10 Things to Know About External Validity

**Abstract**

After months or years under development and implementation, navigating the practical, theoretical and inferential pitfalls of experimental social science research, your experiment has finally been completed. Comparing the treatment and control groups, you find a substantively and statistically significant result on an outcome of theoretical interest. Before you can pop the champagne in celebration of an intervention well done, a friendly colleague asks: “But what does this tell us about the world?”

**1. What is external validity?**

External validity is another name for the generalizability of results, asking “whether a causal relationship holds over variation in persons, settings, treatments and outcomes” (Shadish, Cook and Campbell 2002). A classic example of an external validity concern is whether traditional economics or psychology lab experiments carried out on college students produce results that are generalizable to the general public. In the political economy of development, we might consider how a community-driven development program in India might apply (or not) in West Africa, or Central America.

External validity becomes particularly important when making policy recommendations that come from research. Extrapolating causal effects from one or more studies to a given policy context requires careful consideration of both theory and empirical evidence. This methods guide discusses some key concepts, pitfalls to avoid, and useful references to consider when going from a Local Average Treatment Effect to the larger world.

**2. How is this different than internal validity?**

Internal validity refers to the quality of causal inferences being made for a given subject pool. As originally posited by Campbell (1957), internal validity asks, “did in fact the experimental stimulus make some significant difference in this specific instance.” This concept dovetails with the counterfactual approach to causality that experimentalists typically use, which states that if the experimental intervention had not been present, neither would the outcome (more details here (link to Macartan’s guide) in “10 things you need to know about causal inference”).

Before you can extrapolate a causal effect to a distinct population, it is vital that the original Average Treatment Effect be based on a well-identified result. For most experimentalists, random assignment provides the requisite identifying variation, provided no attrition, interference, spillovers, or other threats to inference. For observational studies, additional identifying assumptions are needed, such as conditional independence of the treatment from potential outcomes.

**3. Navigating the trade-offs between internal and external validity**

There has been an ongoing debate within the social sciences regarding the relative importance of identifying internally valid results, which by definition apply to a local sample, and generating results that can be extrapolated to broader populations of interest. This discussion has typically been focused on observational empirical work rather than experiments, because randomized interventions are by nature relatively likely to be well identified (assuming no inference, etc.). Nonetheless, it is helpful to be familiar with this discussion when considering design trade-offs that inevitably crop up in resource-limited interventions. That both sides of the argument include luminaries of econometrics illustrate the importance of the topic.

On one side of the argument fall advocates of ‘identification first’, who argue that with internally valid results, a study simply does not contribute useful information, regardless of whether it is a local or general population or context. As put by Imbens (2013), “without strong internal validity studies have little to contribute to policy debates, whereas [internally valid] studies with very limited external validity often are, and in my view should be, taken seriously in such discussions.”

Others argue that even without full identification of an internally valid result, useful information can be salvaged, especially if it is relevant for important questions that affect a broad context. Manski (2013) writes that “what matters is the informativeness of a study for policy making, which depends jointly on internal and external validity.” With data from a broad but a poorly identified study, Manski argues, bounds on the estimand of interest can be generated that, while not as useful as a precise point estimate, still moves science forward.

**4. Theory and generalization**

Extrapolating a result to a distinct context, outcome, population or treatment is not a mechanical process. As discussed by Samii (2016), relevant theory should be used to guide generalization, taking the relevant existing evidence and making predictions for other contexts in a principled fashion. Theories boil down complex problems into more parsimonious representations, and help to elucidate what factors matter. Just as theory guides the content of interventions and research designs, theoretical propositions can tell you which scope conditions are relevant for extrapolating a result. What covariates matter? What contextual information matters?

**5. How can I determine where my results apply?**

There are two primary means of generalizing results, one based on the covariates of units in the study and the other based on actual experimental manipulation of moderating variables. Exogenous variation in a covariate, such as through direct randomized manipulation, allows for a causally identified heterogeneous effect. Observing how a treatment effect varies over a non-randomized pre-treatment variable can describe treatment effect heterogeneity but not pin down whether the heterogeneity is due to that variable or an unobserved factor (Gerber and Green 2012, chapter 9).

Because generalization is primarily a prediction exercise, asking where we can expect a causal relationship similar to one observed locally, extrapolating heterogeneous effects based on similar covariates is often reasonable, provided theory does not indicate sources of confounding (Bisbee, Dehejia, Pop-Eleches and Samii 2015). Nonetheless, the strongest evidence for the generalizability of a result comes from a well-identified interaction between an exogenous moderator and the treatment, then projected across the covariate profile of a target population. Indeed, with some strong assumptions extrapolation can provide as good or better results than carrying out a second experiment in situ (Bisbee et al. 2015). This can often be best carried out using machine learning, although linear regression also performs reasonably well (Kern, Stuart, Hill and Green 2016).

