Olsen, R. B., & Orr, L. L. (2016). On the "where" of social experiments: Selecting more representative samples to inform policy. 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*, 61–71.

4

On the "Where" of Social Experiments: Selecting More Representative Samples to Inform Policy

Robert B. Olsen, Larry L. Orr

Abstract

Most social experiments are conducted in samples of sites that are not formally representative of the population of policy interest. These studies may produce impact estimates that are unbiased for the sample but biased for the population from which the sample was selected. Recent research has estimated the bias associated with nonrandom inclusion. Although some research has focused on solutions to the problem at the design stage or the analysis stage, research on ways to address this problem is still sparse. This paper provides four recommendations to help researchers obtain more representative samples in impact studies. The fundamental challenge is that, in most impact studies, sites are not required to participate if selected. Therefore, obtaining a sample that adequately represents the population of policy interest can be difficult, and the resulting impact estimates may suffer from external validity bias. The recommendations in this chapter address this challenge to help researchers obtain more representative samples when obtaining a perfectly representative sample is not possible. The recommendations, which are based on standard survey sampling methods, demonstrate that researchers can take practical steps to obtain impact estimates that are more generalizable from the study sample to the broader population of policy interest—and therefore more relevant for informing policy decisions. © 2016 Wiley Periodicals, Inc., and the American Evaluation Association.

Impact evaluations based on prospective research designs, including randomized experiments, usually select a sample in two stages. In the first stage, the evaluators select a sample of "sites" defined by geography (e.g., counties) or local administrative units (e.g., welfare offices, school districts). In the second stage, they select units (individuals or groups of individuals, such as classrooms or families) within each site.

In most impact evaluations, the sites are not required to participate in the evaluation. If some sites refuse to participate, the sites included in the study may not be representative of the population from which they were selected and hence of the population affected by the policy decisions likely to be influenced by the study's findings. Anticipating the challenges in obtaining cooperation from sites, researchers rarely even try to recruit a representative sample of sites and instead select a convenience or purposive sample of sites from the outset that at best is designed to match the population of policy interest on a small number of observed characteristics.

Nonrandom site selection can potentially lead to biased impact estimates for the broader population(s) from which participating sites were selected. Olsen, Bell, Orr, and Stuart (2013) provide a formal model for purposive site selection, define external validity from purposive site selection as the difference between the average impact in a purposive sample and the true average impact in the population of policy interest—defined as the population that would be affected by the policy decision(s) the study is intended to inform—and derive a mathematical expression for this bias. This expression shows that this bias, which Olsen et al. refer to as "external validity bias," arises if (a) the probabilities of sites participating in the evaluation vary, (b) treatment effects vary, and (c) the correlation between those two quantities is nonzero. In other words, external validity bias becomes a problem when there is treatment effect heterogeneity and the factors that influence whether sites participate also influence the magnitude of the impact. Because the existence of these conditions cannot be known in advance (or, in most cases, after the fact), studies that select sites nonrandomly begin with the risk that the resulting impact estimates will be biased.

The only sure way for an evaluation to avoid external validity bias is for it to obtain a random sample of sites. Doing so requires sampling sites randomly and obtaining cooperation from all or a random sample of these sites. In this chapter, we focus on the former—the selection of sites. However, the concluding section reflects on the latter as well as other approaches to improving the external validity of impact evaluations.

Recent empirical research has attempted to quantify the magnitude of the external validity bias from purposive site selection. Bell, Olsen, Orr, and Stuart (2016) estimate the bias from conducting a hypothetical evaluation of the Reading First education program in a purposive sample to be approximately 0.10 standard deviations: This is twice the (admittedly arbitrary) threshold set by the What Works Clearinghouse for acceptable levels of attrition bias in randomized experiments (What Works Clearinghouse, 2014). It is also twice as large as the bias from using a differences-in-differences model, whose internal validity many researchers would question. Furthermore, Allcott (2015) and Allcott & Mullainathan, 2012 find evidence of external validity bias, which they call "partner selection bias," in evaluations of the impact of energy conservation programs.

