What does it mean to learn across studies?  
Explanation and Generalization

19 May 2016

# Overview

For more than half a century, the statistical analysis of large data sets has prevailed as the preferred approach to the study of political phenomena. One term, *a quest to generalize*, has been the primary motivation underlying the rise of large-N quantitative analysis. Przeworski and Teune made an especially influential statement in their *Logic of Comparative Social Inquiry* (xxxx), which was written as a critical reaction to the qualitative case study approach that had dominated disciplinary research for decades. (need quote)

Closely intertwined with the rise of survey data, generalization came to be associated with two distinct notions, a distinction that often became blurred. First, generalization refers to the process whereby a researcher uses an observed random sample to learn about an unobserved population. It is no exaggeration to assert that this statistical discovery has shaped modern social scientific research more than any other. One need do no more than reflect on the contributions of studies using data sets such as the National Election Studies and the World Value Surveys. Second, generalization refers to the idea that findings based on many cases are more generalizable than findings based on few. Thus the rise of cross-national studies. Indeed, scholars who use cross-national data might argue that they succeed in meeting both conceptions of generalization.

We argue that the aspiration to generalize, in both senses, has been misdirected. In the first instance, and perhaps surprisingly, inferring from a random sample to a population is vulnerable to the strong possibility that some cases in the sample will be weighed more heavily than others in regression models, thus undermining the very idea that researchers can safely infer from the sample to population from which the sample is drawn (Aronow and Samli 2016). In the second instance, and our primary focus, the idea that an empirical study can and should be generalizable misses an essential point: no study, including cross-national studies that use data collected from many countries, is generalizable.

Humans, not empirical studies, generalize. Cognitive psychologists have demonstrated that efforts to explain causes people to “learn more effectively and generalize more readily to novel situations” (Williams and Lombrozo 2010). Explaining causes people to interpret observations in terms of unifying themes and patterns—lest we use the term, theories? Of course, this explanation-generalization nexus will be especially operative amongst, although not limited to, trained social scientists. The important point: generalization, using what they already know, describes what social scientists *do* when they try to explain something new and unknown.

If all the above is true, then to criticize some studies as inherently not generalizable, and thus not holding equal status to large-data-set, quantitative studies, misses the mark. To the contrary, all types of study—case studies, randomized experiments, quasi-experiments, and observational studies based on a convenient collection of cases—have a potential to enhance scholars’ explanations and understandings of political phenomena. At the same time, each suffers its unique limitations. (AN INTERESTING QUESTION—WHICH WOULD WE PREFER—A LARGE N STUDY IN WHICH THE AUTHOR MAKES LITTLE EFFORT TO EXPLAIN, AND THUS FALLS SHORT OF GENERALIZATION, OR A SMALL N STUDY IN WHICH THE AUTHOR OFFERS A RIGOROUS AND COHERENT EXPLANATION THAT, ALMOST BY DEFINITION, WILL BE GENERALIZABLE?)

# Intellectual Context

We are not the first to engage the topic of generality. In suggesting that completing a single study often will not suffice, Rosenbaum (2010) contends that generalization often arises when a researcher conducts multiple studies, rather than a single study:

If you do not know the effects of the treatment on the individuals in your study, you are not well-positioned to infer the effects on individuals you did not study who live in circumstances you did not study. Randomization addresses internal validity. In actual practice, [generalization] is often addressed by comparing the results of several internally valid studies conducted in different circumstances at different times. (Rosenbaum 2010, 56–57)

We would only add that multiple studies will be most effective when conducted serially, such that each subsequent study addresses a matter not conclusively addressed in the preceding study. Some of the most convincing efforts we know follow this very pattern. Notable examples include the following: Festinger’s research on cognitive dissonance, which began as an exploration of the beliefs and attitudes of public housing residents; Snow’s coordinated sequence of explorations to identify the source of cholera; Wand et al.’s research into the effects of using the butterfly ballot in cc County, Florida on vote totals in the 2000 presidential election; and Iyengar and Kinder’s many sequential laboratory experiments designed to understand the effects of media on what people think about and how they think about it. None of these research endeavors would have succeeded by using a single, large-N data set.

There have also been recent efforts to *estimate* the effects of an experimental or observational intervention on unobserved individuals. For example, Cole and Stuart (2010) … DuGoff, Schuler, and Stuart (2014) work on the case of learning about causal effects for the underlying population when one has a randomly sampled population and a matched observational design. Stuart et al. (2011) and Hartman et al. (2015) develop methods for extrapolating or forcasting treatment effects from an estimate within one randomized and observational study to individuals who are not a part of the experimental pool, but for whom investigators have observed background characteristics that overlap with the background characteristics of those in the experimental pool.

And the EGAP metaketa initiative is currently fielding groups of related randomized field experiments in an effort to learn more about underlying questions about political accountability and voter information (for example) <http://egap.org/metaketa>.

This essay does not provide a method or technique, but we hope, reminds social scientists how to talk and think about what we gain from one study or several, and how, in general, we ought to think about learning across studies of different kinds.

To summarize this paper in two sentences, we would say: “Studies do not generalize. Only humans generalize.”

