . . Wu, H. (2018). Evaluating the replicability of social science experiments in nature and science between 2010 and 2015. *Nature Human Behaviour*, *2*(9), 637–644. Retrieved from `https://doi.org/10.1038/s41562-018-0399-z` doi: 10.1038/s41562-018-0399-z

Cole, S. R., & Stuart, E. A. (2010). Generalizing evidence from randomized clinical trials to target populations: The actg 320 trial. *American Journal of Epidemiology*, *172*(1), 107-15.

Cunningham, S. (2021). *Causal inference: The mixtape*. Yale University Press.

de la O, A., Green, D. P., John, P., Goldszmidt, R., Lenz, A.-K., Valdivia, M., . . . Hyde, S. (2021). Fiscal contracts? a six-country randoized experiment on transaction costs, public services, and taxation in developing countries. , *Working paper*. Retrieved from `\url{https://nikhargaikwad.com/resources/De-La-O-et-al_2021.pdf}`

Deaton, A. (2010). Instruments, randomization, and learning about development. *Journal of economic literature*, *48*(2), 424–55.

Deaton, A., & Cartwright, N. (2018). Understanding and misunderstanding randomized controlled trials. *Social Science & Medicine*, *210*, 2–21.

Dehejia, R., Pop-Eleches, C., & Samii, C. (2021). From local to global: External validity in a fertility natural experiment. *Journal of Business & Econonic Statistics*, *39*(1), 217-243.

Dunning, T. (2016). Transparency, replication, and cumulative learning: What experiments alone cannot achieve. *Annual Review of Political Science*, *19*, 541-563.

Dunning, T., Grossman, G., Humphreys, M., Hyde, S., Mcintosh, C., Nellis, G., . . . Sircar, N. (2019). Voter information campaigns and political accountability: Cumulative findings from a preregistered meta-analysis of coordinated trials. *Science Advances*, *5*(7).

Egami, N., & Hartman, E. (2020). Elements of external validity: Framework, design, and analysis. *Design, and Analysis (June 30, 2020)*.

Evidence in Governance and Poltics. (2022). Retrieved 23 September 2022, from `https://egap.org/about/`

Falleti, T. G., & Lynch, J. F. (2009). Context and causal mechanisms in political analysis. *Comparative Political Studies*, *42*(9), 1143-1166.

Fariss, C. J. (2014). Respect for human rights has improved over time: Modeling the changing standard of accountability. *American Political Science Review*, *108*(2), 297–318.

Ferraz, C., & Finan, F. (2008). Exposing corrupt politicians: The effects of brazil's publicly released audits on electoral outcomes. *Quarterly Journal of Economics*, *123*(2), 703-745.

Flache, A., & Macy, M. W. (2013). The weakness of strong ties: Collective action failure in a highly cohesive group. , 19–44.

Frege, G. (1948). Sense and reference. *The philosophical review*, *57*(3), 209–230.

Gailmard, S. (2021). Theory, history, and political economy. *Journal of Historical Political Economy*, *1*(1), 69–104.

Gerber, A., & Malhotra, N. (2008). Do statistical reporting standards affect what is published? publication bias in two leading political science journals. *Quarterly Journal of Political Science*, *3*(3), 313-326.

Gerber, A. S., & Green, D. P. (2012). *Field experiments: Design, analysis and interpretation*. New York: W. W. Norton & Company.

Giere, R. N. (2010). *Scientific perspectivism*. University of Chicago press.

Goldberger, A. S. (1972). Structural equation methods in the social sciences. *Econometrica*, 979–1001.

Goodman, N. (1978). *Ways of worldmaking* (Vol. 51). Hackett Publishing.

Hacking, I. (1990). *The taming of chance*. Cambridge University Press.

Hacking, I. (2006). *The emergence of probability: A philosophical study of early ideas about probability, induction and statistical inference*. Cambridge University Press.

Hamming, R. W. (1980). The unreasonable effectiveness of mathematics. *The American Mathematical Monthly*, *87*(2), 81–90.

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

Heckman, J. J. a. (2010). Comparing iv with structural models: What simple iv can and cannot identify. *Journal of Econometrics*, *156*(1), 27-37.

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

Huber, J. D. (2017). *Exclusion by elections: Inequality, ethnic identity, and democracy*. New York: Cambridge University Press.

Imbens, G. 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, G. W., & Angrist, J. D. (1994). Identification and estimation of local average treatment effects. *Econometrica*, *62*(2), 467–475.

Imbens, G. W., & Wooldridge, J. M. (2009). Recent developments in the econometrics of program evaluation. *Journal of economic literature*, *47*(1), 5–86.

