Bell, S. H., & Stuart, E. A. (2016). On the “where” of social experiments: The nature and extent of the generalizability problem. In L. R. Peck (Ed.), *Social experiments in practice: The what, why, when, where, and how of experimental design & analysis. New Directions for Evaluation*, *152*, 47–59.

3

# On the “Where” of Social Experiments: The Nature and Extent of the Generalizability Problem

*Stephen H. Bell, Elizabeth A. Stuart*

## Abstract

*Although randomized experiments are lauded for their high internal validity, they have been criticized for the limited external validity of their results. This chapter describes research strategies for investigating how much nonrepresentative site selection may limit external validity and bias impact findings. The magnitude of external validity bias is potentially much larger than what is thought of as an acceptable level of internal validity bias. The chapter argues that external validity bias should always be investigated by the best available means and addressed directly when presenting evaluation results. These observations flag the importance of making external validity a priority in evaluation planning.* © 2016 Wiley Periodicals, Inc., and the American Evaluation Association.

Most social experiments seeking to measure the impact of government programs study nonrandomly selected individuals or geographic locations. This tendency to sample nonrepresentatively from the population of policy interest may skew measured effects of the studied intervention away from its true impact in the population. The potential for impact skewing has long been recognized in the program evaluation literature by those concerned about the generalizability, or “external validity,” of research findings (see e.g., Cronbach, Gleser, Nanda, &

Rajaratnam, 1972; Green & Glasgow, 2006; Humphreys, Weingardt, & Harris, 2007; Julnes, 2011; Olsen, Bell, Orr, & Stuart, 2013; O'Muircheartaigh & Hedges, 2014; Tipton et al., 2014). In the statistics literature, there has also been increasing interest in and attention to methods for estimating population treatment effects (Bareinboim & Pearl, 2013; Hartman, Grieve, Ramsahai, & Sekhon, 2015; Kern, Stuart, Hill, & Green, 2016; Olsen & Orr, Chapter 4; Stuart, Cole, Bradshaw, & Leaf, 2011).

However, there has been surprisingly little formal investigation of how large external validity bias may be in typical evaluations, and little has been done to address this threat in the practice of social program impact evaluation.

This omission may be rectified by identifying ways to measure the bias produced in rigorous impact evaluations by nonrandom sample selection, and in particular by the inclusion of a nonrepresentative set of geographic sites in which the research is conducted. This chapter provides evaluators with four tools for investigating how much nonrepresentative site selection may bias impact findings. It addresses the external validity bias that may exist when trying to generalize impact estimates from a rigorous evaluation to a target population of interest.[1] Recently Olsen et al. (2013) have shown that external validity bias arises through the combination of three circumstances: (a) true impacts that vary from one site to another, (b) differing probabilities of inclusion in the study for these distinct sites, and (c) impact magnitudes that correlate with the probability of site inclusion. Formally, these conditions require three statistical parameters to be nonzero: the variance in site inclusion probabilities, the variance of impacts across sites, and the correlation between these two factors.

One might ask whether the typical practice of including a nonrandom set of sites in impact evaluations poses much threat to reliable findings, because tests for variation in impact magnitude across sites in the literature on large-scale social experiments have rarely produced statistically significant results. There are, however, some instances in which treatment effect heterogeneity by site has been demonstrated conclusively (e.g., see Greenberg, Meyer, Michalopoulos, & Wiseman, 2003; Hamilton, Brock, & Farkas, 1995; Nisar, 2010). Also, even where statistical tests cannot confirm that observed variation in measured impact across sites must be caused by something more than sampling variability, impact heterogeneity may exist but not be detectable from available sample sizes, which tend to be small for individual sites.[2] In light of these circumstances, we believe evaluators

[1] In reality, there may be multiple target populations of interest. This discussion assumes that one population of interest has been determined, but the analyses described could be repeated for multiple target populations.

