Methods for Policy Analysis

Rebecca A. Maynard
Editor

Authors who wish to submit manuscripts for all sections except Book Reviews should do so electronically in PDF format through Editorial Express.

EXTERNAL VALIDITY IN POLICY EVALUATIONS THAT CHOOSE SITES PURPOSIVELY

Robert B. Olsen, Larry L. Orr, Stephen H. Bell, and Elizabeth A. Stuart

Abstract

Evaluations of the impact of social programs are often carried out in multiple sites, such as school districts, housing authorities, local TANF offices, or One-Stop Career Centers. Most evaluations select sites purposively following a process that is nonrandom. Unfortunately, purposive site selection can produce a sample of sites that is not representative of the population of interest for the program. In this paper, we propose a conceptual model of purposive site selection. We begin with the proposition that a purposive sample of sites can usefully be conceptualized as a random sample of sites from some well-defined population, for which the sampling probabilities are unknown and vary across sites. This proposition allows us to derive a formal, yet intuitive, mathematical expression for the bias in the pooled impact estimate when sites are selected purposively. This formula helps us to better understand the consequences of selecting sites purposively, and the factors that contribute to the bias. Additional research is needed to obtain evidence on how large the bias tends to be in actual studies that select sites purposively, and to develop methods to increase the external validity of these studies. © *2012 by the Association for Public Policy Analysis and Management.*

INTRODUCTION

In recent years, a great deal of attention has been paid to the internal validity of impact evaluations—that is, whether the evaluation yields an unbiased estimate of the impact of the policy or program in the studied sites. Much less attention has

Journal of Policy Analysis and Management, Vol. 32, No. 1, 107–121 (2013)

Published by Wiley Periodicals, Inc. View this article online at wileyonlinelibrary.com/journal/pam
DOI:10.1002/pam.21660

been paid to external validity—that is, whether the evaluation yields an unbiased estimate of the impact of the program on the population or populations of policy interest.

Impact evaluations of social programs are often carried out in multiple *sites*, such as school districts, housing authorities, local TANF offices, or One-Stop Career Centers. Ideally, in each impact evaluation, sites would be selected randomly from the population of interest, and all selected sites would agree to participate. Under this scenario, the resulting sample would be formally representative of the population from which it was selected. As a result, the study findings would have high external validity in the sense that they provide unbiased estimates of the impacts of the policy or program for the population of policy interest.

In practice, most impact evaluations select sites purposively using a nonrandom process. In these evaluations, the samples are not designed to represent a well-defined population in the formal statistical sense. Furthermore, in many of these evaluations, some or many of the selected sites choose not to participate in the evaluation. Therefore, in the typical impact evaluation, the resulting impact estimates may not be generalizable to the primary population of policy interest.

For some evaluations, this may not be a problem. In particular, efficacy trials are typically designed to assess whether the program can work under favorable circumstances, not whether the program does in fact work under more typical conditions. Therefore, efficacy trials are not designed to produce results that are generalizable to any population of policy interest.

However, many impact evaluations are clearly intended as effectiveness studies to inform policy decisions. Furthermore, many impact evaluations that are not clearly designated as either efficacy trials or effectiveness studies may be perceived to provide the best available evidence on the effect of a program or intervention. When these studies are based on purposive samples of sites, the generalizability of the study findings to broader populations of interest is uncertain.

In evaluations of government programs and policies, the primary population of policy interest (hereafter referred to as "the population of interest" for simplicity) generally consists of all individuals who are potentially affected by the policy decision that the study findings are intended to inform. If these individuals are spread across multiple sites, the population of interest covers all sites that include these individuals (e.g., all cities with homeless people or all welfare offices that provide child-care assistance to welfare mothers).

To inform policy decisions, policymakers need evidence on the impacts of the program or policy on the population of interest, however it is defined. When evaluations produce evidence from samples that are not representative of the population of interest, there is a disconnect between the evidence needed for policy and the evidence produced by our evaluations. This disconnect can lead evaluations to provide misleading evidence to policymakers when the impacts in study sites differ substantially from the impacts in the other sites in the population of interest.

It is important to recognize that purposive site selection can only lead to misleading impact estimates for the population of interest if the impacts vary across sites (as we show formally later in this paper). To the best of our knowledge, no comprehensive synthesis of the evidence on treatment effect heterogeneity for social policy interventions has been conducted. At the same time, there are plenty of examples of research studies that have found variation in impacts across sites. For example, Greenberg et al. (2003) found significant variation in impacts across sites in two large-scale evaluations of welfare-to-work programs: the National Evaluation of Welfare-to-Work Strategies (Freedman et al., 2000; Hamilton & Brock, 1994) and the Greater Avenues for Independence evaluation (Riccio, Friedlander, & Freedman, 1994). In addition, two recent studies of charter schools found significant variation in impacts across schools (Gleason et al., 2010; Nisar, 2010).

In this paper, we begin by defining what we mean by purposive site selection and documenting its prevalence in social experiments. Then we present a conceptual model for purposive site selection—that is, the process by which purposive samples are chosen. This model allows us to derive a formula for the bias that can result from purposive site selection when impacts vary across sites in the population. We conclude with recommendations for future research, including research designed to estimate the magnitude of the bias, to help the field gain an understanding of the limitations of impact evaluations that select sites purposively.