**6. Strategic behavior can scuttle your extrapolations**

Extrapolating a local result to a distinct context can prove challenging even with a compelling covariate profile to which you want to generalize effects. A randomized experimental manipulation in a local area is “partial equilibrium effect,” meaning that within general equilibrium we might expect different results even where the covariate profile matches. In short, strategic dynamics, including compensatory behavior or backlashes, outside the local context of an experimental intervention can complicate efforts to generalize a result.

Sometimes causal relationships only work when they are applied to some people; for example, imagine a job skills program that functions very well (as compared to those who did not receive it), what would happen if it were extended to all workers? Even if there are positive effects across all participants, there could be reduced or no average effects as higher skilled jobs are already filled by the first batch and the second batch is forced to remain in their previous jobs, now overqualified.

**7. Don’t confuse external validity with construct validity or ecological validity**

Internal and external validity are not the only ‘validity’ concerns that can be leveled at experimental work, and though relevant, they are also distinct. Ecological validity, as defined by Shadish, Cook and Campbell (2002) concerns whether an intervention appears artificial or out of place when deployed in a new context. For example, does an information workshop in a rural town carried out by experimenters resemble the kinds of information sharing that the population may experience in regular life? Similarly, if the same workshop were held in a large city, would it appear out of place?

Construct validity considers whether a theoretical concept being tested in a study is appropriately operationalized by the treatment(s). If your experiment is testing the effect of anger on political reciprocity and you are in fact manipulating fear or trust in your treatment, construct validity may be violated. Both construct and ecological validity are relevant for generalizations, and thus useful for making claims about external validity.

**8. Extrapolation across treatments and outcomes**

While much of this guide has implicitly focused on porting a given treatment to a new place or time, external validity also considers variations in treatments and outcomes. That is, imagine we did the same experiment on the same sample, but with a variation on the treatment, would we predict the local causal effect to be similar? Similarly, can we predict if a given treatment will produce the same or different causal effects on a different outcome? Rather than considering the features of subjects, extrapolation in this case requires thinking through, aided by theory, the characteristics of the treatments or outcomes and make reasonable predictions.

**9. Replication is important**

No single study represents the final word on a scholarly question. Following the logic of Bayesian updating, additional evidence in favor or against a given theory allows the scientific and policy community to update their beliefs about the validity of causal relationship.

Replication of studies is an important part of this: scholars should replicate studies in contexts that look very different, but also in some contexts that look very similar. The former allows us to identify local causal relationships that can be triangulated with existing evidence and generalized as appropriate. At the same time, it is important to directly replicate existing studies under conditions that are as close as possible to the original in order to verify that local effects one may be interested in extrapolating are indeed reliable. The Open Science Collaboration (2015) found, for example, that when reproducing 100 major psychology experiments, only 39 percent of the effects replicated the original statistically significant results.

**10. Thinking beyond average effects**

The frequentist statistical paradigm, which is employed by many experimentalists, emphasizes means and average effects, because we have the most well-developed, consistent and unbiased estimators for this central tendency. Even for small sub-groups, we are still estimating mean treatment effects. Deaton (2009) notes that this means that “a trial might reveal an average positive effect although nearly all of the

population is hurt with a few receiving very large benefits.”

We can move beyond group averages to make individual level predictions based on covariates, but again these are in expectation; we know that a person will not be exposed to a program 10,000 times, they will participate (or not) just once. When thinking about causal relationships of interest, it is important also to consider time: do things we learn about the past extend to the future? How does an individual’s potential outcomes change over time? When making decisions about the policy relevance, generalizability of results, these factors can help scholars determine a reasonable level of uncertainty and help policy makers adjust accordingly.

*Originating Author*

Renard Sexton

October 2016

**References**

Bisbee, James; Rajeev Dehejia; Cristian Pop-Eleches & Cyrus Samii. (2016). “Local Instruments, Global Extrapolation: External Validity of the Labor Supply-Fertility Local Average Treatment Effect.” *Journal of Labor Economics*

Campbell, D. T. (1957). Factors relevant to the validity of experiments in social settings. *Psychological bulletin*, *54*(4), 297.

Deaton, A. S. (2009). *Instruments of development: Randomization in the tropics, and the search for the elusive keys to economic development* (No. w14690). National Bureau of Economic Research.

Gerber, A. S., & Green, D. P. (2012). *Field experiments: Design, analysis, and interpretation*. WW Norton.

Imbens, G. (2013). Book Review Feature: Public Policy in an Uncertain World: By Charles F. Manski. *The Economic Journal*,*123*(570), F401-F411.

Imbens, G. W. (2010). Better LATE than nothing. *Journal of Economic Literature*, *48*.

Kern, H. L., Stuart, E. A., Hill, J., & Green, D. P. (2016). Assessing methods for generalizing experimental impact estimates to target populations. *Journal of Research on Educational Effectiveness*, *9*(1), 103-127.

Manski, C. F. (2013). Response to the review of ‘public policy in an uncertain world’.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716.

Samii, Cyrus. (2016). “Causal Empiricism in Quantitative Research.” *Journal of Politics* 78(3):941–955.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton, Mifflin and Company.