Article

Evaluation Review
1-31

DOI: 10.1177/0193841X13516324
erx.sagepub.com
SAGE

Stratified Sampling Using Cluster Analysis: A Sample Selection Strategy for Improved Generalizations From Experiments

Elizabeth Tipton[1]

Abstract

Background: An important question in the design of experiments is how to ensure that the findings from the experiment are generalizable to a larger population. This concern with generalizability is particularly important when treatment effects are heterogeneous and when selecting units into the experiment using random sampling is not possible—two conditions commonly met in large-scale educational experiments. **Method:** This article introduces a model-based balanced-sampling framework for improving generalizations, with a focus on developing methods that are robust to model misspecification. Additionally, the article provides a new method for sample selection within this framework: First units in an inference population are divided into relatively homogenous strata using cluster analysis, and

[1] Department of Human Development, Teachers College, Columbia University, NY, USA

Corresponding Author:
Elizabeth Tipton, Department of Human Development, Teachers College, Columbia University, 525 W 120th St, Box 118, NY 10027, USA.
Email: tipton@tc.columbia.edu

then the sample is selected using distance rankings. **Result:** In order to demonstrate and evaluate the method, a reanalysis of a completed experiment is conducted. This example compares samples selected using the new method with the actual sample used in the experiment. Results indicate that even under high nonresponse, balance is better on most covariates and that fewer coverage errors result. **Conclusion:** The article concludes with a discussion of additional benefits and limitations of the method.

Keywords
cluster analysis, experimental design, external validity, model-based sampling, stratified sampling, treatment effect heterogeneity

In the social, educational, and medical sciences, evaluations of interventions are typically conducted using randomized experiments. Randomized experiments are preferred since they have high internal validity, ensuring that the treatment effect estimated within the experiment is the causal effect of the treatment. This random assignment to treatment conditions, however, does not help when generalizations about the effect of the treatment for units *not* in the experiment are desired. Since experiments very rarely select units using probability sampling from a well-defined population (Shadish, Cook, and Campbell 2002), any generalization must typically be based on qualitative judgments regarding how similar a particular population of interest is to the composition of units in the experiment (Cornfield and Tukey 1956).

Recently, statisticians have begun developing new methods for improving generalizations from completed experiments. Stuart et al. (2011) introduced propensity score-based methods for quantitatively evaluating the degree of similarity between a population and an experimental sample, while Hedges and O'Muircheartaigh (2011) developed a method for adjusting the estimate and standard errors to account for these differences using a propensity score poststratification estimator. Tipton (2013) further developed the assumptions necessary for causal generalization using propensity score methods and properties of the poststratification estimator. Furthermore, Tipton showed that these propensity score-based methods perform best when there is no coverage error. Coverage errors arise when particular segments of the population do not have relevant comparison units "like" them in the sample used in the experiment.

In contrast to these retrospective approaches, this article provides a new framework and method for sample selection in experiments that improve causal generalizations prospectively. The goal of this framework is for the sample selected for inclusion in the experiment to be compositionally similar to the inference population on a variety of important covariates that possibly explain variation in potential treatment effects. The approach can be used broadly and does not require (nor preclude) random sampling. To achieve these goals, the method uses cluster analysis techniques to classify the population into nearly homogenous strata and then provides a simple distance-based approach for selecting units within each stratum into the experiment. The method is similar to that proposed by Tipton et al. (2014) but differs in that here eligibility criteria differentiating between the inference population and the units eligible to be in the study are not required.

Since in most practical cases the units selected into an experiment are in fact clusters or aggregates of individuals—for example, schools or school districts—the method developed here is a version of a stratified cluster sampling design. The goal is to first divide the clusters (e.g. schools) into strata using cluster analysis methods and then to select clusters (e.g. schools) within each of these strata into the experiment. Note that the language used here can be confusing, since the word *cluster* is used differently in the field of cluster analysis than in the fields of experimental design and survey sampling. In order to clearly differentiate, we use the word cluster throughout to mean a group of aggregated individuals, for example, a school or school district. This follows practice common in the design and analysis of large-scale experiments. We then refer to the groups of these clusters that are created using cluster analysis methods as *strata*, since these will be used for the creation of a stratified sampling plan for generalization.

Overall, the article is organized as follows. In the first section, we frame the problem of sample selection in the model-based sampling literature and introduce the goals for our approach. In the second section, we develop a stratified sample selection method using cluster analysis to meet these goals. In the third section, we apply and evaluate this method using an example. Finally, in the last section, the article concludes with a discussion of additional benefits and extensions to the method.

Generalizations From Experiments

The Role of Models in Generalization

In developing a method for sample selection, it is helpful to begin by reviewing what would happen in an ideal study aimed at causal generalization. In this ideal study, first a well-defined inference population P of size N

would be carefully enumerated and defined. For example, in a large-scale educational experiment, this might be a list of the 65,134 regular middle schools in the United States obtained from the Common Core of Data (National Center for Education Statistics). Second, a sample S of n sites would be randomly selected from this population. For example, in an educational experiment, the sample would typically include between 20 and 60 schools or districts, depending on the study design (Spybrook 2012). Third, within these n schools, units would be randomly assigned to treatment conditions. Depending upon the design of the study, this randomization might happen at the cluster level (e.g., schools) or at a lower level (e.g., classrooms or students). This dual randomization would ensure that both the site and the treatment selection processes were *ignorable* or *noninformative*, where an ignorable selection process is one in which unobserved covariates have no effect on the conditional distribution of outcomes given the observed covariates (Rubin 1976; Smith and Sugden 1988). The result would be an "experiment within a survey," which would clearly enable causal generalizations (Smith and Sugden 1988; Shadish, Cook, and Campbell 2002).

In practice, this dual randomization procedure is generally infeasible (Bloom 2005; Rubin 1974; Shadish, Cook, and Campbell 2002). In fact, Olsen et al. (2013) found that random site selection was implemented in only 7 of the 273 experiments reported in the *Digest of Social Experiments* (Greenberg and Schroder 2004). Instead, study designers and analysts often choose only one level of randomization, resulting in either a probability survey or an experiment (Fienberg and Tanur 1987; Imai, King, and Stuart 2008). In experiments, while treatment is assigned randomly, the typical practice is to select a convenience sample of n sites, where here by "convenience" we mean without clear reference to a well-defined population (Shadish, Cook, and Campbell 2002).

Given the infeasibility of random site selection, in experiments causal generalizations are typically made through the use of statistical models. To see why models are necessary, note that we can decompose a population average treatment effect (PATE) as follows (Imai, King, and Stuart 2008),

$$\mathrm{PATE} = (n/N)\ \mathrm{SATE}\ +\ (1 - n/N)\ \mathrm{NATE},$$

where the sample average treatment effect (SATE) and nonsample average treatment effect (NATE) are the average treatment effects for those in the sample and not in the sample, respectively, the sample includes n units and there are N total units in the population. While we can estimate the SATE directly from the n units in the experiment, estimating the average treatment effect for units *not* in the experiment (NATE) is difficult. If the sites were selected using probability sampling, the NATE could be estimated by taking into account the selection probabilities for the units in the sample. Since the sites in an experiment are typically not selected randomly, the NATE must be predicted based upon a model relating the sample to the population.

In experiments, when generalizations beyond the sample of *n* sites are of interest, the results are typically analyzed using random effects models (Kirk 1995; Raudenbush 1993); these multilevel models are generalizations of the analysis of variance models typically used to analyze single-site experiments. These super-population models make generalizations through use of random effects (with an assumed distribution), not based upon selection or random assignment probabilities. Specifically, two super-population models are common: the hierarchical model (for cluster-randomized designs) or the random block model (for multisite trials). Whether estimated using traditional hierarchical linear model methods (Raudenbush and Bryk 2002) or the Neyman model (Schochet 2013), the average treatment effect estimated is therefore considered generalizable to schools or districts that are "similar" to those in the study. These "bottom-up" generalizations are difficult since what is meant by similar is typically vaguely defined.

Recent work in the analysis of experiments has focused on how to improve generalizations within this super-population framework. The methods begin by carefully defining an inference population of interest, using a population frame like the Common Core of Data or a state longitudinal data system (e.g., Stuart et al. 2011; Tipton 2013). An important feature of these population frames is that they both enumerate all units in the population and include a rich set of covariate information on these units. Next, the *N* schools in the population and the *n* schools in the sample used in the experiment are compared on a set of *p* covariates using a propensity score (Rosenbaum and Rubin 1983). These covariates are selected so as to meet a *sample selection ignorability condition*; here, ignorability is met if this set includes all covariates associated with both the site selection process and the treatment effect variability (Stuart et al. 2011; Tipton 2013). Propensity score methods are then used to reweight the sample and the population so that the two groups are *balanced* on this set of *p* covariates, where balanced means that the means (and higher order moments) of these covariates are similar for the two groups. These reweighting estimators include subclassification or poststratification estimators used in combination with the multilevel models given earlier (e.g., Hedges and O'Muircheartaigh 2011). The result is an estimate of the average treatment effect given for a well-defined population based upon a super-population model.

