Field Experiments, Theory, and External Validity

Anna Wilke[*] *and Macartan Humphreys*[†]

July 2019

Abstract

Critics of field experiments lament a turn away from theory and criticize findings for weak external validity. In this chapter, we outline strategies to address these challenges. Highlighting the connection between these twin critiques, we discuss how structural approaches can both help design experiments that maximize the researcher's ability to learn about theories and enable researchers to judge to what extent the results of one experiment can travel to other settings. We illustrate with a simulated analysis of a bargaining problem to show how theory can help make external claims with respect to both populations and treatments and how combining random assignment and theory can both sharpen learning and alert researchers to over-dependence on theory.

[*] `amw2229@columbia.edu` Columbia University. We thank Donald Green, James Fearon, Alan Jacobs, and Richard Nielsen for their generous comments on earlier versions of this paper. This paper is written as a contribution to the *Handbook of Research Methods in Political Science and International Relations* edited by Luigi Curini and Robert J. Franzese, Jr.; we thank them for their guidance.

[†] `mh2245@columbia.edu` Columbia University and WZB Berlin Social Science Center.

Contents

1 Introduction

In a recent study, Yeh et al. (2018) report on the first randomized control trial that tests whether there are health benefits from wearing parachutes when jumping out of airplanes and helicopters. The authors find no evidence of benefits and their estimate of no effect is very precise. A potential weakness of the study is that, in order to protect human subjects, the airplane and helicopter used for the trial were small, stationary, and grounded. For critical readers this detail might raise flags about whether we can generalize from this study to other applications of interest.

Of course, Yeh and colleagues meant their study as a joke. It is *obvious* that you cannot learn anything about effects in realistic situations from this experiment. We think, however, that it is not entirely obvious why it is obvious. The experiment lacked external validity, perhaps, because the "target" applications of interest involve planes that are in flight and far from the ground. The sample of jumps is not *like* the population of jumps out of an aircraft that we are interested in. Yet we might not have that concern were the plane in the air but not in flight—for example if it were attached to a large crane, even though in this case too, the study would not be like the target applications of interest. Why would we be less concerned? The reason is that our determination of "likeness" depends on a prior model of how things work—in this case that falling at speed is a key mechanism that is interrupted by parachutes and that speed depends on the height from which you jump. So, a theory is used to assess external validity. This chapter is about such theories and the role they play in assessing external validity.

We are motivated by persistent concerns with field experimental approaches, which have seen an explosion in political science in the last fifteen years (Grose 2014; Druckman et al. 2011). Core concerns repeatedly raised are that experimental results lack external validity and are disconnected from theory. As in the parachute example, these two concerns are deeply interrelated. Put differently, the charge is that political scientists are doing parachute experiments from stationary planes without knowing it.

In the following sections, we describe these concerns in more detail and review strategies to address them. We begin by introducing a stylized example that we will draw on throughout the chapter and by clarifying our usage of the terms theory and external validity. We then review challenges that arise from a weak connection to theory and limited external validity and describe approaches, including recent innovations, to address these. Finally, we extend our example to illustrate opportunities and risks of parametric structural estimation with experimental data, an approach that links estimation closely to theoretical models and that, in principle, allows for inferences that go beyond those available with design-based approaches alone.

2 Preliminaries

2.1 A running example

Throughout this chapter, we will make use of a stylized example of a fictitious field experimental study that seeks to explain bargaining outcomes. For concreteness, we imagine a study of the bargaining process between taxi drivers and customers who negotiate taxi fares. See Michelitch (2015) and Castillo et al. (2013) for existing studies of this kind.

The study aims to explain why some taxi customers have to pay more to their drivers than others. We start with a simple theory which says that taxi fares are determined by three variables: whether the customer makes the first offer, how many offers and counteroffers can be made before bargaining breaks down (the number of bargaining rounds), and the behavioral "type" of the customer.

Whether the customer makes the first offer might be thought of as linked to the customer's identity, assertiveness, or skills. Skilled negotiators, for example, may be more likely to approach the driver with an offer. How long bargaining can continue may depend on contextual factors such as the competitiveness of the taxi market. Behavioral types can vary with some customers being rational decision makers who seek to maximize their gains from the bargaining process, and other customers seeking to follow established fairness norms. For simplicity, we presume that taxi drivers are always of the first type.

With this theory in mind, we imagine running a field experiment that focuses, first, on the effect of making the first offer. Suppose we recruit a sample of taxi customers, provide each of them with the same sum of money and ask them to negotiate a taxi fare for a certain distance. Taxi customers can keep whatever they do not spend on the ride. Say, we randomly assign half of taxi customers to approach the driver with an offer ("move first"), while the other half is instructed to ask the driver for a price ("move second"). Additionally we will imagine being able to control for how long bargaining can continue. We will assume players have common knowledge over endowments and types.

This example is certainly contrived but it has the advantage of being easily connected to simple and well understood theoretical models. This makes the example useful for demonstrating core ideas and for walking through the development and use of structural models.

2.2 What we mean by theory

Critics worry that a disconnect from theory limits the types of inferences that can be drawn from experimental research (Deaton and Cartwright 2018; Harrison 2014; Card, DellaVigna, and Malmendier 2011; Huber 2017).

What is theory? Conscious that the term is used to mean very different things in different research communities,[1] we will make use of a quite simple notion of theory: by theory we

[1] In their review of the role that theory plays in experimental studies published in top economics journals, for example, Card, DellaVigna, and Malmendier (2011) classify an experimental study as one that draws

will mean a set of general claims about the causal relations among a set of variables, from which claims about a case can be deduced under the supposition that the theory is true. For instance, if theory T says that all taxi customers who make the first offer pay less, then the specific claim “this customer will pay less if she makes the first offer” follows from the theory. A theory of this form may itself be deduced from a deeper theory, for example a claim about equilibrium offers in a class of bargaining games (of which we take the taxi game to be an instance).

Drawing on work by Pearl (2009) and the treatment in Humphreys and Jacobs (2017) we think it useful to distinguish between different “levels” of theory that differ in the specificity with which they describe the relations between variables.

Level 1: A statement of which variables are causally related to which other variables. An example might be a non-parametric (structural) equations model (NPSEM) which consists of three elements:[2]

- A set of unobserved background variables U
- A set of possibly observable variables V that are endogenous in the sense that they are determined by other variables in the model (their “parents”),
- A causal structure that specifies relations of conditional independencies between variables.

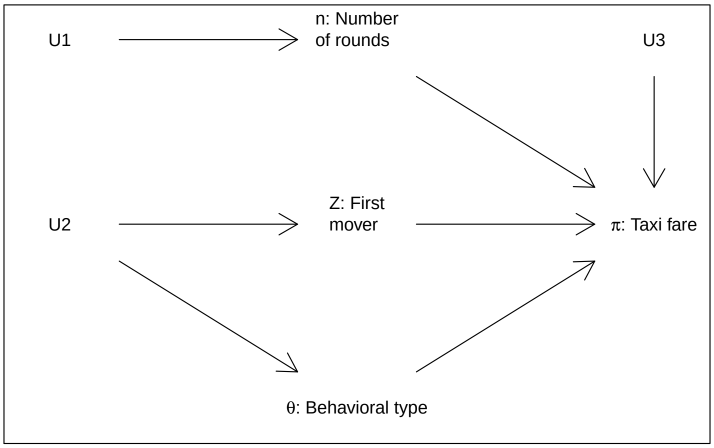

Figure 1: A directed acyclic graph (DAG) for a non-parametric bargaining model

A convenient way to display a Level 1 model is a directed acyclic graph (DAG). Figure 1 shows the DAG for the Level 1 version of our example theory of bargaining. In addition to the relationships between observed variables, the DAG makes explicit the role of unobserved

on theory only if the study explicitly includes mathematical expressions. Other accounts appear to have a less stringent view that encompasses informal statements about causal relationships derived from prior knowledge (Huber 2017).

[2] For a discussion of related alternatives see for example Robins (2003).

factors that are determined outside the model. There are unobserved factors U_3 that affect the taxi fare and unobserved factors U_2 which affect both, whether the customer makes the first offer and the customer's behavioral type. The Level 1 version of our theory also contains information about which variables do not affect each other. For example, the number of rounds for which bargaining may continue is determined by unobserved factors U_1, but these unobserved factors do not have a direct impact on any of the other variables in the model.

Note that although the theory involves many substantive claims it has no implications for effect sizes or even whether or not variables interact with each other to produce outcomes.

Level 2: Rather than simply specifying the qualitative structure of the causal relationships between variables, a theory may contain more quantitative statements about the functional form of these relationships. These might be statements about marginal effects, for example: "Making the first offer in taxi bargaining reduces the price that the customer has to pay." Alternatively, a Level 2 model may contain fully specified functional relations, $f = f_1, f_2, ..., f_n$, one for each endogenous variable V_i, that map values that can be taken by the parents PA_i of V_i and the set of unobserved exogenous variables U_i into values of V_i.

For instance, a parametric bargaining model might add to the Level 1 model the following functional relationship:

$$\pi_i = \theta_i(z_i\omega + (1 - z_i)(1 - \omega)) + (1 - \theta_i)\phi + u_3$$

where $\omega = \sum_{t=2}^{n}(-1^t)\delta^{t-1}$ is the taxi price predicted by the Rubinstein (1982) bargaining solution under the assumption that taxi bargaining follows an alternating offers protocol with t possible bargaining rounds, that customers and drivers act rationally and that the customer gets to make the first offer. This representation normalizes the "pie" that the customer and driver are bargaining over to 1. In the context of our example experiment, we can treat ω as the share of her endowment that a rational customer ends up paying to the driver. δ stands for a discount factor (see section 6 for more details). This Level 2 version of our theory implies that rational customers ($\theta_i = 1$) pay the price implied by the Rubinstein bargaining solution. This price depends on whether customers go first ($z_i = 1$) or second ($z_i = 0$). Non-rational customers ($\theta_i = 0$) insist on giving the driver some share ϕ of their endowment, irrespective of whether they go first or second. The last term, u_3, is a random disturbance.

This theory is not well motivated, but, as is well known, the behavior of rational types can be derived from a more complex level 2 model that fully specifies the extensive form of the bargaining game and allows for a richer characterization of optimal behavior (e.g. offers and responses in every period). We provide this motivation below.

Unlike a Level 1 theory, the Level 2 theory *does* make claims about the values of endogenous variables given the values of exogenous variables. However, this is not enough for the theory to make claims about the average effects of treatments.[3]

[3] The reason is that average effects of one variable can depend upon levels of another variable which depends on the distribution of exogenous variables. Consider the effect of "going first." This effect depends on the customer's behavioral type. Since the theory does not specify the distribution of these types in the

Level 3: A still more fully specified model might add assumptions about how the set of unobserved exogenous variables U is distributed. A theory that specifies all of these elements is called a **probabilistic causal model** by Pearl (2009).[4]

To continue with our example, a Level 3 model might add $\theta_i \sim \text{Bernoulli}(p = .5), Z_i|\theta_i = 1 \sim \text{Bernoulli}(p = .6), Z_i|\theta_i = 0 \sim \text{Bernoulli}(p = .4), n \sim \text{Binomial}(10, p = .5), U_3 \sim \text{Beta}(\alpha = 2, \beta = 2)$. Since we now have information about how the relevant variables are distributed, this model allows us to make statements about average causal effects.

Note that a Level 3 model implies a Level 2 model and a Level 2 model implies a Level 1 model, but the converse is not true. Note also that starting from a Level 1 model one could use data to develop a Level 2 or a Level 3 model.

2.3 What we mean by external validity

The external validity critique is that findings from field experiments may not tell us much about processes in contexts other than the one where the findings were generated (Lucas 2003). Many discussions of threats to external validity focus on whether the study population is sufficiently similar to a *target* population of interest. This framing gives rise to what might be the more useful term, "target validity." Can we make claims to a particular target of inference (Westreich et al. 2018)? Cronbach and Shapiro (1982) describe four dimensions along which studies may differ from their inferential targets: units, treatments, outcomes, and settings (collectively, "UTOS"). Using the language of causal models, we can distinguish three of these four dimensions. Units and contexts are equivalents in this language.

Let W denote an observed or unobserved collection of background variables, Y and V outcomes, and Z and X treatments. The question about units and contexts is:

- What can we infer about the distribution $\Pr(Y = y|Z = z, W = w')$ from our knowledge of $\Pr(Y = y|Z = z, W = w)$? For instance, what does an experiment in a place with competitive taxi markets tell us about the effect of moving first in taxi bargaining when taxi competition is low?

The treatments question is:

- What can we infer about the distribution $\Pr(Y = y|Z = z)$ from our knowledge of $\Pr(Y = y|X = x)$? For instance, what does an experiment on the effect of moving first in taxi bargaining on prices tell us about the effect of an increase in taxi competition on prices?

The outcomes question is:

population, we would not be able to make statements about the average effect of going first based on this model alone.

[4] One could imagine these distributions being added either to a non-parametric or a parametric causal model.

