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

7 June 2016

# Overview: The Quest to Generalize

One of the primary motivations underlying the rise of large-N quantitative analyses as a tool for advancing the study of politics has been a quest to generalize. Przeworski and Teune (1970) made an especially influential statement in their *Logic of Comparative Social Inquiry*, which was written as a critical reaction to the qualitative case study approach that had dominated disciplinary research for decades.

(need quote about replacing proper names with general statements, the need for large-N comparison)

Closely intertwined with the rise in the use of survey data in political science, generalization came to be associated with two distinct notions, a distinction that often became blurred. First, generalization referred to the process whereby a researcher uses an observed random sample to learn about an unobserved population. The idea that a mean calculated in a sample may be an unbiased estimator of a unobserved population quantity motivated discipline-wide efforts to make random sampling a standard over quota sampling or convenient case choices: the substantive impact of research based on the American National Election Studies and World Values Surveys arises in part from this methodological insight. Second, generalization refers to the idea that findings based on many cases teach us more about the unknown than findings based on few. The idea of studying nearly all the countries in the world, or all of the wars fought, via cross-national studies seems to have been justified by the idea that statistical precision enhances the scientific value of 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 even if the statistical foundations underlying claims of unbiased and consistent and precise estimators were correct. In the first instance, and perhaps surprisingly, inferring about counterfactual causal effects from a random sample to a population using a regression model is vulnerable to the strong possibility that some cases in the sample will be weighed more heavily than others, thus undermining the very idea that researchers can safely infer from the sample to population from which the sample is drawn (Aronow and Samii 2016). In the second instance, even a descriptive inference (like a mean of support for some candidate) calculated from an excellent and large random sample at of a well-defined population at one point in time does not, in fact, perfectly target the same population quantity the next day: history operates, humans and politics change one another, the population sampled today will be different tomorrow, you can't survey the same river twice. In the third instance, and our primary focus here, 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 without human action.

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, Lombrozo, and Rehder 2010). Explaining causes people to interpret observations in terms of unifying themes and patterns—lest we use the term, to theorize? Of course, this explanation-generalization nexus will be especially operative amongst, although not limited to, trained social scientists when it comes to studying human behavior in the social, political, and economic world. The important point here is that, 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 as contributors to scientific understanding. This leads us to a question that we will return to in the conclusion: which would we prefer, a large survey study with a random sample describing a well-defined population with a simple number or a small survey study of an idiosyncratic and convenient sample in which multiple implications of a theory are assessed and compared to one another? If we could have multiple studies, how would they vary? What would be the point of valuing multiple studies over, say, one excellent study?

# 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 Palm Beach 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 more on them). Also, DuGoff, Schuler, and Stuart (2014) show how one may estimate average counterfactual 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 forecasting counterfactual 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) (see <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."

## Paper Plan

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. Finally, we identify some common approaches to observation and discuss the potential and unique contributions and limitations of them.

# Explanation and Understanding

## What does it mean to explain?

Most scientific explanations either describe a logic or mechanism producing some specific 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).[[1]](#footnote-27)

## Why explain?

A good explanation enables us to easily and quickly develop expectations about the unobserved. Perhaps a good explanation does a good job predicting what will happen when we confront the next set of previously unobserved events, relations, and units. 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. This does not imply that one must formally test hypotheses, but that the point of observation is to reflect on expectations about partially or unobserved relationships or quantities: and this is the definition of a hypothesis for us. 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" (cites). 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 (1993) on interpretable comparisons? Rosenbaum and Fisher on elaborate theories with many observable implications?).

# Types of Studies

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 “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. [Distinguish between generalizability and external validity here?]

For example, studies falling under the potential outcomes framework would appear to dismiss, or, at best, give short shrift to, the very goal that motivated the efforts to create disciplinary capacity to move from in-depth 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. Only recently are political scientists invoking 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.

Here we describe a few common research designs, and articulate how each of them may help political scientists accomplish their central task of explanation. We use our discussion of randomized experiments to distinguish between the concepts of "external validity" and “generalizability”.

# Generalization and External Validity in Randomized Experiments

Scholars routinely use the terms "generalization" and “external validity” interchangeably but we think that they differ. We have already explained that “generalization” is an act undertaken by human actors and does not inhere in any given research design.

What about external validity?

Consider an experiment in which red and green M&M candies were randomly assigned to political scientists Methodology Society Business Meeting at APSA and where the outcome measured was whether a given individual said, "Yum." What would we learn from this study? It seems trivial. What do we mean by trivial? Randomization guarantees that our comparison is unconfounded and provides what Kinder and Palfrey (1993) would call an “interpretable comparison.” Why does it feel like this study would be a waste of time?

Consider another randomized experiment in which a large random sample of adults are assigned to watch negative campaign ads versus positive ads, and in which the turnout of those adults is measured. Say that this experiment discovers no effects of viewing negative ads on turnout. Yet, another experiment randomly assigned negative ads to media markets, and finds a negative effect on turnout at the media market level. In the first experiment, we have an effect on a combination of people people who would not have chosen to watch the ads given the chance and people who would have chosen to watch the ads. In the second experiment, we assume that most of those watching the ads were those who found them attractive — so that the effect of the negative ads was negative among those who chose to watch them. Is the first study, where the treatment effect is unrealistic, bad science? Or a study from which we cannot or ought not learn?[[2]](#footnote-34)

While scholars could use either study as a source of learning and prediction, neither study is especially externally valid because we have trouble linking what the study is showing with the theoretical interests of social scientists.[[3]](#footnote-35)

In the first case, neither the action nor the population is of general social scientific interest. We have spent no effort to develop expectations or theories that have implications for the situation. Thus, even if the randomization guarantees interpretable comparisons and thus an internally valid study, we might say, in a short hand kind of way, "It doesn't add to knowledge." But really mean something more like, “We have no previous language game within which this new study has meaning.”