Retrospective generalizations, like those given earlier, are helpful in that they shift generalizations from vaguely defined to well-defined populations. However, as Tipton (2013) shows, sometimes the effectiveness of these methods is limited, particularly when there are coverage errors. Coverage errors arise when there exist segments of the population for whom there are no similar units in the experiment; this, we argue, is the lasting effect of the bottom-up generalization approach. In this article, we focus instead on developing a method for site selection that makes these generalizations "top-down."

Instead of carefully defining the inference population *after the study is completed*, our goal here is to *begin the study* with a well-defined inference population and then design a site-selection process that makes model-based generalizations possible.

Balanced Sampling

Just as in the ideal study, the goal here is to begin by first carefully defining and enumerating an inference population P of size N and then to select the n units in the sample S strategically, using a more formal sample selection plan. The goal is for the sample selection process to be *noninformative* or *ignorable* (Rubin 1976) by which we mean that the resulting sample and population are compositionally similar on the set of covariates that explain treatment effect variability (Stuart et al. 2011; Tipton 2013).

The approach we develop here builds on theory and results found in *model-based sampling* in the sample survey literature (Valliant, Dorfman, and Royall 2000). Model-based sampling is a purposive sampling alternative to design-based random sampling methods; while random selection can be used within the model-based framework, it is not required. The model-based sampling approach, while not as commonly used in survey sampling, has much in common with the model-based super-population approach commonly used to analyze experiments (Fienberg and Tanur 1987; Rao 2005). As such, this method will allow inferences from a sample to super-population using the same random effects approach currently used in the analysis of large-scale experiments; the key difference here is that now the definition of the super-population will be clearly and carefully defined.

In order to develop this strategy, we first define the causal effects of interest in the potential outcomes framework. For each unit in the population P, let $W = 1$ if a unit is assigned to the treatment condition. Then assume that for each unit, there exists two potential outcomes,

$Y(1) = Y(W = 1)$ and $Y(0) = Y(W = 0)$, where $Y(1)$ is the unit's potential outcome under treatment and $Y(0)$ is the unit's potential outcome under some specified alternative condition. Now for each unit in the population P, let the *potential treatment effect* $\Delta = Y(1) - Y(0)$. Note that as a result of the Fundamental Problem of Causal Inference, both potential outcomes—and by extension the potential treatment effect—can never be observed for a particular unit (Holland 1986). However, the goal of an experiment is generally to estimate the PATE $\tau^P = \mathrm{E}[\Delta]$, where the expectation is across all units in the inference population P. Finally, let $Z = 1$ if a unit in the population P is selected into the sample S. Since only units in the sample ($Z = 1$) are in the experiment, we focus here on causal impact estimators that are based only on the n units in S.

An important question is under what conditions an estimator T that is unbiased for the SATE, $\tau^S = \mathrm{E}(\Delta|\mathrm{Z}{=}1)$, is unbiased for the PATE, $\tau^P = \mathrm{E}(\Delta)$. Stuart et al. (2011) and Tipton (2013) show that bias arises in relation to covariates that explain variation in potential treatment effects. Since we do not know, a priori, what these covariates are (since we have yet to conduct the experiment), this requires us to propose a model. We might posit a simple model

$$\Delta = \beta_0 + \beta_1 X, \tag{1}$$

where X is a single covariate with known values for all units in the population. For example, based on theoretical or previous empirical findings, we may believe that the effect of a school-based reading intervention (Δ) linearly increases or decreases in relation to last year's school average reading test scores (X).

If in fact the potential treatment effects do vary in relation to X, then by selecting the sample so that the average value of X in the sample and population is the same (i.e., $\mathrm{E}(X|Z = 1) = \mathrm{E}(X)$), an estimator T of the SATE τ^S would be model unbiased for the PATE τ^P. We call this a *balanced sample* since the sample and population are balanced on the covariate X. Importantly, this idea of balance is similar to the goal of balance found in retrospective methods; the key difference is that retrospective balance is achieved through reweighting, while here balance is achieved through the strategic selection of the sample.

The idea of a balanced sample is in many ways similar to the naive sense of a "representative sample" (Royall and Herson 1973), where here by representative we mean that the sample is a like a "miniature" of the population (Kruskal and Mosteller 1980). Importantly, while random sampling will result in a balanced sample *on average*, it is not the only or best method for achieving such balance, particularly when the sample is small. Developing a method that performs well in small samples is of particular concern here since in most cluster-randomized or multisite studies in education, the number of higher level units (e.g. schools or districts) tends to be between 20 and 60 (Spybrook 2012).

Bias-Robust Balanced Sampling

Based on the model 1.1, we might propose to develop a sample selection plan that results in a sample that is *balanced* for the covariate X (i.e., $E(X|Z = 1) = E(X)$). This necessarily leads us to ask: what happens if our model is wrong? For example, suppose that potential treatment effects actually vary in relation to the model

$$\Delta = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 W. \tag{2}$$

For example, it may be that the potential treatment effects vary nonlinearly in relation to school average pretest scores (X) and vary also in relation to the proportion of the school that is minority (W). Now the bias of an estimator T for τ^P can be written as

$$\mathrm{E}(T) - \tau^P = \beta_2\left[E\left(X^2|Z = 1\right) - E\left(X^2\right)\right] + \beta_3[E(W|Z = 1) - E(W)].$$

Clearly, T is only unbiased for τ^P if $E(X^2|Z = 1) = E(X^2)$ and $E(W|Z = 1) = E(W)$. This means that a sample that is balanced on X is only adequate if our former model (1) holds but not under this new model (2).

In model-based sampling, the goal is to select the sample so that the method is bias robust, where bias robust is shorthand for "bias-robust-against-model-failure" (Valliant et al. 2000). In this framework, in causal generalization this requires us to first propose multiple models relating the potential treatment effects to possible moderators, and second, to develop a sample selection plan that guards against selection of the wrong model. Here the key tool guarding against model failure is the selection of a balanced sample. Subsequently, we define this more formally.

Definition: balanced sample (of order R*).* Let $\mathbf{X} = \{X_1, \ldots, X_p\}$ be a set of covariates that might explain variation in the potential treatment effects Δ. A sample S is said to be a *balanced sample (of order R)* if for each covariate X_h (for $h = 1 \ldots p$) and for each power $r = 1 \ldots R$, $E(X_h^r|Z = 1) = E(X_h^r)$, where $E(X_h^r|Z = 1)$ is the average value of X_h^r in the sample S and $E(X_h^r)$ is the average value in the population P

When a balanced sample (of order R) is selected on p covariates, it is easy to show that an unbiased estimator T of the SATE τ^S is unbiased for the PATE τ^P if the true model is of the form

$$\Delta = \delta_0\beta_0 + \sum_{h=1}^{p}\sum_{r=1}^{R}\delta_{hr}\beta_{hr}X_h^r, \tag{3}$$

where for $h = 1 \ldots p$, β_{hr} is the regression coefficient associated with the covariate $X_h{}^r$ and $\delta_{hr} = 1$ if the coefficient β_{hr} is included in the model and zero otherwise. For example, if a sample is balanced (of Order 2), then T is unbiased when the true relationship is linear

$$\Delta = \beta_0 + \beta_{11}X_1 + \beta_{21}X_2 + \ldots + \beta_{\mathrm{p}1}X_{\mathrm{p}}, \tag{4}$$

quadratic,

$$\Delta = \beta_0 + \beta_{11}X_1 + \beta_{21}X_2 + \ldots + \beta_{\mathrm{p}1}X_{\mathrm{p}} + \beta_{12}X_1^2 + \beta_{22}X_2^2 + \ldots + \beta_{\mathrm{p}}X_{\mathrm{p}}^2, \tag{5}$$

when it includes only a subset of the covariates,

$$\Delta = \beta_0 + \beta_{11}X_1 + \beta_{21}X_2, \tag{6}$$

or any other combination of the models. Since we cannot and do not know the true model, in the language of model-based sampling, the goal is to select a sample using a bias-robust *strategy*. In causal generalization, this is a strategy that leads to an unbiased estimate of $\tau^P = E(\Delta)$ under a variety of possible models.