[2] Limited sample sizes especially delimit the potential for statistically significant findings when conducting "difference in differences" tests using statistics with four additive

need to better understand and examine how much external validity bias may exist when trying to make policy decisions on the basis of rigorous impact evaluations. In fact, some initial research on the size of external validity bias indicates that the bias may be as large as the amount of internal validity bias that researchers tend to worry about (e.g., bias equal to 0.10 of a standard deviation of the evaluation's outcome of interest; see Bell, Olsen, Orr, & Stuart, 2016).

In this chapter, we outline four strategies for judging the presence of and magnitude of external validity bias in particular social experiments: (a) comparing sites' baseline characteristics and/or outcomes to those in the target population, (b) comparing impact findings at risk of external validity bias to impacts measured in the corresponding population, (c) directly estimating the Olsen et al. (2013) bias parameters noted previously, and (d) simulating bias under different posited site inclusion mechanisms. To illustrate the first two approaches, we present examples from existing research on experimental evaluations in the elementary education field—from a schoolwide behavior improvement intervention (Stuart, Bradshaw, & Leaf, 2015) and the national Reading First program (Bell et al., 2016). Efforts currently underway (by the authors and their collaborators) using the latter two strategies are also discussed. Future application of all four of these strategies should expand the empirical knowledge base on the extent of external validity bias in experimental evaluations, generally.

We organize our presentation into five parts. The first four sections describe and illustrate each of four strategies for moving the literature from theoretical constructs to empirical assessment of the existence and degree of the external validity bias problem in actual rigorous impact evaluations of social programs. These strategies, described in the previous paragraph, include background and outcome comparisons, impact finding comparisons, bias parameter estimation, and bias simulations. We then conclude with a discussion of next steps for research and implications for evaluation practice.

## Background and Outcome Comparisons

An obvious question arises when evaluating program impacts using data from a nonrandom sample of locations: Do the selected sites "look like" the population the evaluation seeks to investigate? Existing scholarship considers two approaches for understanding sample-versus-population results from rigorous impact evaluations: (a) comparisons of background characteristics and (b) comparisons of outcomes. We describe each of these

components in their variance formulas, such as the estimated treatment-minus-control-group difference in mean outcomes for Site A minus the treatment-minus-control-group difference in mean outcomes for Site B.

techniques here along with their common limitation, unobserved factors that compromise the comparisons.

## Comparisons of Background Characteristics

This approach measures the extent to which a given set of sites mirrors the population of interest in its background characteristics—for example, regarding various measured traits of program participants and local communities. This can be done for individual characteristics one at a time or for a summary measure, such as the probability of participation in the evaluation, expressed as a weighted average of a vector of such characteristics (Stuart et al., 2011; Tipton, 2013). The closer the study sample aligns with the population on such an indicator, the greater one's confidence that the sampled sites well represent the population. Tipton (2014) and Stuart et al. (2011) give guidelines for how close is "close enough" in this regard for reliable generalization from sample to population. A benefit of this approach is that it can be implemented using only covariate data from the sampled sites and the target population; no outcome data are required.

## Comparisons of Outcomes

In addition to comparing background covariates, sometimes a comparison of outcomes can yield insights regarding the potential generalizability of study results. There are two ways this comparison can be implemented; both rely on the idea that, if two groups are similar, their average outcomes under the same treatment condition should be similar. When new interventions are evaluated, one can compare a key outcome measure or measures for "untreated" control group members from the study sites with the same outcome measure or measures—also untreated by the test intervention—taken from data on the full population of sites. The intuition here is that, if the sites in the evaluation do represent the population, their average outcomes absent the treatment should be similar to average outcomes absent the treatment in the population. This is labeled a "placebo test" by Hartman et al. (2015).

Other impact evaluations examine social programs that already treat the entire population of interest but do so experimentally by randomly excluding cases from the intervention to create control groups for measuring impact in some nonrandom subset of locations. Here, one can compare a key outcome or outcomes for cases receiving the intervention in the study sites (from the experiment's "treatment group") with the same outcome measure or measures—also treated by the intervention—taken from data on the full population of sites. The intuition here is that, if the sites in the evaluation do represent the population, their average outcomes with the treatment should be similar to average outcomes with the treatment for the population.