In the second case, the sample reflects an interesting population and the action/intervention produces a difference with expectations or a distinction that matters. However, what happened in one experiment does not reflect what happened in another experiment. And that difference between experiments raises the question about whether, say, a policy banning negative advertisements would raise turnout overall in the short or long run. Whereas, multiple experiments, all showing the same relationship, would naturally lend support to the guess that a policy change would have a positive effect.

Thus, although randomization does provide what Fisher called a "reasoned basis for inference" or what Kinder and Palfrey called “interpretable comparisons” within the context of a given experimental pool or delimited group of observations, different experiments can aid scientific generalization more or less, and thus can aid policy making more or less as well. Experiments that speed learning about simple hypotheses (like “Negative ads depress turnout”) often tend to be those which simulate the process as it actually works outside of the experimental context.

What about accumulation of knowledge? Or learning across studies? We believe that the idea of the benefits of multiple studies describes what philosophers label as confirmation.[[4]](#footnote-36)

In an important although not widely cited article, Mook (1983) makes similar points. He argues that the concepts of internal validity vs. external validity/generalization nexus as social scientists employ it applies nicely to corn and agriculture, where an experiment done on one field may well seem to teach us something very direct about what would happen on another field if the same intervention were deployed. It applies less usefully to social phenomena, where the intent of the research must be to assess the generality of *theoretical* conclusions: the extent to which one theory or explanation developed to explain one situation in a general way also helps to explain other situations. He gives as examples two highly influential experimental studies, Harlow's wire-and-cloth-mother experiments with rhesus monkeys and Milgram’s experiments on obedience to authority. In neither case did the researcher try to simulate conditions as they existed outside the experimental context. In fact, experimental conditions could not have been less like real-world conditions. Harlow, for example, used cloth and wire "mothers," for the baby rhesus monkeys, which, needless to say, do not emulate mothers in jungle conditions. Moreover, the rhesus monkeys that Harlow used in the experiments had grown up in captivity; they did not resemble jungle monkeys in any sense. Harlow was not trying to make general statements about how monkeys interact with their mothers in real settings. Rather, he used his experiments to test a prevailing theory that hunger reduction led baby monkeys (and human beings, for that matter) to cuddle up to their mothers. Holding availability of food constant, he found that baby monkeys favored cloth “mothers” to wire ones.

In short, experiments often can (and should) be used to test prevailing theories and explanations, not to generate findings about how some phenomenon works outside the experimental context. Mook (1983) summarizes as follows:

[When employing experiments], there are a number of things we may be doing. We may be asking whether something *can* happen, rather than typically whether it *does* happen. Second, our prediction may be in the other direction: it may specify something that ought to happen *in the lab*, and so we may go to the lab to see whether it does. Third, we many demonstrate the power of a phenomenon by showing that it happens even under unnatural conditions that ought to preclude it. Finally, we may use the lab to produce conditions that have no counterpart in real life at all, so that the concept of "generalizing to the real world" has no meaning. But even where findings cannot possible generalize and are not supposed to, they can contribute to an *understanding* (our italics) of the processes going on. Once again, it is that understanding which has external validity (if it does)—not the findings themselves, much less the setting and the sample. (page 382)

## 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 is an expectation about a comparison between unobservable 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 Comparative Observational Study On A Convenient Collection of Units

[TO DO]

## A Case Study

[TO DO]

## 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/have a small sample, so this study is not generalizable."

"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.

Certain research designs make it easier to estimate or forecast into the future or into other, as yet unobserved, units. However, research design per se does not determine how we can use one study to update our understanding of a social phenomenon let alone how a creative community of social scientists can apply that understanding to the design of a new study or a policy or to make a judgement about the social and political world.

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.

Gaines, Brian J, and James H Kuklinski. 2011. “Experimental Estimation of Heterogeneous Treatment Effects Related to Self-Selection.” *American Journal of Political Science* 55(3): 724–36.

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.

Kinder, D.R., and T.R. Palfrey. 1993. “On Behalf of an Experimental Political Science.” *Experimental foundations of political science*: 1–39.

Mook, Douglas G. 1983. “In Defense of External Invalidity.” *American psychologist* 38(4): 379.

Przeworski, Adam, and Henry Teune. 1970. “The Logic of Comparative Social Inquiry.”

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.

Williams, Joseph Jay, Tania Lombrozo, and Bob Rehder. 2010. “Why Does Explaining Help Learning? Insight from an Explanation Impairment Effect.” In *Proceedings of the 32nd Annual Conference of the Cognitive Science Society*, 2906–11.

1. [*This website*](http://plato.stanford.edu/entries/scientific-explanation/) seems useful as a starting place to think about explanation. See also (cites). [↑](#footnote-ref-27)
2. Here following Gaines and Kuklinski (2011) on their discussion of the differences in causal effects on people who would choose to be exposed to a treatment versus the people forced to be exposed in standard experimental protocols. [↑](#footnote-ref-34)
3. Add a Note about Gaines and Kuklinski (2011) on experiments with self-selection. [↑](#footnote-ref-35)
4. Note all of the problems that attend this notion; perhaps note the white swan argument.) Each time the original result is found again, the more inclined we will be to assume that the initial finding travels beyond the initial study. [↑](#footnote-ref-36)