A sample selection strategy can be bias robust in two ways. First, the sample becomes more bias robust as the dimension p of X increases. This is because in practice when balance is achieved on a wide variety of covariates it is often approximately achieved on other covariates, including those that may have been omitted from the model (Stuart 2010; Smith and Sugden 1988; Royall and Herson 1973; Brewer 1999; Rubin and Thomas 1996). Second, for a fixed set of covariates X, a sample becomes more bias robust as the order of R increases. For example, a sample which is balanced on not just the first moments but also the second moments of X—a balanced sample of Order 2—will lead to an unbiased estimate whether the true model is linear or quadratic (Valliant, Dorfman, and Royall 2000).

The idea that the sample should be selected using a bias-robust strategy with the goal of achieving balance on both the means and the higher moments of multiple covariates is familiar—it is exactly the post hoc approach used in the propensity score literature (Rosenbaum and Rubin

1983). Propensity scores are commonly used to achieve balance under a variety of models when random processes are not possible (e.g., quasi-experiments; post hoc generalizations from experiments) or when they fail (e.g., attrition in experiments; nonresponse in surveys). While the method we develop here does not use a propensity score approach, the goals of the procedure are the same—to replace a random process with a model and to design the study in such a way that the results do not depend heavily on any one model.

Once the sample of *n* units is selected, the robustness of the sample to model misspecification can be evaluated by comparing the *R* moments of the *p* covariates under study in the realized sample and population. Ideally, these differences will be small, enabling the use of the simple multilevel random effects estimator commonly used in experimental analysis. If differences remain, post hoc methods like regression adjustment or those introduced by Hedges and O'Muircheartaigh (2011) and Tipton (2013) can be used to decrease bias. Importantly, since the sample was selected with an eye toward balance, however, the achieved sample and population are likely to be more similar than if a bias-robust selection method had not been used. The fact that the sample and population are more similar will mean that there will be fewer coverage errors and that, if adjustment is needed, the cost in terms of variance inflation will be smaller (Tipton 2013).

Stratified Sampling as a Tool for Generalization

Defining a Stratified Estimator

Recall that in the balanced-sampling framework, the goal is simply to select the sample so that it is like a "miniature" of a well-defined population. To do so, we begin by positing a model explaining variation in potential treatment effects (e.g., Model 1.1), and then propose alternative models (a bias-robust strategy). To this end, we need a method for sample selection that allows for balance on orders greater than one (i.e., not only on first but also on second or higher moments) and enables *X* to include a large and varied set of covariates. While many possible methods exist, in model-based sampling, the simplest method that achieves this goal is that of stratified sampling with proportional allocation (Valliant, Dorfman, and Royall 2000).

Stratified sampling is already widely used in both survey sampling and in large-scale experiments. In probability sampling, strata are used to reduce the variance of an estimate; since the focus is on variance reduction (not bias), here it is common for the strata to be created on only one or two covariates (Lohr 1999). The use of one or two covariates for strata creation is also sometimes used for nonrandom site selection in experiments. For example, experiments sometimes attempt to include sites in rural and urban locations (i.e., urbanicity) or in regions throughout the country (e.g., northeast, southeast, midwest, west). Here while strata are created in order to improve generalizations, the method and framework for making these generalizations are typically informal. In contrast, in this article we propose to create strata with the goal of reducing bias through the inclusion of many covariates—hopefully all covariates that explain treatment effect heterogeneity—and in order to balance the sample and population on both means and higher order moments. In contrast to prevailing practice, our goal is to explicitly state, develop, and evaluate a bias-robust strategy for sample selection aimed at generalization.

If a simple Model (1) is of interest, in order to create a stratified sample with proportional allocation, three steps would be involved. First, the values of the covariate X would be divided into strata. If X is categorical with k categories, then k strata could be naturally created. If X is continuous, many possible strategies for strata creation could be used; one of the simplest is to define the $j = 1 \ \ldots \ k$ strata so that each contains an equal portion (i.e., $w_{pj} = N/k = N_j/N$) of the population units (Cochran 1968). Second, under *proportional allocation*, stratum j would be allocated $n_j = w_{pj} \times n$ units in the sample. This means that in stratum j, n_j/N_j units would need to be selected into the experiment. Third, in stratum j, the sample would need to be selected so that $E(X|Z = 1, j = j) = E(X|j = j)$, which is to say that the stratum-specific mean of X is the same in the population and the sample.

The main benefit of using a stratified sampling approach with proportional allocation is that the resulting sample is *self-weighting*. This means that the sample and population are *balanced* on the covariate X, since

$$E(X|Z = 1) = \Sigma\left(n_j/n\right)E(X|Z = 1,j = j) = \Sigma w_{pj}E(X|\, j = j) = E(X). \tag{7}$$

The fact that the sample is self-weighting means that no additional adjustments are needed and that the usual multilevel random effects model can be used for estimating the average treatment effect and making generalizations. Since the standard estimator can be used, this means that the sample selection process does not impact the power analysis (used to determine the sample size n). As a result, the issues of statistical power and generalization can be separated, which is logistically helpful in designing the experiment. It is also helpful that stratified sampling with proportional allocation is conceptually easy to understand and explain, making it appealing in the policy context where many of the results of large-scale experiments are used and interpreted.

When X is a single covariate, k strata can be easily created leading to a balanced sample. However, when the dimension of $\mathbf{X}$ is large and when $\mathbf{X}$ includes both categorical and continuous covariates, strata creation becomes more difficult. For example, if each covariate takes only two values, this would lead to 2^p unique strata, which can easily be larger than the total sample size n. For this reason, a method for stratification is needed that allows for the selection of balanced and bias-robust samples with fewer strata.

In the survey sampling literature, the problem of stratified sampling under a multivariate and continuous $\mathbf{X}$ has been addressed through the use of *cluster analysis*. These methods were pioneered by Day and Heeler (1971) and Golder and Yeomans (1973) who used cluster analysis to classify units in a pricing experiment and city wards in a study of seat belt use, respectively. Since 1980 cluster analysis methods have been used to create strata for the Current Population Survey (CPS) and other demographic surveys administered by the U.S. Census Bureau (Mansure and Reist 2010; Murphy 2008) as well as in surveys by census agencies around the world (e.g. Northern Ireland Statistics and Research Agency [2001] 2002). Additionally, in education research clustering methods have been used to classify schools (e.g. ETS 2008; Sleegers, Bergen, and Giesbers 1994; Lipson et al. 2004), and in the analysis of experimental data clustering methods have been developed to combine similar treatment groups (e.g. Cox and Spjotvoll 1982; Scott and Knott 1974; Tukey 1949) and to analyze subgroup average treatment effects in experiments (e.g., Peck 2005; Peck et al. 2012). Here, we propose to use cluster analysis to define strata for a stratified sampling design aimed at improving generalizations from experiments.

Cluster Analysis Method

The goal of cluster analysis is to divide units in the inference population into strata so that units in the same stratum are more similar than units in different strata. We focus here on the k-means partitioning method, though other methods are available (Everitt et al. 2011). The basic idea of k-means clustering is to create k strata and then to assign each unit to one stratum so that a measure of similarity is maximized. There are two main steps involved in a cluster analysis, and we briefly review these here.

Choosing a Distance Metric. In this article, we use cluster analysis to group units into strata that are as close to homogeneous as possible. This means that a measure of *distance* or *similarity* is required to define "close." There are two common distance measures that we argue are useful for our purposes, though others are available (Everitt et al. 2011). The decision to use one of these metrics over the other will largely depend on the type of covariates included in **X** and on information or assumptions regarding the importance of each covariate.

When all of the covariates in **X** are continuous, the distance metric most commonly used is the weighted Euclidean distance,

$$d^e_{ii'} = \sqrt{\sum\nolimits_{h=1}^{p} w_h (X_{ih} - X_{i'h})^2}, \tag{8}$$

where each covariate X_h has weight w_h, and X_{ih} and $X_{i'h}$ are the values of the hth covariate for units i and i'. One option for weights is to set $w_h = 1$ for all covariates h, in which case $d^e_{ii'}$ is the Euclidean distance; this gives the most weight to covariates with the largest variances. Alternatively, if there is no information regarding which covariate is a more important or better predictor of treatment effect heterogeneity, then the obvious solution is to use inverse-variance weights, where $w_h = 1/V(X_h)$. In this framework, the weighted covariates have a common variance of one, and each covariate contributes equally to the distance metric (though other weighting methods are possible).

