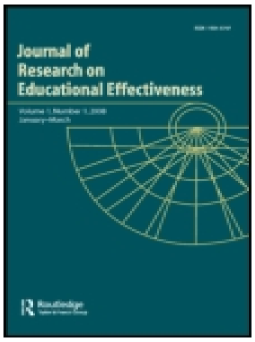

Journal of Research on Educational Effectiveness

ISSN: 1934-5747 (Print) 1934-5739 (Online) Journal homepage: https://www.tandfonline.com/loi/uree20

Sample Selection in Randomized Experiments: A New Method Using Propensity Score Stratified Sampling

Elizabeth Tipton , Larry Hedges , Michael Vaden-Kiernan , Geoffrey Borman , Kate Sullivan & Sarah Caverly

To cite this article: Elizabeth Tipton , Larry Hedges , Michael Vaden-Kiernan , Geoffrey Borman , Kate Sullivan & Sarah Caverly (2014) Sample Selection in Randomized Experiments: A New Method Using Propensity Score Stratified Sampling, Journal of Research on Educational Effectiveness, 7:1, 114-135, DOI: 10.1080/19345747.2013.831154

To link to this article: https://doi.org/10.1080/19345747.2013.831154

Journal of Research on Educational Effectiveness, 7: 114–135, 2014
Copyright © Taylor & Francis Group, LLC
ISSN: 1934-5747 print / 1934-5739 online
DOI: 10.1080/19345747.2013.831154

Routledge
Taylor & Francis Group

Sample Selection in Randomized Experiments: A New Method Using Propensity Score Stratified Sampling

Elizabeth Tipton
Columbia University, New York, New York, USA

Larry Hedges
Northwestern University, Evanston, Illinois, USA

Michael Vaden-Kiernan
SEDL, Austin, Texas, USA

Geoffrey Borman
University of Wisconsin–Madison, Madison, Wisconsin, USA

Kate Sullivan and Sarah Caverly
SEDL, Austin, Texas, USA

Abstract: Randomized experiments are often seen as the "gold standard" for causal research. Despite the fact that experiments use random assignment to treatment conditions, units are seldom selected into the experiment using probability sampling. Very little research on experimental design has focused on how to make generalizations to well-defined populations or on how units should be selected into an experiment to facilitate generalization. This article addresses the problem of sample selection in experiments by providing a method for selecting the sample so that the population and sample are similar in composition. The method begins by requiring that the inference population and eligibility criteria for the study are well defined before study recruitment begins. When the inference population and population of eligible units differs, the article provides a method for sample recruitment based on stratified selection on a propensity score. The article situates the problem within the example of how to select districts for two scale-up experiments currently in recruitment.

Keywords: Experimental design, scale-up, sampling, stratified, recruitment

Randomized experiments are the strongest tools available for answering questions about the average *causal* effect of an intervention. Similarly, studies using probability (random) sampling from well-defined inference populations are the strongest tools available for making inferences (generalizations) about populations. The advantages of both random assignment and random sampling stem from their simplicity and the fact that they permit inferences based on well-defined probability distributions (based on either a random assignment model or a sampling model). The simplicity of model-free inference is compromised in either case by imperfections in design or execution (such as nonresponse), requiring model based adjustments to support inferences.

Address correspondence to Elizabeth Tipton, Columbia University, Teachers College, 525 West 120th Street, New York, NY 10027. E-mail: tipton@tc.columbia.edu

Although randomized experiments have become increasingly common in education, random selection of units (random sampling) into the experiment is not; in fact, recent work suggests that as few as 3% of social experiments have used this dual randomization process (Olsen, Orr, Bell, & Stuart, 2012). The fact that random sampling is not commonly used has two important effects: first, the purposive selection process used is usually not well documented; second, because purposive selection doesn't require the population of interest to be well defined, it is difficult to know to whom the experimental results generalize. This problem is only made more complex when eligibility criteria for the experiment, which impose additional limits on the possible sample, are added into the mix.

Although the problem of causal generalization applies to all experiments, the problem arises most profoundly for experiments aimed at scale-up. The focus of scale-up experiments is typically on determining the causal effect of an intervention in more common classroom or school circumstances and with a broader population (Schneider & McDonald, 2007). The Institute for Education Sciences (IES; 2011) has funded 11 of these Goal IV Scale-Up experiments in the last 8 years. For example, in 2009 and 2010, IES funded scale-up studies of the Open Court Reading (OCR) and Everyday Math (EM) programs. Because both curricula are widely used and have limited evidence of efficacy, these two studies are aimed at determining what the average effects may be in more common, "real-world" circumstances when both curricula are disseminated and implemented at scale.

The focus of this article is on developing a method for sample selection in experiments—particularly scale-up experiments—that results in a sample that is *representative* of a well-defined inference population, where by representative we mean that the sample is like a miniature of the population (Kruskal & Mosteller, 1979). The method we provide does not require random sampling but instead uses propensity score methods to select a sample that is compositionally similar to a well-defined population of interest. By compositionally similar we mean that the sample and population are balanced on a medium to large number of continuous covariates that are believed to explain variation in treatment effects. Quite similar methods to achieve balanced samples are used in the observational studies literature (e.g., Cook, 1993; Rosenbaum & Rubin, 1983). A similar approach is also used in demography where it is called standardization (i.e., standardizing a population composition so that comparisons between desired quantities in two different populations are not confounded with difference in population composition), in economics to form index numbers (see, e.g., Kitagawa, 1964), and to decompose effects to isolate the impact of changes in population composition (see, e.g., Oaxaca, 1973). The same idea is used as a tool for analysis in survey research (see, e. g., Kalton, 1968; Rosenberg, 1962; Valliant, Dorfman, & Royall, 2000). It is also a basic tool in the analysis of missing data (see, e.g., Groves, Dillman, Eltinge, & Little, 2002; Little & Rubin, 2002; Rubin, 1976). It is important to note that most of these methods focus on post hoc adjustments to achieve balance; in contrast, this article focuses on methods for improving balance prospectively, through the design of the sample selection process.

