Abstract

The external validity of empirical studies is a topic that has received recent attention due to the rise of popularity of the treatment effect model of empirical analysis and the so-called "crisis of reproducability." I review the current debate and put it in a historical context of statistics and econometrics. Further, I relate the historical evolution of this problem in econometrics to recent advances in the machine learning fields of domain adaptation and causal discovery. The exposition of this connection opens up grounds for further research and the development of new methods targeted at policy prediction based on empirical studies.

Introduction

Randomized control trials, or natural experiments that replicate them, have become the official gold standard of empirical work in economics and many related fields. This focus on identification via experimental or quasi-experimental methods has lead to the so-called "identification police" effect in many discussions of applied economics, where identification of a causal effect has become the primary qualification for a "successful" empirical study.

In response to this trend, several prominent economists have raised an alarm bell to dampen the party of the "randomistas," emphasizing that cleanly identified counterfactual analysis is not the end goal of econometrics and that external validity should be considered equally important (Shadish, T. D. Cook, and D. T. Campbell 2002,Heckman 2008, Deaton 2010, Manski 2013, Deaton and Cartwright 2018).

One concern is that this focus on counterfactual identification comes at the detriment of the other parts of the process of scientific induction needed to build knowledge that can be applied by policy makers. Heckman (2008) worries that "this emphasis on randomization or its surrogates, like matching or instrumental variables, rules out a variety of alternative channels of identification of counterfactuals from population or sample data." Similarly, Deaton and Cartwright (2018) worry that "the lay public, and sometimes researchers, put too much trust in RCTs over other methods of investigation."

Banjerjee and Snowberg (2016) echo the need for more formal systems for the external validity of experimental studies, saying "it is our belief that creating a rigorous framework for external validity is an important step in completing an ecosystem for social science field experiments, and a complement to many other aspects of experimentation."

Manski (2013) voices the concern that experimental studies have tended to "be silent" on the question of external validity and that "from the perspective of policy choice, it makes no sense to value one type of validity above the other. What matters is the informativeness of a study for policy making, which depends jointly on internal and external validity."

Applied research in economics, if it is meant to be "applied" to inform policy making, must therefore deal with external validity one way or another. Internal validity, identifying counterfactual effects, is not enough. Many top researchers are concerned about a lack of a rigorous framework for external validity in much applied research.

Following up on these theoretical concerns, a small but growing literature has sprung up around empirically proving that results from prominent RCT studies do not extrapolate to new contexts (Pritchett and Sandefur 2015, Gechter 2015, Allcott 2015, Bisbee et al. 2017, Rosenzweig and Udry 2019).

The purpose of the current work is to review the history of external validity and the process of empirical inference in economics and statistics.... ???

Definitions

Critiques against the randomistas have focused on their preference for internal validity and tendency to completely ignore external validity. It is worth defining these terms. The most agreed upon definition seems to come from Shadish, T. D. Cook, and D. T. Campbell 2002, where they update their taxonomy of validities from T. Cook and D. Campbell 1979:

> **Statistical Conclusion Validity:** The validity of inferences about the correlation (covariation) between treatment and outcome.
>
> **Internal Validity:** The validity of inferences about whether observed covariation between A (the presumed treatment) and B (the presumed outcome) reflects a causal relationship from A to B as those variables were manipulated or measured.
>
> **Construct Validity:** The validity of inferences about the higher order constructs that represent sampling particulars.
>
> **External Validity:** The validity of inferences about whether the cause-effect relationship holds over variation in persons, settings, treatment variables, and measurement variables.

Many authors use the term external validity to refer to what Shadish, Cook, and Campbell separate into construct and external validity. As both

are involved in generalization, and both are external to the particulars of the sample, I will follow that abuse of terminology and refer to all the challenges of both as "external validity."

What, then, is meant by the term "inference?"

Following from Aristotelian tradition of logic, inference is the process of reasoning and can be broken into two parts: reasoning from particulars to generals (induction) and reasoning from generals to particulars (deduction). Deduction is often recognized by Aristotle's syllogisms, such as:

> All men are mortal
> Socrates is a man
> Therefore, Socrates is mortal

Once one has induced a general law ("all men are mortal"), one can deduce facts that might otherwise not yet be apparent ("Socrates will die"). There is, of course, a natural contradiction in this process: how can one know that all men are mortal, if one did not already know that Socrates will die? In other words, how can one ever claim that "all men are mortal" until one has seen every man die?