Alternatively, when **X** includes both continuous covariates and categorical or dummy variables, any Euclidean-based measure does not perform as well. There are various solutions to this problem. One solution that is commonly used and easily implemented is to use a general similarity (distance) measure proposed by Gower (1971),

$$d^g_{ii'} = \sum\nolimits_{h=1}^{p} w_{ii'h} d_{ii'h} \Big/ \sum\nolimits_{h=1}^{p} w_{ii'h}, \tag{9}$$

where $d_{ii'h}$ is the similarity between units i and i' on the covariate X_h. Note that this measure of distance, $d_{ii'h}$, can differ for different variable types. For dummy and categorical variables, $d_{ii'h} = 1$ if the two units i and i' have the same value and 0 otherwise. For continuous variables, it is standard to use

$$d_{ii'h} = 1 - \frac{|X_{ih} - X_{i'h}|}{R_h}, \tag{10}$$

where |.| indicates absolute value and R_h is the range of observations for the covariate X_h. Using distance measures defined this way ensures that $d_{ii'h} \in [0,1]$ for covariates of all types. Additionally, this general distance measure allows for missing data. For example, the weights can be determined so that $w_{ii'h} = 0$ if the outcome of the X_h is missing for either or both of units i and i'. Again, other weighting schemes can also be used, particularly if information on the importance of particular covariates in explaining potential treatment effect heterogeneity is available.

Determining Strata. In *k*-means clustering, once a set of covariates **X**, a distance metric, and the number of strata *k* have been defined, optimization algorithms are used to classify units into these *k* strata so that the distance metric is minimized.

One difficulty with this procedure is that the number of strata *k* must be determined in order to generate the strata. One solution to this is to use other cluster analysis methods—for example, hierarchical cluster analysis—to explore the structure of the population. A second solution is to generate and evaluate the strata created using different values of *k*, where $k = 1, 2, \ldots q$ for some maximum number of strata *q*. While it is obvious that $q \leq n$, it may also be desirable to choose a manageable number of strata as the maximum, such as $q = 10$ or 20. After determining *q*, for each value of *k*, the optimization algorithm is used to divide the population of units into *k* strata.

Once results have been generated for various values of *k*, the results are then compared to determine which value of *k* is best for the particular population and experiment. One common evaluation method is simply to partition the total variability in the covariates in **X** into the total variability within clusters ($\sigma_w{}^2$) and the variability between clusters ($\sigma_b{}^2$). From this, a measure of the between-cluster variability, $\rho_k = \sigma_{bk}{}^2/(\sigma_{wk}{}^2 + \sigma_{bk}{}^2)$, can be calculated for each number of clusters *k*. Importantly, as ρ_k approaches 1, most of the variation is between strata, leading to a balanced sample of increasingly higher orders. After calculating ρ_k for $k = 1 \ldots q$, the values can be compared visually via an *elbow graph*, where on the x-axis are values of *k* and on the y-axis are values of ρ_k. Since the value of ρ_k monotonically increases with *k*, the statistical criteria commonly used for selecting *k* is based on when the rate of change for each additional stratum slows; this is akin to how a scree plot is used in principal components analysis (Timm 2002). Other criteria for determining the number of clusters can be found in Milligan and Cooper (1985) and include the cubic clustering criteria (CCC; Sarle 1983), the pseudo-F (PSF), and the pseudo-T-square (PTS).

In practice, choosing the optimal number of strata involves both statistical and practical criteria. Statistically speaking, the ideal number of strata can be large, since this results in more homogeneous strata and a more bias-robust sample. Practically, however, it can be difficult to achieve an adequate sample if some of the strata are too small (since the response rate for recruitment in experiments is often small). Additionally, the amount of resources (in terms of time, people, money) aimed at recruitment can be small, which leads to a desire for fewer strata.

Sample Allocation, Selection, and Evaluation

Once the number of strata k has been selected, a strategy for sample selection within each cluster must be developed. In this section, we detail the steps involved in this process.

Allocate the Sample to the Strata. In each of the k-clusters developed in the previous section, there are N_j units in the population, where $N_1 + N_2 + \ldots + N_k = N$. Using proportional allocation, the sample is then allocated so that $n_j = [(N_j/N)n]$, where we use [.] to signify that each value must be rounded to the nearest integer.

Calculate and Rank Within Stratum Distances. Since the overall goal is to select a balanced sample (of some order R), a method for sample selection within each stratum is needed. The goal is to simply select a balanced sample (of Order 1) within each stratum. This means selecting a sample such that in each stratum $j = 1 \ldots k$, $E(X_h|Z = 1, j = j) = E(X_h|\, j = j)$ for each covariate X_h in $\mathbf{X}$ and where $Z = 1$ indicates units in the sample. One method for meeting this goal is to select the n_j units randomly. This would ensure that on average (over repeated samples) units in stratum j would be balanced on both the observed covariates (in $\mathbf{X}$) and the unobserved covariates (not in $\mathbf{X}$). While this certainly increases the bias robustness of the method by guarding against having the wrong model, in any particular sample this method may not result in balance, particularly when the sample size n_j is small. This is particularly important since the multilevel random effects model used for generalization makes inferences in reference to the realized sample.

An alternative method for selection is as follows. First, in each stratum, calculate $E(X_h|\, j = j)$ for each covariate X_h. Then for each of the $i = 1 \ldots N_j$ units in stratum j, a measure of distance is calculated. One strategy is to use the weighted Euclidean distance

$$d_{ij} = \sqrt{\sum\nolimits_{h=1}^{p} w_h \left(X_{ijh} - \mathrm{E}(X_h|\mathrm{j} = j)\right)^2}, \tag{11}$$

where w_h is the weight given to covariate X_h, $E(X_h|j = j)$ is the average value of the hth covariate in stratum j, and X_{ijh} is the value of the hth covariate for unit i in stratum j. Thus, each unit i in stratum j has one total combined distance measure d_{ij}. Again, just as discussed previously, different weights could be used, particularly if it is hypothesized that one covariate matters more in terms of explaining potential treatment effect heterogeneity than another covariate. Note that to the degree that the strata are homogenous (having small within-stratum variances), we would expect these distances d_{ij} not to vary by much.

Based on these within-stratum distance measures, each of the N_j population units within a particular stratum j can be ranked from smallest to largest. This ranked list can then be used for selecting the n_j units into the experiment. For example, a recruiter might start at the top of the list with the unit ranked "1," then if the unit does not agree to be in the experiment, move on to the unit ranked "2," and so on until n_j units agree to be in the experiment.

Nonresponse (refusals). As noted earlier, it is assumed that many units will not agree to be in the experiment. For example, after ranking units within a stratum, it is possible that the first unit to successfully enter the experiment is not "1" but instead "14." Here the concern is that the n_j units that agree to be in the sample are different than the N_j units in the population on either the covariates in **X** or other covariates (related to nonresponse). The method we propose here guards against both of these problems. To see this, first note that if the strata are completely homogenous on **X**, then each of the N_j units in the same stratum can be considered "replicates"; in this case, all of the rankings would be identical (since all of the distances would be zero). In practice, although the strata will not be completely homogenous (resulting in rankings), when the strata are sufficiently different, the absolute differences between the N_j units within each stratum will be small. This means that differences between the units ranked "1" and "14" within the same stratum will generally be smaller than differences between any two units from different strata.

The second concern is that units that refuse to take part in the experiment are different than those that agree to be in the experiment. This is a question regarding an omitted variable X^*. Note that differences in relation to X^* will only cause bias if the potential treatment effects are a function of X^*

(conditional on the other **X** values). This concern suggests that a bias-robust approach to selecting **X** is particularly helpful here—both through the inclusion of a large variety of covariates and through a clear statement and evaluation of models; together this makes a discussion of the effects of possible omitted variable bias on generalization possible.

Evaluation of the Sample. In practice, it is unlikely that the n_j "best" units (those with the smallest distances d_{ij}) in each stratum will all agree to be in the study. As discussed earlier, when the strata are sufficiently homogenous, the inclusion of later ranked units will typically not be problematic. Regardless, once the sample is selected, it is important to evaluate the degree of balance between the final sample S and population P. To do so, balance (of Order 1) can be assessed in each stratum j by comparing $E(X_h|Z = 1, j = j)$ to $E(X_h|j = j)$ for each covariate X_h. Then the overall balance of the final sample S and the population P can be evaluated by comparing $E(X_h{}^r|Z = 1)$ to $E(X_h{}^r)$ for various values of r. In order to evaluate whether this degree of balance is adequate, both substantive criteria regarding the importance of each covariate can be used and statistical criteria like standardized mean differences or t-tests. Finally, if large residual differences are detected, these can be reduced through post hoc strategies like regression adjustment or reweighting (e.g., Hedges and O'Muircheartaigh 2011).