The particular problem that this article addresses is how to achieve balance between the sample and population on a medium to large number of continuous covariates under two constraints—random sampling is not feasible and at least some portion of the units in the population are not eligible to be in the experiment. We develop this theory in relation to the particular problems of recruiting school districts in the OCR and EM scale-up studies just introduced. Although we refer throughout to units or school districts, the same method could be used to select schools, hospitals, community centers, or any other type of sites typically found in large-scale evaluation studies.

The article is organized as follows. In the first section, we review common practice in experiments and the current literature on generalization. In the second section, we introduce the definitions and assumptions needed for the method generally and with respect to the OCR and EM studies. In the third section, we begin by introducing the method for stratified selection on the propensity score, and then illustrate the method in the OCR and EM cases. We conclude with a discussion of when this strategy works well and common issues and pitfalls.

CURRENT PRACTICE AND METHODS

In a typical experiment, units are recruited using a process that is both informal and purposive, and that operates within a set of constraints. First, as a result of power analyses, budgetary issues, and concerns with internal validity, not all units in the population may be considered eligible to be in the experiment (Boruch, 1996). For example, the eligibility criteria may include location (e.g., a certain state), school size, or current curricula used. Units that do not meet these eligibility criteria can be thought of as having zero probability of being in the experiment. Second, researchers are typically interested in gathering the sample quickly and inexpensively; this means that they often target units for recruitment that have a large payoff. This might include large school districts (bringing many schools to the study), school districts they have worked with in the past (a higher likelihood of participation), or school districts close by others that have already agreed to be in the study. This means that whereas all of the eligible units have some probability of being in the experiment, some units have a larger probability than others. Third, in an informal nod toward generalizability, researchers may attempt to get units in different regions or from different levels of urbanicity. Overall, these three concerns—eligibility, targeting, and generalizability—lead to a process that is iterative, informal, and not well documented or evaluated. In contrast to prevailing practice, the goal of this article is to develop a formal process for targeting units for participation in an experiment that takes into account eligibility criteria and leads to the ability to generalize to a well-defined population.

To develop this method, we turn to recent work on improving generalizations from completed experiments through a new application of propensity score matching methods (Hedges & O'Muircheartaigh, 2011; Stuart, Cole, Bradshaw, & Leaf, 2011; Tipton, 2013). These publications develop retrospective methods for matching the units in a completed experiment to those in a well-defined population via a sampling propensity score. Although these methods can easily be used retrospectively, theoretical and empirical results from Tipton (2013) suggest that retrospective generalizations are limited by two factors. First, it is common for some portion of the population of interest to not have any units in the experiment "like" them, which results in a *coverage error*. This means that it is impossible to get a bias-free estimate of the population average treatment impact. Second, when the composition of units in the experiment and population differ largely, any reweighting approach will lead to large increases in standard errors. In some cases, the standard errors are so large as to make the estimate of the treatment impact practically useless.

In this article we provide a prospective version of this propensity score method that at worst reduces and at best eliminates both coverage error and variance inflation problems. The method we develop is stratified selection, where the strata are defined in relation to a propensity score. The propensity score relates the composition of units in the inference population to the units eligible to be in the experiment. By stratifying on the propensity score, a more diverse sample can be targeted for selection than is usual in common practice. When the common support for the propensity score (relating the eligible units to the inference population) includes all units in the population, coverage errors are reduced or eliminated, as units of all types are targeted for inclusion in the experiment. In addition, by stratifying on the propensity score, recruitment targeting practices and resources can be differentially allocated to the strata, leading to increased participation rates. The goal of the method is for the achieved sample to be compositionally similar to the inference population.

INFERENCE POPULATIONS AND SAMPLING FRAMES

Definitions, Assumptions, and Goals

Generalizations from experiments are often bottom-up; that is, the results from an experiment are commonly said to generalize to units "like" those in the study, even when what is meant by "like" is ill defined (Cornfield & Tukey, 1956). In contrast, the method we provide here is top-down. It begins with a well-defined population and guides recruitment so that the sample can be selected to enable the desired generalizations, given assumptions that are explicitly stated.

To begin to develop a method for generalization, the following three groups need to be defined:

1. *An inference population* (P) of size N must first be well defined. For example, the population could be all districts containing elementary schools in Illinois or the states in the southwest. In addition, this method requires that there exists a *population-frame* such as a census or administrative data system. This frame must enumerate the units in the population.
2. *The population of eligibles* (E) of size M must be well defined. This population is based on any constraints or inclusion criteria, including power-analysis, financial, or practical. Examples of these constraints include (a) the power analysis dictates that only schools with at least 40 students in first grade should be included; (b) for financial reasons, it may be preferable to include only those schools close to major metropolitan areas; or (c) for practical and scientific reasons, only schools in which the curriculum under study is not currently being used are eligible for the study. It may be the case that the inference population is the same as the eligible population, that is, $E \equiv P$, but this is not required. The dataset enumerating the eligible units is called the *sampling frame*, as it is from the eligibles that the sample will be selected.
3. The size n of the sample (S), where S is a subset of E ($S \subset E$), must be determined a priori (e.g., based upon commonly available methods for power analysis).