This problem forms the foundation of David Hume's "Problem of Induction":

> "As to past Experience, it can be allowed to give direct and certain information of those precise objects only, and that precise period of time, which fell under its cognizance: but why this experience should be extended to future times, and to other objects, which for aught we know, may be only in appearance similar, this is the main question on which I would insist" (Hume).

There are two distinct problems Hume raises, that of generalizing from one object to another (from seeing some men die to assuming all men have died) and that of generalizing from the past to the future (because all men have died, thus, all men will die). He goes on to explain that this extrapolation is only valid under strong assumptions as to the "uniformity of nature." One must assume nature is uniform in such a way as to enable extrapolation from one object to another or from the past to the future.

R.A. Fisher framed his statistical techniques in terms of "inductive inference." In the introduction to The Design of Experiments (1935), he explicitly frames his canonical book and its techniques in the terms of logicians such as Hume, arguing for the possibility of induction via statistical methods:

> "it is possible to draw valid inferences from the results of experimentation... as a statistician would say, from a sample to the population from which the sample was drawn, or, as a logician might put it, from the particular to the general." (Fisher)

Fisherian statistical inference is a tool in the process of induction that seeks to address Hume's problem of extending experience with one object to that of similar objects. In particular, the "similar objects" to which conclusions are extended are not just similar in appearance, but rather have a distinct relationship to the experienced objects: they are the population from which the experienced objects represent a sample.

Fisher's methodologies for significance testing relate to drawing conclusions about populations given a sample. The validity of the use of such techniques in a study falls squarely under the category of "statistically conclusion validity" in the taxonomy of Shadish, Cook, and Campbell.

Fisher's theory of experiments (in particular, the RCT) addresses internal validity and causality. The connection between the RCT and the identification of a causal relationship comes straight from John Stuart Mill's (1843) "method of difference" for causal discovery:

> "If an instance in which the phenomenon under investigation occurs, and an instance in which it does not occur, have every circumstance save one in common, that one occurring only in the former; the circumstance in which alone the two instances differ, is the effect, or cause, or a necessary part of the cause, of the phenomenon."

Fisher (1935) clearly had this in mind when he defended randomization, claiming that holding all other variables constant was not feasible, and thus, holding them to the same distribution, by making the assignment of treatment independent of those variables, was desirable (Rosenbaum 2005). It should be noted, however, that for the subjects of interest to Fisher, the difference between something being a "cause" or "a necessary part of the cause" was not especially important. What Fisher was essentially interested in was interventional prediction: what is the effect when one makes individual changes to the growing conditions of these plants.

Holland (1986) begins his landmark paper, before formulating the Rubin-Neyman causal model, by framing his goal:

> "Others are concerned with deducing the causes of a given effect. Still others are interested in understanding the details of causal mechanisms. The emphasis here will be on measuring

> the effects of causes because this seems to be a place where statistics, which is concerned with measurement, has contributions to make."

The success of Fisher's framework of randomization and the Rubin-Neyman causal model comes down to this razor focus in purpose: they make no claims to discover all the causes of a given effect, to discover the mechanism of the cause, or even to separate between a cause or a necessary part of a cause. They allow us to reason about counterfactuals: what would have happened, on average and in the past, had we treated our entire population rather than a randomized part of a randomized sample drawn from that population. It operationalizes Mill's method of differences, creating a pathway to internal validity that is achievable in the real world.

The counterfactual causal model, however, provides no framework for generalizing from a specific population to a more general one or from the past to the future (Heckman 2008). The "effects of causes" is a black-box methodology for causal identification (Heckman 2008). It does not require one to answer the question "why" and it is therefore inherently context specific: it asks, "what happened when I did X in context D?" In Fisher's line of work (agriculture), none of those shortcomings were problematic. He was able to randomly sample from the exact population (seeds of grain) to which his inferences needed to generalize.

The success of Fisher's randomization in his field of agriculture, and the subsequent success of the Rubin-Neyman causal model in epidemiology, is in no way incidental. These are fields that deal with encapsulated biological units where agreed-upon scientific theory tells us that the relationships in these systems will be invariant to a wide degree of changes that happen in the world from one year to the next or from one country to the next.