Example

In order to illustrate the implementation of this method and its benefits compared to the conventional bottom-up approach to generalization, in this section we present a reanalysis of an experiment evaluating a middle-school mathematics program, SimCalc. The original study included 73 schools that, while selected with an eye toward generalization, did not use any formal method for doing so (Roschelle et al. 2010). In order to better generalize from these schools to the population of noncharter schools serving seventh graders in Texas ($N = 1{,}713$), Tipton (2013) reanalyzed these data using a propensity score subclassification estimator based on 26 covariates from the state academic excellence system. Here we present a new analysis in which we ask, what would happen if we could go back in time and instead collect the sample using the sample selection method developed here? How different would this sample be from the sample actually collected in experiment?

All analyses presented here were conducted in the statistical program **R** (R Development Core Team 2012), though clustering algorithms are available in most statistical packages. Since the 26 covariates in Table 1 includes one covariate ("rural") that is binary, we use the measure of distance provided by Gower (1971), which can be calculated using the function **daisy** in the **cluster** package (Maechler et al. 2005). For numbers of classes $k = 1 \ldots 20$, we then used the **kmeans** function to create different segmentations of the population. For each of these 20 iterations, we saved the measures of the proportion of variance between classes. The results of this analysis are presented in the elbow plot in Figure 1.

Figure 1 is an elbow plot illustrating the proportion of variance between strata. Note that at first, adding strata dramatically increases ρ_k, but eventually these changes become smaller. Based on this figure, we selected the $k = 9$ strata solution, since for this number over 80% of the total variation in the 26 variables in **X** is between strata. In Table 1, we give the population means by stratum for each of the 26 covariates. These reveal some interesting differences between strata. For example, almost all schools located in rural counties are found in Strata 3, 5, and 6; of these, Stratum 3 contains smaller schools with larger proportions of at-risk, mobile, and low-achieving students. In contrast, Strata 4, 8, and 9 include no schools located in rural counties; of these, Stratum 9 includes schools that have a larger proportion of minority, at-risk, and economically disadvantaged students, while Stratum 8 includes schools with higher academic achievement.

Table 1 also includes, for each stratum, the proportion of the population ($w_{pj} = N_j/N$), the number of units in the sample allocated using proportional allocation with rounding (n_j), and the number of units in the actual experiment (n_{ej}). This reveals that while the actual experiment represented the population fairly well, it did greatly overrepresent schools from Stratum 1 and underrepresent schools in Stratum 3. The fact that the experiment did not include any schools from Stratum 3 is an example of a *coverage error*.

In order to evaluate the overall balance and bias robustness that would be achieved using this approach, in Table 2 we report the $E(X_h^r)$ for $r = 1,2,3$ for each of the 26 covariates in the population. We then use our distance-based method to order the units in each stratum j, where lower ranks indicate smaller distances from the stratum mean for the population. Since we do not have a sense of which covariate is likely to have the largest impact, we set the weights so that $w_h = 1/V(X_h)$, therefore weighting each covariate equally. In order to illustrate the usefulness of the method under both ideal and high nonresponse, we include two possible samples. The first sample selects the n_j highest ranked schools in each stratum; this is the "ideal" sample. The second ("nonresponse") sample instead assumes that the first 50 units in each stratum refused to participate in the experiment and that

Table 1. Comparison of Population Means by Variable and Stratum.

	Population Mean									
Description	Overall	Stratum 1	Statum 2	Stratum 3	Stratum 4	Stratum 5	Stratum 6	Stratum 7	Stratum 8	Stratum 9
Teacher tenure (mean years)	7.09	7.58	7.08	8.73	6.84	***7.95***	6.85	7.04	6.65	**6.36**
Teacher experience (mean years)	11.58	10.43	12.44	13.93	11.78	***14.09***	12.19	**9.61**	11.38	10.06
Teacher-student ratio	12.70	14.33	10.84	10.50	13.49	10.31	**9.59**	13.37	***14.73***	14.01
Teachers that are African American (%)	8.39	9.47	2.50	9.54	3.20	**1.00**	4.20	41.69	5.45	***11.41***
Teachers that are Hispanic (%)	14.72	***49.13***	5.37	17.05	4.04	**3.03**	25.52	26.48	4.96	10.49
Teachers in the school (total)	39.87	55.20	24.84	**9.31**	40.16	19.60	18.77	54.26	***57.59***	53.00
Teachers in first year of teaching (%)	8.32	8.79	8.19	5.68	7.32	**5.03**	11.93	***12.41***	6.26	10.08
Teachers with 1–5 years experience (%)	28.01	33.45	24.24	23.49	26.97	**17.81**	23.28	***36.65***	29.18	33.18
Teachers with >20 years experience (%)	20.25	17.42	21.87	***28.74***	20.11	27.01	22.91	16.87	19.11	**16.02**
Students in disciplinary alternative education programs (%)	3.10	4.44	2.78	1.46	2.53	1.28	3.42	4.90	**1.16**	***4.70***
Seventh grade retention (rate)	1.83	1.64	1.23	***12.12***	1.17	0.91	2.28	3.05	**0.46**	1.74
Students who are mobile (%)	19.23	19.41	15.73	***83.72***	14.73	14.07	19.05	26.35	**9.17**	19.17
Students in school who are in seventh grade (%)	31.21	35.80	27.96	**9.79**	35.21	19.60	19.93	***35.91***	35.62	37.57
Students in seventh grade (total)	190.40	282.09	85.65	**8.49**	206.27	45.07	46.52	267.12	***309.51***	287.52
Students who are African American (%)	11.79	6.42	7.83	20.38	8.70	**3.57**	8.63	***36.84***	8.88	18.62
Students who are Hispanic (%)	40.27	***86.00***	32.53	47.10	21.62	18.27	52.65	58.73	**17.30**	43.51
Students who are LEP (%)	7.54	17.39	4.39	8.04	3.14	**2.58**	8.82	***19.47***	2.72	7.07
Students who are economically disadvantaged (%)	53.64	81.56	54.52	49.42	36.69	42.54	69.26	***86.26***	**17.89**	55.63
Students who are at risk (%)	43.47	56.39	42.58	***86.47***	33.27	30.03	50.84	63.81	**18.89**	45.83
Students proficient in seventh grade reading (%)	81.90	79.78	86.79	**2.21**	90.79	93.08	55.29	70.62	***95.95***	84.89
Students proficient in seventh grade math (%)	72.79	69.97	75.08	**0.75**	83.59	86.44	41.67	58.29	***92.93***	73.07
Students proficient in Grades 3–11 math (%)	73.60	69.10	74.41	**13.60**	82.28	85.54	66.15	57.36	***92.25***	71.89
Students proficient in Grades 3–11 all (%)	63.29	55.76	62.27	**21.95**	72.63	75.77	53.99	43.27	***86.57***	60.68
Students with commended performance, Grades 3–11, math (%)	19.61	14.86	16.24	**2.12**	23.32	26.32	13.37	9.05	***42.49***	16.84
Students with commended performance, Grades 3–11, reading (%)	8.71	6.19	6.89	**1.61**	10.48	11.44	4.86	3.30	***21.81***	7.10
County of school is rural	0.33	0.03	**0.99**	0.16	**0.00**	**0.99**	0.95	0.02	**0.00**	**0.00**
Proportion of population (w_{pj})	1	0.14	0.17	0.04	0.19	0.10	0.05	0.08	0.09	0.15
Sample size required using proportional allocation (n_j)	73	10	12	3	14	8	4	6	6	11
Sample size observed in completed experiment (n_{ej})	73	15	13	0	13	6	4	2	9	11

Note. LEP = limited English proficiency. Bold–italics and dark gray values indicate those are the maximum value across strata for the particular covariate, while boldface and light gray indicate the row minimums.

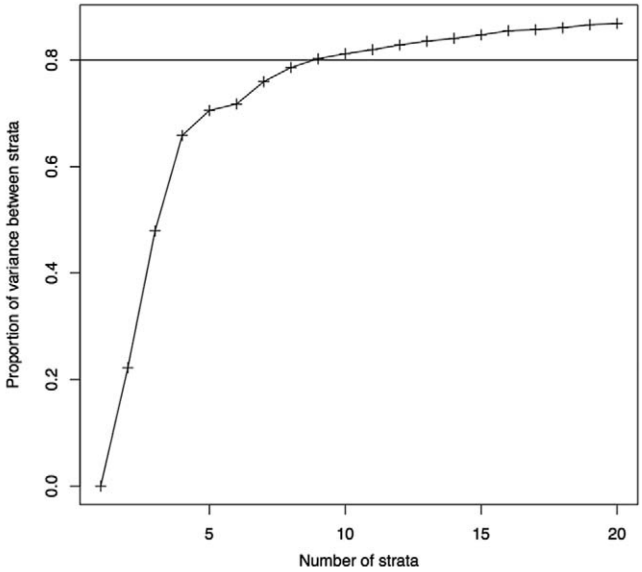

Figure 1. Elbow plot of the proportion of total variance between strata.

from there the next n_j schools agreed. Note that this would mean that in total 450 schools refused participation before any agreed to participate, which corresponds to a nonresponse rate of at least 83%. In each sample of schools, we then calculated $E(X_h{}^r|Z = 1)$ for $r = 1,2,3$ for each covariate. We also calculated these moments for the actual sample used in the experiment.