- What can we infer about the distribution $\Pr(V = v|X = x)$ from our knowledge of $\Pr(Y = y|X = x)$? For instance, what do we learn from the impact of moving first in taxi bargaining on taxi fares about the effect of moving first on the duration of taxi negotiations?

3 Concerns Around Theory and External Validity

3.1 Why worry about theory?

We can distinguish at least four worries that result from weak connections between theory and experimental research.

Two relate to the applicability of findings and two, more fundamentally, to the broader orientation of research.

Undefined scope: Units and Settings. As in the parachute example, without theory, researchers will not have the information needed to think through which units and settings are "like" those in study sites in order to licence the transportation of results.[5] The concern might be less severe if experiments use random samples from the target population of interest and produce homogeneous results within strata. But if such conditions do not hold, then absent theory, researchers may be at a loss about what broader inferences to make.

Undefined scope: Treatments and Outcomes. It is often not feasible to experimentally evaluate all relevant versions of a particular intervention or to examine all outcomes of interest. For example, field experiments tend to involve small-scale versions of interventions that are to be rolled out on a larger scale at a later stage. In such cases general equilibrium effects may mean that effects at one scale are very different from effects at another. Inferences then require extrapolation to conditions not studied by an experiment. Or, for example, a study might confirm that information provided in a particular way changes attitudes of voters but not have data on voter behavior. Making inferences to voting requires understanding how attitudes affect voting (for those for whom information affects attitudes).

Experimental results themselves do not give a handle on how to extrapolate out of sample or make inferences to other relations, but theory might.

These two worries illustrate the connection between theory and external validity. Weak connection to theory leaves open the question of whether and how results can be generalized.

The next two concerns relate to the questions being asked.

Restrictive estimands. Experimental research largely limits itself to estimating variations of the average treatment effect. By the magic of the linearity of expectations, random assignment lets one estimate average differences between outcomes in treatment and control conditions by looking at the differences in average outcomes in treatment and control groups. But many questions of interest are not summaries of average treatment effects. Consider, for

[5] See Mesquita and Tyson (2019) for a treatment of "commensurability" which can be used to assess whether research designs capture quantities that are of theoretical interest.

example, questions about so-called "causes of effects" as opposed to "effects of causes." A field experiment randomly assigns individuals to go first in bargaining. Randomization does not, in general, let researchers answer the question: "Knowing that individual A made the first offer, what is the probability that she would have paid a higher price had she not made the first offer?" Murtas, Dawid, and Musio (2017) show that although this quantity is not identified—and is largely ignored in experimental analysis—bounds around this probability can be tightened by drawing on theoretical knowledge about the variables that mediate the relationship between the treatment and the outcome of interest (even if such mediators are not observed).

The point of empirical work is to learn about theories. The point of research, critics argue, is to understand *how* things work. Thus the *goal* is to develop theory and so engaging in research while ignoring theory is missing the point. Experimentalists, in this view, are too often satisfied with "black-box" accounts that do not go beyond relations between variables. Deaton (2010), for instance, advocates for a more explicit focus on the usage of causally identified empirical work to test hypotheses that are derived from a lower-level theory in order to learn about the validity of the theory itself.

Subsequently, we will explore in more detail some of the ways in which researchers have responded to these worries.

3.2 Why worry about external validity?

Some of the concerns around external validity apply to all research. In general, empirical work does not ever draw on a random sample of the units to which inferences might be made—if only because we want to make inferences to future events from past events. The problem of making inferences from sample to populations is common to all case study research and the challenge of making inferences from the study of one treatment to another or one outcome to another is not specific to experiments.

However there are ways in which experiments are somewhat more vulnerable to the external validity critique than observational work. The core of the criticism is that, through experimentation, researchers alter the world to make it amenable to study and, in doing so, they create distance to target environments.

Experiments are strange. Insofar as experimenters control the conditions of an experimental study, they risk orchestrating environments that differ from the target environment in ways that researchers might not be aware of.

Experiments may, by design, hold relevant variables constant at atypical levels. For example, suppose we are interested in both the effect of going first in taxi bargaining and the effect of increasing the number of possible bargaining rounds. Say, we randomly assign taxi customers to go first and also vary for how long bargaining can continue. This experiment allows us to estimate the effect of additional bargaining rounds among customers who have been assigned to go first. Yet, this effect may differ from the effect among those who actually choose to go first outside our experiment. Voluntary first movers may be systematically different from other customers. A design that randomizes who goes first provides no information about who would have gone first outside the experiment and, as a consequence, makes it impossible to obtain effect estimates among this group without further assumptions.

In addition, *how* variables are controlled can also matter. The process of random assignment of a treatment in itself may have consequences. For instance, voters might not reward a politician for patronage if they know that patronage was distributed at random, precisely because rewards normally result from the information communicated by a transfer rather than by the transfer itself (see also the treatment in Mesquita and Tyson (2019)). Note that this is an exclusion restriction violation.

Randomization bias. Very often, neither the sites nor the subjects of experimental studies are randomly selected. Especially for experiments that depend on partnerships with government or other actors, site selection tends to be contingent on the willingness of partners to participate. Yet, governments that are willing to assess anti-corruption interventions, for example, may be fundamentally different from those that are not. The problem is akin to that in medical studies where subjects who are willing to participate tend to be those most likely to stick to regimens. See Allcott (2015) for an example of how bias can arise from non-random site selection.

Sample focused inferential strategies. External claims from experiments can sometimes be rendered difficult by the fact that common approaches to analyzing experimental data implicitly assume that the target of inference is the sample estimand, even if this is not always explicitly stated.

Randomization inference, for instance, can be used to calculate exact p-values — but only under the assumption that all variability comes from assignment processes within the sample and not from the selection of the sample itself. Similarly, at least under the assumption of constant effects, clustering standard errors at the level of treatment assignment can produce confidence intervals with the correct coverage. Yet, this approach implicitly assumes that study units have not been randomly sampled in a clustered way. Suppose for example, that a study randomly samples a set of schools and assigns an intervention on the classroom level. In this case, clustering might have to be performed at levels above the level of assignment—though this is not common practice (Abadie et al. 2017).

These are ways in which the tools of experimental analysis tend to orient researchers towards sample inference, even though, in principle, approaches that focus on population inference can certainly be employed with experimental data.

4 The Place of Theory in Experimental Research

In practice, experimentalists have employed a range of strategies to combine theory and experimental work (or not). We discuss six of those.

4.1 Strategy 1: Push Back – Experimentation obviates the need for theory

A first response to the critique that experimental research tends to be atheoretical is unapologetic. The absence of theory is a strength. As characterized by Heckman (1991), the ability to dispense with theory was, if anything, a motivation to engage in experimentation. That you can find out whether democratic institutions cause growth without having to assume a model of human behavior is remarkable, and to be celebrated The identification of average causal effects is not possible from observational data without a model that tells you which variables you should or should not condition on (Pearl 2009). An experiment removes much of this model dependence.[6]

This response leaves open the question of how experiments can be used to learn about theories. Moreover, despite the remarkable ability of experiments to identify average treatment effects under minimal assumptions, there are quantities that cannot be learnt from an experiment. Both topics are discussed in more detail below.

4.2 Strategy 2: Use theory as a helper for inference

One strategy for bringing experiments together with theory is to use theories to draw inferences that could not be drawn based on experimental data alone. Consider the following examples:

1. Transportation of results to other settings. Suppose, we conduct an experiment at a site A and we would like to use our treatment effect estimates to learn about the treatment effect at site B. Unless site A was randomly chosen from some population of sites that also contains site B, our experimental data alone do not speak to whether and how transportation from A to B is possible, unless additional assumptions are made. We discuss in section 5.6 how a causal model can help answer these questions and provide an example in section 6.

2. Predicting the effects of other treatments. Many causes of theoretical interest cannot be experimentally manipulated. Similarly, it is often prohibitively expensive to evaluate all policy-relevant variations of an intervention with an experiment. One solution is to make use of a structural causal model that serves as the basis for the extrapolation of estimates from a single, possibly small-scale experiment to the effect of a different intervention or of the same intervention at a larger scale. Todd and Wolpin (2006), for example, use a structural causal model to extrapolate estimates of the effect of a randomized school subsidy program in Mexico to the effects of similar programs with different subsidy schedules. See section 6 for an example.

3. Inferences to unidentified causal quantities. There are causal quantities that cannot be estimated without additional assumptions, even when the treatment of interest has been randomized. One such quantity is the probability of a non-zero causal effect for a specific

[6] Even with an experiment, however, the identification of average causal effects cannot be achieved completely without assumptions. Experiments must invoke some form of the Stable Unit Treatment Value Assumption (SUTVA) (Gerber and Green 2012).

subject, also referred to as a "causes of effects" question. Suppose we have randomly assigned some subjects but not others to receive a letter that encourages them to turn out to vote. And suppose that we have found that the letter treatment increases turnout. What could we say about the probability that the decision to turn out by a subject who did receive a letter was indeed caused by the letter? While it is not possible to estimate this probability, it is possible to find an upper and a lower bound. Dawid, Humphreys, and Musio (2019) show how these bounds can sometimes be tightened if we can measure the values of mediators through which the letter treatment affects vote choice, say, subjects' views on whether voting is a civic duty. In short, knowing that X causes Y through M and observing M may improve our inferences about whether X caused Y for a particular unit.

4.3 Strategy 3: Use theory as a helper for design

Apart from helping with inference, theories can provide guidance for various aspects of experimental design. For example:

Site selection. Causal models may help us decide where to run an experiment. We may want to choose, for instance, a site that allows results to be transported to as many other settings as possible. As we discuss in section 5.6, causal models provide guidance on the extent to which an experiment conducted in one setting will be informative for treatment effects in other settings. Alternatively, if our aim is to test a causal model (see section 4.5), we may consider contexts for which the model's predictions differ from the predictions of alternative models.

Treatments. Causal models may help researchers decide which treatments to implement. If the aim is to test a causal model, we would obviously like to randomize causes relevant to the model. If the aims is to estimate the parameters of a structural model (see section 4.6), additional treatments can sometimes help with the identification of model parameters other than the effect of the treatment itself (DellaVigna 2018).

Sampling. In section 5.3, we describe a way of transporting treatment effect estimates from a setting A to a setting B by calculating weighted averages of estimates within subgroups. Crucial for our ability to use this strategy is that our subject pool in setting A contains enough subjects in each subgroup to estimate treatment effects within these groups. Prior knowledge of a causal model that tells us the dimensions along which treatment effects vary can help us design a sampling strategy that achieves this goal.

Random Assignment. The same considerations affect the way in which we assign units to treatment conditions. Being able to estimate effects in a subgroup requires a sufficient number of treated and untreated subjects in this subgroup. We can fix the number of treated subjects in each subgroup by assigning treatment within blocks. Causal models thus help with the design of blocking schemes. Especially in the presence of spillovers, they can also help decide how to allocate units across experimental conditions in order to maximize statistical power (Bowers et al. 2018).

Measurement. A causal model can help assess which covariates need to be measured in setting A and B in order to be able to transport treatment effect estimates from A to B. A causal model can also give us guidance on which variables we need to measure if we would like to bound our estimates of the probability of causation (see point 3 in section 4.2).

4.4 Strategy 4: Use experiments as building blocks of theories

A fourth approach is to think of the ability of experiments to identify average treatment effects as an opportunity for inductive learning about theories. Being able to claim that X causes Y already establishes a theory of sorts. Beyond that, researchers sometimes stitch experimental results together to form more elaborate theories.

Imagine that we run our example experiment that randomizes whether customers make the first offer in taxi bargaining. Moreover, suppose that we also find a way to randomly vary how many offers and counter-offers can be made before bargaining breaks down. For the sake of argument, assume that the customers in our experiment are a random sample from the population of interest. From this experiment, we could obtain estimates of the average treatment effect of going first for different numbers of bargaining rounds and of the average treatment effect of additional bargaining rounds conditional on whether the customer goes first. Even without prior knowledge of the causal model presented in section 2.2, we could use these estimates to "piece together" the functional relationship between moving first, the number of bargaining rounds and the expected taxi fare in the population of interest.

Unfortunately, things become more complicated when causal models are slightly more complex. Consider an effort to establish that X causes Y through M by stitching together estimates of the effect of X on M and the effect of M on Y (Imai et al. 2011; Green, Ha, and Bullock 2010).

Figure 2 shows two problems one might run into.

1. Even if a treatment X has an average treatment effect on a variable M and M has an average treatment effect on Y, M might not be a mediator of the relationship between X and Y for any *units.* For example, as in the upper panel of Figure 2, X may affect M among some set of units and M may affect Y among another set units, but these sets may not overlap.

2. Even if a treatment X has no average treatment effect on a variable M, and M has no average effect on Y, M may mediate a relationship between X and Y for every *unit.* This could arise, as in the lower panel of Figure 2, if the effects of X on M and of M on Y have opposite signs in two subpopulations and therefore offset each other.