For example, it is implicitly assumed that the growth of corn will not be affected by its expectations of how it will be cut down and processed during the harvest. Corn grows according to the properties of the soil and the sun it receives. Its growth will be independent of its destination in corn flakes or arepas, conditional on those properties. Opioids will inhibit pain in humans, regardless of their political ideology, faith in their governmental institutions, or their love of Shark Tank.

These arguments are not made explicitly, but their validity is absolutely necessary to enable the application of their findings: they create laws defined in relation to the entire range of contextual changes that one might want for prediction- and decision-making and that allows the laws to be

applied deductively in a wide array of useful situations.

While the validity of these arguments is taken for granted in contained biological processes, this is simply not the case in the social sciences. This is why Shadish, Cook, and Campbell lay out 19 different threats to construct and external validity of social science experiments. In the case of growing corn, the threats either simply do not apply or are implicitly neutralized through basic scientific understanding of plant growth.

In the case of economics, the outcomes are regularly based on individuals decisions to consume, work, study, invest, or move. The treatments are regularly subjected to a gauntlet of mediating factors and interacting variables that are highly context-dependent and correlated across individuals in a single place and time. The population for which one wants to draw actionable inference is in the future and it is fundamentally different from the population the sample from drawn from. Arguments for external validity must be made explicit, the internal validity of a study is rarely sufficient for scientific inference in the social sciences.

The Danger of Informal Inference

It might be argued that it is, and should be, up to the sound judgement and expert opinion of the policy maker to determine if a given counterfactual analysis should extrapolate to their context or not. Empirical studies only need to provide internal validity, according to this argument, as the extrapolation is done by experts who know their target domain.

While incorporating expert domain knowledge can only help predictions, we can differentiation between using a formal statistical framework and using an informal framework of judgement. John Stuart Mill, in his — –, reflects on these difference and the dangers of inferring without formal frameworks.

He begins by denying that the only process of inference consists of separately applying induction (particulars to generals) and then deduction (general law to particular context):

> "All inference is from particulars to particulars: General propositions are merely registers of such inferences already made, and short formulae for making more: The major premise of a syllogism, consequently, is a formula of this description: and the conclusion is not an inference drawn from the formula, but an inference drawn according to the formula: the real logical antecedent, or premise, being the particular facts from which

> the general proposition was collected by induction." (John Stuart Mill)

In other words, in the process of creating the general law, one has created a series of particular laws, and only once assured that all the particular laws are valid can one be assured of the more general law. He goes on to caution against the direct reasoning of particulars to particulars because it is informal and we are likely to bring our own biases into the process and make mistakes:

> "In reasoning from a course of individual observations to some new and unobserved case, which we are but imperfectly acquainted with (or we should not be inquiring into it), and in which, since we are inquiring into it, we probably feel a peculiar interest; there is very little to prevent us from giving way to negligence, or to any bias which may affect our wishes or our imagination, and, under that influence, accepting insufficient evidence as sufficient." (John Stuart Mill)

John Stuart Mill argues that formal procedures, such as that implied by the framework of induction and deduction, allow individuals to avoid these biases. In his world, there was no formal procedure for reasoning from particulars to particulars, but there was for reasoning from particulars to general. As such, he recommends the latter as a way to avoid biases of "wishes" and "imagination."

In an ideal world, the research community has discovered a set of general laws that are invariant to all contexts and the policy maker can apply them, via the process of deduction, to get the desired result in their circumstance. But what if that general law has not yet been discovered? Then the policy maker must look at individual studies (particulars) and attempt to infer a prediction for the result of a similar policy in their context. This is exactly the situation that John Stuart Mill has warned is fertile ground for bias and imagination.

Current State of the Art in Dealing with External Validity

Shadish, T. D. Cook, and D. T. Campbell 2002 lay out, along with their taxonomy of validities, a taxonomy of "threats" to each type of validity. While not a formal framework, per say, by creating the taxonomy they invite researchers to address each threat in turn. If a researcher is able to argue that all the threats to external validity have been addressed and protected against, then one might consider the work of generalizing to be finished.

In a sense, they provide a checkbox of things that are worth worrying about. Unfortunately, echoing again Manski's ((**Manski2008**) concern, researchers in applied economics have not systematically adopted a system of discussing threats to external validity.