### Unobserved Factors

Background comparisons and outcome comparisons share a similar limitation when used as just described to gauge external validity bias. They cannot tell researchers whether study sites differ from the population on factors (a) absent from the background characteristics and/or outcome measures examined that (b) affect the magnitude of intervention impacts. What might these factors be? Many of the unmeasured determinants of site inclusion may be idiosyncratic to particular evaluations and only evaluable through researchers' knowledge of how a given study came to include a particular set of sites. The usefulness of the comparison approaches just described depends in part on whether researchers understand from their own involvement in study set-up the ways in which included sites differ from the population.

Other determinants of site inclusion may not be idiosyncratic but exert similar influences across many evaluations. Sometimes these factors are measured and sometimes they are not. If such site inclusion factors can be identified from theory, then the research team for a given study could judge whether the background characteristics and outcome measures it is able to compare, population versus sample, encompass those (or at least the great preponderance of those) influences. Unfortunately, relatively little is known at present about the factors that influence site participation in rigorous impact evaluations and simultaneously moderate treatment effect magnitude. Stuart et al. (2016) suggest that future work "further investigate whether program effects vary [by site] (and across which factors, both individual and contextual) and account for those factors when assessing external validity" (p. 483). Some work already exists or is currently underway among evaluation researchers to quantify cross-site variation in impact magnitude and the factors that associate with impacts using data from rigorous, randomized multisite impact evaluations. Examples include Bloom, Hill, and Riccio (2005) on welfare-to-work programs and Institute of Human Development and Social Change (2015) concerning the effects of the Head Start program.

If a consistent set of correlates of impact magnitude emerge from these and related studies, future evaluations can benefit from collecting the same measures (in their particular program contexts) for all sites in the population of interest to use for diagnosing the potential for external validity bias among the subset of sites actually included in the evaluation. The "unobserved factors" problem in background characteristic checks of population and sample similarities would be ameliorated as a result.

## Impact Finding Comparisons

The remaining three strategies for assessing the external validity of rigorous impact evaluations only indirectly address the reliability of the investigator's own particular evaluation. They focus instead on examining whether other experiments—or hypothetical simulated experiments—produce or would produce externally valid impact estimates for a population of interest and, if not, by how far do they fail.

The first of these strategies extends the notion of doing *outcome* comparisons between the population and study sites (see previous discussion) to the idea of doing *impact* comparisons between the population and the study sites. It does so by paralleling the "design replication" (also known as the "within-study comparison") strand of the evaluation literature, in which researchers gauge the *internal* validity of quasiexperimental impact evaluations by comparing the impact estimates they produce to internally valid experimental estimates for the same sites (e.g., Bifulco, 2012; Cook, Shadish, & Wong, 2008; Fraker & Maynard, 1987; LaLonde, 1986). In the external validity context, the parallel design replication strategy gauges bias by comparing impact estimates between sample and population, in the special case where both estimates exist with adequate internal validity. In such circumstances, one does not *need* the sample-based finding to determine the effect of the studied intervention: the population finding of impact does so in a superior and acceptable fashion. One can use what is learned from such a comparison to determine if external validity bias is likely to be a threat in other, similarly constructed experiments for which population impact estimates are not obtainable—just as is true of the now widespread practice of deciding whether a given nonexperimental impact evaluation is adequately protected from internal validity bias by looking at the performance of its analysis approach in other studies where experimental impact measure with high internal validity can be used as a benchmark (e.g., Cook et al., 2008; Steiner, Cook, Shadish, & Clark, 2010).

In the one known application of this strategy of direct measurement of external validity bias against a rigorous benchmark, Bell et al. (2016) compare impact estimates for a population of policy interest to impact estimates from various evaluation samples to see how well or poorly the sites that real-world impact evaluation samples have actually included represent the population in terms of the ultimate goal of the research, getting the impact measure right. Bell et al. (2016) examine a setting in which rigorous estimates of the impact of an elementary school reading intervention can be constructed for all public schools in nine states. Importantly, the analysis uses lists of the school districts from those states that are actually included in 11 rigorous (mostly randomized) impact evaluations recently conducted by the U.S. Department of Education (DOE) as their hypothetical study samples. Combining "population" data from the nine states with subsets defined by the 11 sets of districts included in actual evaluations yields 11 sample-versus-population checks of external validity bias based on estimated intervention impacts. These checks are reflective of the ways 11 different rigorous DOE evaluations obtained participating sites—whether by convenience or local willingness to conduct random assignment or some other means.