The broader take away is that there are limits to the extent to which experiments enable us to learn about causal models from empirical observation alone. Even if we were able to randomize all variables in a model, we would not be able to recover many causal quantities of theoretical relevance. Mediation estimands are not the only quantities which pose such problems. Other examples include median treatment effects and "causes of effects" questions. As we have described in section 4.2, prior theoretical knowledge can help strengthen the inferences that can be drawn about such quantities from experimental data.

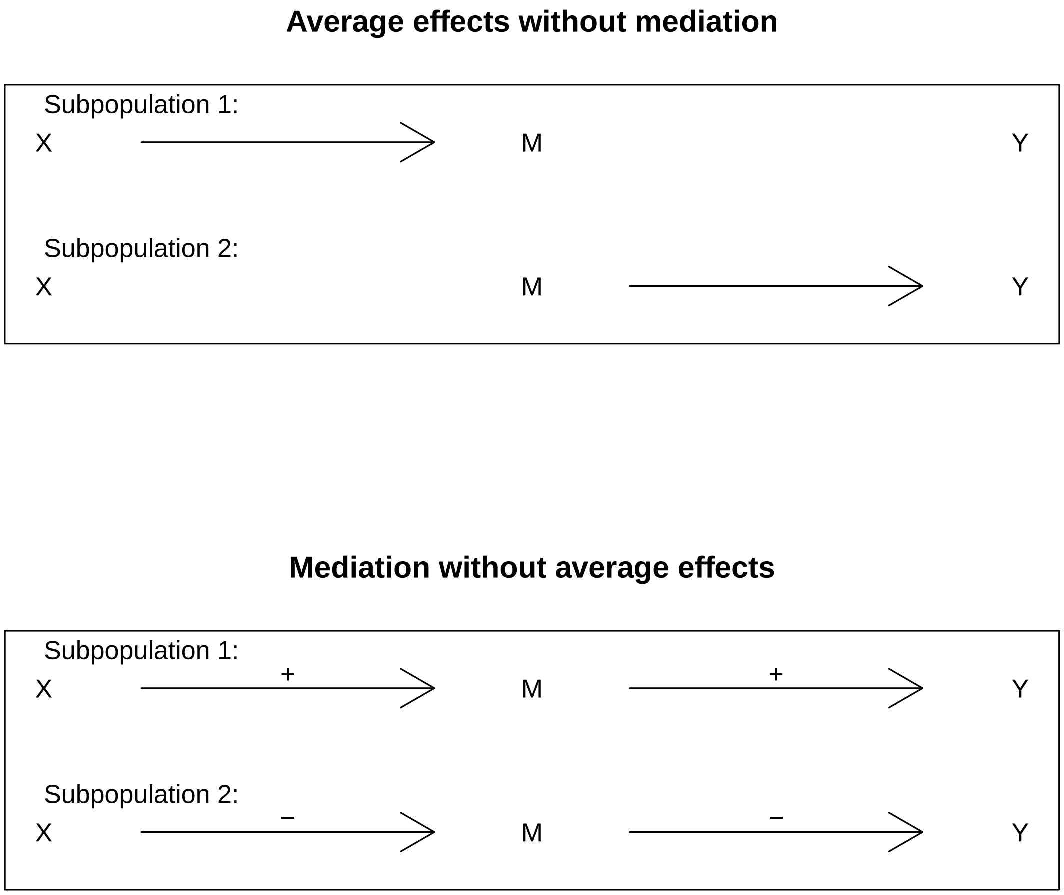

Figure 2: The hazards of trying to stitch experimental results together to form a theory

4.5 Strategy 5: Use experiments to put theories to the test

Rather than seeking to *generate* theories, experimentation could focus on "selecting" theories. Indeed, for many, the falsification of theories is what science is all about—at least in principle (Lakatos 1970).

Experiments are useful for theory testing, because they allow for valid statistical tests of hypotheses about causal effects. Of course, testable implications of theories can take many forms. Our theory about bargaining, for example, implies a correlation between the number of bargaining rounds and whether a customer moves first once we condition on the fare that the customer has to pay.[7] One advantage of implied causal relationships, however, is that they are often consistent with a smaller set of alternative theories. For example, imagine an alternative model according to which the number of bargaining rounds has a causal effect on whether the customer moves first. Both our model and the alternative model are consistent with the finding that the number of bargaining rounds and the customer's first-mover status are correlated conditional on the taxi fare, but only one of them is consistent with the finding that the number of bargaining rounds causes the customer to move first. In short, being able to test hypotheses about causal effects is essential for our ability to empirically distinguish causal models.

The perhaps most common way of combining field experiments and theory is in line with this aim of theory testing and involves the following steps:

Step 1: Derive claims about marginal effects from a causal model, for example: "Whether the customer makes the first offer has a non-zero (positive or negative) effect on the price that the customer has to pay."[8]

Step 2: Design and implement a field experiment to test these claims.[9]

Step 3: Use experimental data to test one or more null hypotheses of no effect against the relevant alternative hypotheses.[10]

Recent work has pushed further on what can be done with testing. In particular, work in the tradition of Fisher (1935) and Rosenbaum (2002, 2010) has explored the potential of using experiments to test a set of more elaborate causal models against each other (see Bowers, Fredrickson, and Panagopoulos 2013). More elaborate here means that these models specify a functional relationship for treatment and outcome for every unit conditional on one or more model parameters. For example, one of the simplest such models entails that going first adds the same constant τ to the taxi price for all customers. A more complex model

[7] The reason is that the taxi fare is a "collider" on the path from the number of rounds to whether the customer moves first (Pearl 2009)

[8] In practice, the models that are invoked vary in their complexity from several informal statements about hypothesized causal relationships (see e.g. Chong et al. 2014; Olken 2010) to fully specified decision-theoretic or game-theoretic models from which empirical predictions are derived in the form of comparative statics (see e.g. Blattman and Annan 2016; Avdeenko and Gilligan 2015).

[9] See section 4.3 for examples of how the causal model that one would like to test may affect design choices.

[10] In practice, researchers tend to either test a null hypothesis of no average treatment effect, if they rely on the Neyman (1933) tradition of hypothesis testing, or the sharp null hypothesis of no treatment effect for any unit, if they follow the Fisherian (1935) approach. See chapter 42 in this volume for more on the differences between these approaches.

may specify that the taxi price is given by $\pi_i = \alpha_i + \tau^{'} z_i + \beta s_i$, where $\tau^{'}$ is a constant effect of going first, s_i is the number of other customers in the same shared taxi who go first, β is a constant marginal effect of each additional first mover in a taxi and α_i is the price that the customer pays when she and all other taxi passengers go second. Fisherian inference proceeds by hypothesizing specific values for the parameters in the model (e.g. $\tau^{'} = -0.2$ and $\beta = -0.1$) and testing the null hypothesis that the assumed model and parameter values are correct. The same test is performed for a grid of parameter vectors for any given model and, also, for different models. For example, one could compare the p-value associated with a constant additive treatment effect model with $\tau = -0.5$ to the p-value associated with the more complex model with $\tau^{'} = -0.2$ and $\beta = -0.1$.

In a recent application, Ichino, Bowers, and Fredrickson (2013) use this approach to test two different agent-based models of how party agents who aim to rig the voter registration process react to the placement of observers at randomly selected voter registration centers. This paper demonstrates how the close connection between the causal model of interest and the null hypotheses that are being tested facilitates learning about the causal model. It also shows, however, that the wholesale rejection of one model for another is often impossible, since different models can imply the same distribution of outcomes for some values of their respective parameters.

4.6 Strategy 6: Use experiments to estimate structural models

An approach that has been the norm in economics in the past (Heckman 1991) and seems in the process of making a comeback (DellaVigna 2018) is to use field experiments for the estimation of structural models. In section 6, we demonstrate some of the advantages and pitfalls of this approach using our stylized example theory of bargaining. Here, we highlight the main points.

1. Building a structural model. The basis for structural estimation is typically a decision-theoretic or game-theoretic model from which the equations that link exogenous and endogenous variables in the model can be derived. A key step towards the specification of a model that can be estimated is the modeling of heterogeneity. For example, below we derive the taxi fare that a customer pays as a deterministic function of three variables (the customer's behavioral type, the number of possible bargaining rounds and whether the customer gets to make the first offer) and of the customer's discount factor, a parameter that we seek to estimate. Naturally, it seems unrealistic that every subject in a real-world experiment would behave exactly in accordance with these functions. The DAG in Figure 1 represents this idea through the arrow that points from the unobserved variable U_3 into π. Structural models explicitly incorporate such heterogeneity, typically by allowing for random shocks to the utility of players, for random implementation errors or for heterogeneity in model parameters (DellaVigna 2018). Usually, these models assume that this heterogeneity follows a particular distribution.

2. Identification. In order to estimate them unambiguously, the parameters of a structural model need to be identified, i.e. the distribution of data should only be implied by one set of parameter values. If more than one set of parameter values can generate the same distribution of data, then we cannot distinguish true and false parameters from each other even with infinite data and even if the true data generation process is indeed the assumed model. Sometimes, additional assumptions about functional forms or distributions are required for the purpose of identification. In section 6 we discuss how experiments can aid the identification of structural models.

3. Estimation and extrapolation. The estimation of the model parameters can be performed using one of various methods including Maximum Likelihood Estimation (MLE), Generalized Method of Moments (GMM) or Bayesian approaches. A key advantage of structural estimation is that the resulting estimates, together with the model, can subsequently be used for various extrapolations (e.g. towards the effects of the same treatment in other settings or the effects of different treatments) that go beyond what could be learned from experimental data alone.

4. Theory dependence and cross-validation. Whether such inferences will be misleading depends, of course, on the extent to which the model itself is a good approximation to reality. Crucial to the extrapolation of results to other settings is often the assumption that the parameters of a model are "structural" in the sense that they do not vary across settings (Acemoglu 2010). For example, we estimate a discount factor which captures the extent to which individuals value the future relative to the present. In order to predict treatment effects for other settings, we need to assume that the players in these other settings have the same discount factor. Yet, individuals may not necessarily place the same value on the future in all contexts.

Researchers typically assess the sensitivity of their estimates to alternative assumptions and cross-validate their models in various ways. If we believe that certain parameters are indeed structural, one way to validate a model is to compare the resulting estimates of, say, the discount factor to estimates of the same parameters from other studies. Another possibility is to make predictions for other settings and compare those to actual data from these settings. See Martinez, Meier, and Sprenger (2017) and DellaVigna et al. (2017) for examples. In section 6, we show how the ability of experiments to obtain unbiased estimates of average treatment effects can sometimes be helpful for cross-validating a structural model.

5 Six Strategies to Address External Validity Concerns

Below, we review six strategies that researchers have used to addressed the issue of external validity. While some of these approaches treat external validity as an empirical question, most of them draw, in some way or the other, on prior theoretical knowledge to determine whether and how the results from an experiment can be generalized.

5.1 Strategy 1: Push Back – Researchers ***should*** focus on sample effects

The first response is to ignore concerns over external validity and limit claims to sample effects. One justification for this might be that empirical research in social science should be about testing general theoretical propositions and not about estimating effects. As long as a proposition applies to a sample, results from the sample can be declared consistent or inconsistent with the proposition even if they are not in other ways representative of a population.

5.2 Strategy 2: Claim a bellwether

A second response is to assert that the case studied is in some way especially informative for other cases. Researchers often claim that their experiment is "ideal" in some sense. This might mean that it is typical in some way, or that it is atypical in an informative way.

Say that we choose to run an experiment at a site that has some value on background variable X, and that we have good reason to believe that the causal effect of interest is decreasing continuously in X. Then:

1. finding a positive effect in a location with a high value of X is informative for the claim that effects are in general positive in the target population;
2. finding a negative effect in a location with a high value of X is not very informative for the claim that effects are in general negative in the target population;
3. finding a positive effect in a location with a low value of X is not very informative for the claim that effects are in general positive in the target population;
4. finding a negative effect in a location with a low value of X is informative for the claim that effects are in general negative in the target population;
5. finding a positive (negative) effect in a location with a typical (e.g. modal) value of X is informative for the claim that effects are in general positive (negative).

Claims 1 and 4 are strong because the case is atypical. Claims 2 and 3 are weak because the case is atypical. Claim 5 is strong because the case is typical.

X here could be any characteristic of the experiment including properties of the treatment. For example, we may expect treatment effects to increase in the intensity of the intervention or decrease in its scale.

Critically, claiming inference from unusualness (or typicality) depends on prior beliefs about the distribution of effects as a function of selection criteria, here X. Justification or at least articulation of these beliefs is needed to assess these claims. Pre-existing theoretical knowledge can help in this regard. A more empirical approach might be to elicit beliefs about the *ranking* of effects across a set of cases, including the case at hand.

Note that, given these considerations, the *ex post* informativeness of an experiment is not the same as its *ex ante* informativeness. You may not be wise spending resources to search for a positive treatment effect in a setting with a high value of X, for example, since the chances of finding such an effect are small. But if you find it, such a result is highly informative.[11]

5.3 Strategy 3: Exploit variation within studies

A third approach is to use variation in effects *within* a study site to justify claims that go beyond the study site.