Site Selection in Impact Evaluations

In general terms, there are three approaches to selecting sites for an evaluation: (1) selecting all possible sites, (2) selecting a random sample of sites, and (3) selecting a purposive sample of sites. Purposive sampling is "a method by which units are selected to be in a sample by a deliberate method that is not random" (Shadish, Cook, & Campbell, 2002, p. 511). Purposive samples are presumably selected with some goal or goals in mind, goals that at least in some instances stem from the specific objectives of the research. For example, if the evaluator or evaluation sponsor wants to compare the impacts for one group of sites to the impacts for a different group of sites (such as urban to rural), the evaluator may select equal numbers of sites from both groups. In other cases, purposive sampling may seek to ensure that a range of different types of sites is represented in the sample.[1]

We define the term *purposive site selection* to include what is sometimes called *convenience sampling*, where the primary objective is to minimize the costs incurred or the time necessary to meet the evaluation's sample size requirements. In our experience, purposive sampling is often implemented as a form of *stratified convenience sampling*: researchers may select the distribution of sites that they hope to achieve (e.g., X urban sites and Y rural sites), but subject to those constraints, include the sites that are most convenient to include in the study (e.g., those that were easiest to persuade to participate in the study). In this paper, we use the term purposive site selection to cover both stratified and unstratified convenience sampling.

Purposive site selection has been used and defended since the earliest large-scale social experiments. For example, in the final report on the New Jersey Income Maintenance Experiment (Watts, Peck, & Taussig, 1977), the authors explained the operational necessity of implementing the test in a small number of sites and defended a *test bore* strategy of "examining a discrete number of purposely chosen and distinctive samples from which a complete composite can eventually be formed" (p. 445).

Over the ensuing decades, the test bore approach—purposive selection of a small number of sites and often a small fraction of all potential sites—became the accepted standard in evaluation. To assess the prevalence of purposive site selection, we reviewed the *Digest of Social Experiments* (Greenberg & Shroder, 2004). The *Digest* attempted to capture every social experiment designed to determine how one or more policy interventions affected individual or household market behavior (e.g., employment and earnings; consumption of food, energy, housing, or health care; or receipt of government transfers). Of the 273 evaluations included in the *Digest*, all but seven appear to have included a purposive sample of sites. In addition, a substantial majority included sites that covered 10 or fewer localities, which in many cases is a small fraction of all localities that are or could be implementing the

[1] Later in this paper, we consider how these objectives can be met through stratified random sampling of sites.

program.[2,3] In summary, most social experiments covered in the *Digest* have used much the same approach that was used in the income maintenance experiments: purposive selection of a small number of sites, often aiming for diversity of sites, but seldom choosing sites in a way that would allow rigorous generalization.

We should note that some evaluations select all eligible sites into the evaluation sample (which of course yields a representative sample of eligible sites). The *Digest* identified five such evaluations: the National Job Corps Study, the Washington State Intensive Applicant Employment Services Evaluation, the Delaware Dislocated Worker Pilot Program, and two evaluations conducted outside of the United States.[4]

In contrast, random site selection is rare. The *Digest* clearly identifies only two social experiments as having selected sites randomly: (1) the Food Stamp Employment and Training Program Evaluation (Puma et al., 1990), and (2) the Pennsylvania Reemployment Bonus Demonstration (Corson et al., 1991). We know of three other social experiments that selected sites randomly, were included in the *Digest*, but were not clearly identified by the *Digest* as having selected sites randomly: (1) the National Evaluation of the Upward Bound Program (Seftor, Mamun, & Schirm, 2009), (2) the Evaluation of Ohio's Learning, Earning, and Parenting Program (Bos & Fellerath, 1997), and (3) the Evaluation of Florida's Project Independence (Kemple, Friedlander, & Fellerath, 1995). Although there may be other evaluations that selected sites randomly, the combination of our experience and the evidence from the *Digest* suggests to us that random site selection is very rare in social experiments.

Because the *Digest* includes studies that are typically at least 10 years old, it is reasonable to ask whether random site selection has become more common in the last decade. Our experience suggests not. We are only aware of three recent multisite impact evaluations that have selected a random sample of sites: (1) the Head Start Impact Study (Puma et al., 2010), (2) the Benefit Offset National Demonstration (Stapleton et al., 2010), and (3) the Impact Evaluation of Upward Bound's Increased Emphasis on Higher-Risk Students (cancelled less than one year after sites were selected).

In our experience, evaluators and evaluation sponsors often justify selecting sites purposively instead of randomly to reduce the costs of the evaluation. To control costs, researchers may focus on sites that offer the largest study sample or the least resistance to participating in the study. In addition, for random assignment evaluations, evaluators may focus on sites with more eligible applicants than the program can serve; otherwise, the evaluation may bear the costs of recruiting additional applicants. Furthermore, the benefits of random site selection may be unclear when the number of sites is small or the selected sites may choose to opt out the evaluation. In either case, the sample of participating sites may not resemble the population from which it was selected, even if sites were initially selected randomly. Regardless of whether these factors provide an adequate justification for purposive site selection, they likely contribute to the fact that few evaluations select sites randomly in the

[2] Our calculations suggest that of the 231 evaluations conducted in the Unites States and reported by the *Digest*, 177 included sites in 10 or fewer localities (e.g., cities or counties) and 28 included sites in 11 or more localities. The remainder of the evaluations were conducted throughout one region of a single state (in an unspecified number of localities), statewide in a single state, statewide in 2–10 states, or, in the case of the National Job Corps Study, nationwide.

[3] The number of sites included in an evaluation is heavily influenced by the evaluation budget. Therefore, our discussion here should not be misconstrued as a criticism of evaluations that are based on relatively few sites.