Banerjee and Duflo (**Banerjee2014**) discuss concerns to experimental validity and generalization. In describing the concerns, they do not choose to use the taxonomy or terminology of Shadish, cook, and Campbell, instead reinventing a select few of the ideas, indicating a lack of impact the taxonomy of threats has had on the work of lead researchers today. As a primary tool to address external validity of RCTs, they emphasize the need for replication studies. Indeed, they posit what can be thought of as asymptotic theory of external validity from RCTs:

> "If we were prepared to carry out enough experiments in varied enough locations, we could learn as much as we want to know about the distribution of the treatment effects across sites conditional on any given set of covariates."

Asymptotic theory can be comforting, it's nice to know that the road we are on leads somewhere. However, the key term "any given set of covariates" must be restricted for this to be actionable in the finite lifespan of our species. The definition of those covariates comes down to defining the uniformity of nature assumption posed by Hume. If the covariate set is infinite, then we assume nothing regarding the nature of uniformity, and induction is impossible.

How, then, do we define those covariates? Which parts of nature must be uniform and which parts are allowed to change? This will answer the following practical question for any policy maker attempting to learn from previous studies: how many studies are enough studies? And if the answer to that question depends on where one intends to implement a new policy, then how can one relate the target context to the context of the studies? These questions are left unanswered.

Athey and Imbens (**Athey2017**) similarly reflect on the recent concerns over external validity quoted previously by the likes of Manski, Deaton, and Heckman. In an article aptly titled *The State of Applied Econometrics: Causality and Policy Evaluation*, they lay out three main recent advances in addressing external validity.

The first advance is that of addressing the concerns about the Local Average Treatment Effect (LATE) measured by instrumental variables (IV) (...). The concerns relate to the generalization of the local effect caused by

the variation in the instrument, to the global effect on the entire population (who were potentially not touched by the variation in the instrument in the study).

The second advance is that of addressing concerns regarding the local nature of regression discontinuity designs (...). These techniques all involve various methods to test whether the effect is only present locally at the discontinuity, or whether the effect is likely to extend to the rest of the sample.

Both of these techniques can be thought of addressing "statistical validity" in the validity taxonomy of Shadish, Cook, and Campbell, that of drawing inferences to the population from the sample. They do not address construct or external validity, they do not provide any information to the policy maker who is considering making a decision in another context.

The third advance is more salient to the major thrust of external validity: that of combining observational and experimental results.

Banjerjee and Snowberg (2016) create a formal system of "speculation" to create "falsifiable claims" of research studies as an attempt to codify the process of generalization and external validity that can be attached to any empirical study from the counterfactual paradigm. The falsifiable claims would presumably be used by other researchers to test the assumptions necessary for external validity of the original findings to hold true.

(TODO: something on meta-analysis methods – Meager)

The most obvious alternative to the treatment effect approach, which addresses the problem of external validity, is that of structural economics (Heckman 2008).

(TODO: some shortcoming of structural models).

It is worth reviewing the history of structure in economics and seeing how they thought about the problem of external validity.

The Origins of Structure

Economists during the same period as Fisher took a different approach to conceptualizing and thinking about the inference they were doing. Ragnar Frisch, theorizing about macro-dynamic analysis, set out several key ideas relating to the structure of a system and the autonomy of a structure (Frisch 1995). For Frisch, the "structure" of a system was all the characteristics of the phenomena that could be quantitatively described. In his macrodynamic systems, the structure is defined by a set of functional (simultaneous) equations. He then poses the question: what would happen

to the system due to an arbitrary change in a single variable? To do so could imply a different "structure" than the one which the equations describe, requiring a different set of equations altogether to describe the new system.

> "But when we start speaking of the possibility of a structure different from what it actually is, we have introduced a fundamentally new idea. The big question will now be in what directions should we conceive of a possibility of changing the structure?... To get a real answer we must introduce some fundamentally new information. We do this by investigating what features of our structure are in fact the most autonomous in the sense that they could be maintained unaltered while other features of the structure were changed... So we are led to constructing a sort of super-structure, which helps us to pick out those particular equations in the main structure to which we can attribute a high degree of autonomy in the above sense. The higher this degree of autonomy, the more fundamental is the equation, the deeper is the insight which it gives us into the way in which the system functions, in short, the nearer it comes to being a real explanation. Such relations form the essence of 'theory'."

This concept, that of "the essence of theory" being the discovery of some relationship that is autonomous and invariant to a great degree of changes we can imagine performing to a system, is taken up by Trygve Haavelmo (1944), who writes that:

> "The principal task of economic theory is to establish such relations as might be expected to possess as high a degree of autonomy as possible."