The simplest approach is to identify relevant dimensions along which the study site differs from the target site and to demonstrate that there is no heterogeneity in effects along these dimensions within the study at hand.

If there is heterogeneity along such dimensions, one possibility is to use weighting or propensity score subclassification estimators to estimate effects for the target population using variation in the study population. See Kern et al. (2016) for an assessment of several such approaches.

The general idea is to identify a set of strata across which effects vary, estimate effects for each of these strata in the study population, and take a weighted average of the resulting estimates where the weights correspond to the share of subjects in each stratum in the target population.

For intuition, say you undertake a study in location A. You want to make a claim for target site B. In site A, one third of subjects are young and two thirds are old. In site B, it is the reverse. The strategy is to estimate the effect separately for young and old subjects in A and then calculate a weighted average of these group-specific estimates using the proportion of young and old subjects in site B. Let $\hat{\tau}_O^A$ and $\hat{\tau}_Y^A$ be the respective treatment effect estimates among old and young people in setting A. Suppose we learn from study A that $\hat{\tau}_Y^A = 1$ and $\hat{\tau}_O^A = \frac{1}{4}$. The overall treatment effect estimate in A is given by $\hat{\tau}^A = \frac{1}{3} \times 1 + \frac{2}{3} \times \frac{1}{4} = \frac{1}{2}$. For site B, we estimate the treatment effect to be $\hat{\tau}_B = \frac{2}{3} \times 1 + \frac{1}{3} \times \frac{1}{4} = \frac{3}{4}$.

This approach depends on many assumptions, formalized in Pearl and Bareinboim (2014) and Tipton (2013). Most obviously the support of B must be a subset of the support of A—that is, for any stratum in B for which one wants to generate estimates there should be a corresponding stratum in A. You cannot estimate effects for, say, a co-ed school based on estimates obtain from an experiment in a boys-only school.

More substantively, you need to be willing to assume that the distribution of effects within each stratum is the same in the two populations. This claim is most easily made when the study sample is itself a random draw from the study population. Yet even in such an ideal case, one has to worry about how spillovers work not just in the study population but also in the target population (Tipton 2013).

Finally, this approach has implications for how one can design the study sample in A to facilitate inference to target B (see Tipton and Peck 2017).

[11] This logic links to case selection criteria that are used in qualitative research (Van Evera 1997).

5.4 Strategy 4: Exploit variation across studies

A fourth strategy is to design studies so that they can feed into meta-analyses that seek to make broader claims. A common approach is to think of a superpopulation of effects that might obtain across studies and contexts.

$$\tau_j \sim f(\mu, \sigma)$$

The interest is in learning about the average superpopulation effect, μ, but also the variation in effects σ. A single study drawn randomly from a population can give an unbiased, but noisy, estimate for μ and says nothing about σ. Multiple comparable studies can provide tighter estimates of both μ and σ.

Here, the external gains from a study operate through complementarities with other studies. Yet, up until recently, only few topics in political science had generated a large enough set of comparable experiments to allow for a meta-analysis. Initiatives that encourage the coordinated implementation of experiments on the same topic in various contexts—such as the "Metaketa" projects—seek to address this problem (Dunning et al. 2018).

Meta-analyses can also include a more systematic analysis of treatment effect heterogeneity across studies. Vivalt (2019), for example, finds that experiments implemented by governments tend to have smaller effects than those implemented by non-governmental actors. Ultimately, such results can become useful for theory development.

5.5 Strategy 5: Cross validation

A fifth approach used in Dunning et al. (2018) and Coppock, Leeper, and Mullinix (2018) treats external validity as an empirical question.[12] Dunning et al. (2018) ask: do research consumers *in fact* update inferences for out of sample estimands? They gather results from multiple related experiments and assess whether exposing research consumers to the findings makes them update their beliefs about effects in studies they have not been informed about, and whether updating goes in the right direction. Coppock, Leeper, and Mullinix (2018) assess empirically whether the results from online samples are *in fact* consistent with what we know from representative samples (and vice versa). They find that they are and attribute this to low effect heterogeneity. One could engage in a similar kind of exercise using a single study by analyzing whether results estimated on a non-random sub-sample correspond to those in the rest of the study. Results will depend on the degree of effect heterogeneity.

This empirical approach gives, in some sense, an iron clad answer to the question of external validity. Yet, it has an obvious shortcoming: *checking* external validity requires already having an answer to the target estimand. One can make external claims, and check whether the claims are correct, but this establishes the claim to external validity only in cases in which it is not needed. Put differently, you do not know whether the claim for external validity is itself externally valid outside of the test set.

[12] See also Pritchett and Sandefur (2015), Vivalt (2019), Dehejia, Pop-Eleches, and Samii (2015) and Bisbee et al. (2017).

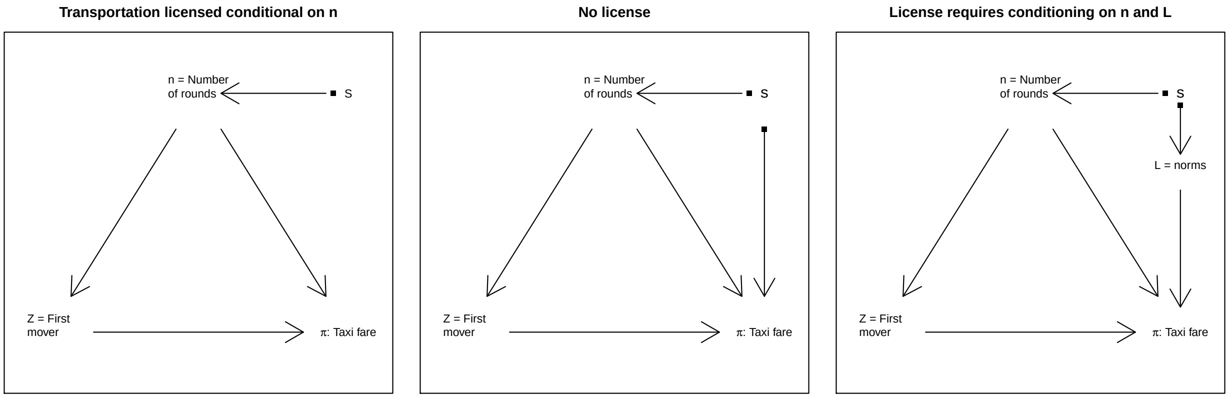

Figure 3: Three selection graphs

5.6 Strategy 6: Formal transportation

The last approach we consider uses causal models to formally justify—or "licence"—claims to external validity. This is an example of the use of theory discussed in section 4.2. Pearl and Bareinboim (2014) develop a framework in which researchers provide a causal model and the associated DAG for their study population and then represent the ways in which the target population differs from the study population as a set of "selection" nodes that are the origins of arrows that point into nodes in the original DAG. This representation of differences makes it possible to assess whether there is a weighting strategy that allows inference to the target population. The assessment is, of course, conditional on the model.

To illustrate, imagine first that two sites share a common causal model relating the taxi fare, the number of bargaining rounds, and whether the customer makes the first offer. Say, for instance, data on taxi bargaining is generated in Kenya and one wants to make inferences to Somalia. Figure 3 displays the corresponding DAG.

We are interested in the effects of the customer being the first mover on the taxi price. We know that the average effect of moving first depends on how long bargaining can continue before negotiations break down. In the first graph in Figure 3, the selection variable "S" characterizes the differences between Kenya and Somalia suggesting that there is a difference in bargaining rounds between these sites. Perhaps the taxi market is more competitive in Kenya and, as a consequence, drivers and customers are more likely to separate if they cannot reach an agreement after a small number of offers and counter-offers. If this is an accurate characterization of the differences between these sites, then results can be transported using a strategy much like that described in section 5.3: re-weight the estimates from Kenya using information on propensities for bargaining to break down after a specific number of rounds for a given customer-driver interaction. The ability to use this strategy requires –beyond the model being right– (a) that the range of bargaining rounds in Somalia is also present in Kenya and (b) that we have data about the distribution of this variable in both contexts. There are thus gains from implementing experiments in places with wide variation and in gathering data about distributions of variables out of sample.

The implications of this framework extend further, however. Say that the differences between

Kenya and Somalia are those depicted in panel 2. In this case, these differences extend to how order of play and bargaining protocol affect outcomes. Average effects could be different in these sites which would prevent extrapolation even with complete data. In graph theoretic terms, the reason is that there is no set of variables that you can condition on that "d-separate" the selection variable from the outcome.

Say, finally, that we have a model that "explains" the differences in effects via some specified intermediate variable, as in the third panel. The selection graph in the third panel can be interpreted as saying that the (conditional) effect of moving first on the taxi price is different in Somalia because there are different norms in Somalia that moderate this effect. If so, we can now get good estimates of effects in Somalia by conditioning not just on the likelihood of bargaining breaking down after a specific number of rounds but also on a measure of norms. We re-weight by finer strata. Formally conditioning on norms now separates the selection node S from the outcome.

This logic captures the core elements of Theorem 2 in Pearl and Bareinboim: Let Z denote a stratum. The strata-specific causal effect of X on Y is transportable from one graph to another if conditioning on Z produces independence between the outcome Y and the set of selection nodes S.

The gains from this framework are twofold. First, it becomes possible to state justifications for transportation in terms of independence relations between variables, which are statements that can then be assessed. Second, given a justification, there is clarity on the set of variables for which data should be gathered to calculate stratum level causal effects.

That said, to our knowledge, the framework cannot be used to extrapolate from findings on the effects of one treatment to the effects of another, which, in principle, is possible with parametric models (see sections 4.6 and 6).

As far as we know, the Pearl and Bareinboim (2014) framework has not yet been used within political science.

6 Illustration of a Parametric Structural Model Connecting Theory and Experimentation

We now return to our example experiment on taxi bargaining and use it to walk through the logic and the payoffs of estimating a simple parametric structural model.

Recall that, in this example, we are interested in why some taxi customers end up paying more than others. According to our Level 1 theory illustrated in Figure 1, there are three variables that directly affect our outcome of interest: the number of rounds for which bargaining can continue, whether the customer gets to make the first offer, and the customer's type. We begin by further developing this theory into a Level 2 theory in the form of a parametric structural model. We then illustrate how we would turn our Level 2 theory into a Level 3 theory and estimate it using simulated data. Finally, we highlight four possible benefits of the approach.

6.1 Setting up and estimating a structural causal model

We will assume that the taxi bargaining process can be captured by a standard alternating offers bargaining game with complete information (Rubinstein 1982). A customer and a driver take turns in making offers of how to divide a pie. We think of the pie as the unit endowment that we provided to the customer. For simplicity, we assume that the size of the pie is known to both players. Whoever moves first makes a suggestion on how to divide the pie. For example, if the customer moves first, she may say "I will pay you 0.2 and keep 0.8 for myself." The second player decides whether to accept or reject this offer. If the second player accepts, each player receives the share of the pie that the offer allocated to her (e.g. the customer pays 0.2 to the driver and keeps 0.8 for herself). If the second player rejects, the game moves to the next round and the second player gets to make an offer which the first player can accept or reject. If no agreement is reached, bargaining ends after n rounds and both players receive a payoff of 0. We might imagine, for example, that customers have to pay the endowment back to us if they do not secure a ride.

To capture preferences about time, we assume that players discount their payoffs at a rate of $0 \leq \delta \leq 1$: If they reach an agreement in round 1 and a player receives an amount x, the player will value this amount at x. If the same agreement is reached in period 2, the player will value the amount at δx; in period 3 she will value the amount at $\delta^2 x$ etc. The discount factor represents the idea that customers and drivers are impatient and value the pie less the longer bargaining continues.

The standard solution is found via backwards induction for a finite number of bargaining rounds n. Suppose that $n = 1$, i.e. the player who goes first gets to make a take-it-or-leave-it offer. Since both players know that there will be no second round, the first mover will take the entire pie for herself. The second player will accept, since she will receive a payoff of 0 anyway, irrespective of what she does. Let's consider $n = 2$, i.e. there can be one offer and a counter-offer. We already know that, in the second round, the second-mover will be able to take the whole pie of 1 for herself. (Again, the other player will receive 0 irrespective of whether she accepts or rejects the offer.) Seen from the first period, the second mover knows that she can achieve $\delta \times 1$ by rejecting the first mover's offer. In the first period, the first mover will therefore offer exactly δ to the second mover and the second mover will accept. Letting π_n^j denote the equilibrium share paid by player j in an n-round game, the first mover in a two-round game gives $\pi_2^1 = \delta$ to the second mover and keeps $1 - \pi_2^1 = 1 - \delta$ for herself. The same logic can be applied to $n = 3$, $n = 4$ etc.[13]

In the infinite version of the game, the optimal solution involves offering an amount so that the receiver is indifferent between accepting and moving on to the next round of the infinite game in which she would offer the same amount. That is we seek an offer π_∞^1 such that $\pi_\infty^1 = \delta(1 - \pi_\infty^1)$ which implies $\pi_\infty^1 = \delta/(1+\delta)$.