[4] The *Digest* includes 42 experimental and quasi-experimental evaluations that were conducted outside of the United States.

hopes of obtaining a representative sample—and that most evaluations instead rely on purposive samples.

Contribution of This Paper

The main contributions of this paper are the development of a conceptual model of purposive site selection and the derivation of a formal expression for the bias that can result when evaluations select sites purposively. This paper is not the first to address external validity of study results from an unrepresentative sample. For example, Shadish, Cook, and Campbell (2002) and Imai, King, and Stuart (2008) provide conceptual frameworks for thinking about external validity and the biases that may exist when trying to generalize study findings to target populations. In addition, Cole and Stuart (2010) and Weisberg, Hayden, and Pontes (2009) provide expressions for external validity bias in settings where individuals are randomized and impacts may vary across individuals.[5]

However, although other papers derive ex post expressions for the bias in terms of the realized differences between the purposive sample and the population, this paper provides an a priori expression for the bias by treating purposive site selection as a sampling process. This allows us to show that the bias from purposive site selection results from a selection process that favors certain types of sites over others, but that fails to account for this key feature of the site-selection process in the analysis. In particular, in evaluations that select sites purposively, evaluators typically favor sites based on site-level characteristics that are available to the evaluators and accept sites that agree to participate based on a process we cannot observe. We formally demonstrate that the process by which sites are included in purposive samples can produce biased impact estimates—that is, impact estimates that are systematically different from the impacts that policymakers would like to know. Because most impact evaluations of social programs begin by selecting a sample of sites, the paper's findings are broadly relevant to the field of rigorous program evaluation.

We stress that other factors may also limit external validity in any particular study. For an excellent discussion of the threats to validity, see Shadish, Cook, and Campbell (2002). We focus here on external validity bias due to purposive site selection, and leave investigation of other factors that may limit generalizability to future research.

A FORMAL MODEL OF PURPOSIVE SITE SELECTION

Our conceptual model of purposive site selection in multisite impact evaluations is based on the proposition that any purposive sample of sites can usefully be conceptualized as a random sample of sites from some well-defined population, for which the sampling probabilities are unknown and vary across sites. We use this model to derive a formal expression for the bias that results from selecting sites purposively.

Conceptual Model

In our experience, the process of selecting and recruiting sites for multisite impact evaluations has four key steps. First, the evaluation may specify its population of

[5] Weisberg, Hayden, and Pontes (2009) are particularly concerned with examining the consequences for generalizability of the strict inclusion and exclusion criteria of many clinical trials.

interest. In other cases, the population of interest may be left unspecified—either because the population of interest is obvious (e.g., all program participants, in evaluations of ongoing programs) or because there may be multiple populations of interest to the evaluation sponsor (e.g., in evaluations of interventions that may be adopted voluntarily by local entities). Second, the evaluation defines what a site is and may specify eligibility criteria for whether a site can be included in the evaluation, or at least describe the types of sites it aims to recruit. Third, the evaluation selects a sample of sites to recruit and invites them to participate in the evaluation. Fourth, these sites must decide whether or not to participate. The selection and recruiting process typically continues until the evaluation meets its sample size requirements (or decides to conduct the evaluation with a reduced sample).

In our conceptual model, we treat purposive site selection as a process. Instead of focusing on the outcome—the sample actually selected—we conceptualize the selection process as a random process with well-defined but unknown probabilities. More specifically, we assume that for any evaluation, for each site in the population of interest, there exists a well-defined probability of inclusion in the evaluation. Like all probabilities, the probability for each site falls between 0 and 1, inclusive. However, unlike formal probability sampling, the probabilities are unknown even to the researchers who selected or recruited the sample.

In this model, we define a site's probability of inclusion as the proportion of *replications* of the site inclusion process in which the particular site would be included in the evaluation sample (i.e., in which the site would both be chosen and agree to participate). We define a replication as a hypothetical execution of a site inclusion process, which is defined by certain fixed parameters, but also includes some variable or random elements. The fixed parameters of the inclusion process may include the universe of potential sites and the target number of sites to be included; the variable elements of the inclusion process may include the procedures used to recruit eligible sites and time-varying factors that influence sites' willingness to participate, including the personality traits of site-level decisionmakers and political factors that influence their decisions. Under this conceptual model, the particular sites included in the evaluation can vary across replications.

Although this conceptual model may seem restrictive, it is in fact sufficiently general to allow for any kind of inclusion process. At one extreme, it allows for a perfectly deterministic site inclusion process (e.g., 60 eligible schools with an inclusion probability of 1 and 940 eligible schools with an inclusion probability of 0). At the other extreme, it allows for a perfectly random process (e.g., all 1,000 eligible schools with a 6 percent chance of inclusion in the sample).

Most importantly, our model allows for more realistic situations in which some sites in the population of interest have zero probabilities of inclusion and other sites have positive but varying inclusion probabilities. For example, for a hypothetical random assignment evaluation of after-school programs, the probability of inclusion may be zero for sites that lack enough excess demand to conduct random assignment, small but positive for oversubscribed sites that serve a small number of children (e.g., those located in rural areas), and larger for oversubscribed sites that serve a large number of children (e.g., those located in urban areas).

One way of understanding our conceptual model is by analogy. Our conceptual model for purposive sampling is analogous to the conceptual models behind Donald Rubin's theory of missing data (e.g., Rubin, 1976) and James Heckman's theory of sample selection bias (e.g., Heckman, 1976). Both of these models consider the absence of particular units from the analysis sample as having a probabilistic component. Our model can thus be thought of as a special case of more general models that have played a prominent role in evaluation research.