In order to compare the ideal and nonresponse samples, as well as the "actual" sample used in the experiment to the population, in Table 2 we report the percentage of absolute bias, where

$$\% \text{ absolute bias} = \left|E(X_h^r|Z = 1) - \mathrm{E}\left(X_h^r\right)\right|/E\left(X_h^r\right), \tag{12}$$

for each $r = 1,2,3$, and where depending upon the column $Z = 1$ indicates the actual sample or the proposed sample based on the method developed here. Bolded values indicate variables for which the balance achieved is better using either the ideal or nonresponse sample selection strategies developed here than that achieved in the actual experiment. This balance is better for 19 or more variables (of 26) in terms of first and second moments and 14 or more in terms of third moments. Additionally, the maximum absolute relative difference is 0.253 for the ideal sample (over all 26 covariates) and 0.297 for the nonresponse sample, compared to 0.695 for the actual experiment. In general balance is better for both Orders 1 and 2 than Order 3, indicating that the ideal and nonresponse samples lead to approximately balanced samples of Order 2.

Since there are some imbalances remaining, post hoc propensity score methods could be used for further adjustments. In these methods, the achieved sample of 73 schools is compared to the population of 1,713 schools on the 26 covariates using a logistic regression to estimate the propensity score. Comparing the empirical densities of propensity score logits indicates how well the results might generalize (Stuart et al. 2011). When these densities are similar, generalizations are easier since reweighting approaches lead to large reductions in bias with only small increases in variance; when they differ, particularly when the region of common support is small (indicating coverage errors), reweighting approaches are less effective, leading to estimates with remaining bias and larger variance inflations (Tipton 2013).

In Figure 2, we compare the empirical densities of the propensity scores in the sample and population for each of the three samples under comparison (actual, ideal, and nonresponse). As the shapes of the densities and the lines marking the population quintiles indicate, the samples selected using methods from this article are more similar to the inference population than the actual sample used in the experiment. Importantly, the longtail in the actual sample indicates a large coverage error (Tipton 2013); this longtail is not present in either of the samples produced using the methods developed here, making additional adjustments using post hoc methods more effective.

Discussion and Conclusion

The purpose of this article has been twofold. First, it proposes a framework for site selection in large-scale randomized experiments (model-based sampling) that can be used in conjunction with the super-population models commonly used to estimate PATEs. Second, it provides a method for selecting a self-weighting bias-robust balanced sample within this framework through the creation of strata based on cluster analysis methods. In this section, we conclude by briefly discussing the benefits, limitations, and possible extensions of the method.

Table 2. Comparison of Balance of Orders 1,2,3 for the Population, Planned Sample, and Completed Experiment.

	Balance (Order 1, $E(X_h)$)				Balance (Order 2, $E(X_h^2)$)				Balance (Order 3, $E(X_h^3)$)			
Description	Population	% Bias using planned sample (ideal)	% Bias using planned sample (non-response)	% Bias using completed experiment	Population	% Bias using planned sample (ideal)	% Bias using planned sample (non-response)	% Bias using completed experiment	Population	% Bias using planned sample (ideal)	% Bias using planned sample (non-response)	% Bias using completed experiment
Teacher tenure (mean years)	7.1	**0.01**	**0.02**	0.04	57.0	**0.09**	**0.09**	0.11	513.2	0.22	**0.19**	0.20
Teacher experience (mean years)	11.6	**0.001**	**0.029**	0.05	143.1	**0.037**	**0.087**	0.12	1,874.4	**0.102**	**0.164**	0.18
Teacher–student ratio	12.7	**0.027**	**0.036**	0.04	170.29	**0.036**	**0.073**	0.08	2,387.90	**0.029**	0.120	0.10
Teachers who are African American (%)	8.4	**0.139**	**0.204**	0.70	328.68	**0.286**	**0.013**	0.93	19,685.58	**0.396**	**0.243**	0.98
Teachers who are Hispanic (%)	14.7	**0.181**	**0.145**	0.47	794.20	**0.405**	**0.145**	0.67	57,712.39	**0.556**	**0.125**	0.80
Teachers in the school (total)	39.9	0.078	**0.023**	0.08	2,163.31	**0.045**	**0.009**	0.08	138,601.57	0.045	**0.034**	0.04
Teachers in first year of teaching (%)	8.3	0.108	0.147	0.05	123.39	0.393	0.182	0.09	2,650.02	0.661	0.068	0.01
Teachers with 1–5 years experience (%)	28.0	**0.024**	0.074	0.03	934.47	0.112	0.130	0.01	35,259.71	0.232	0.224	0.04
Teachers with > 20 years experience (%)	20.3	**0.024**	**0.063**	0.13	526.68	**0.167**	**0.191**	0.27	17,006.58	**0.359**	**0.366**	0.44
Students in disciplinary alternative education programs (%)	3.1	**0.073**	**0.048**	0.10	19.36	**0.172**	**0.065**	0.18	179.69	0.474	0.249	0.14
Seventh grade retention (rate)	1.8	**0.253**	0.297	0.29	34.81	**0.688**	0.862	0.82	2,217.74	**0.902**	0.988	0.98
Students who are mobile (%)	19.2	**0.043**	**0.016**	0.23	589.22	**0.150**	**0.071**	0.58	32,368.22	**0.316**	**0.131**	0.86
Students in school that are in seventh grade (%)	31.2	**0.024**	**0.022**	0.12	1,156.94	**0.039**	**0.007**	0.20	47,020.63	**0.134**	**0.020**	0.25
Students in seventh grade (total)	190.4	**0.069**	**0.027**	0.18	60,538.08	**0.049**	**0.051**	0.17	22,992,240.24	0.208	0.159	0.10
Students who are African American (%)	11.8	**0.121**	**0.131**	0.57	393.08	**0.364**	**0.192**	0.87	19,693.03	**0.525**	**0.172**	0.97
Students who are Hispanic (%)	40.3	**0.012**	**0.063**	0.17	2,511.52	**0.024**	**0.090**	0.29	189,543.89	**0.065**	**0.085**	0.38
Students who are LEP (%)	7.5	**0.050**	**0.031**	0.25	158.82	**0.334**	**0.169**	0.42	5,976.16	0.642	**0.005**	0.41
Students who are economically disadvantaged (%)	53.6	0.038	**0.008**	0.03	3,467.40	0.096	**0.013**	0.02	248,724.52	0.152	0.051	0.01
Students who are at risk (%)	43.5	0.035	**0.011**	0.07	2,220.43	**0.036**	**0.058**	0.15	129,401.04	**0.035**	**0.130**	0.27

(continued)

Table 2. (continued)

	Balance (Order 1, $E(X_h)$)				Balance (Order 2, $E(X_h^2)$)				Balance (Order 3, $E(X_h^3)$)			
Description	Population	% Bias using planned sample (ideal)	% Bias using planned sample (non-response)	% Bias using completed experiment	Population	% Bias using planned sample (ideal)	% Bias using planned sample (non-response)	% Bias using completed experiment	Population	% Bias using planned sample (ideal)	% Bias using planned sample (non-response)	% Bias using completed experiment
Students proficient in seventh grade reading (%)	81.90	**0.002**	**0.009**	0.05	7,145.30	**0.010**	**0.012**	0.04	629,832.50	**0.025**	**0.014**	0.04
Students proficient in seventh grade math (%)	72.79	**0.013**	**0.017**	0.04	5,762.90	0.038	0.019	0.02	468,041.73	0.064	0.023	0.01
Students proficient in Grades 3–11 math (%)	73.60	0.020	**0.005**	0.02	5,707.96	0.043	**0.004**	0.01	453,311.23	0.066	0.004	0.00
Students proficient in Grades 3–11 all (%)	63.29	0.016	**0.003**	0.01	4,278.33	0.034	0.006	0.00	301,184.65	0.055	0.012	0.01
Students with commended performance, Grades 3–11, math (%)	19.61	0.054	0.045	0.04	516.45	0.149	**0.048**	0.10	16,812.24	0.262	**0.010**	0.22
Students with commended performance, Grades 3–11, reading (%)	8.71	**0.003**	**0.009**	0.02	120.03	**0.105**	**0.007**	0.13	2,213.91	**0.248**	**0.019**	0.30
County of school is rural	0.33	**0.015**	0.067	0.04	0.33	**0.015**	0.067	0.04	0.33	**0.015**	0.067	0.04

Note. LEP = limited English proficiency. % bias is the absolute difference from the population, standardized by the population value. Boldface values indicate the smallest bias when comparing the planned sample to the completed experiment. The “ideal” planned sample selects the first n_j units from each stratum (see Table 1), while in each stratum in the “nonresponse” planned sample, the first 50 schools refused participation and the next n_j units agreed.