Note that, in this model, agreement is always reached in the first period irrespective of the number of *possible* bargaining rounds n.

[13] In the general solution for $n > 1$ possible rounds, a customer who moves first pays $\pi_n^1 = \sum_{t=2}^{n}(-1^t)\delta^{t-1}$, where t indexes the bargaining round. The price paid by a customer who moves second π_n^2 is just $1 - \pi_n^1$.

	Rational Customers		
	$\boldsymbol{\pi_n^1}$	$\boldsymbol{\pi_n^2}$	$\boldsymbol{\tau_n = \pi_n^1 - \pi_n^2}$
$\mathbf{n = 1}$	0	1	-1
$\mathbf{n = 2}$	δ	$1 - \delta$	$2\delta - 1$
$\mathbf{n = 3}$	$\delta(1 - \delta)$	$1 - \delta(1 - \delta)$	$2\delta(1 - \delta) - 1$
$\mathbf{n = \infty}$	$\frac{\delta}{1+\delta}$	$1 - \frac{\delta}{1+\delta}$	$2\frac{\delta}{1+\delta} - 1$
	Behavioral Customers		
	$\boldsymbol{\pi_n^1}$	$\boldsymbol{\pi_n^2}$	$\boldsymbol{\tau_n = \pi_n^1 - \pi_n^2}$
$\mathbf{n = 1}$	$\frac{3}{4}$	$\frac{3}{4}$	0
$\mathbf{n = 2}$	$\frac{3}{4}$	$\frac{3}{4}$	0
$\mathbf{n = 3}$	$\frac{3}{4}$	$\frac{3}{4}$	0
$\mathbf{n = \infty}$	$\frac{3}{4}$	$\frac{3}{4}$	0
	Population With Share $\boldsymbol{q}$ of Behavioral Customers		
	$\mathbf{E}(\boldsymbol{\pi}_\mathbf{n}^\mathbf{1})$	$\mathbf{E}(\boldsymbol{\pi}_\mathbf{n}^\mathbf{2})$	$\boldsymbol{\tau_n} = \mathbf{E}(\boldsymbol{\pi}_\mathbf{n}^\mathbf{1} - \boldsymbol{\pi}_\mathbf{n}^\mathbf{2})$
$\mathbf{n = 1}$	$q\frac{3}{4}$	$q\frac{3}{4} + (1 - q)$	$-(1 - q)$
$\mathbf{n = 2}$	$q\frac{3}{4} + (1 - q)\delta$	$q\frac{3}{4} + (1 - q)(1 - \delta)$	$(1 - q)(2\delta - 1)$
$\mathbf{n = 3}$	$q\frac{3}{4} + (1 - q)\delta(1 - \delta)$	$q\frac{3}{4} + (1 - q)(1 - \delta(1 - \delta))$	$(1 - q)(2\delta(1 - \delta) - 1)$
$\mathbf{n = \infty}$	$q\frac{3}{4} + (1 - q)\frac{\delta}{1+\delta}$	$q\frac{3}{4} + (1 - q)\left(1 - \frac{\delta}{1+\delta}\right)$	$(1 - q)\left(2\frac{\delta}{1+\delta} - 1\right)$

Table 1: Equilibrium prices for first (π_n^1) and second (π_n^2) moving customers and average treatment effect (τ_n) of the customer moving first for games with up to n rounds.

The top panel of Table 1 summarizes the price that a rational customer should pay according to this model depending on the number of rounds n and on whether the customer moves first (π_n^1) or second (π_n^2).

We are interested in the effect, τ_n, of being able to make the first offer on the price paid in an n round game. This effect simply equals the difference between π_n^1 and π_n^2 (see the last column of Table 1).

So far, we have only considered the behavior of rational players. Yet, according to our causal model, customers—though not drivers—can be of different types. In this stylized example, we assume that there are some behavioral customers who always pay $\frac{3}{4}$ of the pie to the driver and keep $\frac{1}{4}$ for themselves, irrespective of the number of bargaining rounds or whether they go first or second. As can be seen in Table 1, the treatment effect of going first on the price paid is always 0 for behavioral customers.

We assume that a customer's behavioral type is not observed by the researchers and that the share of behavioral types in the population of customers is q. Given a share q of behavioral types, we first ask: what would be the average treatment effect of the customer going first on the taxi fare predicted by the theory? As displayed at the bottom of Table 1, the population-level average of the price paid by customers who move first or second is just a weighted average of the price paid by rational customers and the price paid by behavioral customers. Accordingly, since the average treatment effect among behavioral customers is zero, the average treatment effect in the population is just the proportion of rational customers $(1 - q)$ times the predicted treatment effect among rational customers.

Subsequently, we focus on $n = 2$ and $n = \infty$. Note from the table that $\tau_2 > \tau_\infty$ for $\delta > 0$,[14]

[14] $\tau_2 > \tau_\infty \leftrightarrow 2\delta - 1 > 2\delta/(1 + \delta) - 1 \leftrightarrow \delta > 0)$

though τ_2 and τ_∞ may differ in sign and τ_∞ may be larger in absolute value. In the extreme case when δ is close to 1, τ_2 is close to $(1 - q)$ and τ_∞ is close to 0.

6.1.1 Taking the model to data

Our Level 2 model makes predictions about average taxi fares as a function of whether the customer moves first, the number of bargaining rounds and two model parameters, the discount factor δ and the share q of behavioral customers in the population. Suppose we have run our taxi bargaining experiment that randomly assigns customers to make the first offer in a place where $n = 2$ or $n = \infty$. How can we use these data to get estimates of the model parameters, δ and q and estimates of treatment effects τ_n?

We do so by using the model's equilibrium predictions to motivate a data generating process that describes the probability of observing any data given a set of parameter values. We then rely on Maximum Likelihood Estimation to find estimates of δ and q. The same could be achieved through various other estimation methods, including Generalized Method of Moments (GMM) or Bayesian approaches.[15]

An obvious challenge in using real world data to estimate the parameters of a model is that the real world might produce data that are inconsistent with an overly simple model. For this reason it is generally necessary to allow for a stochastic component that can render all data *possible*, even if improbable. In the DAG in Figure 1, the idea that observed taxi fares may deviate from the predictions in Table 1 is captured by the node U_3. We can think of U_3 as representing factors other than the number of bargaining rounds, whether the customer moves first and the customer's behavioral type that also affect the price a customer pays.

What could be sources of U_3 in the context of our model? There are multiple possibilities here. In his review of structural estimation in behavioral economics, DellaVigna (2018) discusses the three most common ways in which researchers incorporate heterogeneity in their models. First, it is sometimes assumed that the actors in the model receive an unobserved utility shock which generates heterogeneity in how they behave. Second, we can imagine that there is heterogeneity in some of the parameters of the model. Imagine, for example, that instead of a fixed number, δ is a random variable that follows, say, a beta distribution in the population of interest. Individuals would then be heterogeneous in their δ which, in turn, would result in heterogeneity in the way the customer's endowment is divided. A final possibility is to assume that individuals make implementation errors.

This last approach corresponds well with the way in which we will introduce heterogeneity here. Imagine that players attempt to act in accordance with the model but make random

[15] We will use MLE, partly because it is commonly used to estimate models that are more complex than ours. To estimate this stylized example, we could also rely on a GMM approach which would not require assumptions about the distribution of unobserved exogenous variables in our model which will be necessary to estimate the model via MLE. Our structural model provides us with two "moment conditions." For $n = 2$, we have $2(E(\pi_2^1) + E(\pi_2^2) - 1) - q = 0$ and $\frac{E(\pi_2^1)+3E(\pi_2^2)-3}{4(E(\pi_2^1)+E(\pi_2^2))-6} - \delta = 0$. In a nutshell, a GMM approach would entail (i) replacing $E(\pi_2^1)$ and $E(\pi_2^2)$ with their empirical counterparts, i.e. the average of shares received by first movers (treatment group) and second movers (control group) and (ii) choosing q and δ such that the right-hand side of these expressions becomes as close to 0 as possible.

mistakes when dividing the endowment. As a consequence, the prices that get paid are not always the ones predicted by the model.

To estimate our model via MLE, we need to be specific about the distribution of these implementation errors. Here, we assume that the price paid by a given customer is a draw from a beta distribution that is centered on whatever price the model predicts for this customer. The beta distribution is a natural choice in this case, since it is defined on the interval $[0, 1]$, just like the shares of the endowment that customers pay as a result of the taxi bargaining process. The distribution has two parameters, α and β, that control its shape. The mean of the beta distribution can be expressed as a function of these parameters, $\mu = \frac{\alpha}{\alpha+\beta}$. In turn, we can write the two parameters as $\alpha = \kappa\mu$ and $\beta = \kappa(1-\mu)$. In this way, we can model μ, the mean of the distribution, as a function of the variables in our model. Specifically, μ will depend on

- customer i's treatment status z_i, where $z_i = 1$ if the customer has been randomly assigned to go first and $z_i = 0$ if the customer has been randomly assigned to go second,
- customer i's (unobserved) behavioral type t_i, where $t_i = 1$ if the customer is a behavioral type and $t_i = 0$ if the customer is rational and
- the number of rounds $n \in \{2, \infty\}$ for which bargaining can continue.

Thus κ enters the model as a new parameter that describes the variance of the distribution of prices but not its mean. It is of some substantive interest in that it captures how close behavior is to the Level 2 model predictions.

Using this parametarization and the predictions of the model, we can write down the following joint probability distribution for the observed prices paid by N customers in our experimental subject pool (where we now use subscript i to denote individuals):

$$\begin{aligned}
\text{Prob}(\boldsymbol{\pi}|q,\delta,\kappa) &= \prod_{i=1}^{N} q\text{Beta}\left(\pi_i|\frac{3}{4}\kappa, \frac{1}{4}\kappa\right) + (1-q)\text{Beta}(\pi_i|\mu_i\kappa, (1-\mu_i)\kappa) \qquad (1)\\
\mu_i &= \begin{cases} z_i\delta + (1-z_i)(1-\delta) & \text{for } n = 2 \\ z_i\frac{\delta}{1+\delta} + (1-z_i)\left(1-\frac{\delta}{1+\delta}\right) & \text{for } n = \infty \end{cases}\\
z_i &= 1 \text{ with prob. } p = \frac{1}{2}
\end{aligned}$$

From this function we can extract the likelihood function directly. A couple of features are noteworthy: First, this data generating process reflects that customers are randomly assigned to go first; hence $z_i = 1$ with probability $\frac{1}{2}$ for all units. Relatedly, we do not need to know a customer's behavioral type t_i in order to compute the likelihood, which is fortunate as we do not observe this characteristic. Instead, the joined probability distribution is simply a mixture between a beta distribution with mean $\frac{3}{4}$ and a beta distribution with mean equal to the predicted price paid by rational customers, conditional on whether they go first or second. The weight on each of these distributions is given by the share of behavioral types q.

6.1.2 Estimation and inferences

Now that we have equation 1 in hand, we use standard MLE approaches to find the values for q, δ and κ that maximize the likelihood of the data we observe in our experiment.

Once our analysis is set up it is easy to inquire into how well this procedure works. Suppose our model is correct. We can then find out how well we recover unbiased estimates of the model parameters using the following steps: 1) simulate data using draws from the data generating process described above, 2) perform the estimation and 3) compare the estimates to the parameter values assumed during the simulation. By repeating these steps many times, we can figure out whether our estimates are on average correct and how variable they are.

In practice, we use the `R` package `DeclareDesign` for this exercise (Blair et al. 2018). We provide all code for the declaration and diagnosis of this model in supplementary material.

The simulations involve (a) imagining that we have run an experiment that randomly assigns taxi customers to make the first offer, either in a setting where bargaining can continue for two rounds ($n = 2$) or in a setting where bargaining can continue indefinitely ($n = \infty$) (b) using a maximum likelihood estimator to estimate δ, κ and q in either setting (c) drawing inferences regarding treatment effects using the parameter estimates. This last step is done by referring to the last column of Table 1. For example, with estimates of $\hat{q} = 1/2$ and $\hat{\delta} = 4/5$, we would predict a treatment effect of $\hat{\tau}_\infty = \frac{1}{2}\left(2\frac{\frac{4}{5}}{1+\frac{4}{5}} - 1\right) = -\frac{1}{18} \approx -0.06$ for a setting where bargaining can continue indefinitely.

For each estimand-estimator pair we then report the expected bias (which simply equals the difference between the estimand and the mean estimate across all simulations).[16]

In Table 2, we report the estimates of model parameters and in Table 3, we report estimates of average treatment effects, comparing those generated from parameter estimation to those generated using a simple differences-in-means estimator.

We see that we recover unbiased estimates of q and δ, irrespective of whether we use data from a setting with $n = 2$ or $n = \infty$. The estimates of κ are slightly biased. Moreover, our estimates of model parameters allow us to recover unbiased estimates of average treatment effects.