Izzo, F., Dewan, T., & Wolton, S. (2020). Cumulative knowledge in the social sciences: The case of improving voters' information. *Available at SSRN 3239047*.

Jerven, M. (2013). *Poor numbers*. Cornell University Press.

Keane, M. P. (2010). Structural vs. atheoretic approaches to econometrics. *Journal of Econometrics*, *156*, 3-20.

Kern, H. L., Stuart, E. A., Hill, J., & Green, D. P. (2016). Assessing methods for generalizing experimental impact estimates to target populations. *Journal of Research on Educational Effectiveness*, *9*, 103-127.

King, G., Keohane, R. O., & Verba, S. (1994). *Designing social inquiry: Scientific inference in qualitative research*. Princeton university press.

Kitchin, R. (2014). *The data revolution: Big data, open data, data infrastructures and their consequences*. Sage.

Kripke, S. A. (1972). *Naming and necessity*. Springer.

Laudan, L. (1977). *Progress and its problems: Towards a theory of scientific growth*. University of California Press.

Leamer, E. E. (1983). Let's take the con out of econometrics. *The American Economic Review*, *73*(1), 31–43.

Lowe, M. (2021). Types of contact: A field experiment on collaborative and adversarial caste integration. *American Economic Review*, *111*(6), 1807-1844.

Manski, C. F. (1995). *Identification problems in the social sciences*. Harvard University Press.

Mayo, D. G. (1996). *Error and the growth of experimental knowledge*. University of Chicago Press.

Mousa, S. (2020). Building social cohesion between christians and muslims through soccer in post-isis iraq. *Science*, *369*(6505), 866-870.

Munger, K. (2021). Temporal validity. Retrieved from `https://osf.io/4utsk/`

Neyman, J. (1923). Sur les applications de la theorie des probabilites aux experiences agricoles: essai des principes (masters thesis); justification of applications of the calculus of probabilities to the solutions of certain questions in agricultural experimentation. excerpts english translation (reprinted). *Statistical Science*, *5*, 463–472.

Olson, M. (1965). *The logic of collective action.* Harvard University Press.

Open Science Collaboration. (2015). Estimating the reporducibility of psychological science. *Science*, *349*(6251), 1-8.

Pearl, J., & Bareinboim, E. (2011). Transportability of causal and statistical relations: A formal approach. In *Twenty-fifth aaai conference on artificial intelligence.*

Pearl, J., & Bareinboim, E. (2014). External validity: From do-calculus to transportability across populations. *Statistical Science*, *29*(4), 579–595.

Porter, T. M. (1996). *Trust in numbers.* Princeton University Press.

Putnam, H. (1974). Meaning and reference. *The journal of philosophy*, *70*(19), 699–711.

Putnam, H. (1981). *Reason, truth and history.* Cambridge University Press.

Raffler, P., Posner, D. N., & Parkerson, D. (2020, October). *Can citizen pressure be induced to improve public service provision?* (Working paper, Harvard University)

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, *66*(5), 688.

Rust, J. (2010). Comments on "structural vs. atheoretic approaches to econometrics" by michael keane. *Journal of Econometrics*, *156*, 21-24.

Ryle, G. (1949). *The concept of mind.* University of Chicago Press 2002.

Salmon, W. C. (1984). *Scientific explanation and the causal structure of the world.* Princeton University Press.

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

Schelling, T. C. (1960). *The strategy of conflict.* Harvard university press.

Shadish, W., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference.* Houghton Mifflin Boston, MA.

Slough, T. (2022). Phantom counterfactuals. *American Journal of Political Science,forthcoming.* Retrieved from `http://taraslough.com/assets/pdf/phantom\textunderscorecounterfactuals.pdf`

Slough, T., Rubenson, D., Levy, R., Rodriguez, F. A., del Carpio, M. B., Buntaine,

M. T., . . . Zhang, B. (2021). Adoption of community monitoring improves common pool resource management across contexts. *Proceedings of the National Academy of Sciences*, *10.1073*, 1-10.

Slough, T., & Tyson, S. A. (2021). *Conceptual replication under external validity.* (Working paper, New York University)

Slough, T., & Tyson, S. A. (2022). External validity and meta-analysis. *American Journal of Political Science*, *Forthcoming*.

Venn, J. (1888). *The logic of chance: An essay on the foundations and province of the theory of probability, with especial reference to its logical bearings and its application to moral and social science, and to statistics*. Macmillan.

Wilke, A., & Humphreys, M. (2020). Field experiments, theory, and external validity. In *Sage handbook of research methods in political science and international relations* (pp. 1007–35). SAGE London.

Woodward, J. (2005). *Making things happen: A theory of causal explanation*. Oxford university press.

Acknowledgements

Thanks to...