External Validity Bias

In this subsection, we derive a mathematical expression for the bias that results from selecting sites purposively and then using standard methods to obtain a pooled impact estimate. First, let us formally establish the parameter of interest in multisite impact evaluations as the average impact in the population of interest. We assert that in most evaluations, the main parameter of interest is either the average impact across all sites in the population of interest or the average impact across all individuals in this population (where the latter is simply a weighted average of the former). To derive a formal expression for the bias, we focus on the former impact. This would be a key parameter of policy interest if individual sites can choose whether to adopt the intervention, or if policy decisions are made at a higher level for all sites in the population, but the number of program participants per site is the same across all sites in the population. Equation (1) defines the parameter of interest as Δ:

$$\Delta = \frac{1}{K}\sum_{s=1}^{K}\Delta_s \tag{1}$$

where K equals the number of sites in the population and Δ_s is the impact in site s for $s = 1, \ldots, K$.[6]

Suppose that J sites are included in the evaluation, where $J < K$, and the J sites included in the evaluation are a subset of the K sites in the population of interest. Equation (2) defines the pooled impact estimator that is often computed in multisite evaluations based on purposive site selection, which is just a simple average of the site-level impact estimates from the sites included in the evaluation:

$$\hat{\Delta}_{pooled} = \frac{1}{J}\sum_{j=1}^{J}\hat{\Delta}_j \tag{2}$$

where j subscripts the J sites included in the evaluation sample and $\hat{\Delta}_j$ is the impact estimate in site j.

An alternative way of expressing this estimator is the following:

$$\hat{\Delta}_{pooled} = \frac{1}{J}\sum_{s=1}^{K}\hat{\Delta}_s I_s \tag{3}$$

where I_s equals 1 if site s from the population was included in the evaluation and equals 0 otherwise.

The bias of the estimator in equation (3) equals the expected difference between this estimator and the average impact shown in equation (1):

$$Bias = E\left(\hat{\Delta}_{pooled} - \Delta\right) = E\left(\hat{\Delta}_{pooled}\right) - \Delta \tag{4}$$

The expectation in equation (4) is defined across replications of a given evaluation design. The evaluation design to be replicated includes both a specific process for selecting sites and a specific methodology for estimating impacts in each site that

[6] If the impacts vary across individuals in site s, we can think of Δ_s as the average impact in site s.

could potentially be included in the evaluation. The methodology for estimating impacts includes both the process for selecting the study sample in each included site, and for evaluations based on random assignment and many quasi-experimental methods, a process for assigning sample members to groups. The pooled impact estimate will vary across replications for two reasons: (1) the sites selected for the evaluation will vary across replications, and (2) for each site, the individuals included in the treatment and control or comparison groups will vary across replications. The expected value of the impact estimate, $E(\hat{\Delta}_{pooled})$, is defined as the limit of the simple average of the pooled impact estimates across replications of the evaluation as the number of replications approaches infinity.

Substituting equation (3) into equation (4), and moving the expectation inside the summation, yields equation (5):

$$Bias = E\left(\frac{1}{J}\sum_{s=1}^{K}\hat{\Delta}_s I_s\right) - \Delta = \left[\frac{1}{J}\sum_{s=1}^{K}E(\hat{\Delta}_s I_s)\right] - \Delta \tag{5}$$

To simplify this expression, we assume that for each site s, $\hat{\Delta}_s$ and I_s are statistically independent across replications of the evaluation—that is, $E(\hat{\Delta}_s|I_s) = E(\hat{\Delta}_s)$ and $E(I_s|\hat{\Delta}_s) = E(I_s)$. (Note that we are assuming an impact estimate could be computed in every site in the population of interest for each replication of the evaluation, and inclusion in the study only affects whether the impact estimate is observed and included in the pooled impact estimate.) A sufficient condition for the independence assumption to hold would be if the variation in each variable across replications were purely random and uncorrelated with all other variables. Intuitively, we would expect the independence assumption to hold if, (a) for each site s, the inclusion of the site in the evaluation (I_s) does not affect the site's impact estimate ($\hat{\Delta}_s$)—as distinct from the true impact in the site (Δ_s)—(b) the site's impact estimate does not affect the site's inclusion in the evaluation, and (c) there are no other factors that affect both and thereby create a statistical dependence between the two variables.

By assuming that $\hat{\Delta}_s$ and I_s are independent for each site s, we arrive at equation (6):

$$Bias = \left[\frac{1}{J}\sum_{s=1}^{K}E(\hat{\Delta}_s)E(I_s)\right] - \Delta \tag{6}$$

The expected value of the impact estimate in site s ($E(\hat{\Delta}_s)$) can be expressed as the sum of the true impact in site s (Δ_s) and the bias in the impact estimator or methodology for estimating the impact in site s (b_s), which we refer to as the internal validity bias in estimating the causal effect of the intervention in site s. In addition, the expected value of the 0–1 site inclusion indicator for site s ($E(I_s)$) is P_s by definition. Substituting these values into equation (6), we arrive at equation (7):

$$Bias = \left[\frac{1}{J}\sum_{s=1}^{K}(\Delta_s + b_s)P_s\right] - \Delta \tag{7}$$

The first term in equation (7) can be decomposed into two parts: the portion attributable to the actual site-level impacts in each site (Δ_s) and the portion attributable to the internal validity bias in the individual site-level impact estimates

(b_s), as shown in equation (8):

$$Bias = \left(\frac{1}{J}\sum_{s=1}^{K}\Delta_s P_s\right) + \left(\frac{1}{J}\sum_{s=1}^{K} b_s P_s\right) - \Delta \tag{8}$$