Table 2: Estimation of model parameters using the correct model

Estimand	Estimator	Estimand Value	Bias
δ	MLE_2	0.80	-0.00
δ	MLE_∞	0.80	0.00
κ	MLE_2	6.00	0.04
κ	MLE_∞	6.00	0.03
q	MLE_2	0.50	-0.00
q	MLE_∞	0.50	-0.00

[16] In the appendix, we also report the root-mean-square error (RMSE).

Table 3: Estimation of average treatment effects using difference-in-means (DIM) and parameter estimation (MLE)

Estimand	Estimator	Estimand Value	Bias
τ_2	MLE_2	0.30	-0.00
τ_2	DIM_2	0.30	0.00
τ_∞	MLE_∞	-0.06	-0.00
τ_∞	DIM_∞	-0.06	-0.00

Why, in this case, are we able to obtain unbiased estimates of the model parameters of interest? To see the intuition, recall that, according to our model (see Table 1), the average prices paid by first and second moving customers in a setting with $n = 2$ are given by the following functions: $E(\pi_2^1) = q\frac{3}{4} + (1-q)\delta$ and $E(\pi_2^2) = q\frac{3}{4} + (1-q)(1-\delta)$. It turns out that, if we knew the values of $E(\pi_2^1)$ and $E(\pi_2^2)$, these two equations could, in most cases, be solved for the unique values of q and δ. For example, if $E(\pi_2^1) = \frac{1}{2}$ and $E(\pi_2^2) = \frac{3}{4}$, then we know that q must equal $\frac{1}{2}$ and δ must equal $\frac{1}{4}$.[17] An experiment that randomly assigns customers to go first or second provides us with unbiased estimates of $E(\pi_2^1)$ and $E(\pi_2^2)$. As a consequence, we are able to recover estimates of q and δ.

Even though this result seems encouraging, one may wonder why we would want to estimate the parameters of our model in the first place. After all, as can be seen in Table 3, using a simple difference-in-means estimator yields unbiased estimates of the average treatment effect of moving first irrespective of whether we perform our experiment in a context with $n = 2$ or $n = \infty$. What more can we learn from this exercise about why some taxi customers pay higher prices than others? In the next four sections, we demonstrate what we see as some of the potential benefits of combining experimental data and structural estimation.

6.2 Benefit 1: Theory allows for answers to a wider set of questions

A key benefit of the structural model is that it allows us to answer a more varied array of questions than can typically be addressed by design-based inference, in particular questions regarding different settings, different treatments, and different outcomes.

Our causal model suggests that the average treatment effect of moving first may vary with the number of rounds for which bargaining can continue. As can be seen in Table 3, the estimates that we obtain in a setting with $n = 2$ potential bargaining rounds will not generalize to one in which bargaining can continue indefinitely.

To what extent can the structural model help us? Imagine that we conduct our experiment in a setting where bargaining continues for $n = 2$ rounds and we obtain estimates of q and δ.

[17] In general, $q = 2(E(\pi_2^1) + E(\pi_2^2) - 1)$ and $\delta = \frac{E(\pi_2^1)+3E(\pi_2^2)-3}{4(E(\pi_2^1)+E(\pi_2^2))-6}$. Note that for the special case of $E(\pi_2^1) = \frac{3}{4}$ and $E(\pi_2^2) = \frac{3}{4}$, the system implies $q = 1$ and is consistent with any value of δ. Moreover, there are values of $E(\pi_2^1)$ and $E(\pi_2^2)$ for which this system of equations has no solution. For example, there is no combination of $0 \leq q \leq 1$ and $0 \leq \delta \leq 1$ that can produce $E(\pi_2^1) = 1/2$ and $E(\pi_2^2) = 1/4$.

We now want to extrapolate to treatment effects in a setting where bargaining can continue indefinitely. We can again estimate such treatment effects by consulting the equations in Table 1, using estimates from data derived in one setting ($n = 2$) to make claims about another ($n = \infty$). Thus, provided that our model is correct, we can use an experiment conducted in a single setting to predict effects for a variety of other settings.

Table 4 shows estimates based on this approach for our model.

Table 4: Extrapolation

Estimand	Estimator	Estimand Value	Bias
τ_2	MLE_2	0.30	-0.00
τ_2	MLE_∞	0.30	0.00
τ_∞	MLE_2	-0.06	-0.00
τ_∞	MLE_∞	-0.06	-0.00

Beyond generalization across settings, the model also helps us predict the effects of alternative treatments. The equations in Table 1 imply, for example, that the average treatment effect of a change that allows bargaining to continue for $n = 3$ rounds instead of $n = 2$ rounds on the average price paid by a customer who moves first is $-(1-q)\delta^2$. Based on our estimates of $\hat{q} = 1/2$ and $\hat{\delta} = 4/5$, we would thus predict that such a change reduces the price paid by a customer who makes the first offer by $\frac{8}{25}$.

Giving more structure to our model has considerably expanded the set of inferences that we would be able to draw from a single experiment. Of course, these inferences are only warranted if our structural causal model is the correct one. This may well be an unrealistic assumption. Among other things, it entails that the structural parameters δ and q are indeed structural in the sense that they do not vary across settings. If the proportion of behavioral customers in the population varies across settings, for example, our predictions will not be correct. We will show below how reliance on the wrong model can lead to biased inferences and how randomized experiments can sometimes draw attention to flaws in a model.

6.3 Benefit 2: Theory provides pointers to better design

Sticking for the moment with the assumption that our stylized model is correct, we can use the model to draw conclusions for how to design our hypothetical experiment.

Consider, for example, settings in which only take-it-or-leave-it offers are possible, i.e. $n = 1$. We can see from Table 1 that neither $E(\pi_1^1)$ nor $E(\pi_1^2)$ depends on δ.

An experiment in a setting where $n = 1$ would thus allow us to estimate q but not δ. Without an estimate of δ, we will not be able to generalize to other settings or treatments in the manner described above.

Similarly, not all equations for $E(\pi_n^1)$ and $E(\pi_n^2)$ can be easily solved for q and δ. In the case of $n = 3$, for example, one set of estimates of $E(\pi_3^1)$ and $E(\pi_3^2)$ can be consistent with two different solutions for q and δ. In order to maximize the inferences that can be drawn, the experiment should be conducted in a setting that enables us to uniquely identify both q and δ.

The model also has implications for case selection if we are interested in maximizing power. Suppose, for example, that we believe it likely that δ is close to 1 and that the share of behavioral types q is not very large prior to running our experiment. We would then expect $|\tau_2| > |\tau_\infty|$, which suggests that we will be better powered to detect a treatment effect if we run our experiment in a setting where bargaining must end after two rounds than in a setting where bargaining can continue indefinitely. While this has the feel of letting design considerations determine estimands rather than the the other way around, greater power here may also imply better ability to estimate the same fundamental parameters.

6.4 Benefit 3: Experimental data make it possible to learn more from theory

The approach we have described combines theory and experimental data. Yet, given that we are doing so well in terms of parameter estimation, one may wonder whether it is even necessary to run an experiment.

As it turns out, random assignment played an important role in this case. To show why, we change the data generating process in our simulation to the following:

$$\begin{aligned}
\text{Prob}(\boldsymbol{\pi}|q,\delta,\kappa) &= \prod_{i=1}^{N} q\text{Beta}(\pi_i|\frac{3}{4}\kappa,\frac{1}{4}\kappa) + (1-q)\text{Beta}\left(\pi_i|\mu_i\kappa,(1-\mu_i)\kappa\right) \qquad (2)\\
\mu_i &= \begin{cases} z_i\delta + (1-z_i)(1-\delta) & \text{for } n=2 \\ z_i\frac{\delta}{1+\delta} + (1-z_i)(1-\frac{\delta}{1+\delta}) & \text{for } n=\infty \end{cases}\\
z_i &= 1 \text{ with prob. } p = 0.8t_i + 0.5(1-t_i)\\
t_i &= 1 \text{ with prob. } q
\end{aligned}$$

In line with our causal model we assume that, in the absence of a randomized experiment, the probability of going first is related to a player's behavioral type. Specifically, we implement a data generating process according to which behavioral types are more likely to go first ($p = 0.8$) than rational players ($p = 0.5$).

While we change the data generating process in our simulation, we do not make any changes to the estimation strategy, i.e. we still choose q, δ and κ to maximize the likelihood of our observed data according to equation (1). Table 5 contains the resulting estimates of the model parameters and average treatment effects. We also include treatment effect estimates obtained from a difference-in-means estimator. In the absence of random assignment, all our estimates are biased. Intuitively, without random assignment, we are unable to achieve unbiased estimates of π_n^1 and π_n^2, since the relationship between the treatment (going first) and the outcome (the price that a customer pays) is confounded by an unobserved variable, the customer's type. Without being able to estimate π_n^1 and π_n^2, however, we cannot uncover τ_n and neither q and δ. Of course, this problem could be solved by conditioning on the customer's type if we could observe it.

Even with the correct model, randomization can thus help to identify parameters of interest, which can subsequently be used for extrapolation. This point gains importance in more complex models where identification of the parameters of interest is more of an issue. See DellaVigna (2018) and Card, DellaVigna, and Malmendier (2011) for more on how additional treatments can help identify parameters of a structural model.

Table 5: Diagnosis of design with correct model but no randomization

Estimand	Estimator	Estimand value	Bias
τ_2	DIM_2	0.30	0.11
τ_2	MLE_2	0.30	0.11
τ_2	MLE_∞	0.30	0.12
τ_∞	DIM_∞	-0.06	0.08
τ_∞	MLE_2	-0.06	-0.03
τ_∞	MLE_∞	-0.06	0.04
δ	MLE_2	0.80	-0.01
δ	MLE_∞	0.80	0.13
κ	MLE_2	6.00	-0.09
κ	MLE_∞	6.00	-0.04
q	MLE_2	0.50	-0.22
q	MLE_∞	0.50	0.01

6.5 Benefit 4: Experimental data can help improve theory

So far, we have assumed the model is right. But of course we know the model is wrong (Box 1976). The question is how to assess the consequences of relying on a model that is incorrect and how to react if the model is misleading. One way to do so is to take a model seriously, confront it with data, and then step back to see whether the theory did violence to the data. Poor fit can be suggestive of the need to improve a model (Browne and Cudeck 1993; Gelman, Meng, and Stern 1996).

For example, imagine we did not consider the possibility of behavioral types and instead erroneously assumed that all customers acted rationally, i.e. $q = 0$. Essentially, this means ignoring the bottom two panels of Table 1 and considering the top panel only. Table 6 displays the results of changing our estimation strategy accordingly (we use MLE' rather than MLE to denote the new estimators). The first thing to note is that, because we have returned to the scenario where we randomly assign a customer to make the first offer, the difference-in-means estimator recovers unbiased estimates of the average treatment effects of going first for both cases, $n = 2$ and $n = \infty$. Our estimates of δ, however, are biased irrespective of whether we rely on data from an experiment in the $n = 2$ or $n = \infty$ context. Moreover, we naturally do not obtain any estimates of q, since we assume $q = 0$. It is thus not surprising that our predictions for treatment effects in other settings based on our estimates of δ are severely biased. Yet, we do not only obtain biased estimates of the treatment effect in a *different* context. Our prediction of the treatment effect for the *same* context is also slightly biased. Based on our estimate of δ obtained from an experiment run in a place where $n = 2$, for example, we would predict a treatment effect of 0.26 for this same place. Yet, the actual treatment effect, which we recover using the difference-in-means estimator, is 0.3.

Table 6: Diagnosis of design with incorrect model (assume $q = 0$)

Estimand	Estimator	Estimand value	Bias
τ_2	DIM_2	0.30	0.00
τ_2	MLE'_2	0.30	-0.04
τ_2	MLE'_∞	0.30	0.52
τ_∞	DIM_∞	-0.06	-0.00
τ_∞	MLE'_2	-0.06	-0.17
τ_∞	MLE'_∞	-0.06	0.01
δ	MLE'_2	0.80	-0.17
δ	MLE'_∞	0.80	0.11
κ	MLE'_2	6.00	-3.83
κ	MLE'_∞	6.00	-3.05
q	MLE'_2	0.50	

This pattern shows how random assignment can help us detect problems with our theoretical model. Knowing that we have an unbiased estimate of the average treatment effect allows us to use this estimate to validate the prediction of our theoretical model. The larger the gap between our experimental estimates and our predictions, the smaller our confidence that our theoretical model is correct.

A similar strategy is used by Todd and Wolpin (2006) who evaluate the effects of a randomized school subsidy program in Mexico. They fit a structural model using households who did not receive the subsidy and validate the model by comparing the predicted effect of the subsidy program to the experimental estimates. Similarly, DellaVigna et al. (2016) use not only new but also existing experimental results to validate their model. Specifically, the authors develop a structural model of voting based on the idea that individuals derive pride from telling others that they voted or face costs when lying about whether they voted or not. The authors rely on an experiment with several randomized treatments to estimate the parameters of the model. Subsequently, they compare the effects predicted by their model to the results of one new and various existing get-out-the-vote campaigns.