Note that it is always true that $S \subset E$, but that $E \subset P$ is not required (though, as we develop later, generalizations require fewer assumptions when $E \subset P$). By explicitly defining each of these before recruitment begins, units can be carefully targeted and, as we show later, resources for recruitment can be allocated accordingly. When key assumptions hold, this ensures that a sample S is selected that is representative of the population P.

By selecting a sample S that is compositionally similar to the population P, the n units in the experiment can be used to get an unbiased estimate of the *population average treatment effect* for the population P (when assumptions introduced below hold). We focus here on the average treatment effect as this is by far the most commonly used estimand and the one of interest in efficacy and effectiveness trials. This estimand can more formally written as

$$\Delta = \frac{1}{N}\sum_{i=1}^{N}\Delta_i$$

where N is the number of units in the inference population P and Δ_i is the unit-specific treatment effect for units $i = 1 \ldots N$. The average treatment effect Δ is typically estimated using the n units in the sample and a multilevel model that accounts for the study design (e.g., random block design or cluster randomized design; Raudenbush & Bryk, 2002), and the impact of the experiment is evaluated by comparing this estimate to its standard error. Imai, King, and Stuart (2008) and Olsen et al. (2012) have shown that when the sample is not representative of the population and when site-specific treatment effects vary, the sample based estimate of Δ is biased. Our goal here, therefore, is to develop a strategy for selecting the n units in the sample S so that the sample is compositionally similar to the N units in the inference population P, thereby leading to a less biased and more precise estimate of the population average treatment effect.

In the statistically ideal world, after defining these three groups (P, E, S), the sample would be selected from the population using probability-sampling methods commonly used in surveys (Kish, 1987; Lohr, 1999). Probability sampling is ideal because it ensures that the sample and population are balanced on all observed and *unobserved* covariates. In contrast, nonrandom methods like the one we propose here focus on achieving balance on a set of observed covariates; this means the resulting estimator is unbiased only when certain assumptions, developed next, hold. Although ideal, because probability sampling is often infeasible in experiments, we focus instead on the development of a method for *balanced* selection. The goal of the method is to select a sample $S \subset E$, such that S and P are balanced on a set of p covariates, $\mathbf{X} = \{X_1, \ldots, X_p\}$. Here by balanced we mean that $E_S(\mathbf{X}) \approx E_P(\mathbf{X})$, where the first expectation is with respect to the sample S and the second is with respect to the population P. Note although we focus here on first moments, the concept of balancing is not limited to balancing only on first moments; polynomials or quantiles of particular covariates can also be included in $\mathbf{X}$, which is equivalent to balancing on second and higher moments. For a more detailed introduction to the goals and purpose of balanced sampling for generalization, see Tipton (in press).

An important question is how to choose these covariates for balanced selection. The following two assumptions provide general criteria.

(A1) *Unconfounded Sample Selection* (Stuart et al., 2011). The population P and sample S should be balanced on a set of covariates $\mathbf{X} = \{X_1, \ldots, X_p\}$ that contains all those that explain variation in treatment effects.

(A2) *Unconfounded Eligibility.* Let $\mathbf{G} = \{G_1, \ldots, G_{\mathrm{m}}\}$ contain all the covariates used to define the population of eligible units E. When E and P are not identical it is required that $\mathbf{G} \cap \mathbf{X} = \emptyset$. This is to say that $\mathbf{X}$ cannot contain any covariates that define the set of eligibles E.

The validity of the method proposed here depends on meeting these assumptions. If A1 is not met, it means that despite the fact that S and P are balanced on the $\mathbf{X}$ covariates, there exist other covariates that explain variation in treatment effects but upon which S and P are not balanced (leading to bias). Similarly, if A2 is not met, it means that there exist units in the population that have no comparison units in the experiment; this leads to bias when these differences are in relation to variables that moderate the treatment effects. For example, if only schools with at least 40 first graders are eligible for the study, but the population contains schools that also have less than 40 first graders, then A2 requires us to assume that the treatment effects do not vary in relation to the size of the first-grade class. If A2 cannot be met, then one option is to redefine P so that it can.

The two assumptions just stated are sufficient to permit unbiased estimation of the average treatment effect Δ in the inference population using the methods described in this article; this result is an extension to sample selection ignorability more formally developed for the post hoc case in Tipton (2013). Unfortunately it is impossible to test if these assumptions have been met before the experiment has been conducted, and even then, most tests of differential treatment impacts will be underpowered (Spybrook & Raudenbush, 2009). The fact that these assumptions are difficult to test may lead researchers to believe that this method is not useful. To this we argue three points. First, in current practice these same assumptions are being implicitly made but not stated or evaluated. This method simply formalizes this process, allowing the assumptions to be documented and debated, both before the study begins and later in study reports. Second, we argue that the usefulness of this method does not hinge on obtaining perfect balance between S and P. Current practice often leads to situations in which S and P are at most balanced on one or two crude categorical variables. Even if perfect balance is not attained, this method can substantially reduce bias by achieving balance on a larger (if not exhaustive) set of covariates that impact variation in treatment effects. Third, when exact balance is not achieved, residual imbalances can be statistically adjusted by using a post hoc subclassification estimator (Hedges & O'Muircheartaigh, 2011; Tipton, 2013).