---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

Our discussion proceeds as follows. We first ask, what goal(s) are social scientists trying to achieve? Explanation and understanding, we argue, are the ultimate goals of all social scientific studies. Next, we discuss the relationship between observation and explanation, and, especially, how observation can enhance explanation. (theory???) Finally, we identify some common approaches (research designs?) and discuss the potential and unique contributions and limitations of them. (Jake, please feel free to reject this roadmap, although I tried to stay close to your draft. If it is approximately right, we have a lot of work to do. Right now, I see a lot of gaps. We will not be able to fill them all in before the presentation, to be sure, although we might try to bolster the current discussion, more in the form of filling in.)

------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

What should one study teach us? Most social scientists would agree that different research designs have different strengths. We routinely hear, for example, that the randomized experiment provides an ideal design for teaching us about counterfactual comparisons often called “causal effects”. Or that ethnography provides an ideal design for learning about the social construction of meaning. Or that the large scale cross-sectional comparison, say, among all the current nations of the world, or among the respondents in a large random sample of some population, allows for great “generalizability”. Colleagues tell us that sometimes they must trade clarity of understanding within a specific context (“internal validity”) for broad applicability (“external validity”). In fact, we often hear people ask “Is this study generalizable?”.

In this essay we argue in favor of a reconceptualization of terms like “generalizability” and “external validity”. We suggest that one should not ask “Is this study generalizable?” because, in fact, no study is generalizable. Generalization is an action taken by human beings. In recent years some have argued that a key weakness of case studies, multi-case ethnography, and randomized experiments (in the lab or field) has been “generalizability” or “external validity”. In this essay we suggest that the weaknesses of those studies are shared by *all social science studies* in which scholars observe some part of the social world.

For example, studies falling under the potential outcomes framework would appear to dismiss, or, at best, give short rift to, the very goal that motivated the efforts to create disciplinary capacity to move from indepth examinations of few cases to large-scale cross-sectional studies for more than 50 years: a capacity to make general statements that apply across units. We take this apparent conundrum as an opportunity to distinguish two processes: generalization, where the researcher uses an observed sample to learn about an unobserved population, and confirmation, where the researcher replicates studies to determine whether the original findings or the theory behind them hold across people or settings. Political scientists rarely invoke the term “replication,” even though Popper and many philosophers of science since him have discussed it at length. Instead, they use the term “generalization” to apply to both processes, or to either one of them.

Scholars who use cross-national data might argue that they simultaneously generalize and confirm. They generalize because they estimate individual-level relationships across a number of countries, and, in so doing, they also confirm (or not). Unfortunately, such studies suffer the very problem of potentially biased estimates that experiments and quasi-experiments are designed to overcome.

# Intellectual Context

We are not the first to engage with these issues. For only one example, By suggesting that research design alone does not allow clear learning from any given study, some point out that external validity arises across multiple studies. For example, Rosenbaum (2010) points out that the idea of learning from one study to another is a function of multiple studies not any single study:

If you do not know the effects of the treatment on the individuals in your study, you are not well-positioned to infer the effects on individuals you did not study who live in circumstances you did not study. Randomization addresses internal validity. In actual practice, external validity is often addressed by comparing the results of several internally valid studies conducted in different circumstances at different times. (Rosenbaum 2010, 56–57)

There have also been recent efforts to *estimate* the effects of an experimental or observational intervention on unobserved individuals. For example, Cole and Stuart (2010) … DuGoff, Schuler, and Stuart (2014) works on the case of learning about causal effects for the underlying population when one has a randomly sampled population and a matched observational design. Stuart et al. (2011) and Hartman et al. (2015) develop methods for extrapolating or forcasting treatment effects from an estimate within one randomized and observational studies to individuals who are not a part of the experimental pool, but for whom investigators have observed background characteristics that overlap with the background characteristics of those in the experimental pool. Aronow and Samii (2016) shows . Also ao2014generalizing ….

And the EGAP metaketa initiative is currently fielding groups of related randomized field experiments in an effort to learn more about underlying questions about political accountability and voter information (for example) <http://egap.org/metaketa>.

This essay does not provide a method or technique, but we hope, reminds social scientists how to talk and think about what we gain from one study or several, and how, in general, we ought to think about learning across studies of different kinds.

To summarize this paper in two sentences, we would say: “Studies do not generalize. Only humans generalize.”

# What do social scientists want to do?

Social scientists seek to explain.

## What does it mean to explain?

Most scientific explanations either describe a logic or mechanism producing some specify class of outcome or produce a unifying account of many different kinds of outcomes. For example, consider explanations for the speed of a ball rolling down a hill. In one form of explanation, we can explain the speed of a ball rolling down a hill with an equation representing how the parts of the explanation fit together — say, our explanation only needs the mass of the ball and the angle of the slope given an assumption of a smooth hill. Why does the ball roll as fast as it does at the bottom of the hill? Because the ball is this heavy and the slope is this inclined. Another form of explanation would “explain” the speed of the ball by saying, “The ball rolls at that speed because of gravity.” (with perhaps a more elaborated account of how gravity affects the falling speed of relatively small objects on planets). ([This website seems useful as a starting place to think about explanation][http://plato.stanford.edu/entries/scientific-explanation/])

## Why explain?

A good explanation enables us to develop expectations about the unobserved. An explanation of the rolling speed of a ball should help us develop expectations about other balls, in other places, and other circumstances. A good explanation need not offer precise forecasts — the model of a smooth ball on a smooth incline may not offer precise predictions about a boulder on a mountain top. But, an explanation that links mass and incline to the speed of a ball might encourage us to direct our observations at the mass of the boulder and the incline of the mountain as we go about developing an expectation for the speed of the boulder.

