During the past 60 years, the emergence of large-N data sets and statistical techniques to analyze them has fostered a legend that is now deeply embedded in social scientific folklore. According to this legend, which we label *causal generalization*, the statistical analysis of quantitative data collected on a large number of cases provides a vehicle by which social scientists can make causal claims that apply widely, across a variety of contexts. Causal generalizationis not only desirable, it is doable.

Recent scholarship has already put a dent in the “causal” part of the causal generalization legend. Today’s scholars take more seriously than ever the admonition that “correlation does not equal causation.” The replacement of a regularity, or associational, conception of causality with a counterfactual one has driven the change. Any serious social science research design course now includes readings on the perils that accompany the statistical analysis of observational, or nonexperimental, data. In the words of Paul Rosenbaum, a leader in the potential outcomes movement, “Passive observation of a natural population followed by regression analysis is often unsuccessful as an approach to inference about treatment effects…” (1999, p. 259). Even when the statistical analysis of quantitative data might be successful, it remains vulnerable to criticisms that it is not, due largely to the possible existence of unobserved influences.

This chapter challenges the second aspect of the legend, that the statistical analysis of large-N data sets fosters generalization by generating findings that apply to all of the cases, or contexts, included in the sample. (FN—There is another notion of generalization, which has been challenged and is not our focus (see Aronow and Samli 2016). We argue, to the contrary, that the estimates generated by large-N statistical studies more than likely fail to apply to any of them.

Even more fundamentally, early advocates of such studies unwittingly changed the meaning of generalization as philosophers of science have historically defined, and continue to define, it. Whereas political scientists have ridden the generalization banner, and arguably, congratulated themselves for their crowning achievement, philosophy of science scholars have continued to address “the generalization *problem*” as though this crowning achievement never occurred. They do not see the success that social scientists see. In their eyes, large-N studies hold no advantage over other types of study, including qualitative, experimental, quasi-experimental, and convenience-sample-based observational when it comes to generalization.

To conclude that large-N statistical studies fail in the quest for generalization does not answer the question: Is generalization a feasible goal? A desirable goal? None other than Donald Campbell, with whom generalization, via the term “external validity,” is most strongly associated, concluded late in his career that it likely is not. Factors that interact with even randomly assigned experimental treatments, he argues, vary across contexts, almost ensuring that the estimated treatment effects will vary as well. In her recent work, Nancy Cartwright adopts and extends a similar theme.

Although we agree with Campbell’s conclusion, we take what he views as a major obstacle to comparing treatment effects across contexts as an opportunity for scholars to make more compelling and informative comparisons both within and across contexts, to dig more deeply,than they have done to date. The question of whether findings from one context travel to another context remains. The additional questions are: What are the contextual factors that potentially interact with the treatment? Do differences in these factors explain why treatment effects differ across these contexts?

To succeed in answering the additional questions requires a combination of

informed theoretical deduction, much in the spirit of philosophy of science

scholar Nancy Cartwright, and in-depth, shoe leather investigations of

contexts, much in the spirit of statistician David Freedman. Cartwright

urges applying deduction to infer the theoretically relevant aspects of

context that might differ across contexts, and thus cause the treatment

effects to vary; Freedman provides a means, Sherlock-Holmes-like detective

work, as the means by which to construct fine-grained measures of these

factors. In essence, answering the additional questions will require bringing

more evidence to bear than researchers have typically done.

Our discussion proceeds as follows. First, we review and critique the logic underlying the use of large-N studies in the pursuit of generalization. Even when accepting the conception of generalization associated with the use of large-N data sets, these studies fail.

Second, we distinguish how the notion of generalization underlying large-N quantitative studies differs from that found in the writings of philosophers of science. Since several terms related to the notion of generalization have been used interchangeably, we try to clear some underbrush along the way by distinguishing among generalization, external validity, replication, and portability. We settle on portability as the term that most aptly describes the philosophy of science conception that we adopt throughout this paper.

The statistical analysis of many cases, we argue, has masked the task that social scientists face when trying to assess portability.

Third, building on the work of Cartwright and Freedman, we propose a strategy by which to begin to identify contextual factors whose presence or absence are essential to the portability of a treatment effect. We also identify the challenges associated with this task, not the least being the complex nature of causality.