The findings in Bell et al. (2016) indicate that the kind of school districts typically included in large-scale education evaluations have *smaller* impacts than the average impact for the entire population of interest. The average absolute "error" in the sample-based estimates equals 0.10 of the standard deviation of the outcome of interest (student reading test score) in the population—a degree of external validity bias the authors argue is large relative to several reasonable metrics for judging research reliability taken from the literature on nonexperimental study designs. The evidence to date from this innovative approach to estimating the size of external validity bias calls into question the policy relevance of impact evaluations not based on a statistically representative set of sites, at least in the field of education.

One challenge in using this approach to estimating external validity bias—which we call "population replication" analysis to parallel its "design replication" antecedent from the internal validity literature—is finding appropriate data. Rigorous impact evaluations protected from internal validity bias—that is, those based on random assignment of individual cases within sites—rarely have data for the entire population of interest, or even for a statistically representative probability sample of the population (Olsen & Orr, Chapter 4, provides strategies for getting closer to this goal). The National Job Corps Evaluation (Burghardt et al., 1999) is one exception that included the entire population of interest in an experiment, whereas the National Head Start Impact Study (Puma, Bell, Cook, & Heid, 2010) is another exception that enrolled a random sample of Head Start sites in a randomized experiment. These studies should be seen as top candidates for further direct measurement of external validity bias in nonrepresentative sites.

However, even in these studies, another challenge to estimating the size of external validity bias exists: programs evaluated by populationwide experiments such as Job Corps provide population estimates of effectiveness but do not tell us what sites *would have* participated had a less comprehensive method of site recruitment been used—an approach that yields a typical set of nonrepresentative sites. As a result, there is no realistic sample of nonrepresentative sites from which to derive (potentially biased) impact estimates under more conventional site inclusion circumstances, for comparison to the available population (or population-representative) impact.

At root, the problem is that any particular impact study will have *either* population-representative impact data *or* a nonrepresentative sample of sites, but not both. Here, it is instructive to consider how Bell et al. (2016) overcame this problem by combining three ingredients. The first was extensive data on a large set of localities that constituted a plausible population of policy interest, data that included a rich set of background variables, longitudinal outcome information, and indicators of when different sites in the database implemented the intervention of interest (Reading First). The second ingredient was lists of sites that actually participated in other randomized evaluations of educational interventions, with which to simulate different nonrepresentative evaluations of the focal intervention (Reading

First). Bell et al. (2016) were able to test whether the sites assembled for these other evaluations would have performed well in measuring the effects of Reading First. Third, the recipe required a highly refined, but admittedly not infallible, method for estimating impacts without random assignment (comparative interrupted time-series) because random assignment had not been applied to the population data set. These three ingredients can perhaps be combined in other settings to obtain further empirical estimates of the size of external validity bias in rigorous impact evaluations.

## Estimation of Bias Parameters

As noted earlier, Olsen et al. (2013) have shown that external validity bias arises in multisite impact evaluations if three conditions hold simultaneously: (a) true impact magnitude varies by site, (b) the probability of study inclusion varies by site, and (c) the impact magnitude correlates with the probability of site inclusion. This formalization creates the opportunity to assess external validity bias by empirically estimating these three parameters for a particular program and evaluation context, then combining them using the formula in Olsen et al. (2013) to calculate the magnitude of external validity bias. If a technique for parameter estimation of this sort can be found, such estimation can be repeated across many evaluation and program contexts where impact evaluation involves a nonrepresentative set of sites.