Eligibility Issues in Sample Selection

Before proceeding with the example, one final issue can arise that is important to address. In some instances, the population of eligible units E is a subset of the inference population P. These are the easiest situations for generalization. In other instances, the population of eligible units E and the inference population P do not intersect at all. This means that all units in the inference population have zero probability of being in the experiment! We call this second type of generalization a *synthetic* generalization for this reason. Generalizations of this type are common in science and practice. For example, most psychology research takes place on college sophomores, and much of early drug research takes place only on nonhumans, yet the results of both are generalized much more broadly. These situations are easy to interpret through our framework. For example, in drug research, when mice or chimpanzees are used instead of people, in our framework this amounts to assuming that the set of covariates ($\mathbf{X}$) that explains the effectiveness of the drug does not vary in relation to species (a covariate in $\mathbf{G}$). In the psychology example, this means we must assume the set of covariates ($\mathbf{X}$) that explain variability in psychological processes does not include age or education level ($\mathbf{G}$). As our example will illustrate, this same type of problem occurs in effectiveness trials when schools which already use an intervention are included in the inference population but cannot be eligible for the study. The benefit of our approach here is that it requires these assumptions to be articulated, discussed, and debated, which in current practice rarely occurs.

Example: OCR & EM

To situate this work, throughout we focus on the particular case of how to select the sample for two scale-up studies—OCR and EM. These studies were funded by IES in 2009 and 2010, respectively, and are currently in recruitment. OCR is a core reading program for elementary school students, which Borman, Dowling, and Schneck (2008) recently evaluated employing a multisite cluster-randomized design involving 27 OCR classrooms and 22 business-as-usual classrooms across five schools. Multilevel analyses of classroom-level effects of assignment to OCR revealed statistically significant treatment effects on all three of the Comprehensive Test of Basic Skills (fifth edition), Terra Nova literacy posttests. The OCR effect sizes were $d = 0.16$ for the Reading Composite, $d = 0.19$ for Vocabulary, and $d = 0.12$ for Reading Comprehension. In addition, the curriculum is built on research-based practices cited in the National Reading Panel (2000) report and emphasizes phonemic awareness, phonics, fluency, vocabulary, and text comprehension. The EM curriculum is a comprehensive, reform-based mathematics curriculum for elementary school students. It has been researched and developed for almost three decades, used on a wide scale, and has shown promising evidence of program efficacy (Slavin & Lake, 2007; What Works Clearinghouse, 2007, 2010).

In an effort to maximize sample size and statistical power, a decision was made to combine these two trials in terms of design, recruitment, sampling frame, and study samples. In this two-stage recruitment plan, first school districts would be recruited into the study, and second within each district four elementary schools would be recruited for the study. Two of these schools would be randomly assigned to receive the OCR program at one grade level and the other two would receive the EM program at a different grade level. The opposing schools would act as the statistical controls. This design has been used in other studies, though is certainly not widespread; a key benefit is that every school in the study receives an intervention (e.g., Borman et al., 2005). With the design in place, the three groups (P, E, S) could be defined.

Inference Populations. Despite their combined design, the goal of this study is to provide separate estimates of the average causal effects of OCR and EM. This means that two separate inference populations have to be defined. The OCR population is defined to be the population of school districts in the 48 contiguous states and Washington, DC, that currently purchase any amount of the OCR program; the EM population is similarly defined. The focus on current users was made for two reasons: First, it is expected that future users will be similar to current users; second, it would allow for comparisons to be made between the treatment effects estimated in the experiment and those estimated from observational studies of current users.

Information made available from McGraw-Hill (MGH) enabled us to carefully define this user population for the 3-year period from 2008 to 2010. Of those districts purchasing the OCR and/or EM programs from MGH, 30% purchased only the OCR program, 56% purchased only the EM program, and 14% purchased both programs. The OCR population therefore consisted of the 44% of districts purchasing the OCR program, whereas the EM population consisted of the 70% of districts purchasing EM program.[1] From here forward we use P_{OCR} and P_{EM} to denote these two populations.

Population of Eligibles. The population of eligible districts for this study was determined based on two criteria: first, sales history and second, a power analysis. Districts passed the first eligibility criterion if they had not purchased either the OCR or EM programs in the previous 3 years (2008–2010). Of importance, this means that the inference populations (which consist of current users) and the population of eligibles do not intersect. Eligible units were defined this way because including current users in the experiment would jeopardize the internal validity of the study.

After meeting the first criterion, districts were then evaluated based on the availability of schools meeting selection requirements set forth by the power analysis for the study. These stipulated that districts must include at least four elementary schools with at least 44 students in each of grades kindergarten through fifth grade.

Based on these criteria, the eligible population of school districts for both the OCR and EM studies included 675 school districts across the country. It is important to note that although the inference populations differ for the two studies, as a result of the study design the population of eligible units is the same.

Sample Size. A power analysis was conducted, which determined that 15 districts would need to be recruited into the combined scale-up study, for a total of 60 schools. Therefore the goal of this article is to develop a method for selecting 15 school districts out of the 675 districts that are eligible and that best represent the OCR and EM user populations. That is, the goal is to develop a procedure for selecting a sample so that S and P_{OCR} are balanced, as are S and P_{EM}. Because the same sample must be used for two populations, several unique problems arise; we address these later.

Based on assumption A1, 11 covariates were selected that could potentially explain treatment effects for both OCR and EM. These covariates were selected based on the hypothesis that the effectiveness of the OCR and EM programs across districts might vary in relation to district resources (e.g., expenditures, location), family and community resources and context (e.g., labor force participation, location), and student characteristics (e.g.,% of ELL students). The full set of covariates is listed in Table 1. Note that although the same covariates were selected here for both OCR and EM, this is not a requirement of the method. These covariates were selected from the Common Core of Data (National Center for Education Statistics, 2011), which lists information on every school district in the United States. Table 1 also includes the means and standard deviations for the eligible districts and for the OCR and EM populations.