Very little research considers solutions to the problems that arise from nonrandom site selection into experimental evaluations. Recent work (Tipton, 2013a; Tipton & Peck, 2016) has proposed a design-based approach that ensures that the nonrandom sample of participating sites successfully recruited for an experiment matches the population of interest on a set of observed characteristics. However, this approach may still produce impact estimates with external validity bias if unobserved site-level characteristics influence both site-level decisions about participating and the impact of the intervention in their sites. Other research has focused on analysis-based solutions once an unrepresentative set of sites is obtained; these solutions involve post-hoc statistical corrections for differences between the sample and the population on characteristics observed for both the sample and the population (e.g., Cole & Stuart, 2010; Kern, Stuart, Hill, & Green, 2016; Stuart, Cole, Bradshaw, & Leaf, 2011). Bell and Stuart (Chapter 3) discuss such approaches. Here, we focus on design-based methods.

Contribution of the Chapter

In this chapter, we make four recommendations for how to obtain more representative samples for impact evaluations. These recommendations are designed to obtain samples that are similar to the population of interest on both observed and unobserved characteristics. Obtaining a sample that is similar to the population on unobserved characteristics, as well as observed characteristics, is important in settings where the combination of theory and empirical evidence leaves uncertainty about the factors driving impact variation across sites. In this context, many of the site-level factors influencing site-level impacts are likely to be unobserved (because no one thought to collect data on those factors).

To obtain more representative samples in experimental impact evaluations, we recommend a strategy that relies heavily on random selection. The remainder of this chapter more fully articulates our proposed strategy, and each of the four recommendations that comprise this strategy.

Recommendation 1: Identify the Population of Policy Interest

Social experiments are conducted to inform policy decisions. Remarkably, many, if not most evaluations do not even attempt to specify the policy decision they are intended to inform or the population that would be affected by that policy decision. We view clear identification of this population (which we refer to as the "population of policy interest" and Tipton et al., 2014, refer to as the "inference population") as an essential first step in sample selection. As we will see, this is not always straightforward, a fact that may explain why many evaluators simply skip this step. Failure to identify the population of interest, however, especially when coupled with nonrandom sample selection, leaves the user of evaluation results with almost no guidance as to the applicability of the results beyond the evaluation sample itself.

To identify the population of policy interest, evaluators may wish to ask themselves the following questions:

- What policy decisions do we hope to inform with the results of our experiment?
- Whom would those policy decisions affect (i.e., which individuals, businesses, schools, states, offices)?
- Which group or groups of potentially affected individuals or entities are of greatest interest to policymakers?

One way to answer these questions is to focus on the population of interest to the evaluation sponsor that funded the study. Although other entities also may use the results to inform policy decisions that they face, the primary goal of an experiment is typically to inform a policy decision faced by the experiment's sponsor. Therefore, in this chapter, we focus on the population of interest to the study's research sponsor, given the sponsor's goals in conducting the impact evaluation.[1]

Identifying the population(s) of interest requires communication with the evaluation sponsor and careful thinking about the goal(s) of the experiment. If an evaluation is designed to inform the decision of whether to keep or eliminate an existing program, the population of interest could be current and future participants in places where the program currently operates. This is probably the population of policy interest in most evaluations of established federal programs. For example, the population of interest for the National Job Corps Study (Schochet, Burghardt, & McConnell, 2008) was probably all Job Corps participants nationwide, whereas the population of interest for the Head Start Impact Study (Puma, Bell, Cook, & Heid, 2010) was probably all Head Start participants nationwide. However, the population of interest in these studies might be defined more broadly to include all *eligible* individuals, including eligible nonparticipants, because they are the intended beneficiaries of the program.

[1] At the same time, we recognize that research funded by one sponsor may be used by a separate entity to inform policy decisions that it faces. For example, the state of California may fund an experiment to inform a state policy decision, but the state of Massachusetts may be interested in using the results to inform the same or similar policy decisions faced by state policymakers in Massachusetts.

If the evaluation is designed to inform the policy decision of whether to expand a program, the population of interest may include only individuals who are *not* currently participating in the program. For example, the population of interest might be eligible nonparticipants in places where the program is not currently operating who could potentially participate if the program were expanded, or eligible nonparticipants in existing program communities who might participate if program targeting were broadened, program outreach were more aggressive, or program funding were increased in existing program sites.

Evaluations are sometimes categorized into "effectiveness" and "efficacy" studies. Effectiveness studies are explicitly intended to measure the effects of the intervention on the population of interest. Efficacy studies are conducted to determine whether an intervention has a positive effect under favorable conditions, and the results are interpreted as evidence of whether an effectiveness study is warranted.