With respect to the first of the three parameters, the variance of impact magnitude across sites is currently being derived by evaluators in a range of social programs contexts, as referenced earlier. Probabilities of site inclusion for actual experimental evaluations that do not select sites on a probability or populationwide basis can also be estimated regularly, at least to the extent that nonrepresentative sample are drawn based on *measured* background characteristics (see Stuart, Bell, Ebnesajjad, Olsen, & Orr, 2016, as one example). However, the third parameter—the population correlation between these two factors across sites—may not be as easy to obtain, at least not in a statistically precise way and for the right population. Unlike the estimation of the variability in site-specific impacts or in the site-specific probabilities of inclusion in experimental evaluations, the estimation of this correlation requires researchers to assemble data for individual sites showing both a numerical measure of impact magnitude and the numerical probability of inclusion in a nonrepresentative impact evaluation.

The second of these quantities is neither known nor knowable in a strict sense: all we know about an individual site is that it either was or was not included in a given evaluation. However, researchers can model the probability of participating in an evaluation by fitting models of participation as a function of observed background characteristics of program participants and community settings, which can at least give an estimate of the probability of participation. With this and site-specific impacts available on the same set of sites, the correlation between these measures can easily be calculated, as done by Allcott and Mullainathan (2012).

Unfortunately, the parameter needed in the Olsen et al. (2013) bias formula is the *population* correlation between impact and the probability of inclusion, not the *sample* correlation in a data set from a nonrepresentative experiment. From where can one obtain site-specific impact estimates for all sites in the population—and in particular, the sites *not* included in the current randomized impact evaluation? To obtain this information, it may be necessary to again return to a data set in which low-internal-validity-bias impacts can be calculated using rigorous nonexperimental methods, for individual sites across an entire population of sites. Having done so, one should consider whether calculating the Olsen et al. (2013) bias parameters adds value to the exercise—or if instead the more direct approach of measuring external validity bias in Bell et al. (2016) fully capitalizes on situations in which impacts for *all sites* in a population can be calculated. Future work should also consider how sampling error in the estimation of the three bias parameters magnifies statistical uncertainty in the resulting bias estimate once a multiplicative product of the three parameters is formed. We also encourage further thought as to whether other types of data sets or estimation strategies might be brought to bear on the challenge of quantifying the three key parameters of the Olsen et al. (2013) external validity bias formula.

## Bias Simulations

A final strategy for estimating external validity bias involves simulation of the impact findings that would arise in hypothetical evaluations from different simulated nonrepresentative site selection processes. Starting from a large set of sites in which randomization of individuals to treatment or control groups provides internally unbiased impact estimates, researchers could "play" with the site-specific impact estimates pretending that various hypothetical evaluations had been conducted in which some of the available sites are included and others left out. Kern et al. (2016) use this strategy to examine how well statistical methods estimate population treatment effects under a range of different sample selection mechanisms.

Such simulations might seem to have the potential for greatly advancing our knowledge of the consequences of nonrepresentative site selection, because so many various site inclusion scenarios could be run from so many various multisite randomized evaluation data sets. One must ask, though, how convincing exercises of this sort can be made to be in two respects: how well will they mimic site inclusion mechanisms that actually arise in real world experiments? and how meaningful will be the simulation results regarding generalizability to whole populations when those results themselves derive from a nonrepresentative set of sites? This exercise would become more convincing if the collection of "building block" sites for simulating different hypothetical estimates encompassed a meaningful population in its own right, such as in the elementary reading intervention example described previously or the Job Corps or Head Start Impact Study studies also referenced earlier. In these situations, simulated scenarios of site selection could reach across the full diversity of sites potentially included in future evaluations. The exercise would also become more convincing if site selection for the simulations were guided by actual patterns of site inclusion in real-world evaluations—a goal that can be pursued using approaches described earlier for modeling the types of sites that tend to participate in rigorous impact evaluations.