Finally, note that in order to meet A2, we must assume that the district average treatment effects are not functions of the number of elementary schools in the district or the size of these schools. We must also assume that the decision of districts to choose to purchase the OCR or EM programs is either fully a function of the 11 covariates or is a function of a covariate that is not associated with treatment effect heterogeneity. If either A1 or A2 have not been met, then the treatment effect estimated in the experiment will not be unbiased for the OCR and EM populations.

[1] Although for proprietary reasons we do not include the actual numbers of districts purchasing these programs from MGH, in the actual analyses these numbers were used.

Table 1. Covariates used for matching the sample to the OCR and EM populations

		Population of Eligibles		EM Population		OCR Population	
Category	Covariate	M	SD	M	SD	M	SD
Community	Educational Attainment						
	% Grade 8 or lower	9.9	7.6	7.7	5.9	11.4	9.5
	% <HS grad	15.5	6	13.5	6	16.3	6.2
	% HS grad	36.8	9.1	40.9	11.8	39	10.5
	% Postsecondary	37.9	16.4	37.8	18.4	33.3	17
Census area financials	% labor force	64.6	6.5	64.9	7.1	62	7.8
	Median income (overall)	54,046.2	18,114.2	54,452	21,061	47,940	19,042.5
	% 5–17-year-olds in poverty	13.6	9.1	11.9	9.2	16.7	11
District	Urbanicity of districts						
	% Urban	24.3	42.9	7.8	26.9	9.2	28.8
	% Rural	13.9	34.6	23.2	42.2	26	43.9
	% Suburban	34.7	47.6	29.2	45.5	20.8	40.6
	% Town	30.8	33.9	19	39.2	18.9	39.1
	Geographic location						
	% Northeast	21.2	40.9	27.5	44.7	16.1	36.8
	% Midwest	17.3	37.9	40.5	49.1	16.2	36.8
	% South	30.8	46.2	15.8	36.4	35.5	47.9
	% West	30.7	46.1	16.2	36.8	32.3	46.8
	District expenditures per student	11,691.4	3,692.1	13,506.2	8,670.7	13,529.6	10,638.4
Student	Average number of student in the district	12,322.9	11,650.6	6,605.4	2,1858	8,596.4	26,324.1
	Race/ethnicity of district						
	% White	52.2	28.7	73.4	26.5	58.2	31.2
	% Black/African American	22.9	24.8	9.6	15.6	19.5	25.2
	% Hispanic	1.1	4.3	2.2	9.8	3.7	13.4
	% other	7	9.8	5.3	11.1	7.2	14.7
	% students ELL	10.3	12.2	4.3	8.1	8.7	13
	% students F/RL	44.2	22.2	36	21.9	43.8	23.2

Source: Common Core of Data for 2010.

Note. OEM = Open Court Reading; EM = Everyday Math; ELL = English language learner; F/RL = free/reduced-price lunch.

SAMPLE SELECTION: PROPENSITY SCORE STRATIFIED SELECTION

Method for Sample Selection

Having defined P, E, S, and the covariate vector $\mathbf{X}$, in this section we develop a method for choosing a sample. The method we propose is a variation of stratified sampling, where stratification is based on the propensity score.

Stratified Selection. Imagine the simple case in which the goal is to balance S and P on two categorical covariates, each with q_1 and q_2 categories respectively, for a total of $k = q_1 \times q_2$ distinct covariate values. Define N_{j}, $j = 1, \ldots, k$ to be the number of units in stratum j of the inference population and let $N = N_1 + \cdots + N_{\mathrm{k}}$ be the total number of units in the inference population. For example, experiments commonly aim to recruit schools or districts including various degrees of urbanicity (e.g., urban, rural, suburban) and in various regions (e.g., NE, *S*, W, MW). In this simple case, stratified selection could be used to ensure that units from each of the k covariate categories were represented. To see this, the stratified estimator T_{S} for the population average treatment effect can be written,

$$T_S = \sum_{j=1}^{k} w_{pj} T_j, \tag{1}$$

where the subscripts $j = 1, \ldots, k$ define the strata, $w_{\mathrm{pj}} = N_{\mathrm{j}} / N$ are the weights each stratum receives based on the inference population, and T_{j} is an estimator of the treatment effect in stratum j. For example, T_{j} might be the simple difference in means estimator, $\bar{Y}_{T_j} - \bar{Y}_{C_j}$, where $\bar{Y}_{Tj}$ and $\bar{Y}_{Cj}$ are the treatment and control group means in the jth stratum.

An important question is how to allocate the sample to the strata. Define $n_{\mathrm{j}}, j = 1, \ldots,$ k to be the sample size in the j^{th} stratum. One method, which we advocate here, is to use *proportional allocation*, which is to let $n_{\mathrm{j}} = w_{\mathrm{pj}}$ x n. In proportional allocation, the sampling ratio $n_{\mathrm{j}} / N_{\mathrm{j}} = n/N$ is the same for all strata, and as a result, the sample is self-weighting. This means that when $T_j = \bar{Y}_{Tj} - \bar{Y}_{Cj}$, the estimator T_{S} is equal to the estimator $T = \bar{Y}_T - \bar{Y}_C$, where $\bar{Y}_T$and $\bar{Y}_C$ are the (equally weighted) means of the treatment and control sample units across all strata. Thus the weighted mean of stratum mean differences is just the (unweighted) difference between the treatment and control group means.