Defining the population of interest is absolutely essential in effectiveness studies. Effectiveness studies are used to determine whether some population is (or would be) better off with the program than without. However, even efficacy studies that assess whether the program or intervention has impacts under favorable conditions would benefit from having a well-defined population of interest. It would be useful for evaluators conducting efficacy trials to clarify the conditions thought to be favorable (e.g., sites that demonstrate a particular capacity believed to be important for implementing the intervention with high fidelity) and to define the population of interest to include sites in which these favorable conditions are present. Defining the population of interest would help policymakers understand the conditions under which an intervention has been shown to work (or not work); it would also help to identify conditions under which the intervention has not yet been tested, to inform future research.

Recommendation 2: Develop a Sampling Frame

Once the population of interest is defined, evaluators need a sampling frame from which to select sites. Although a sampling frame may not be necessary to identify a convenience sample of volunteers, it is essential for any type of sampling from the larger population (random or otherwise).

In many evaluations, a sampling frame of all sites in the population may be readily available. For example, in evaluations of federal grant programs, the federal agency sponsoring the evaluation will be able to provide a list of all current grantees (if the list is not publicly available online). In evaluations of educational programs, the Common Core of Data maintained by the National Center for Education Statistics at the U.S. Department of Education provides a census of public schools and school districts that can be used to construct a sampling frame when the population of interest consists of some subset of public school students. Similar sampling frames exist in other policy areas.

In other evaluations, there may be no ready-made listing of the universe of possible sites to serve as a sampling frame, and the researchers may need to construct the list themselves (e.g., see Tipton and Peck, 2016, for a discussion of this issue in the context of welfare evaluations). In such cases, the sampling frame need not include all eligible sites in the population of interest (because constructing such a frame could be prohibitively expensive). Evaluators can instead randomly select geographic areas or political jurisdictions, identify all eligible sites within those areas, and select from those sites, in a process similar to multistage probability sampling designs commonly used in survey research. This is effectively the same as, and in terms of bias risk statistically equivalent to, having the entire universe in the enumerated frame.

For example, suppose that evaluators were conducting an evaluation of home health care providers. The evaluators could select a random sample of counties—perhaps oversampling large counties—canvas the selected counties to identify all of the home health care providers located within them, and select a random sample of the service providers that were identified in these counties. The resulting sample could be weighted using the selection probability to represent the full population of home health care providers nationwide.

Recommendation 3: Select Sites Randomly

As noted earlier, the only way to be sure that the evaluation sites represent the larger population of interest is to select them randomly from that population. Selecting sites randomly will not guarantee that the resulting sample will be representative of the population if sites can opt out of the study. However, it will guarantee that within sampling error, the sites *recruited* to participate will be statistically equivalent to the population on both observed and unobserved factors. This seems to us a major advantage over starting with a sample of unknown external validity. Random selection of sites allows us to weight sites by the inverse of their selection probability, as we do in surveys, and to reweight for nonparticipation of selected sites, as we do to correct for nonresponse in surveys.

In selecting sites randomly, researchers could select a simple random sample from the sampling frame of eligible sites. Simple random sampling ensures that there are no systematic differences between the sample of selected sites and the population of sites from which they were selected.

However, in practice, stratified random sampling—that is, stratifying the sites into groups and then selecting a simple random sample from each group—has an important advantage over simple random sampling: It ensures that the composition of the sampled sites is identical to that of the population from which it was selected on each of the stratifying variables.

By removing chance differences on these factors between the sample and the population, stratified random sampling can improve both the face validity of the sample for estimating average impacts in the population and the statistical precision of the analysis.

To obtain more precise estimates of the average impact in the population, strata should be defined based on factors that the evaluator expects to be associated with the impacts of the intervention.[2] For example, if the effects of the intervention might be expected to vary with the size of the local program, then the sampling frame of potential sites could be stratified by size (e.g., dividing sites into large and small or large, medium, and small before selecting a random sample within the strata). Selecting stratifiers that are likely to be associated with the impacts of the intervention requires theory—or at least some hard thinking—and clearly benefits from empirical evidence. In the absence of either theory or evidence, evaluators may want to consider conducting a cluster analysis to identify a smaller number of groups that are similar on a larger number of variables (Tipton, 2013a; Tipton & Peck, 2016). Finally, the maximum number of strata is constrained by the number of sites to be selected (see Orr, 1999, for more guidance on the optimal number of sites).