He then goes on to consider a distinction between the "invariance" of a relationship under hypothetical changes in structure versus the "persistence" of a relationship under observed changes in structure:

> "...if we always try to form such relations as are autonomous with respect to those changes that are in fact *most likely to occur*, and if we succeed in doing so, then, of course, there will be a very close connection between actual persistence and theoretical degree of autonomy."

This implies a connection between autonomy and the *type* of changes to which it is invariant *with respect to*. This point is made even more explicitly by Leonid Hurwicz (1966). Similar to his predecessors, his model consists of a system of equations that constrain the state of the world, given a history of states. He calls this system of equations a "behavior pattern." He states that:

> "A great deal of effort is devoted in econometrics and elsewhere to attempts at finding the behavior pattern of an observed configuration... But do we really need the knowledge of the behavior pattern of the configuration?... It will be approached here from the viewpoint of prediction... That is, the word 'need' in the above question will be understood as 'need' for purposes of prediction."

He then goes on to define what he calls a "structural form" as one which is identified and identical across all possible behavior changes that one *needs* to predict within. He stresses that:

> "The most important point is that the concept of structure is relative to the domain of modifications anticipated."

Thus, there is an inherent and irrevocable connection between what we consider a "law" and the degree of changes we require the law to persist across. The law is defined in relation to those changes. Folowing Hume, the performance of induction is always connected to a specific assumption about the uniformity of nature. Additionally, the exact way in which nature must be uniform, for a particular system under study, is defined by the predictions we need to make with the discovered relations in that system.

Invariant Conditionals

Robert Engle, building on the work of Frisch and Hurwicz regarding invariant/autonomous structures, creates a statistical definition of what he terms "super exogoneity." A good review of this historical evolution of autonomy can be found in Aldrich 1989.

Super exogoneity is defined by Engle as follows. Consider a model in which the "structure" (the functional form) of a relationship between outcome y and variable z is parameterized by "structural" parameters $\lambda_1, \lambda_2 \in \lambda$. If the joint density implied by the model can be factorized as follows:

$$P(y, z, \lambda) = P(y|z, \lambda_1)P(z|\lambda_2)$$

and the conditional, $P(y|z, \lambda_1)$ remains invariant to changes in the marginal $p(z)$ (presumably caused by changes in it's generating process, parameterized by λ_2), then z is super exogenous. This definition allows super exogoneity to be refuted by data:

> "It is clear that any assertion concerning super exogeneity is refutable in the data for past changes in $D(z, |X_{t-1}, \lambda_2)$ by examining the behavior of the conditional model for invariance when the parameters of the exogenous process changed...However, super exogeneity for all changes in the distribution of z, must remain a conjecture until refuted, both because nothing precludes agents from simply changing their behavior at a certain instant and because only a limited range of interventions will have occurred in any given sample period."

Writing a system in terms of super exogenous variables is akin to finding Frisch's "super-structure." This is the part of the system that stays invariant to a set of allowable modifications, parameterized by λ_2. This is the part of the system that must be known in order to make policy predictions, when λ_2 includes the changes in the policy, whose total causal effect (in the sense of Pearl 2000) is transmitted to the output variable y through z. The existence of such an invariant super-structure is a necessary prerequisite to successfully predict the effects of policy and therefore to successfully inform policy choice from empirical data.

It should be clear that, even if such a super-structure exists, it is possible that z is either fully or partially unobserved. How will that effect the invariance? Consider again the model with a slight modification (we drop λ_1 for ease of notation, as we are only interested in the conditions under which the conditional distribution is invariant and thus λ_1 is static):

$$P(y, z_1, z_2, \lambda) = P(y|z_1, z_2)P(z_1|z_2, \lambda_2)p(z_2|\lambda_3)$$

Where we consider a latent variable, z_2, which is independent of the policy modification of interest, λ_2, and whose generating process is controlled by structural parameter λ_3. We can marginalize out z_2 by fixing $\lambda_3 = \ell$:

$$P(y, z_1, z_2, \lambda_2, \lambda_3 = \ell_3) = P(y|z_1, \lambda_3 = \ell_3)P(z_1|\lambda_2, \lambda_3 = \ell_3)$$

Thus, our conditional is still invariant to modifications to λ_2, but it is fixed as regards the structural parameter λ_3 which determined the distribution of the latent variable z_2. If the latent variable, z_2, was also effected

by the structural parameters λ_2, then one would have to fix that as well in order to get an invariant conditional:

$$P(y|z_1, \lambda_2 = \ell_2, \lambda_3 = \ell_3)$$

It is worth considering the implications of the existence of latent variables in the super-structure that are evident in this simple exercise of probability algebra:

1. If the latent variable is not independent of the set of modifications to which a relationship should be invariant ($P(z|\lambda_2) \neq P(z)$), then the relationship cannot be determined generally.

