External Validity: Generalizing Evidence in Policy Formulation

Jeffrey Smith

University of Wisconsin Madison

January 15, 2023

This note summarizes my plans for a paper on external validity intended for a special issue of *Evaluation Review*. External validity in the evaluation context captures the extent to which the impact estimates obtained in one study, or several studies, generalize to populations other than those studied. External validity often competes, at least implicitly, with internal validity, which captures the plausibility of causal claims regarding the impact estimate as applied to the study population. Imbens (2013) provides a fine discussion of this study design tradeoff in the context of the evaluation literature in economics. External validity matters for the systematic accumulation of evaluative knowledge and for the application ("transportation" in the terminology of some) of impact estimates from one study to external populations.

This paper goes beyond existing surveys of the literature, such as the one aimed at educational researchers in Tipton and Olsen (2018), in four major ways. First, it will outline the so-called "structural" approach to external validity in economics, in language that non-economists can understand. It will contrast its virtues and vices to those of the design-based approaches typically featured in the program evaluation literature, especially the part of it that lies outside of economics. The structural approach focuses on the estimation of policy-invariant parameters that characterize responses to entire classes of interventions. A leading example includes the use of the wage and income elasticities of labor supply to estimate the effects of tax and transfer policies ranging from the Negative Income Tax in the 1970s to the expanded Child

Tax Credit in the 1990s. Wolpin (2013) makes the case for the structural approach at length and at a much more technical level than I will adopt.

Second, my paper will integrate research on design-based approaches to external validity from multiple disciplines. At present, like many methodological literatures within the broad tent of program evaluation (e.g. the literature on matching and weighting estimators), the literature on external validity proceeds to a great degree independently in a variety of disciplinary contexts, with only occasional and imperfect attempts at communication between them. This paper will provide some potential examples of the intellectual and substantive payoff to venturing across disciplinary boundaries. For example, readers operating in the educational evaluation world might benefit from the wise discussions of moderator confounding and moderator common support in Hotz, Imbens and Mortimer (2005) and from the novel applications of applied Bayesian methods in Meager (2019), both drawn from the economics literature. They might also benefit from the insights offered by Findley, Kikuta, and Denly (2021) in the political science literature. More prosaically, my paper will add value simply by citing papers from many disciplines in one place, as an aid to those who wish to explore and a wake-up call to those who do not know what they are missing.

Third, much of the extant literature focuses primarily on external validity in dimensions related to participant characteristics (e.g. demographics, education, etc.) and to program context (e.g. local unemployment rate). I plan to review (and remark on) this literature, but also to include explicit separate discussions of external validity in time (including but not limited to over the business cycle) and in program space. The latter concerns issues like implementation quality differences, variation in the mix of services provided within some well-defined set (as in many active labor market programs), variation in the assignment rule that matches specific participants

to specific services, variation in organizational form, and so on. This discussion will draw in part on the recent literature on optimal treatment coding in programs offering heterogeneous treatments, such as VDW. It also relates to the literature on site effects, e.g. Bell et al. (2016), though I have in mind a broader discussion that nests variation in impacts among sites.

Fourth, I will devote more attention to the question of moderator selection. As Tipton and Olsen (2018) point out, when thinking about design-based reweighting strategies for generalizing existing evidence to new populations, one wants to weight on moderators, which is to say on observed variables that capture variation in treatment effects. Buhl-Wiggers et al. (2023) note that the existing literature in education has little to say about how to find moderators, instead opting for a set of usual suspects that typically explain very little of the available treatment effect variation. The same pattern holds in every other evaluation literature I know well. I will follow the lead of Buhl-Wiggers et al. (2023) and highlight the role of advances in applied theory, i.e. on models of effect moderation, and in measurement in improving our knowledge of how, when, and why treatment effects vary and thereby our ability to generalize in credible ways.

The paper will also reference (i.e. mention and provide pointers to the relevant literature), but not explore in depth, two additional types of external validity. One considers generalization from local average treatment effects, say from an experimental evaluation with imperfect compliance, to more general parameters such as the average treatment effect on the treated. See e.g. Black et al. (2022) for discussion. The other, a current preoccupation of the development economics literature (in addition to having made some headway into the "implementation science" literature) concerns scale-up, where one can frame concerns about scale-up as concerns about the external validity to a full-scale program of impact estimates obtained from a small demonstration program. List (2022) provides a book-length treatment of this topic.

The paper will conclude with a summary, accompanied by a synthesis that aims to distill the key insights of the paper for evaluation practitioners, and suggestions for future research for more academically-inclined readers.

References

Bell, Steven, Rob Olsen, Larry Orr, and Elizabeth Stuart. 2016. "Estimates of External Validity Bias when Impact Evaluations Select Sites Non-Randomly." *Educational Evaluation and Policy Analysis* 38(2): 318-335.

Black, Dan, Joonhwi Joo, Robert LaLonde, Jeffrey Smith, and Evan Taylor. 2022. "Simple Tests for Selection: Learning More from Instrumental Variables." *Labour Economics* 79: 102237.

Findley, Michael, Kyosuke Kikuta, and and Michael Denly. 2021. "External Validity." *Annual Review of Political Science* 24: 365-393.

Buhl-Wiggers, Julie, Jason Kerwin, Juan Muñoz-Morales, Jeffrey Smith, and Rebecca Thornton. 2023. "Some Children Left Behind: Variation in the Effects of an Educational Intervention." *Journal of Econometrics*, forthcoming.

Hotz, V. Joseph, Guido Imbens, and Julie Mortimer. 2005. "Predicting the Efficacy of Future Training Programs Using Past Experiences at Other Locations." *Journal of Econometrics* 125(1-2): 241-270.

Imbens, Guido. 2013. "Book Review Feature: *Public Policy in an Uncertain World* by Charles F. Manski" *Economic Journal* 123: F401-F411.

List, John. 2022. *Voltage*. Currency.

Meager, Rachel. 2019. "Understanding the Average Impact of Microcredit Expansions: A Bayesian Hierarchical Analysis of Seven Randomized Experiments." *American Economic Journal: Applied Economics* 11(1): 57-91.

Tipton, Elizabeth and Rob Olsen. 2018. "A Review of Statistical Methods for Generalizing From Evaluations of Educational Interventions." *Educational Researcher* 47(8): 516-524.

Wolpin, Kenneth. 2013. *The Limits of Inference without Theory (Tjalling C. Koopmans Memorial Lectures)*. MIT Press.