Subgroup Analyses

When treatment effects do in fact vary, a single PATE is clearly inadequate for summarizing the effectiveness of a program or an intervention. One solution is to additionally report subgroup average treatment effects. A benefit of our site-selection approach is that a separate average treatment effect can be calculated for each of the k strata. This strategy is similar to the method for subgroup creation proposed by Peck (2005) in general, and, when the covariates contained in **X** also predict treatment compliance (or other post-random-assignment groupings), to the methods proposed by Peck (2003), (2013) and Bell and Peck (2013).

Eligibility Problems

It is possible that some of the N units in the population P are not eligible for inclusion in the experiment. For example, certain schools may already use the program under study, or when resources are limited, travel outside a particular area may be infeasible. The goal of bias-robust sample selection is still useful here, particularly since it requires study planners to determine whether the reasons for ineligibility explain variations in treatment effects (warranting inclusion in **X**). For example, if only large schools are eligible to be in an experiment (because of the design of the study), then generalizations to all schools in the population are only warranted if treatment effects do not vary in relation to school size. Otherwise, the population needs to be redefined to focus only on large schools. Second, stratified sample selection using cluster analysis can also be used here. Again, the strata would be created based on the N units in the population P, though only eligible schools or sites would be included in the ranked list for sample selection. As a result of ineligibility, however, the sampling fractions n_j/N_j may differ by stratum. Tipton et al. (2014) includes a full discussion of sample selection under eligibility constraints, including an alternative bias-robust approach using propensity score methods.

Nonresponse Analysis

One of the biggest practical concerns in sample selection in experiments is the fact that many sites will not agree to be in the experiment. A benefit of this approach is that by articulating at the outset a set of covariates (**X**) that possibly explain treatment effect variability (thereby leading to bias), information on sites that refuse to be in the study can be tracked and later

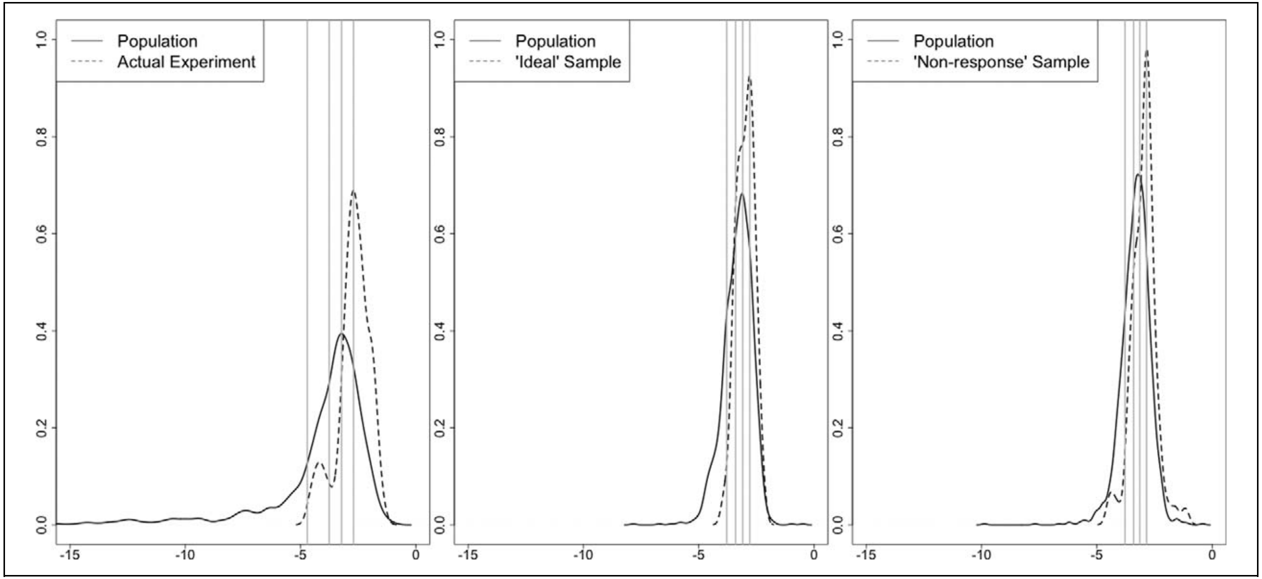

Figure 2. Comparison of propensity score logit densities for three samples versus population.

compared to those that accept and to the inference population of interest using available nonresponse analysis methods (e.g., Little and Rubin 1987). Additionally, this method requires designers to plan for refusals and even allows for the collection of additional information on those who refuse or accept.

Relationship to Post Hoc Methods

The goals and framework for balanced sampling have much in common theoretically with the goals of propensity score matching for post hoc adjustments for generalization. Even when using the stratified sampling method introduced here, differences may remain between the achieved sample and the population, particularly when nonresponse is high. However, as illustrated in the example, these remaining imbalances can be more easily adjusted if generalization is planned for; this is both because any remaining imbalances are typically smaller (Table 2) and since coverage errors are greatly reduced (Figure 2), both of which make reweighting procedures more effective.

Random Selection and Design-Based Inference

Given the infeasibility of random sampling in experiments, this article has provided a method for site selection that is strategic, model based, and nonrandom. However, the method we develop for site selection—stratified sampling with proportional allocation based on cluster analysis—can also be used in a design-based framework with probability sampling.

Data Frame Concerns

A potential weakness of the stratified selection method developed here is that it requires an available sampling frame. This results in two limitations to our approach. First, like the post hoc methods available for generalization, this sampling frame needs to include a rich set of covariates on all units in the population, where here the covariates that matter are those that explain variation in treatment effects. Since most censuses of schools and districts focus on demographics, using any model-based approach could result in omitted variable bias. An important feature of this method is that it requires a thoughtful discussion of the benefits and limitations of the achieved sample for generalizations and to whom, where, and under what conditions or assumptions these generalizations are most warranted.

Second, the fact that the sampling frame must enumerate all N units in the population means that the method we have developed here is most useful when the sampling units are aggregates, since data on aggregates (e.g., schools) are typically publically available, while data on individuals are not. However, while the strata creation method developed here may not be practical when the unit of analysis is the individual, we argue that the bias-robust framework is still useful. Stratified sampling is only one method for creating a balanced sample in the bias-robust framework. Future research should investigate the practicality of methods that do not require such detailed population frames—for example, quota sampling and respondent-driven sampling (e.g., Smith 1983; Watters and Biernacki 1989).

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The author received an NSF grant 1118978.

References

Bell, S. H., and L. R. Peck. 2013 "Using Symmetric Predication of Endogenous Subgroups for Causal Inferences about Program Effects under Robust Assumptions: Part Two of a Method Note in Three Parts." *American Journal of Evaluation* 34:413–26.

Bloom, H. S. 2005. *Learning More from Social Experiments: Evolving Analytic Approaches*. New York: Russell Sage Foundation.

Brewer, K. R. W. 1999. "Design-based or Prediction Based Inference? Stratified Random vs Stratified Balanced Sampling." *International Statistical Review* 67:35–47.

Cochran, W. G. 1968. "The Effectiveness of Adjustment by Subclassification in Removing Bias in Observational Studies." *Biometrics* 24:295–313.

Cornfield, J., and J. W. Tukey. 1956. "Average Values of Mean Squares in Factorials." *The Annals of Mathematical Statistics* 27:907–49.

Cox, D. R., and E. Spjotvoll. 1982. "On Partitioning Means into Groups." *Scandinavian Journal of Statistics* 9:147–52.

Day, G. S., and R. M. Heeler. 1971. "Using Cluster Analysis to Improve Marketing Experiments." *Journal of Marketing Research* 8:340–47.

ETS (Educational Testing Service). 2008. "Access to Success: Patterns of Advanced Placement Participation in US High Schools." ETS Policy Information Report. Accessed March 19, 2012. http://www.ets.org/Media/Research/pdf/PIC-ACCESS.pdf.

Everitt, B. S., S. Landau, M. Leese, and D. Stahl. 2011. *Cluster Analysis*. 5th ed. Wiley Series in Probability and Statistics. London: John Wiley.

Fienberg, S. E., and J. M. Tanur. 1987. "Experimental and Sampling Structures: Parallels Diverging and Meeting." *International Statistical Review* 55:75–96.