If we multiply and divide each term in equation (8) by K—the number of sites in the population—and use standard formulas for the covariance between two variables,[7] we arrive at equation (9):

$$Bias = \left[\frac{K}{J}(\sigma_{\Delta P} + \mu_\Delta \mu_P)\right] + \left[\frac{K}{J}(\sigma_{bP} + \mu_b \mu_P)\right] - \Delta \tag{9}$$

where μ_Δ is defined as the population mean of the site-level impacts, μ_P is defined as the population mean of the site inclusion probabilities, $\sigma_{\Delta P}$ is defined as the population covariance between the site-level impacts and the site inclusion probabilities, μ_b is defined as the population mean of the bias in the site-level impact estimates, and σ_{bP} is defined as the population covariance between the bias in the site-level impact estimates and the site inclusion probabilities.

It is important to recognize that (a) the population mean of the site-level impacts is, by definition, the parameter of interest established in equation (1) (i.e., $\mu_\Delta \equiv \Delta$), and (b) the population mean of the site-level inclusion probabilities is, by construction, equal to the fraction of all sites in the population to be included in the evaluation (i.e., $\mu_P = J/K$).[8]

This allows us to express the bias of the pooled impact estimator as follows:

$$\begin{aligned} Bias &= \left[\frac{K}{J}\left(\sigma_{\Delta P} + \Delta\frac{J}{K}\right)\right] + \left[\frac{K}{J}\left(\sigma_{bP} + \mu_b\frac{J}{K}\right)\right] - \Delta \\ &= \left[\frac{\sigma_{\Delta P}}{J/K}\right] + \left[\frac{\sigma_{bP}}{J/K}\right] + \mu_b \\ &= \left[\frac{\rho_{\Delta P}\sigma_\Delta\sigma_P}{J/K}\right] + \left[\frac{\rho_{bP}\sigma_b\sigma_P}{J/K}\right] + \mu_b \end{aligned} \tag{10}$$

where σ_Δ is the standard deviation of impacts across sites in the population, σ_P is the standard deviation of the site inclusion probabilities across sites in the population, $\rho_{\Delta P}$ is the correlation between the site-level impacts and the site inclusion probabilities across sites in the population, ρ_{bP} is the correlation between the bias in the site-level impact estimates and the site inclusion probabilities across sites in the population, and σ_b is the standard deviation of the bias in the site-level impact estimates across sites in the population.

Now we define cv_P as the coefficient of variation in the site-level inclusion probabilities, where the coefficient of variation is defined as the standard deviation divided by the mean. Because the mean probability of inclusion in the population

[7] The covariance between any two variables x and y (σ_{xy}) in a discrete population can be expressed as $\sigma_{xy} = \frac{1}{N}\sum_{j=1}^{N}(x_j - \mu_x)(y_j - \mu_y)$ or $\sigma_{xy} = (\frac{1}{N}\sum_{s=1}^{K} x_j y_j) - \mu_x\mu_y$, where μ_x and μ_y are defined as the population means of x and y, respectively.

[8] To show why this is true, note that $\frac{J}{K} = \frac{1}{K}\sum_{s=1}^{K} I_s$. Taking the expectation of both sides over an infinite number of replications of the site-selection process, $E(\frac{J}{K}) = E(\frac{1}{K}\sum_{s=1}^{K} I_s) = \frac{1}{K}\sum_{s=1}^{K} E(I_s) = \frac{1}{K}\sum_{s=1}^{K} P_s = \mu_P$. However, $\frac{J}{K}$ is a constant. Therefore, $\frac{J}{K} = \mu_P$.

μ_P is equal to J/K, the bias in the pooled impact estimator can be expressed as a function of the coefficient of variation in the site inclusion probabilities (cv_P):

$$Bias = \rho_{\Delta P}\sigma_\Delta cv_P + \rho_{bP}\sigma_b cv_P + \mu_b \tag{11}$$

Equation (11) effectively decomposes the bias in the pooled impact estimator into three terms. The first term ($\rho_{\Delta P}\sigma_\Delta cv_P$) is the external validity bias attributable to selecting a potentially unrepresentative sample of sites, but weighting them equally in the analysis. The last term is the internal validity bias attributable to using a biased impact estimator in some or all sites ($\mu_b \neq 0$). The middle term ($\rho_{b,P}\sigma_b cv_P$) is the bias that can result from the interaction between internal validity bias and the inclusion probabilities. In particular, the pooled impact estimate can be biased if the inclusion probabilities vary across sites ($cv_P > 0$), the internal validity bias in the impact estimates vary across sites ($\sigma_b > 0$), and the site inclusion probabilities are correlated with the internal validity bias in estimating the impact in individual sites ($\rho_{b,P} \neq 0$).

From this point forward, to focus on the external validity bias that arises from selecting sites purposively, we assume that the internal validity bias for each site is zero ($b_s = 0 \forall s$, which implies that $\mu_b = 0$ and $\sigma_b = 0$). Under this assumption, we obtain a simple expression for the external validity bias in multisite impact evaluations that select sites purposively but weight them equally in the analysis:

$$Bias_X = \rho_{\Delta P}\sigma_\Delta cv_P \tag{12}$$

Equation (12) shows that the external validity bias from purposive site selection depends on three factors: the variance of impacts across sites in the population of interest (σ_Δ), the coefficient of variation in inclusion probabilities across sites in the population (cv_P), and the correlation between site-level impacts and the site inclusion probabilities in the population ($\rho_{\Delta,P}$). If all three of these factors are nonzero, then the external validity bias from purposive site selection will be nonzero, and the magnitude of the bias will depend on the magnitude of the three factors. However, if any of the three factors equals zero, the bias will be zero. In other words, the external validity bias from purposive site selection will be zero if (1) the impact is the same in all sites, (2) the probability of being included in the sample is the same in all sites (i.e., as if the sample were a simple random sample), or (3) impacts and site inclusion probabilities vary across sites in the population, but they are uncorrelated with each other—that is, the site inclusion process does not favor sites with particularly large or small impacts.