Finally, recall that in probabilistic sampling, stratified random sampling is commonly preferable to simple random sampling because it results in fewer "bad" (or unbalanced) samples (Kish, 1987; Lohr, 1999). This is particularly true when the sample size n is small, which is generally the case in cluster-randomized or multisite experiments where n is the number of clusters (in this case, school districts). This is to say that even if we were able to select units into the experiment using probability sampling, because of the small sample size it would be wise to do so using the stratified selection approach we develop here.

Multivariate Stratified Selection. When there are p continuous covariates $\mathbf{X}$ for matching, implementing the stratification estimator can be difficult or impossible. To see why, imagine that the range of each of the $\mathbf{X}$ covariates is divided into two values. This leads to 2^p strata, which—particularly when p is large—can easily lead to empty strata. One way to circumvent this issue is to instead reduce the dimensionality of the problem through the use of a propensity score. Propensity score methods were developed to improve balance between treatment and control groups in observational studies on a large number of covariates (Rosenbaum & Rubin, 1983, 1984). The methods have been applied to the problem of retrospective generalization in experiments, where the goal is to improve balance between the achieved sample and a newly defined inference population (Hedges & O'Muircheartaigh, 2011; Stuart et al., 2011; Tipton, 2013).

Propensity score methods require that there are two groups for matching. In the case we focus on here, in which E and P are not identical, the propensity score can be used to match the population of eligible units to the inference population units. Note that when all or most units in the population are in fact eligible for the study ($E \equiv P$) a different version of this method based on cluster analysis, instead of propensity scores, can be used instead (Tipton, in press). We define the propensity score for the $E \equiv P$ case as follows.

Eligibility Propensity Score

Let F $= E \cup P$ be the union of the eligible population E and the inference population P—a data set that contains both the M eligible units and the N population units. Let $Q = 1$ if a unit is in E. Then the eligibility propensity score is defined to be:

$$g(\mathbf{X}) = Pr(Q = 1 \,|\mathbf{X}), \text{ where } 0 < g(\mathbf{X}) < 1. \tag{2}$$

The requirement that $0 < g(\mathbf{X}) < 1$ means that for every inference population unit there must exist an eligible unit "like" them (at least in terms of values of the covariate vector **X**). In finite samples, this means that the distribution of propensity scores for the eligible units E and the inference population units P must share a *common support*, or, put another way, must *overlap* completely. When this does not occur, it means that there are population coverage problems (Tipton, 2013); this means that there are units in the population with no eligible units "like" them. Therefore the inference population should be redefined so that the condition is met.

An important property of the propensity score is that it is a *balancing score*. This means that given $g(\mathbf{X}) = g_{\text{o}}$, the conditional distributions of **X** for units in P and E are the same. This is especially important here, where the propensity score itself will not have any obvious substantive meaning. Finally, note that propensity scores can be estimated using a variety of strategies, the simplest of which is logistic regression.

Once the propensity score has been obtained, the stratified selection approach given above can be applied. That is, the distribution of the propensity scores in the population P can be divided into $j = 1, \ldots, k$ strata, where each stratum contains N_{j} of the N population units and M_{j} of the M eligible units. Under proportional allocation, because each stratum contains $w_{\text{pj}} = N_{\text{j}} /N$ of the population units, the sample is divided so that each stratum receives $n_{\text{j}} = w_{\text{pj}} \times n$ of the sample units. An important question is how to choose the number of strata k and the stratum boundaries. In observational studies, research suggests that using five strata such that $w_{\text{pj}} = 1/k$ is often close to optimal (Cochran, 1968), though certainly more strata are generally better. The discussion section highlights further the role that strata play here.

Finally, note that although $w_{\text{pj}} = N_{\text{j}} /N = n_{\text{j}} /n$ in the case studied here it is generally true that $w_{\text{pj}} \neq M_{\text{j}} /M$, which is to say that the population of eligible units will be allocated to the strata differently than the inference population units. A good measure of this discrepancy is the stratum specific selection fractions, $f_{\text{j}} = n_{\text{j}} /M_{\text{j}}$. In practice, these f_{j}'s can vary considerably; we discuss the implications of this next.

Within-Stratum Sample Selection. For a given stratum j, once it is determined that n_{j} of the M_{j} units must be included in the sample, an important question is how these n_{j} units should be selected. The statistically ideal method would be to use random sampling without replacement. In this case, the probability of selecting a particular sample of n_{j} units is $n_{\text{j}}!(M_{\text{j}} - n_{\text{j}})!/M_{\text{j}}!$ where ! indicates the factorial function. A difficulty with this approach is that often many units will not agree to be in the experiment, which means the nonresponse rate will be high. If the nonresponse can be assumed to be a function of the covariates in **X** and units in the same stratum are nearly homogenous on **X**, then the nonresponse can be considered missing at random (Little & Rubin, 2002). When the strata are not homogenous on **X**, then the units that agree to be in the experiment may be significantly different from the average unit in the population in the stratum. As a result, the nonresponse would need to be adjusted for, which could be difficult when the sample size in the stratum (n_j) is small (Groves et al., 2009). For this reason, we provide an alternative strategy.