2. If the latent variable is independent of the set of modifications to which a relationship should be invariant ($P(z|\lambda_2) = P(z)$, then marginalizing out the latent variable results in an invariant relationship.

This requirement is proved rigorously in Pearl, Bareinboim, and Mar 2014, where they refer to a relationship as "transportable" if it is invariant across a set of modifications.

Given that z_2 is latent, it might be difficult, in practice, to know whether or not it is independent of λ_2: one cannot directly obtain an empirical estimate of $P(z_2)$. What about going in the reverse direction: if one had empirical data across a set of modifications (different values of λ_2) and one discovered an empirical conditional distribution $P(y|z_1)$ that is invariant across those modifications, does that imply that $P(z_2|\lambda_2) = P(z_2)$?

This method of looking in the data for a conditional distribution that is invariant to "structural" changes that lead to differences in the marginal distribution of its "inputs" has formed the basis of much recent work in statistics and machine learning around causal discovery (**peters2015**, ...).

Domain Adaptation

I will consider the term domain adaptation following Ben-David. There are a set of source domains, S, for which labelled data is available and a target domain, T, for which only unlabelled data is available.

Invariance in Domain Adaptation

Conclusions and Future Research

Nancy Carwright (**Cartwright2016**) presents an example of a policy decision by policy makers in New York City to implement a new program called

Opportunity NYC. The program was modelled after a program in Mexico, Opportunidades, which had proved good results in reducing poverty there. Opportunity NYC was implemented in 2007 and shut down in 2010 due its failure to produce the desired effects.

What went wrong? Should the policy makers in New York simply have known that Mexico is clearly different and no program implemented there should be expected to work in New York? Did the writers of the Opportunidades study have an obligation to investigate and consider the threats to external validity that such a study might have been in danger of?

The purpose of this article was to argue that each of these cases is wrong:

1. Generalization, causal mechanisms, structure, support factors, etc. are all relative to a set of changes to which they must be invariant. One cannot answer the question about external validity without information about the differences between the source and target contexts.

2. The difference between New York and Mexico must be measured over a well-defined covariate space in order to make sense of the question of whether it is "too different." A core part of applied economics research must, therefore, consist of determining, for a given intervention, what part of nature must be uniform. This is akin to determining the super-structure of the events, the deep parameters, or the invariant mechanisms. New research has shown that the discovery of invariant mechanisms, given muti-domain datasets, is feasible. Thus, the data requirement has gone up (more than one study is required to make a prediction in a new context without many further assumptions), but the ability to give a formal answer to the question of whether it will probably work in New York or whether one has no idea, is within our reach and should therefore be pursued.

References

[1] By John Aldrich. "By JOHN ALDRICH* 1. Introduction only to those". In: *Oxford economic papers* 41 (1989), pp. 15–34.

[2] Hunt Allcott. "Site Selection Bias in Program Evaluation". In: (2015), pp. 1117–1165. DOI: 10.1093/qje/qjv015.Advance.

[3] Abhijit Banerjee, Sylvain Chassang, and Erik Snowberg. “Decision Theoretic Approaches to Experiment Desig”. In: (2016).

[4] James Bisbee et al. “Local Instruments, Global Extrapolation: External Validity of the Labor Supply–Fertility Local Average Treatment Effect”. In: *Journal of Labor Economics* 35.S1 (2017), S99–S147. ISSN: 0734-306X. DOI: 10.1086/691280.

[5] Thomas D Cook and D T Campbell. *Quasi-Experimentation: Design and Analysis Issues for Field Settings*. English. Houghton Mifflin, 1979.