Interestingly, a parallel expression for bias has been derived in the survey nonresponse context, where the set of respondents may not be representative of the full population. Brick and Jones (2008) express the bias in the mean of an outcome Y as the product of the coefficient of variation of the probabilities of response, the standard deviation of Y, and the correlation between the response probability and Y (see equation 2 in their paper).

One factor that does not affect the external validity bias is the average impact across sites in the population of interest. Although the variance of the site-level impacts appears in equation (12), the mean of the site-level impacts does not.

In addition, increasing the number of sites in the evaluation does not necessarily reduce the bias, as we might expect. At the extreme, the external validity bias equals 0 when all sites in the population are included in the sample. However, when the study includes a small share of all sites in the population, and all the site inclusion probabilities are less than 1, increasing the number of sites in the sample will not necessarily reduce the external validity bias. For example, if all the site inclusion probabilities are increased by a constant multiplicative factor without increasing

any probabilities above their limit of 1, it can be shown that the sample size will increase by the same factor and the external validity bias will be unaffected (proof available upon request). However, there is no guarantee that the site inclusion probabilities will increase by a constant multiplicative factor if an evaluation increases the number of sites to be included: It will depend on how the site recruiting process is changed to generate a larger sample of sites, and how those changes affect the terms in equation (12). Therefore, there is no necessary relationship between the number of sites included in the evaluation and the magnitude of the external validity bias.

Magnitude of the Bias

While investigating the components of the external validity bias is helpful, knowing the formula for the external bias does not yield any insights into how large the bias from purposive site selection is likely to be—either in the average study or in particular studies. Many papers have provided empirical evidence on the magnitude of a different type of bias—internal validity bias (selection bias) resulting from study designs based on nonexperimental comparison groups (e.g., Bloom, Michalopoulos, & Hill, 2005; Cook, Shadish, & Wong, 2008; Fraker & Maynard, 1987; Glazerman, Levy, & Myers, 2003; LaLonde, 1986). However, to the best of our knowledge, no published papers have provided empirical evidence on the magnitude of external validity bias resulting from purposive site selection.

It is reasonable to ask whether estimates of the external validity bias could be constructed using existing evidence from multisite impact evaluations. Evidence from (the small number of) studies that selected sites randomly could be used to produce unbiased estimates of the variation in impacts across sites in the population (σ_Δ). However, because these studies selected sites randomly and not purposively, they cannot help us to estimate the other two parameters in the bias formula—the variation in the site inclusion probabilities when sites are selected purposively (cv_P) or the correlation between site inclusion probabilities and site-level impacts ($\rho_{\Delta,P}$). Studies based on purposive samples might allow researchers to estimate all three parameters, but these estimates may be biased. Site inclusion probabilities can be estimated if data are available on site-level characteristics for both the purposively selected sites and the broader pool of sites from which included sites were drawn (e.g., by regressing whether or not the site was included in the study—1 for yes and 0 for no—on a set of observed site-level characteristics). However, if some of the factors that influence site inclusion are not observed, estimates of the variation in site inclusion probabilities (cv_P) could be biased. In addition, although purposive samples can be used to estimate the variation in impacts across sites (σ_Δ), these estimates could be biased because the sample is not representative of the population of interest (e.g., if purposively selected sites are more homogeneous than the population as a whole).

In summary, the amount of external validity bias that results from purposive site selection is an empirical question for which we lack empirical evidence. Just as researchers 25 years ago had no evidence on the magnitude of the internal validity bias that would result from a nonexperimental comparison group design, researchers today have no evidence on the consequences of beginning their next multisite impact evaluation by selecting a purposive or convenience sample of sites.

CONCLUDING THOUGHTS AND SUGGESTIONS FOR ADDITIONAL RESEARCH

The lack of any empirical evidence on the typical magnitude of external validity bias from purposive site selection leaves plenty of room for disagreement about how this

potential problem should be handled. Some may argue that in the absence of evidence suggesting the bias is large, researchers should focus their attention on other methodological issues. Others may argue that because the bias could potentially be large, we should design and implement future studies with the specific goal of reducing the bias. At this point, we propose a middle ground and recommend more research in this area. In particular, we make two suggestions for future research.

First, we recommend that future studies focus on producing empirical evidence on the magnitude of the external validity bias that results from purposive site selection. To estimate this bias, researchers could follow an approach that is similar to the studies that have estimated the magnitude of internal validity bias in nonexperimental studies. These studies compare the actual impact estimates from an experiment to a reasonable prediction of what the impact estimates would be if an experiment had not been possible, and researchers had instead selected a nonexperimental comparison group.

To estimate the magnitude of the external validity bias from purposive site selection, researchers could begin with data from one or more studies that were based on a representative sample, compute externally valid impact estimates from the sample, and treat the estimates as the gold standard from the perspective of external validity. Then the study could attempt to predict which sites would have been selected had the study selected sites purposively, estimate impacts for these sites, and compare the estimates to the gold standard.