A good explanation helps us understand the world. And it is this impetus to understand, and perhaps change, the world, drives social scientists to do their work.

A good explanation helps us generalize — that is, a good explanation helps us develop expectations about that which we do not observe.

## How does explanation relate to observation?

Explanations generate ideas and expectations. In the face of many explanations we seek ideas that are testable from among the many inherent in any given explanation. That is, explanations generate hypotheses. The purpose of empirical observation is to try to teach us about an existing explanation via its hypotheses. Sometimes the observations make it difficult for us to believe a given explanation (say, the mass of many different round objects appears unrelated to the speed of rolling down a slope) and thus demand new explanations. Sometimes the observations support the explanation.

This task, in turn, raises the question, what kinds of observation are best suited to aid scholars accomplish their central task of explanation?

# How might observation enhance explanation?

The kinds of observation that are best suited to enhance explanation have long been known to have two characteristics: First, since any given observation is specific to a place and time, and since observations of humans at one place and one moment in time cannot be credibly claimed to represent exactly other humans at another time, one must explain what a given observation “is a case of” (cite to someone talking about case studies? comparative politics?). Second, since ideas describing comparisons tend to be particularly amenable to test, comparisons should reflect as clearly as possible on the hypothesis (cite? perhaps to Kinder and Palfrey on interpretable comparisons?). Both criteria enable a given observation to enable interpretation in terms of a given explanation as clear as possible.

[example or two?]

Here we describe a few common research designs, and articulate how each of them may help political scientists accomplish their central task of explanation.

## A Random Sample Survey: Controlled selection of units helps describe what a sample is a case of

Most survey research involves observing a large sample aiming to represent some well-defined population. A random sample supports arguments in favor of the idea that the observations in a sample are a case of the population. A hypothesis tends to articulate an expectation about a comparison between units. In a survey, a scholar might compare groups of respondents defined on the bases of values of their survey responses (or other values associated with each respondent). Such comparisons may well reflect on the hypothesized comparison, however, such uncontrolled comparisons tend to also contain many other differences between the survey respondents. So, this simple design may not teach us as much about the hypothesis and explanation as we would desire because the observed comparison may be difficult to interpret.

## A Randomized Experiment: Controlled assignment of an intervention helps clarify a comparison

## An Observational Study On A Convenient Collection of Units

## A Case Study

## Summary:

Notice that all of these study types have the potential to help social scientists explain. Notice also that their ability to enable explanation does not hinge on the representativeness of the units under study.

# Discussion and Conclusion: Data do not generalize, Scholars do.

“I only did a case study/lab experiment, so I cannot generalize.”

“If only I had a random sample, then I could generalize.”

If this essay has been compelling, then readers should realize that such common apologies are not necessary. To ask, “do these findings generalize?” is either a vague or misformed question. The act of generalization — i.e. the forming of expectations about circumstances beyond those observed — is a human act. Humans generalize. Humans generalize because they have an explanation for a phenomenon. That is, explanation is a tool for generalization.

Good observation helps us have confidence in existing explanations or demands new explanations. Yet, explanatorily useful observation can arise in many ways.

# References

Aronow, Peter M, and Cyrus Samii. 2016. “Does Regression Produce Representative Estimates of Causal Effects?” *American Journal of Political Science* 60(1): 250–67.

Cole, Stephen R, and Elizabeth A Stuart. 2010. “Generalizing Evidence from Randomized Clinical Trials to Target Populations the Actg 320 Trial.” *American journal of epidemiology* 172(1): 107–15.

DuGoff, Eva H, Megan Schuler, and Elizabeth A Stuart. 2014. “Generalizing Observational Study Results: Applying Propensity Score Methods to Complex Surveys.” *Health services research* 49(1): 284–303.

Hartman, Erin, Richard Grieve, Roland Ramsahai, and Jasjeet S Sekhon. 2015. “From Sample Average Treatment Effect to Population Average Treatment Effect on the Treated: Combining Experimental with Observational Studies to Estimate Population Treatment Effects.” *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 178(3): 757–78.

Rosenbaum, Paul R. 2010. *Design of Observational Studies*. Springer. <http://www.springer.com/statistics/statistical+theory+and+methods/book/978-1-4419-1212-1>.

Stuart, Elizabeth A, Stephen R Cole, Catherine P Bradshaw, and Philip J Leaf. 2011. “The Use of Propensity Scores to Assess the Generalizability of Results from Randomized Trials.” *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 174(2): 369–86.