## Future Research and Implications for Practice

This chapter presents initial thinking on how to gauge the extent of external validity bias in rigorous evaluations of social programs. Until recently, almost no formal attention has been given to potential *external* validity bias in studies whose susceptibility to internal validity bias has been alleviated by using random assignment. Almost nothing is known about the size of external validity bias in the large portfolio of completed experimental impact studies nor about how that bias may affect the usefulness of findings from those evaluations as guides to policy decisions. In fact, almost no attention has been given to understanding the processes through which some sites participate in rigorous evaluations and others do not. We believe that better documentation of these processes, and the resulting samples that emerge, is a crucial first step in understanding how much we need to be concerned about external validity bias. This may include quantitative work such as that described here, as well as qualitative studies to better understand how researchers select sites to participate and how sites decide whether or not to do so.

The premise of all of the methods examined in this chapter is that policymakers have a population in mind when seeking to anticipate the effectiveness of social programs, a population to which they would like to apply results from a rigorous evaluation. We urge policymakers also to ask another question beyond "What impact is measured in an available evaluation?" Of equal importance is the question "Does the available evaluation provide findings close to correct for the population of policy interest?" A key component of nearly all the strategies discussed here for answering the latter question is the combining of information on both the population of interest and on the sites actually included in particular evaluations. Although some insight can be gained by simply comparing background characteristics between the sample and the target population, having outcome data—or, even better, estimated impacts—for the population can provide major additional progress.

We highlight the need for high-quality, expansive data on populations of policy interest to facilitate the use of the methods described here. We also encourage researchers to consider novel ways of combining population data and data from rigorous evaluations to estimate the size of external validity bias. We hope that a broader research base on this topic—a "population replication" literature—can be built up by these efforts, just as the design replication literature assessing internal validity bias in nonexperimental studies has been established over the past 30 years. Obtaining information on the magnitude of the problem constitutes the first crucial step toward ensuring that in the future, the results of rigorous impact evaluations are interpreted and used appropriately for policymaking purposes.

## Acknowledgments

The research reported here was supported in part by the Institute of Education Sciences, U.S. Department of Education, through grant R305D100041 to Abt Associates Inc., and the National Science Foundation, through grant DRL-1335843 to Johns Hopkins University.

### References

Allcott, H., & Mullainathan, S. (2012). *External validity and partner selection bias* (NBER Working Paper No. w18373). Cambridge, MA: National Bureau of Economic Research.

Bareinboim, E., & Pearl, J. (2013). A general algorithm for deciding transportability of experimental results. *Journal of Causal Inference*, *1*, 107–134.

Bell, S. H., Olsen, R. B., Orr, L. L., & Stuart, E. A. (2016). Estimates of external validity bias when impact evaluations select sites non-randomly. *Educational Evaluation and Policy Analysis*, *38*, 318–335. doi:10.3102/0162373715617549

Bifulco, R. (2012). Can nonexperimental estimates replicate estimates based on random assignment in evaluations of school choice? A within-study comparison. *Journal of Policy Analysis and Management*, *31*, 729–751.

Bloom, H. S., Hill, C. J., & Riccio, J. A. (2005). Modeling cross-site experimental differences to find out why program effectiveness varies. In H. S. Bloom (Ed.), *Learning more from social experiments: Evolving analytic approaches*. New York: Russell Sage Foundation.

Burghardt, J., McConnell, S., Meckstroth, A., Schochet, P., Johnson, T., & Homrighausen, J. (1999). *National Job Corps Study: Report on study implementation*. Princeton, NJ: Mathematica Policy Research, Inc. Retrieved from https://wdr.doleta.gov/opr/fulltext/99-jc_implement.pdf

Cook, T. D., Shadish, W. R., & Wong, V. C. (2008). Three conditions under which experiments and observational studies produce comparable causal estimates: New findings from within-study comparisons. *Journal of Policy Analysis and Management*, *27*, 724–750.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York: Wiley.

Fraker, T., & Maynard, R. (1987). The adequacy of comparison group designs for evaluations of employment-related programs. *Journal of Human Resources*, *22*, 194–227.

Green, L. W., & Glasgow, R. E. (2006). Evaluating the relevance, generalization, and applicability of research: Issues in external validation and translation methodology. *Evaluation & the Health Professions*, *29*, 126–153.

Greenberg, D., Meyer, R., Michalopoulos, C., & Wiseman, M. (2003). Explaining variation in the effects of welfare-to-work programs. *Evaluation Review*, *27*, 359–394.