The challenge with taking this approach is predicting which sites would be included in the study if sites had been selected purposively. Fortunately, studies that select sites purposively may provide some evidence on the types of sites that tend to be included in studies when purposive site selection is used and sites may or may not agree to participate. Therefore, one approach to estimating the external validity bias would use data from studies that selected a representative sample of sites and studies that selected sites purposively. In further work on this topic, we plan to pursue this approach.

Second, if future research identifies evaluation settings where the external validity bias due to purposive site selection is (or could be) large, we would encourage future studies to focus on exploring and testing possible approaches to reducing the bias. The most obvious solution is to select sites randomly. If researchers had good reasons to favor certain types of sites over others based on observed characteristics, the study could select a stratified random sample of sites. While one might reasonably doubt the benefits of random site selection in settings where the take-up rate is low—that is, when many selected sites will choose not to participate—whether and how much random site selection can reduce the bias when participation decisions are nonrandom is an empirical question that warrants additional research.

Another approach worth exploring involves devoting additional effort and resources to recruiting sites that initially resist being included in the evaluation, and comparing the impacts in these sites with those in the sites that initially agreed to participate. This suggestion is motivated by studies based on surveys that have invested additional time and resources to interview a random sample of initial nonrespondents with the goal of reducing survey nonresponse bias.[9] We believe this approach would be worth testing in the context of site recruitment, perhaps initially on a small scale, to assess the costs of this approach and the benefits in terms of potential bias reduction.

[9] For example, this approach was used in the Moving to Opportunities experiment (see Orr et al., 2003) and also in a study to describe the outcomes of welfare leavers in Iowa (see Kauff, Fraker, & Milliner-Waddell, 2002; Kauff, Olsen, & Fraker, 2002).

Finally, we recommend exploring and testing the use of observed site-level characteristics to reduce the external validity bias at the analysis stage. In principle, if the site inclusion probabilities were known, sites could be weighted in the analysis to completely eliminate the bias (proof available upon request). In practice, when the site inclusion probabilities are unknown, one may be able to reduce the bias by reweighting the sample of sites to more closely match the population of interest in terms of observed site-level characteristics. This approach has been used extensively and rigorously examined in literatures on (1) reweighting comparison groups to make them more comparable to program participants, via propensity score methods (e.g., Rubin, 2001; Stuart, 2010) and (2) reweighting survey respondents to make them more comparable to the population from which the survey sample was selected (e.g., Brick & Jones, 2008; Little, 1986; Oh & Scheuren, 1983). Reweighting approaches have more recently been applied to improve the external validity of impact estimates from unrepresentative samples (e.g., Haneuse et al., 2009; Pan & Schaubel, 2009; Stuart et al., 2011). We believe this approach is worthy of additional research and testing to assess its ability to reduce external validity bias resulting from purposive site selection.

To recap, this paper lays the groundwork for future research on the generalizability of study findings based on purposive samples of sites, as opposed to random probability samples of sites that formally generalize to a known population. The paper begins with the premise that effectiveness studies conducted to inform policy decisions should estimate the effects of the program or policy for one or more populations of interest to policymakers. We show formally that statistical bias, in the usual sense of the term, will result if impacts vary across sites and the process by which sites are included in the evaluation systematically favors sites with impacts that are either larger or smaller than the average impact for the population of interest. In the paper, we provide a formula for the total bias when estimating a population average treatment effect, isolate the external validity bias due to purposive site selection, and identify the parameters of the site inclusion process that contribute to the external validity bias. Future research is needed to assess the magnitude of the external validity bias resulting from purposive site selection and to test different options for reducing that bias.

ROBERT B. OLSEN is Principal Scientist, Abt Associates, 4550 Montgomery Avenue, Suite 800 North, Bethesda, MD 20814.

LARRY L. ORR is Associate, Institute for Policy Studies, Johns Hopkins University, 4402 Leland Street, Chevy Chase, MD 20815.

STEPHEN H. BELL is Principal Scientist and Senior Fellow, Abt Associates, 4550 Montgomery Avenue, Suite 800 North, Bethesda, MD 20814.

ELIZABETH A. STUART is Associate Professor, Johns Hopkins Bloomberg School of Public Health, 624 N Broadway, 8th Floor, Baltimore, MD 21205.

ACKNOWLEDGMENTS

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305D100041 to Abt Associates Inc. Dr. Stuart's time was partially supported by National Institute of Mental Health Grant Award K25MH083846. The opinions expressed are those of the authors and do not represent views of the Institute of Education Sciences, the U.S. Department of Education, or the National Institute of Mental

Health. We would like to thank Jacob Klerman, Steve Kennedy, Bill Rhodes, Beth Gamse, Cris Price, Howard Rolston, three anonymous reviewers and participants in the Journal Author Support Group at Abt Associates for their helpful comments. The authors alone are responsible for any errors in the paper. Please send questions and comments to Rob Olsen at Rob_Olsen@abtassoc.com.

REFERENCES

Bloom, H. S., Michalopoulos, C., & Hill, C. J. (2005). Using experiments to assess nonexperimental comparison-group methods for measuring program effects. In H. S. Bloom (Ed.), Learning more from social experiments (pp. 173–235). New York: Russell Sage Foundation.

Bos, J. M., & Fellerath, V. (1997). Final report on Ohio's welfare initiative to improve school attendance among teenage parents: Ohio's Learning, Earning, and Parenting Program. New York: Manpower Demonstrative Research Corporation.