Golder, P. A., and K. A. Yeomans. 1973. "The Use of Cluster Analysis for Stratification." *Applied Statistics* 22:213–19.

Gower, J. C. 1971. "A General Coefficient of Similarity and Some of Its Properties." *Biometrics* 27:857–72.

Greenberg, D. H., and M. Shroder. 2004. *The Digest of Social Experiments*. 3rd ed. Washington, DC: The Urban Institute.

Hedges, L. V., and C. A. O'Muircheartaigh. 2011. *Improving Generalizations from Designed Experiments*. Northwestern University.

Holland, P. W. 1986. "Statistics and Causal Inference." *Journal of the American Statistical Association* 81:945–60.

Imai, K., G. King, and E. A. Stuart. 2008. "Misunderstandings between Experimentalists and Observationalists about Causal Inference." *Journal of the Royal Statistical Society, Series A* 171:481–502.

Kirk, R. E. 1995. *Experimental Design: Procedures for the Behavioral Sciences*. 3rd ed. Pacific Grove, CA: Brooks/Cole.

Kruskal, W., and F. Mosteller. 1980. "Representative Sampling, IV: The History of the Concept in Statistics, 1895-1939." *International Statistical Review* 48: 169–95.

Lipson, M. Y., J. H. Mosenthal, J. Mekkelsen, and B. Russ. 2004. "Building Knowledge and Fashioning Success One School at a Time." *The Reading Teacher* 57:534–42.

Little, R. J. A., and D. B. Rubin. 1987. *Statistical Analysis with Missing Data*. New York: John Wiley.

Lohr, S. L. 1999. *Sampling: Design and Analysis*. New York: Duxbury Press.

Maechler, M., P. Rousseeuw, A. Struyf, and M. Hubert. 2005. *Cluster Analysis Basics and Extensions*. R package. Unpublished.

Mansure, K. A., and B. M. Reist. 2010. Evaluating Alternative Criteria for Primary Sampling Units Stratification." JSM Section on Survey Research Methods. 2010 Conference Proceedings. Accessed March 19, 2012. http://www.amstat.org/sections/srms/proceedings/y2010/Files/308700_61432.pdf.

Milligan, G. W., and M. C. Cooper. 1985. "An Examination of Procedures for Determining the Number of Clusters in a Data Set." *Psychometrika* 50:159–79.

Murphy, P. 2008. “An Overview of Primary Sampling Units (PSUs) in Multi-stage Samples for Demographic Surveys.” JSM Conference Proceedings, Section on Government Statistics. Accessed March 19, 2012. http://www.amstat.org/sections/srms/proceedings/y2008/Files/301835.pdf.

NISRA (Northern Ireland Statistics and Research Agency). (2002) 2001. “Census Review and Evaluation: Evaluation of the Northern Ireland Census Coverage Survey.” Accessed 2/1/2012 at http://www.nisra.gov.uk/archive/census/evaluation/CCSEvaluationReportNI.pdf.

Olsen, R. B., L. L. Orr, S. H. Bell, and E. A. Stuart. 2013. “External Validity in Policy Evaluations That Choose Sites Purposively.” *Journal of Policy Analysis and Management* 32:107–21.

Peck, L. R. 2003. “Subgroup Analysis of Social Experiments: Measuring Program Impacts Based on Post Treatment Choice.” *American Journal of Evaluation* 24: 157–87.

Peck, L. R. 2005. “Using Cluster Analysis in Program Evaluation.” *Evaluation Review* 29:178–96.

Peck, L. R. 2013. “On Analysis of Symmetrically Predicted Endogenous Subgroups: Part One of a Method Note in Three Parts.” *American Journal of Evaluation* 34:225–36.

Peck, L. R., I. D’Attoma, F. Camillo, and C. Guo. 2012. “A New Strategy for Reducing Selection Bias in Nonexperimental Evaluations, and the Case of How Public Assistance Receipt Affects Charitable Giving.” *The Policy Studies Journal* 40:601–25.

R Development Core Team. 2012. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. http://www.R-project.org/.

Rao, J. N. K. 2005. “Interplay between Sample Survey Theory and Practice: An Appraisal.” *Survey Methodology* 31:117–38.

Raudenbush, S. W. 1993. “Hierarchical Linear Models and Experimental Design.” In *Applied Analysis of Variance in Behavioral Science*, edited by L. K. Edwards, 459–96. New York: Marcel Dekker.

Raudenbush, S. W., and A. S. Bryk. 2002. *Hierarchical Linear Models: Applications and Data Analysis Methods* (Vol. 1). Newbury Park, CA: Sage.

Roschelle, J., N. Schechtman, D. Tatar, S. Hegedus, B. Hopkins, S. Empson, J. Knudson, and L. P. Gallagher. 2010. “Integration of Technology, Curriculum, and Professional Development for Advancing Middle School Mathematics: Three Large-scale Studies.” *American Educational Research Journal* 47: 833–78. doi:10.3102/0002831210367426

Rosenbaum, P. R., and D. B. Rubin. 1983. “The Central Role of the Propensity Score in Observational Studies for Causal Effects.” *Biometrika* 70:41–55.

Royall, R. M., and J. Herson. 1973. “Robust Estimation in Finite Populations I.” *Journal of the American Statistical Association* 68:880–89.

Rubin, D. B. 1974. “Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies.” *Journal of Educational Psychology* 66:688–701.

Rubin, D. B. 1976. “Inference and Missing Data.” *Biometrika* 63:581–92.

Rubin, D. B., and N. Thomas. 1996. “Matching Using Estimated Propensity Scores: Relating Theory to Practice.” *Biometrics* 52:249–64.

Sarle, W. S. 1983. *Cubic Clustering Criterion*. Tech. Report A-108. Cary, NC: SAS Institute.

Schochet, P. Z. 2013. “Estimators for Clustered Education RCTs Using the Neyman Model for Causal Inference.” *Journal of Educational and Behavioral Statistics* 38:219–38.

Scott, A. J., and M. Knott. 1974. “A Cluster Analysis Method for Grouping Means in the Analysis of Variance.” *Biometrics* 30:507–12.

Shadish, W. R., T. D. Cook, and D. T. Campbell. 2002. *Experimental and Quasi-experimental Designs for Generalized Causal Inference*. Boston, MA: Houghton-Mifflin.

Sleegers, P., S. T. Bergen, and J. Giesbers. 1994. “The Policy-making Capacity of Schools: Results of a Dutch Study.” *Educational Management Administration & Leadership* 22:147.

Smith, T. M. F. 1983. “On the Validity of Inferences from Non-random Samples.” *Journal of the Royal Statistical Society. Series A (General)* 146:394–403.

Smith, T. M. F., and R. A. Sugden. 1988. “Sampling and Assignment Mechanisms in Experiments, Surveys, and Observational Studies, Correspondent Paper.” *International Statistical Review* 56:165–80.

Spybrook, J. 2012. “Detecting Intervention Effects across Context: An Examination of the Power of Cluster Randomized Trials.” *Working Paper*. Kalamazoo: Western Michigan University.

Stuart, E. A. 2010. “Matching Methods for Causal Inference: A Review and a Look Forward.” *Statistical Science* 25:1–21.

Stuart, E. A., S. R. Cole, C. P. Bradshaw, and P. J. Leaf. 2011. “The Use of Propensity Scores to Assess the Generalizability of Results from Randomized Trials.” *Journal of the Royal Statistical Society, Series A*, 174:369–86.

Timm, N. H. 2002. *Applied Multivariate Analysis*. New York: Springer.

Tipton, E. 2013. “Improving Generalizations from Experiments Using Propensity Score Subclassification: Assumptions, Properties, and Contexts.” *Journal of Educational and Behavioral Statistics* 38:239–66.

Tipton, E., L. V. Hedges, M. Vaden-Kiernan, G. D. Borman, K. Sullivan, and S. Caverly. 2014. “Sample Selection in Randomized Experiments: A New

Method Using Propensity Score Stratified Sampling." *Journal of Research on Educational Effectiveness* 7:114–35.

Tukey, J. W. 1949. "Comparing Individual Means in the Analysis of Variance." *Biometrics* 5:99–114.

Valliant, R., A. H. Dorfman, and R. M. Royall. 2000. *Finite Population Sampling and Inference: A Prediction Approach*. Wiley Series in Probability and Statistics. New York: John Wiley.

Watters, J. K., and P. Biernacki. 1989. "Targeted Sampling: Options for the Study of Hidden Populations." *Social Problems* 36:416–30.

Author Biography

Elizabeth Tipton is an assistant professor of applied statistics at Teachers College, Columbia University. Her research focuses on the design and analysis of large-scale experiments, including issues of external validity, and meta-analysis.