[6] Angus Deaton. “Instruments, Randomization, and Learning about Development”. In: *Journal of Economic Literature* 48.2 (June 2010), pp. 424–455. ISSN: 0022-0515. DOI: 10.1257/jel.48.2.424. URL: http://pubs.aeaweb.org/doi/10.1257/jel.48.2.424.

[7] Angus Deaton and Nancy Cartwright. “Understanding and misunderstanding randomized controlled trials”. In: *Social Science & Medicine* 210.October 2017 (2018), pp. 2–21. ISSN: 0277-9536. DOI: 10.1016/j.socscimed.2017.12.005. URL: https://doi.org/10.1016/j.socscimed.2017.12.005.

[8] Ra Fisher. “The Design of Experiments”. In: (May 1935). ISSN: 0002-8312.

[9] Ragnar Frisch. *The Foundations of Econometric Analysis*. Ed. by David F. Hendry and Mary S. Morgan. Cambridge: Cambridge University Press, 1995, pp. 407–419. ISBN: 9781139170116. DOI: 10.1017/CBO9781139170116. URL: http://ebooks.cambridge.org/ref/id/CBO9781139170116.

[10] Michael Gechter. “Generalizing the Results from Social Experiments : Theory and Evidence from Mexico and India”. In: 2008 (2015), pp. 1–50.

[11] Trygve Haavelmo. “The Probability Approach in Econometrics”. In: *Econometrica* 12 (July 1944), p. iii. ISSN: 00129682. DOI: 10.2307/1906935. URL: https://www.jstor.org/stable/1906935?origin=crossref.

[12] James J Heckman. “Econometric Causality”. In: (2008).

[13] Paul W Holland. “Statistics and Causal Inference: Rejoinder”. In: *J Am Stat Assoc* 81.396 (1986), p. 968. ISSN: 0162-1459. DOI: 10.2307/2289069.

[14] Leonid Hurwicz. “On the Structural Form of Interdependent Systems”. In: *Studies in Logic*. Vol. 44. Board of Trustees of the Leland Stanford Junior University, 1966, pp. 232–239. DOI: 10.1016/S0049-237X(09)70590-7. URL: http://dx.doi.org/10.1016/S0049-237X(09)70590-7%20https://linkinghub.elsevier.com/retrieve/pii/S0049237X09705907.

[15] Charles F. Manski. “Public Policy in an Uncertain World”. In: *Public Policy in an Uncertain World* (2013). DOI: 10.4159/harvard.9780674067547.

[16] Judea Pearl. *Causality Second Edition*. 2000, p. 386. ISBN: 0521773628. DOI: citeulike-article-id:3888442.

[17] Judea Pearl, Elias Bareinboim, and M E Mar. “External Validity : From Do-Calculus to Transportability Across Populations”. In: 29.4 (2014), pp. 579–595. DOI: 10.1214/14-STS486. arXiv: arXiv:1503.01603v1.

[18] Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. “Causal inference using invariant prediction : identification and confidence intervals”. In: (2015), pp. 1–51. arXiv: arXiv:1501.01332v3.

[19] Lant Pritchett and Justin Sandefur. “Learning from Experiments when Context Matters”. In: *American Economic Review* 105.5 (May 2015), pp. 471–475. ISSN: 0002-8282. DOI: 10.1257/aer.p20151016. URL: http://eds.b.ebscohost.com.ezproxy.lib.usf.edu/eds/pdfviewer/pdfviewer?vid=5%7B%5C&%7Dsid=aef21225-20b8-458e-842b-e860243ea5f7%7B%5C%%7D40sessionmgr104%7B%5C&%7Dhid=104%20http://pubs.aeaweb.org/doi/10.1257/aer.p20151016.

[20] Mateo Rojas-Carulla, Bernhard Schölkopf, and Richard Turner. “Invariant Models for Causal Transfer Learning”. In: 19 (2018), pp. 1–34.

[21] Paul R. Rosenbaum. “Heterogeneity and causality: Unit heterogeneity and design sensitivity in observational studies”. In: *American Statistician* 59.2 (2005), pp. 147–152. ISSN: 00031305. DOI: 10.1198/000313005X42831.

[22] Mark R. Rosenzweig and Christopher Udry. “External Validity in a Stochastic World: Evidence from Low-Income Countries”. In: (2019).

[23] William R Shadish, Thomas D Cook, and Donald T. Campbell. *Experimental and Designs for Generalized Causal Inference*. 2002. ISBN: 0395615569.