As encouraging as this strategy is, we note that in general there is no guarantee that a model that is wrong will yield observably wrong predictions (though it may still misguide). For example, imagine that the true model is one where a share $q > 0$ of customers behaves non-rationally and that behavioral customers always pay a share of $\frac{1}{2}$ of their endowment (instead of $\frac{3}{4}$). Further suppose again that we as researchers have in mind the wrong model where $q = 0$, i.e. everyone behaves rationally. Table 5 contains the results from a simulation based on these assumptions. Note first that our estimates of δ are, as one would expect, severely biased. Nonetheless, our prediction of the treatment effect in a context with $n = 2$ based on experimental data from the same context is only very slightly biased and the prediction of the treatment effect for $n = \infty$ based on data from this context is not biased at all. Predictions *across* contexts are severely biased, which we would, however, not discover if we conducted only a single experiment in one context.

Table 7: Diagnosis of design with incorrect yet observationally equivalent model

estimand	estimator	Bias	Estimand
τ_2	DIM_2	-0.00	0.30
τ_2	MLE'_2	0.01	0.30
τ_2	MLE'_∞	0.49	0.30
τ_∞	DIM_∞	0.00	-0.06
τ_∞	MLE'_2	-0.15	-0.06
τ_∞	MLE'_∞	0.00	-0.06
δ	MLE'_2	-0.15	0.80
δ	MLE'_∞	0.10	0.80
κ	MLE'_2	-2.62	6.00
κ	MLE'_∞	-0.09	6.00
q	MLE'_2		0.50

Why does a biased estimate of δ yield an accurate prediction of a treatment effect in the same setting? The reason is that a model in which there is a non-zero share of behavioral customers who always pay a share of $\frac{1}{2}$ can generate the same data as a model in which everyone behaves rationally. To see this, note that a model with $q = 0$ and $\delta = 0.8$ would result in the following average prices paid by first and second movers in a context with $n = 2$: $\pi_2^1 = \delta = 0.8$ and $\pi_2^2 = 1 - \delta = 0.2$. The same distribution of shares could also be produced by $q = \frac{1}{4}$ and $\delta = \frac{9}{10}$: $\pi_2^1 = q\frac{1}{2} + (1-q)\delta = \frac{1}{4}\frac{1}{2} + \frac{3}{4}\frac{9}{10} = 0.8$ and $\pi_2^2 = q\frac{1}{2} + (1-q)(1-\delta) = \frac{1}{4}\frac{1}{2} + \frac{3}{4}\frac{1}{10} = 0.2$.[18] In other words, even when we can use unbiased experimental estimates to validate our model, we may not discover that our model is wrong if the wrong model that we have in mind is observationally equivalent to the true model.

7 Conclusion

Even though field experiments have become immensely popular in political science, there are ongoing worries about how much we can actually learn from them. Two interrelated concerns are that experiments enjoy limited external validity and are disconnected from theory.

Throughout this chapter, we have highlighted the connection between these critiques and discussed various ways in which theories can help researchers learn more from their experiments. Examples range from selection diagrams that help assess how results can be trans-

[18] This problem exists for all models in which $\pi_n^1 = 1 - \pi_n^2$ for all parameter values.

ported from one setting to another to structural models that allow for the extrapolation of treatment effects to other settings or even other treatments.

Fundamentally, we think there is scope for researchers to do better on these fronts. Tools already exist. We close, however, with a worry. That theory is a powerful aid to inference is not a new idea. In fact, in some accounts, social experimentation first came up as a method in economics at a time when the dominant mode of inference was structural estimation. Heckman (1991) describes the vision of early experimentalists who were brought up in the structural tradition as one in which the primary goal of an experiment was not the non-parametric identification of an average treatment effect but the estimation of a structural model that could subsequently be used to assess the welfare consequence of numerous other experiments that had not actually been implemented. The turn towards "atheoretical" experimentation was in many ways motivated by concerns about the over-dependence of results on highly parametric and often unrealistic structural models.

Asking experimentalists to accept more theoretical assumptions in exchange for the ability to make broader claims may thus seem like going in circles.

The solution, we think, is to use theories with appropriate skepticism. Like any powerful tool, causal models are easy to misuse and should be handled with care. Such care includes being transparent about theoretical assumptions that drive results and assessing the robustness of conclusions to alternative assumptions. It also means using strategies that let us simultaneously employ theories, question them, and update them.

8 Appendix

8.1 Simulation and estimation when the model is correct

The estimator is declared as a function that returns a tidy data frame:

```r
# Maximum Likelihood Estimator to  estimate kappa (k), delta (d) and q:
structural_estimator <- function(data,
                                 Y = "Y_2_obs",
                                 # function to calculate
                                 # predicted price for rational
                                 # customers given z_i and delta,
                                 # defaults to n=2:
                                 y = function(Z, d) Z*d + (1-Z)*(1-d))

  {

  # Define negative log likelihood as a function of kappa, delta and q
  LL  <- function(k, d, q) {

              m <- with(data, y(Z,d))
              R <- q*dbeta(data[Y][[1]], k*3/4, k*1/4) +
                (1-q)*dbeta(data[Y][[1]], k*m, k*(1 - m))
              -sum(log(R))

              }

  # Estimation
  M <- mle2(LL, method = "L-BFGS-B",
            start = list(k = 2, d = 0.50,  q = 0.50),
            lower = c(k = 1,    d = 0.01,  q = 0.01),
            upper = c(k = 1000, d = 0.99,  q = 0.99))

  # Format output from estimation
  out <- data.frame(coef(summary(M)), outcome = Y)

  names(out) <- c("estimate", "std.error", "statistic",
                  "p.value", "outcome")

  # Use estimates of q and delta to predict
  # average treatment effects (ATEs)

  # Predicted ATE for n=2
  out[4,1] <- (1-out$estimate[3])*(2*out$estimate[2] - 1)
```

```
   # Predicted ATE for n=infinity
  out[5,1] <- (1-out$estimate[3])*(2*out$estimate[2]/
                                     (1+out$estimate[2]) - 1)

  out
}
```

The declaration then includes the estimation function as a design step:

```
# Define parameter values:
N = 500       # Sample size
d = 0.8       # True delta (unknown)
k = 6         # Parameter that affects variance (unknown)
q = 0.5       # Share of behavioral types in the population (unknown)
e = 3/4       # Price paid by behavioral customers (known)
random = TRUE # Switch to control whether
              # first movers are randomly assigned

# Declare the design:
design <-

  # Define the population
  declare_population(N = N,
                     # indicator for behavioral type (norm = 1)
                     norm = rbinom(N, 1, q),
                     # probablity of going first without random assignment
                     p = ifelse(norm == 1, .8, .5)
                     ) +

  # Define mean potential outcomes for n = 2
  declare_potential_outcomes(Y_2 ~ norm*e + (1-norm)*(Z*d + (1-Z)*(1-d))) +

  # Define mean potential outcomes for n = infinity
  declare_potential_outcomes(Y_inf ~ norm*e +
                                        (1-norm)*(Z*d/(1+d) +
                                        (1-Z)*(1-d/(1+d)))) +

  # Define estimands (quantities we want to estimate)
  declare_estimand(ATE_2 = mean(Y_2_Z_1 - Y_2_Z_0),       # ATE n = 2
                   ATE_inf = mean(Y_inf_Z_1 - Y_inf_Z_0), # ATE n = infinity
                   k = k,                                 # kappa
                   d = d,                                 # delta
                   q = q) +                               # q
```

```r
# Declare assignment process (random assignment if random = TRUE)
declare_assignment(prob = if(random) .5 else p, simple = TRUE) +

# Declare revealed potential outcomes
declare_reveal(Y_2,   Z) +
declare_reveal(Y_inf, Z) +

# Get draws from beta distribution given means for n = 2 and n = infinity
declare_step(fabricate, Y_2_obs   = rbeta(N, Y_2*k, (1-Y_2)*k),
                        Y_inf_obs = rbeta(N, Y_inf*k, (1-Y_inf)*k)
             ) +

# Declare estimators

# Difference-in-means for n = 2
declare_estimator(Y_2_obs ~ Z,
                  estimand = "ATE_2",
                  label = "DIM_2") +

# Difference-in-means for n = infinity
declare_estimator(Y_inf_obs ~ Z,
                  estimand = "ATE_inf",
                  label = "DIM_inf") +

# MLE for n = 2
declare_estimator(handler = tidy_estimator(structural_estimator),
                  estimand = c("k","d", "q", "ATE_2", "ATE_inf"),
                  label = "Struc_2") +

# MLE for n = infinity
declare_estimator(handler = tidy_estimator(structural_estimator),
                  Y = "Y_inf_obs",
                  y = function(Z, d) Z*d/(1+d) +  (1-Z)*(1-d/(1+d)),
                  estimand = c("k","d","q","ATE_2", "ATE_inf"),
                  label = "Struc_inf")
```

```
## Warning: Assigning non-quosure objects to quosure lists is deprecated as of rlang 0.3
## Please coerce to a bare list beforehand with `as.list()`
## This warning is displayed once per session.

## Warning: `quo_expr()` is deprecated as of rlang 0.2.0.
## Please use `quo_squash()` instead.
## This warning is displayed once per session.
```

Table 8: Detailed diagnosis of design with correct model

Estimand Label	Estimator	Bias	RMSE	Mean Estimate	Estimand
ATE_2	DIM_2	0.00	0.02	0.30	0.30
ATE_2	Struc_2	-0.00	0.02	0.30	0.30
ATE_2	Struc_inf	0.00	0.06	0.30	0.30
ATE_inf	DIM_inf	-0.00	0.02	-0.06	-0.06
ATE_inf	Struc_2	-0.00	0.01	-0.06	-0.06
ATE_inf	Struc_inf	-0.00	0.02	-0.06	-0.06
d	Struc_2	-0.00	0.01	0.80	0.80
d	Struc_inf	0.00	0.05	0.80	0.80
k	Struc_2	0.04	0.45	6.04	6.00
k	Struc_inf	0.03	0.45	6.03	6.00
q	Struc_2	-0.00	0.04	0.50	0.50
q	Struc_inf	-0.00	0.04	0.50	0.50

8.2 Simulation and estimation with correct model but without random assignment

```
# Change the design to remove random assignment
design_2    <- redesign(design, random = FALSE)
```

Table 9: Detailed diagnosis of design with correct model but no randomization

Estimand Label	Estimator	Bias	RMSE	Mean Estimate	Estimand
ATE_2	DIM_2	0.11	0.11	0.41	0.30
ATE_2	Struc_2	0.11	0.11	0.41	0.30
ATE_2	Struc_inf	0.12	0.14	0.42	0.30
ATE_inf	DIM_inf	0.08	0.08	0.02	-0.06
ATE_inf	Struc_2	-0.03	0.03	-0.09	-0.06
ATE_inf	Struc_inf	0.04	0.04	-0.02	-0.06
d	Struc_2	-0.01	0.02	0.79	0.80
d	Struc_inf	0.13	0.14	0.93	0.80
k	Struc_2	-0.09	0.43	5.91	6.00
k	Struc_inf	-0.04	0.46	5.96	6.00
q	Struc_2	-0.22	0.22	0.28	0.50
q	Struc_inf	0.01	0.04	0.51	0.50

8.3 Simulation and estimation when the model is wrong

```
# Turn random assignment on again
design_3    <- redesign(design_2, random = TRUE)

# New estimator that assumes q = 0
structural_estimator_2 <- function(data,
                                   Y = "Y_2_obs",
                                  # function to calculate
                                  # predicted price for rational
                                  # customers given z_i and delta,
                                  # defaults to n=2:
                                   y = function(Z, d) Z*d + (1-Z)*(1-d)){

  # Define negative log likelihood as a function of kappa and delta
  LL  <- function(k, d) {m <- with(data, y(Z,d))
                         R <- dbeta(data[Y][[1]], k*m, k*(1 - m))
                         -sum(log(R))}
  # Estimation
  M   <- mle2(LL, start = list(k  = 5, d = 0.5))

  # Format output from estimation
  out <- data.frame(coef(summary(M)),  outcome = Y)
  names(out) <- c("estimate", "std.error", "statistic",
                  "p.value",  "outcome")

  # Use estimates of delta to predict
  # average treatment effects (ATEs)

  # Predicted ATE for n=2
  out[3,1] <- 2*out$estimate[2] - 1

  # Predicted ATE for n=infinity
  out[4,1] <- 2*(out$estimate[2]/(1+out$estimate[2])) - 1
  out
}

# Update MLE for n = 2
design_3 <- replace_step(design_3, 11,
            declare_estimator(handler =
                                tidy_estimator(structural_estimator_2),
                              estimand =
                                c("k","d",  "ATE_2", "ATE_inf"),
                              label = "Struc_2_norm")
```

```
          )

# Update MLE for n = infinity
design_3 <- replace_step(design_3, 12,
            declare_estimator(handler =
                                tidy_estimator(structural_estimator_2),
                              Y = "Y_inf_obs",
                              y = function(Z, d)
                                 Z*d/(1+d) + (1-Z)*(1-d/(1+d)),
                              estimand = c("k","d", "ATE_2", "ATE_inf"),
                              label = "Struc_inf_norm")
          )
```