The principal critique of random site selection is that if sites can opt out of the study, random site selection may be of little benefit. We acknowledge that random site selection is not sufficient to eliminate external validity bias. However, it is a *necessary* first step toward producing externally valid impact estimates unless selection into the study is ignorable conditional on observed site-level characteristics (Tipton, 2013b). This condition is unlikely to hold in the real world, where intervention effects vary along dimensions that are not captured by the data collected.

The ability to recruit randomly selected sites will depend heavily on the specific circumstances of the evaluation. In the National Job Training Partnership Act (JTPA) Study, roughly 90% of the sites approached refused to participate, largely for reasons very specific to the JTPA program (Doolittle & Traeger, 1990). The evaluation of the Food Stamp Employment and Training Program (Puma & and Burstein, 1994), on the other hand, achieved an 80% participation rate, resulting in a sample of 53 randomly selected sites, and the National Head Start Evaluation (Puma et al., 2010) included 87 of the 90 eligible grantees that were randomly selected for the study. These efforts demonstrate the feasibility of the approach in at least some circumstances. More generally, the only way to learn how broadly the approach can be applied, and under what conditions, and to develop more effective recruiting techniques, is for evaluators to try to implement random selection of sites. With this chapter, we hope to provide a roadmap for evaluators to begin to do so with greater frequency.

[2] Evaluators may also select stratifiers for other purposes, such as improving statistical power in making subgroup comparisons between different types of sites.

This brief discussion of site recruitment within evaluations highlights the practical obstacles that exist in the field. Because some sites will refuse to participate, researchers must choose backup sites to replace nonparticipating sites. The simplest way to do this is to replace nonparticipating sites with other randomly selected sites from the same sampling frame or stratum. If the number of refusals is high, we also recommend selecting a random subset of the refusing sites for more intensive recruitment efforts, perhaps with additional inducements to participate. The random subset of refusing sites can be "weighted up" to represent all refusing sites to produce a more representative sample with less external validity bias.

Recommendation 4: Set Sample Sizes to Account for Random Site Selection

Evaluators routinely set sample size targets for their evaluations to achieve the desired power for "fixed effects" impact analysis (i.e., analysis of the effects within the sample of participating sites). But policy interest seldom focuses on particular sites. The U.S. Department of Labor may test a new job search approach in Dayton and Kalamazoo, but it is not interested in the impacts in those two cities per se. It is interested in what the effects in those two cities reveal about what the effects might be in the rest of the country. For that purpose, the standard error of the fixed effects estimator gives a biased measure of the uncertainty attached to evaluation estimates because it omits a critical part of the variance of the estimate—the sampling error associated with choosing those two sites rather than any other two (or more) sites in the nation.

To compensate for sampling error when selecting sites randomly, larger samples are needed. As shown in Schochet (2008), the increase in sample size requirements depends entirely on the variation in impacts across sites (see equation 7 in Schochet): The larger the cross-site variation in impacts, the greater the sample required to detect impacts from a random sample of these sites. Selecting sites randomly may increase sample size requirements for one additional reason: The effects of the intervention may be smaller in a random sample of sites than it would be in a purposive sample of sites because purposive selection is often designed to select sites in which the conditions are favorable for positive impacts (as in efficacy trials).

Although larger samples certainly increase the cost of any evaluation, the main alternative and status quo—purposive site selection with smaller samples—has serious limitations that reduce the value of the evaluation findings. Purposive site selection yields evidence that cannot be confidently generalized to the populations of interest to policymakers. Fixed effects analysis, as mentioned earlier, does not properly account for the uncertainty in the estimation—and therefore understates the "risk" associated with making whatever policy decisions the evaluation was designed to inform. In contrast, random site selection is a necessary first step toward producing evidence that can be confidently generalized to the populations of interest to policymakers (those that their policy decisions would affect).

Future Directions

As we asserted earlier, random site selection is a necessary first step to obtaining a representative sample, but it is not sufficient. To be confident that the results from the sample generalize to the population of policy interest requires that the random sample of sites agrees to participate in the study. Except in those cases where sites are effectively required to participate in an evaluation—as when funding agencies include evaluation requirements in grant notifications and agreements—it is probably not realistic to expect all sites to agree to participate. Therefore, it is important to consider additional approaches to produce more externally valid and useful evidence to inform policymakers.