Hamilton, G., Brock, T., & Farkas, J. (1995). *The JOBS evaluation: Early lessons from seven sites*. Washington, DC: U.S. Department of Health and Human Services, Administration for Children and Families, Office of the Assistant Secretary for Planning and Evaluation.

Hartman, E., Grieve, R., Ramsahai, R., & Sekhon, J. S. (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*, 757–778.

Humphreys, K., Weingardt, K. R., & Harris, A. H. S. (2007). Influence of subject eligibility criteria on compliance with national institutes of health guidelines for inclusion of women, minorities, and children in treatment research. *Alcoholism: Clinical and Experimental Research*, *31*, 988–995.

Institute of Human Development and Social Change. (2015). *Secondary analysis of variation in impacts of Head Start center*. New York: New York University. Retrieved from http://steinhardt.nyu.edu/ihdsc/savi

Julnes, G. (2011). Reframing validity in research and evaluation: A multidimensional, systematic model of valid inference. In H. T. Chen, S. I. Donaldson, & M. M. Mark (Eds.), *New Directions for Evaluation: No. 130. Advancing validity in outcome evaluation: Theory and practice* (pp. 55–67). San Francisco, CA: Jossey-Bass. doi:10.1002/ev.365

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*, 103–127.

LaLonde, R. J. (1986). Evaluating the econometric evaluations of training programs with experimental data. *American Economic Review*, *76*, 604–620.

Nisar, H. (2010). Do charter schools improve student achievement? Unpublished manuscript, Department of Economics, University of Wisconsin, Madison. Retrieved from http://www.ssc.wisc.edu/~scholz/Seminar/Charter_School_MPS.pdf

Olsen, R. B., Bell, S. H., Orr, L. L., & Stuart, E. A. (2013). External validity in policy evaluations that choose sites purposively. *Journal of Policy Analysis and Management*, *32*, 107–121.

O'Muircheartaigh, C., & Hedges, L. V. (2014). Generalizing from unrepresentative experiments: A stratified propensity score approach. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *63*, 195–210.

Puma, M., Bell, S., Cook, R., & Heid, C., with Shapiro, G., Broene, P., ... Spier, E. (2010). *Head Start Impact Study: Final report*. Washington, DC: U.S. Department of Health and Human Services, Administration for Children & Families.

Steiner, P. M., Cook, T. D., Shadish, W. R., & Clark, M. H. (2010). The importance of covariate selection in controlling for selection bias in observational studies. *Psychological Methods*, *15*, 250.

Stuart, E. A., Bell, S. H., Ebnesajjad, C., Olsen, R. B., & Orr, L. L. (2016). Characteristics of school districts that participate in rigorous national educational evaluations. *Journal of Research on Educational Effectiveness*.

Stuart, E. A., Bradshaw, C. P., & Leaf, P. J. (2015). Assessing the generalizability of randomized trial results to target populations. *Prevention Science*, *16*, 475–485.

Stuart, E. A., Cole, S., Bradshaw, C. P., & Leaf, P. J. (2011). The use of propensity scores to assess the generalizability of results from randomized trials. *Journal of the Royal Statistical Society, Series A*, *174*, 3969–3386.

Tipton, E. (2013). Improving generalizations from experiments using propensity score subclassification: Assumptions, properties, and contexts. *Journal of Educational and Behavioral Statistics*, *38*, 239–266.

Tipton, E. (2014). How generalizable is your experiment? Comparing a sample and population through a generalizability index. *Journal of Educational and Behavioral Statistics*, *39*, 478–501.

Tipton, E., Hedges, L., Vaden-Kiernan, M., Borman, G., Sullivan, K., & Caverly, S. (2014). Sample selection in randomized experiments: A new method using propensity score stratified sampling. *Journal of Research on Educational Effectiveness*, *7*, 114–135.

*Stephen H. Bell is a vice president and senior fellow at Abt Associates Inc., Social and Economic Policy Division.*

*Elizabeth A. Stuart is a professor at the Bloomberg School of Public Health, The Johns Hopkins University.*