Brick, J. M., & Jones, M. E. (2008). Propensity to respond and nonresponse bias. Metron—International Journal of Statistics, 6, 51–73.

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.

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.

Corson, W., Decker, P., Dunstan, S., & Kerachsky, S. (1991). Pennsylvania reemployment bonus demonstration. Princeton, NJ: Mathematica Policy Research.

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

Freedman, S., Friedlander, D., Hamilton, G., Rock, J., Mitchell, M., Nudelman, J., Schweder, A., & Storto, L. (2000). Evaluating alternative welfare-to-work approaches: Two-year impacts for eleven programs. Washington, DC: U.S. Department of Health and Human Services and the U.S. Department of Education.

Glazerman, S., Levy, D. M., & Myers, D. (2003). Nonexperimental versus experimental estimates of earnings impacts. The Annals of the American Academy of Political and Social Science, 589, 63–93.

Gleason, P., Clark, M., Tuttle, C. C., & Dwoyer, E. (2010). The evaluation of charter school impacts: Final report NCEE 2010–4029. Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Greenberg, D., & Shroder, M. (2004). The digest of social experiments. Washington, DC: The Urban Institute Press.

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

Hamilton, G., & Brock, T. (1994). Early lessons from seven sites. Washington, DC: U.S. Department of Health and Human Services and the U.S. Department of Education.

Haneuse, S., Schildcrout, J., Crane, P., Sonnen, J., Breitner, J., & Larson, E. (2009). Adjustment for selection bias in observational studies with application to the analysis of autopsy data. Neuroepidemiology, 32, 229–239.

Heckman, J. J. (1976). The common structure of statistical models of truncation, sample selection, and limited independent variables. Annals of Economic and Social Measurement, 5, 120–137.

Imai, K., King, G., & Stuart, E. A. (2008). Misunderstandings between experimentalists and observationalists about causal inference. Journal of the Royal Statistical Society, Series A, 171, 481–502.

Kauff, J., Fraker, T., & Milliner-Waddell, J. (2002). Iowa families that left TANF: How are they faring two years later? Washington, DC: Mathematica Policy Research, Inc.

Kauff, J., Olsen, R., & Fraker, T. (2002). Nonrespondents and nonresponse bias: Evidence from a survey of former welfare recipients in Iowa. Washington, DC: Mathematica Policy Research, Inc.

Kemple, J. J., Friedlander, D., & Fellerath, V. (1995). Florida's project independence: Benefits, costs, and two-year impacts of Florida's JOBS program. New York: Manpower Demonstrative Research Corporation.

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

Little, R. J. A. (1986). Survey nonresponse adjustments. International Statistical Review, 54, 139–157.

Nisar, H. (2010). Do charter schools improve student achievement? Unpublished Working Paper. Retrieved June 21, 2012, from http://www.ssc.wisc.edu/~scholz/Seminar/Charter_School_MPS.pdf.

Oh, H. L., & Scheuren, F. J. (1983). Weighting adjustment for unit nonresponse. Incomplete Data in Sample Surveys, 2, 143–184.

Orr, L., Feins, J. D. Jacob, R., Beecroft, E., Sanbonmatsu, L., Katz, L. F., Liebman, J. B., & Kling, J. R. (2003). Moving to opportunity interim impacts evaluation: Final report. Washington, DC: U.S. Department of Housing and Urban Development.

Pan, Q., & Schaubel, D. E. (2009). Evaluating bias correction in weighted proportional hazards regression. Lifetime Data Analysis, 15, 120–146.

Puma, M. J., Burstein, N. R., Merrill, K., & Silverstein, G. (1990). Evaluation of the Food Stamp Employment and Training Program. Alexandria, VA: U.S. Department of Agriculture, Food and Nutrition Service, Office of Analysis and Evaluation.

Puma, M., Bell, S., Cook, R., & Heid, C. (2010). Head Start Impact Study final report. Washington, DC: Administration for Children and Families, U.S. Department of Health and Human Services.

Riccio, J., Friedlander, D., & Freedman, S. (1994). GAIN: Benefits, costs, and three-year impacts of a welfare-to-work program. New York: MDRC.

Rubin, D. B. (1976). Inference and missing data. Biometrika, 63, 581–592.

Rubin, D. B. (2001). Using propensity scores to help design observational studies: Application to the tobacco litigation. Health Services & Outcomes Research Methodology, 2, 169–188.

Seftor, N. S., Mamun, A., & Schirm, A. (2009). The impacts of regular Upward Bound on postsecondary outcomes 7–9 years after scheduled high school graduation: Final report. Washington, DC: Mathematica Policy Research.

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

Stapleton, D., Bell, S., Wittenberg, D., Sokol, S., & McInnis, D. (2010). BOND final design report. Bethesda, MD: Abt Associates Inc.

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. Statistical Science, 25, 1–21.

Stuart, E. A., Cole, S. R., Bradshaw, C. P., & Leaf, P. J. (2011). The use of propensity scores to assess the generalizability of results from randomized trials. The Journal of the Royal Statistical Society, Series A, 174, 368–386.

Watts, H. W., Peck, J. K., & Taussig, M. (1977). Site selection, representativeness of the sample, and possible attrition bias. In H. W. Watts & A. Rees (Eds.), The New Jersey income maintenance experiment (Vol. 3, pp. 441–66). New York: Academic Press.

Weisberg, H., Hayden, V., & Pontes, V. (2009). Selection criteria and generalizability within the counterfactual framework: Explaining the paradox of antidepressant-induced suicidality? Clinical Trials, 6, 109–118.