When nonresponse is expected to be high, another possible strategy for targeting units for recruitment is to use a measure of within stratum distance. That is, for each eligible unit $i = 1, \ldots, M_j$, let d_{ij} be the distance between the unit i and the stratum average for the inference population. This distance could be based on the Euclidean distance between the propensity scores, $d_{ij} = (g_{ij} - g_{\bullet j})^2$ where $g_{\bullet j}$ is the average propensity value in stratum j for the population, or might be based on the Mahalonobis distance $d_{ij} = (\mathbf{X}_{ij} - \mathbf{X}_{\bullet j})' \, \boldsymbol{\Sigma}_j{}^{-1}(\mathbf{X}_{ij} - \mathbf{X}_{\bullet j})$, where $\mathbf{X}_{\bullet j}$ is the matrix of stratum averages for the covariates in **X** in the population, and $\boldsymbol{\Sigma}_j$ is the covariance matrix for **X** in the population. Based on these distances, units in the stratum would then be ranked from most to least desirable for inclusion in the sample. This means recruiters would begin recruitment efforts with the top of the list and work down until n_j units agreed to be in the study. A benefit of this approach is that the list only needs to be created once. In addition, this strategy prioritizes units that are close to average over units that are closer to the stratum boundaries.

Recruitment Resource Allocation. There are two ways that using the stratified selection approach we develop here can help with recruitment in terms of targeting and resource allocation. The first of these is that units in the same stratum are more similar than units in different strata. This means that recruitment materials and approaches can be targeted to the unique needs of the k strata. For example, units in stratum 1 might include mostly urban schools in the northeast, whereas those in stratum 2 include rural schools in the south. This information can allow for targeted recruitment tactics that may be more successful in increasing participation rates.

The second way this approach can help is through resource allocation; examples of resources include financial incentives to the units, overall time devoted to recruitment efforts, or money spent on materials targeting the units. To see how this works, note that the $f_j = n_j / M_j$ selection fractions often vary considerably. Imagine the following simplified problem in which there are only two strata. In Stratum 1, $M_1 = 10$ and $f_1 = 0.20$, whereas in Stratum 2, $M_2 = 100$ and $f_2 = 0.01$. This means that in Stratum 1, two units must be selected out of only 10 eligible units, whereas in Stratum 2, only one unit must be selected out of 100 units. Clearly it would be easier to achieve the required sample in Stratum 2, where 99 units can decline to participate, so long as one unit agrees to do so. This is to say that each eligible unit within Stratum 1 is more essential for recruitment than are the eligible units in Stratum 2.

In light of this, one way to allocate resources is as follows. Let R be the total resources and let R_j be the amount of those resources allocated to stratum j. Taking into account the selection fractions, resources could be allocated so that $R_j = R \times f_j / \Sigma f_j$. This means that strata with larger selection fractions would receive a larger proportion of the resources.

Flexibility in the Approach. Finally, a benefit of this approach is that it is flexible. For example, it may be that in one stratum, for example, the n_j units that agreed to be in the sample were at the bottom of the rankings, indicating that their covariate values differed largely from the stratum mean. When the strata are pooled together, this can create residual imbalances that lead to the situation in which, despite all efforts, the sample S and population P are not well balanced. One way to address this is toward the end of recruitment, when all but a few units have been successfully recruited into the study. Using measures of propensity score difference, these last few units can be targeted to improve the balance of the sample. For example, suppose that all but two units have been recruited and the average propensity score is 0.10 in the sample but 0.13 in the inference population. Then the last units targeted should have average propensity scores equal to $[n \times 0.13 - (n - 2) \times 0.10]/2$, so that when $n = 20$ these last two unit should be targeted to have a propensity score of $[20 \times 0.13 - (20 - 2) \times 0.10]/2 = 0.40$.

Application to OCR & EM

The method proposed here was developed in relation to the problem of how to choose the OCR and EM sample. This example, however, is more complex than the standard sample selection plan, as it involves selecting a sample that achieves approximate balance for two separate populations, P_{OCR} and P_{EM}. We include this as our example, however, because it highlights the flexibility and usefulness of this approach. To make it easier to understand and more applicable to simpler cases, we first develop a selection plan for the P_{OCR} population on its own, and then develop what the plan would be in the two-population case.

Single Population Case (OCR). In the single population case, the problem is to select $n = 15$ districts from $M = 675$ eligible districts so that S and P_{OCR} are balanced on 11 important covariates likely to explain variation in treatment effects (those that meet A1). Recall that the population of OCR users includes districts that purchased any amount of the OCR program from MGH in the 2008–2010 period. To do this, we used the following procedure.

First, we estimated the propensity scores using a simple logistic regression model,

$$logit(g(\mathbf{X})) = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p, \tag{3}$$

where there are $p = 18$ covariates representing the covariates found in Table 1 (i.e., some of the covariates have multiple levels). Using the estimates of $\beta_0, \ldots, \beta_p$ from this model, we calculated the district-specific propensity scores $g_{\text{i}} = g_{\text{i}}(\mathbf{X})$ and logit propensities $l_{OCRi} = logit[g_{\text{i}}]$ for each of the M eligible units and N population units.

Based on the distribution of the propensity score logits, l_{OCRi} in the population P_{OCR}, we defined $k = 3$ strata so that each stratum contained one third of the population and sample cases. We use $k = 3$ strata here for consistency with the two population case treated next; more generally, we would prefer to use more strata, as additional strata lead to greater balance between the sample and inference population. This means that from each stratum, $n_{\text{j}} = 5$ districts will need to be recruited. In addition, for each stratum, we also calculated the stratum selection fraction, $f_{\text{j}} = n_{\text{j}} / M_{\text{j}}$, and the proportion of resources that should be allocated to each stratum based on these, using $f_{\text{j}} / (f_1 + f_2 + f_3)$. Table 2 summarizes this information.