Finally, we offer a concrete illustration of how the strategy might be applied, and identify some difficulties that might arise, using the Gerber et al. study of shaming and voting turnout. (Right example????)

**Whatever Generalization Is, Statistical Analysis of**

**Many Cases It Isn’t**

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 scholars can easily blur. First, generalization refers to the process whereby a researcher uses an observed random sample to learn about an unobserved population. The near-magical discovery that a mean calculated in a sample serves as an unbiased estimator of an unobserved population quantity motivated discipline-wide efforts to make random sampling a standard over quota sampling or convenient case choices. The considerable impact of research based on the American National Election Studies and the World Values Surveys, as well as other large-scale data collections, arises in part from this methodological insight.

Even granting that the statistical foundations underlying claims of unbiased, consistent, and precise estimators are correct, recently reported evidence suggests that inferring about counterfactual causal effects 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. This recent discovery, if true, challenges the very idea that researchers can safely infer from the sample to population from which the sample is drawn (Aronow and Samii 2016). Although this finding will not, and should not, bring the use of random samples to a standstill, it potentially poses serious challenges going forward. (FN-POTENTIALLY FIXABLE?) MORE HERE??

Our focus is the second conception of generalization, the idea that findings based on many cases apply more widely, i.e., better generalize because they take into account more cases, than findings based on few. In Przeworski and Teune’s words (still need a quote; so far I have not found the precisely right one).

QUOTE

In a typical cross-national study, a researcher enters supposed causal factors in a regression to estimate causal effects across countries, or other contexts. To account for differences across contexts that are not measured, the typical solution is to add dummy variables to represent units, for the purpose of capturing each country’s uniqueness. (FN—note emergence of multilevel models to account for error term.) Statistically, this works. From a research design perspective, however, three fundamental problems remain. First, the average treatment effect that a regression coefficient estimates most likely applies to none of the cases included in the analysis. ((More here???) It is a measure of central tendency, and no more.

Second and closely related, regression analyses, as traditionally conducted, fail to account for (identify???) differences in treatment effects across contexts. This is precisely the question that the pursuit of general statements should answer (see following section). Interaction terms between the treatment and dummy variables represents a step in the right direction, although the dummy variables are extremely coarse and therefore not meaningful measures of context.

Third, all of the factors that might interact with the treatment, the contextual factors to which Campbell alluded and what Cartwright and Hardie label “other members of the team,” almost certainly will not be measured, or measured accurately. Someone working with already-collected quantitative data will be constrained to work with what others have generated, which, again, will be coarse measures of context at best. The proper information is best gained, at least bolstered, via up-close observation. Large-N studies mask this need.

This problem immediately grows in significance once one acknowledges that the “other members of the team” are likely to work sequentially, not as independent factors operating at the same time. This observation is the genius of Mackie’s “Causes and Conditions, where he sets forth his INUS conditions. Essentially, he demonstrates, outcomes can occur through different pathways, in which all of the steps in each pathway are necessary, although none is singularly sufficient, and each pathway is sufficient although not necessary. (FN—PS have not picked up on this notion, intuitive though it is. However, to acknowledge it is to open the door to a complex story of cause and effect.----Should also note that this neglect is not a problem only of large-N studies.)

**Terminology**

We have rejected the notion of generalization that underlies large-N studies and argued that

Although terms like generalization, external validity, and replication are now commonplace, scholars have been remarkably inconsistent in their definitions of them. Even then, some terms and the concepts they convey seem more appropriate than others. Here, we first try to explicate the complexities surrounding the use of such terms and then share our rationale for using “portability” throughout the remainder of the paper.

In everyday language, we often criticize someone who generalizes as reckless or irresponsible. So we hear criticisms such as “he is making unfair and unfounded generalizations to support his argument.” Or, we hear that “she spoke in such broad generalities as to render her reasoning meaningless.”

Within the sciences, general statements emanate from theory or a principle.

Often, we accuse people who make reckless and unfounded comments as generalizers.

the two conceptions with the terms “portability” and “generalization,” with the first term referring to the dominant philosophy of science conception. S