In our assessment, three potentially fruitful approaches may produce more externally valid impact evaluations. The first approach involves designing studies to substantially increase sites' motivation to participate. Low take-up rates are arguably the primary challenge in conducting experimental evaluations. Improvements in design or smarter incentives would be welcomed, both to increase the number of randomized experimental evaluations that are conducted and to increase the chances that the resulting samples are representative of the populations from which they were chosen. Advances in behavioral sciences have been exploited to develop low-cost interventions that yield substantial behavioral responses by individuals; behavioral science may hold the key to developing design enhancements, marketing, or financial incentives that would encourage more sites to participate in experimental impact evaluations.

The second approach involves increased use of analytic methods which were designed for other purposes but that may be useful in reducing external validity bias from unrepresentative samples. For example, a nonrandom sample can be reweighted to more closely match the population of interest on observed characteristics (e.g., see Cole & Stuart, 2010; Stuart et al., 2011). To the extent that matching reduces differences between the sample and the population that are associated with impacts, this will reduce external validity bias. Other more sophisticated methods have been developed to relax some of the restrictive assumptions of standard methods (e.g., see the use of Bayesian Additive Regression Trees in Hill, 2011 and Kern, Stuart, Hill, & Green, 2016).

The third potentially fruitful direction would be increased development and application of methods to estimate external validity bias, as discussed in Bell and Stuart (Chapter 5). For example, when the take-up rate in an experiment is low, a quasi-experimental design may be a useful complement because it could be implemented in both the sites that agreed to participate in the experiment and in sites that refused to participate but agreed to implement the intervention (if they had not already). A quasi-experimental design could involve matched comparison sites or individuals. To estimate the external validity bias generated by site refusals, researchers could compare quasi-experimental impact estimates for the sites that agreed to participate in the experiment to quasiexperimental impact estimates for the sites that refused to participate. Kaizar (2011) formalizes this approach and shows how the quasiexperimental and experimental estimates can be combined to reduce external validity bias. The results from this kind of analysis may suggest that the findings from the experiment either overstate or understate the average impacts in the broader population—potentially useful information to policymakers who have to predict the likely consequences of different policy decisions.

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. (2015). Site selection bias in program evaluation *The Quarterly Journal of Economics*, *130*(3), 1117–1165.

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

Bell, S. H., Olsen, R., 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

Cole, S. R., & Stuart, E. A. (2010). Generalizing evidence from randomized clinical trials to target populations: The ACTG-320 trial. *American Journal of Epidemiology*, *172*, 107–115.

Doolittle, F., & Traeger, L. (1990). *Implementing the National JTPA Study*. New York: Manpower Demonstration Research Corporation.

Hill, J. L. (2011). Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, *20*, 217–240. doi:10.1198/jcgs.2010.08162

Kaizar, E. E. (2011). Estimating treatment effect via simple cross design synthesis. *Statistics in Medicine*, *30*, 2986–3009.

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.

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

Orr, L. L. (1999). *Social experiments: Evaluating public programs with experimental methods*. Thousand Oaks, CA: Sage Publications.

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.

Puma, M. J., & Burstein, N. R. (1994). The National Evaluation of the Food Stamp Employment and Training Program. *Journal of Policy Analysis and Management*, *13*, 311–330.

Schochet, P. Z. (2008). Statistical power for random assignment evaluations of education programs. *Journal of Educational and Behavioral Statistics*, *33*, 62–87.

Schochet, P. Z., Burghardt, J., & McConnell, S. (2008). Does Job Corps work? Impact findings from the National Job Corps Study. *American Economic Review*, *98*, 1864–1886.

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. (2013a). Stratified sampling using cluster analysis: A sample selection strategy for improved generalizations from experiments. *Evaluation Review*, *37*, 109–139.

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

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.

Tipton, E., & Peck, L. (2016). A design-based approach to improve external validity in welfare policy evaluations. *Evaluation Review*. doi:10.1177/0193841X16655656

What Works Clearinghouse. (2014). *WWC procedures and standards handbook, version 3.0*. Retrieved from http://ies.ed.gov/ncee/wwc/pdf/reference_resources/wwc_procedures_v3_0_standards_handbook.pdf

Robert B. Olsen is an economist and president of Rob Olsen LLC.

Larry L. Orr is an associate at the Bloomberg School of Public Health, The Johns Hopkins University, and an independent evaluation consultant.