Table 10: Detailed diagnosis of design with incorrect model (assume $q = 0$)

Estimand Label	Estimator	Bias	RMSE	Mean Estimate	Estimand
ATE_2	DIM_2	0.00	0.02	0.30	0.30
ATE_2	Struc_2_norm	-0.04	0.04	0.26	0.30
ATE_2	Struc_inf_norm	0.52	0.52	0.82	0.30
ATE_inf	DIM_inf	-0.00	0.02	-0.06	-0.06
ATE_inf	Struc_2_norm	-0.17	0.17	-0.23	-0.06
ATE_inf	Struc_inf_norm	0.01	0.02	-0.05	-0.06
d	Struc_2_norm	-0.17	0.17	0.63	0.80
d	Struc_inf_norm	0.11	0.12	0.91	0.80
k	Struc_2_norm	-3.83	3.83	2.17	6.00
k	Struc_inf_norm	-3.05	3.05	2.95	6.00
q					0.50

```
# Change true data generating process
# so that behavioral customers pay 1/2

design_4    <- redesign(design_3, e = 1/2)
```

Table 11: Detailed Diagnosis of design with incorrect yet observationally equivalent model

Estimand Label	Estimator	Bias	RMSE	Mean Estimate	Estimand
ATE_2	DIM_2	-0.00	0.02	0.30	0.30
ATE_2	Struc_2_norm	0.01	0.02	0.31	0.30
ATE_2	Struc_inf_norm	0.49	0.49	0.79	0.30
ATE_inf	DIM_inf	0.00	0.02	-0.06	-0.06
ATE_inf	Struc_2_norm	-0.15	0.15	-0.21	-0.06
ATE_inf	Struc_inf_norm	0.00	0.02	-0.06	-0.06

Estimand Label	Estimator	Bias	RMSE	Mean Estimate	Estimand
d	Struc_2_norm	-0.15	0.15	0.65	0.80
d	Struc_inf_norm	0.10	0.10	0.90	0.80
k	Struc_2_norm	-2.62	2.62	3.38	6.00
k	Struc_inf_norm	-0.09	0.36	5.91	6.00
q					0.50

References

Abadie, Alberto, Susan Athey, Guido W Imbens, and Jeffrey Wooldridge. 2017. "When Should You Adjust Standard Errors for Clustering?" National Bureau of Economic Research.

Acemoglu, Daron. 2010. "Theory, General Equilibrium, and Political Economy in Development Economics." *Journal of Economic Perspectives* 24 (3): 17–32.

Allcott, Hunt. 2015. "Site Selection Bias in Program Evaluation." *The Quarterly Journal of Economics* 130 (3): 1117–65.

Avdeenko, Alexandra, and Michael J Gilligan. 2015. "International Interventions to Build Social Capital: Evidence from a Field Experiment in Sudan." *American Political Science Review* 109 (3): 427–49.

Bisbee, James, Rajeev Dehejia, Cristian Pop-Eleches, and Cyrus Samii. 2017. "Local Instruments, Global Extrapolation: External Validity of the Labor Supply–Fertility Local Average Treatment Effect." *Journal of Labor Economics* 35 (S1): S99–S147.

Blair, Graeme, Jasper Cooper, Alexander Coppock, and Macartan Humphreys. 2018. "Declaring and Diagnosing Research Designs." *American Political Science Review*, 1–22.

Blattman, Christopher, and Jeannie Annan. 2016. "Can Employment Reduce Lawlessness and Rebellion? A Field Experiment with High-Risk Men in a Fragile State." *American Political Science Review* 110 (1): 1–17.

Bowers, Jake, Bruce A Desmarais, Mark Frederickson, Nahomi Ichino, Hsuan-Wei Lee, and Simi Wang. 2018. "Models, Methods and Network Topology: Experimental Design for the Study of Interference." *Social Networks* 54: 196–208.

Bowers, Jake, Mark M Fredrickson, and Costas Panagopoulos. 2013. "Reasoning About Interference Between Units: A General Framework." *Political Analysis* 21 (1): 97–124.

Box, George EP. 1976. "Science and Statistics." *Journal of the American Statistical Association* 71 (356): 791–99.

Browne, Michael W, and Robert Cudeck. 1993. "Alternative Ways of Assessing Model Fit." *Sage Focus Editions* 154: 136–36.

Card, David, Stefano DellaVigna, and Ulrike Malmendier. 2011. "The Role of Theory in Field Experiments." *Journal of Economic Perspectives* 25 (3): 39–62.

Castillo, Marco, Ragan Petrie, Maximo Torero, and Lise Vesterlund. 2013. "Gender Differences in Bargaining Outcomes: A Field Experiment on Discrimination." *Journal of Public Economics* 99: 35–48.

Chong, Alberto, Ana L De La O, Dean Karlan, and Leonard Wantchekon. 2014. "Does Corruption Information Inspire the Fight or Quash the Hope? A Field Experiment in Mexico on Voter Turnout, Choice, and Party Identification." *The Journal of Politics* 77 (1): 55–71.

Coppock, Alexander, Thomas J Leeper, and Kevin J Mullinix. 2018. "Generalizability of Heterogeneous Treatment Effect Estimates Across Samples." *Proceedings of the National Academy of Sciences* 115 (49): 12441–6.

Cronbach, Lee J, and Karen Shapiro. 1982. *Designing Evaluations of Educational and Social Programs*. Jossey-Bass,

Dawid, Philip, Macartan Humphreys, and Monica Musio. 2019. "Bounding Causes of Effects with Mediators." *arXiv Preprint arXiv:1907.00399*.

Deaton, Angus. 2010. "Understanding the Mechanisms of Economic Development." *Journal of Economic Perspectives* 24 (3): 3–16.

Deaton, Angus, and Nancy Cartwright. 2018. "Understanding and Misunderstanding Randomized Controlled Trials." *Social Science & Medicine* 210: 2–21.

Dehejia, Rajeev, Cristian Pop-Eleches, and Cyrus Samii. 2015. "From Local to Global: External Validity in a Fertility Natural Experiment." National Bureau of Economic Research.

DellaVigna, Stefano. 2018. "Structural Behavioral Economics." National Bureau of Economic Research.

DellaVigna, Stefano, Attila Lindner, Balázs Reizer, and Johannes F Schmieder. 2017. "Reference-Dependent Job Search: Evidence from Hungary." *The Quarterly Journal of Economics* 132 (4): 1969–2018.

DellaVigna, Stefano, John A List, Ulrike Malmendier, and Gautam Rao. 2016. "Voting to Tell Others." *The Review of Economic Studies* 84 (1): 143–81.

Druckman, James N, Donald P Green, James H Kuklinski, and Arthur Lupia. 2011. *Cambridge Handbook of Experimental Political Science*. Cambridge University Press.

Dunning, Thad, Guy Grossman, Macartan Humphreys, Susan D Hyde, Craig McIntosh, and Gareth Nellis. 2018. "Metaketa I: Information, Accountability, and Cumulative Learning." Cambridge University Press.

Fisher, Ronald Aylmer. 1935. "The Design of Experiments."

Gelman, Andrew, Xiao-Li Meng, and Hal Stern. 1996. "Posterior Predictive Assessment of Model Fitness via Realized Discrepancies." *Statistica Sinica*, 733–60.

Gerber, Alan S, and Donald P Green. 2012. *Field Experiments: Design, Analysis, and Interpretation*. WW Norton.

Green, Donald P, Shang E Ha, and John G Bullock. 2010. "Enough Already About ?Black Box? Experiments: Studying Mediation Is More Difficult Than Most Scholars Suppose." *The Annals of the American Academy of Political and Social Science* 628 (1): 200–208.

Grose, Christian R. 2014. "Field Experimental Work on Political Institutions." *Annual Review of Political Science* 17: 355–70.

Harrison, Glenn W. 2014. "Cautionary Notes on the Use of Field Experiments to Address Policy Issues." *Oxford Review of Economic Policy* 30 (4): 753–63.

Heckman, James J. 1991. "Randomization and Social Policy Evaluation." National Bureau of Economic Research Cambridge, Mass., USA.

Huber, John D. 2017. *Exclusion by Elections: Inequality, Ethnic Identity, and Democracy.* Cambridge University Press.

Humphreys, Macartan, and Alan Jacobs. 2017. "Qualitative Inference from Causal Models." *Draft Manuscript (Version 0.2). Retrieved November* 27: 2017.

Ichino, Nahomi, Jake Bowers, and Mark M Fredrickson. 2013. "Ethnicity and Electoral Fraud in New Democracies: Modelling Political Party Agents in Ghana."

Imai, Kosuke, Luke Keele, Dustin Tingley, and Teppei Yamamoto. 2011. "Unpacking the Black Box of Causality: Learning About Causal Mechanisms from Experimental and Observational Studies." *American Political Science Review* 105 (4): 765–89.

Kern, Holger L, Elizabeth A Stuart, Jennifer Hill, and Donald P Green. 2016. "Assessing Methods for Generalizing Experimental Impact Estimates to Target Populations." *Journal of Research on Educational Effectiveness* 9 (1): 103–27.

Lakatos, Imre. 1970. "Falsification and the Methodology of Scientific Research Programmes." *Criticism and the Growth of Knowledge* 4: 91–196.

Lucas, Jeffrey W. 2003. "Theory-Testing, Generalization, and the Problem of External Validity." *Sociological Theory* 21 (3): 236–53.

Martinez, Seung-Keun, Stephan Meier, and Charles Sprenger. 2017. "Procrastination in the Field: Evidence from Tax Filing." UC San Diego Working Paper.

Mesquita, Ethan Bueno de, and Scott A. Tyson. 2019. "The Commensurability Problem: Conceptual Difficulties in Estimating the Effect of Behavior on Behavior." *Working Paper.*

Michelitch, Kristin. 2015. "Does Electoral Competition Exacerbate Interethnic or Interpartisan Economic Discrimination? Evidence from a Field Experiment in Market Price Bargaining." *American Political Science Review* 109 (1): 43–61.

Murtas, Rossella, Alexander Philip Dawid, and Monica Musio. 2017. "New Bounds for the Probability of Causation in Mediation Analysis." *arXiv Preprint arXiv:1706.04857.*

Neyman, Jerzy, and Egon S. Pearson. 1933. "On the Problem of the Most Efficient Tests of Statistical Hypotheses." *Philosophical Transactions of the Royal Society of London* 231 (694?706): 289?337.

Olken, Benjamin A. 2010. "Direct Democracy and Local Public Goods: Evidence from a Field Experiment in Indonesia." *American Political Science Review* 104 (2): 243–67.

Pearl, Judea. 2009. *Causality.* Cambridge university press.

Pearl, Judea, and Elias Bareinboim. 2014. "External Validity: From Do-Calculus to Transportability Across Populations." *Statistical Science*, 579–95.

Pritchett, Lant, and Justin Sandefur. 2015. "Learning from Experiments When Context Matters." *American Economic Review* 105 (5): 471–75.

Robins, James M. 2003. "Semantics of Causal Dag Models and the Identification of Direct and Indirect Effects." *Highly Structured Stochastic Systems.* Oxford University Press New York, NY.

Rosenbaum, Paul R. 2002. *Observational Studies.* New York, NY: Springer.

———. 2010. *Design of Observational Studies.* New York, NY: Springer.

Rubinstein, Ariel. 1982. “Perfect Equilibrium in a Bargaining Model.” *Econometrica: Journal of the Econometric Society*, 97–109.

Tipton, Elizabeth. 2013. “Improving Generalizations from Experiments Using Propensity Score Subclassification: Assumptions, Properties, and Contexts.” *Journal of Educational and Behavioral Statistics* 38 (3): 239–66.

Tipton, Elizabeth, and Laura R Peck. 2017. “A Design-Based Approach to Improve External Validity in Welfare Policy Evaluations.” *Evaluation Review* 41 (4): 326–56.

Todd, Petra E, and Kenneth I Wolpin. 2006. “Assessing the Impact of a School Subsidy Program in Mexico: Using a Social Experiment to Validate a Dynamic Behavioral Model of Child Schooling and Fertility.” *American Economic Review* 96 (5): 1384–1417.

Van Evera, Stephen. 1997. *Guide to Methods for Students of Political Science.* Cornell University Press.

Vivalt, Eva. 2019. “How Much Can We Generalize from Impact Evaluations?”

Westreich, Daniel, Jessie K Edwards, Catherine R Lesko, Stephen R Cole, and Elizabeth A Stuart. 2018. “Target Validity and the Hierarchy of Study Designs.” *American Journal of Epidemiology.*

Yeh, Robert W, Linda R Valsdottir, Michael W Yeh, Changyu Shen, Daniel B Kramer, Jordan B Strom, Eric A Secemsky, et al. 2018. “Parachute Use to Prevent Death and Major Trauma When Jumping from Aircraft: Randomized Controlled Trial.” *Bmj* 363: k5094.