As Table 2 shows, Stratum 3 in this study has a selection fraction of $f_3 = 0.167$, which translates into selecting five districts for the experiment from a pool of 30 eligible districts. This selection fraction is considerably larger than that in Stratum 1, which is $f_1 = 0.011$, where only five districts need to be selected from a pool of 465 eligible districts. As a result, Stratum 3 should receive roughly 77% of the recruitment resources.

Table 2. Sample allocation plan for the separate OCR study

OCR Stratum	Number of Eligible Districts (M_j)	Number of Districts Needed in the Sample (n_j)	Number of Districts in OCR Population (N_j)	Sampling fraction (f_j)	Proportion of Resources (R_j/R)
1	465	5	724	0.011	0.049
2	124	5	724	0.040	0.185
3	30	5	725	0.167	0.765

Finally, within each of the three strata, we calculated the stratum averages on the propensity score logits,

$$L_j = \frac{1}{N_j}\sum_{i=1}^{N_j} l_{ij}, \tag{4}$$

where the average is for districts in P_{OCR} only. For each unit in the stratum we then calculated the distance $d_{ij} = (l_{ij} - L_{\square j})^2$ and ranked the units from smallest to largest distance. In the two-population case we address next, we show an example of one of these ranked lists. In practice, Table 2 and a ranked list for each stratum would be given to recruiters as tools to guide recruitment efforts.

Two-Population Case (OCR & EM). The two-population case is only slightly more complex than the one population case. Here two separate propensity score models were run using the logistic regression model given in Equation 1. The first of these compared E with P_{OCR}, whereas the second compared E with P_{EM}. This results in each of the eligible units having a vector of estimated logits, $\boldsymbol{l}_i = (l_{OCRi}, l_{EMi})$.

Just as in the single population case, for each of the populations separately, the logits were then divided into $k_{OCR} = k_{EM} = 3$ strata based on the distributions of the l_{OCRi} and l_{EMi} in the respective inference populations. When the $M = 675$ logit pairs $\boldsymbol{l}_i$ for E are plotted (bivariate) this reveals that there are $k = 9 = 3 \times 3$ total strata. These 675 districts and the nine strata are shown in Figure 1. It is important to note that the figure includes only the eligible units, not the population units. We do this to simplify the graphic.

Because $n = 15$ and $k_{OCR} = k_{EM} = 3$, each of the OCR and EM marginal strata were allocated 15/3 = 5 districts for recruitment into the study. Using iterative proportional fitting (with rounding), the sample was allocated to the nine bivariate strata (Deming & Stephan, 1940). The iterative proportional fitting routine took into account the number of eligible units in each of the nine strata. Table 3 shows the number of districts that must be recruited relative to the number of eligible units for each of these nine bivariate-strata. Note that one stratum does not contain any eligible units. This table also includes a measure of the proportion of resources that should be contributed to recruitment efforts in each stratum, based on $f_j / \Sigma f_j$. We order the strata in this table accordingly.

Finally, within each of the eight nonempty strata, districts were ranked from most to least desirable for inclusion in the sample. To do this, for each eligible district $i = 1, \ldots, M_j$ in stratum j, we used the (squared) Euclidean distance $d_{ij} = d_{OCRij} + d_{EMij}$, where $d_{OCRij} = (l_{OCRij} - L_{OCR\bullet j})^2$ and $d_{EMij} = (l_{EMij} - L_{EM\bullet j})^2$. Here $L_{OCR\bullet j}$ is the average value of l_{OCRi}

Table 3. Example of ranked output for Stratum 1 in OCR and EM combined experiment

ID	EM EM Subclass	OCR Subclass	Rank	State	Agency Name	EM logit distance (l_{EMij})	OCR logit distance (l_{OCRij})	EM distance (d_{EMij})	OCR distance (d_{OCRij})	Combined distance (d_{ij})
3619950	3	3	1	NY	BEDFORD CENTRAL SCHOOL DISTRICT	2.84	2.08	11.35	2.48	13.83
3628560	3	3	2	NY	SYOSSET CENTRAL SCHOOL DISTRICT	3.02	1.90	16.47	1.34	17.81
3609990	3	3	3	NY	EAST SYRACUSE-MINOA CENTRAL SCHOOL DISTRICT	3.09	1.94	18.82	1.54	20.36
3904517	3	3	4	OH	ZANESVILLE CITY	3.10	1.96	19.39	1.66	21.05
1905790	3	3	5	IA	BURLINGTON COMM SCHOOL DISTRICT	3.31	1.79	28.10	0.88	28.98
1907710	3	3	6	IA	CLINTON COMM SCHOOL DISTRICT	3.33	1.85	29.00	1.13	30.14
2104800	3	3	7	KY	PIKE COUNTY	2.77	3.02	9.71	23.22	32.93
3613740	3	3	8	NY	HARRISON CENTRAL SCHOOL DISTRICT	3.32	2.27	28.66	4.41	33.07
3904384	3	3	9	OH	DAYTON CITY	3.45	2.38	35.84	5.83	41.67
901770	3	3	10	CT	GROTON SCHOOL DISTRICT	3.63	2.93	47.60	19.70	67.31
3904516	3	3	11	OH	YOUNGSTOWN CITY SCHOOLS	3.72	2.92	54.02	19.35	73.38
2201170	3	3	12	LA	ORLEANS PARISH	3.31	3.92	28.13	91.71	119.84

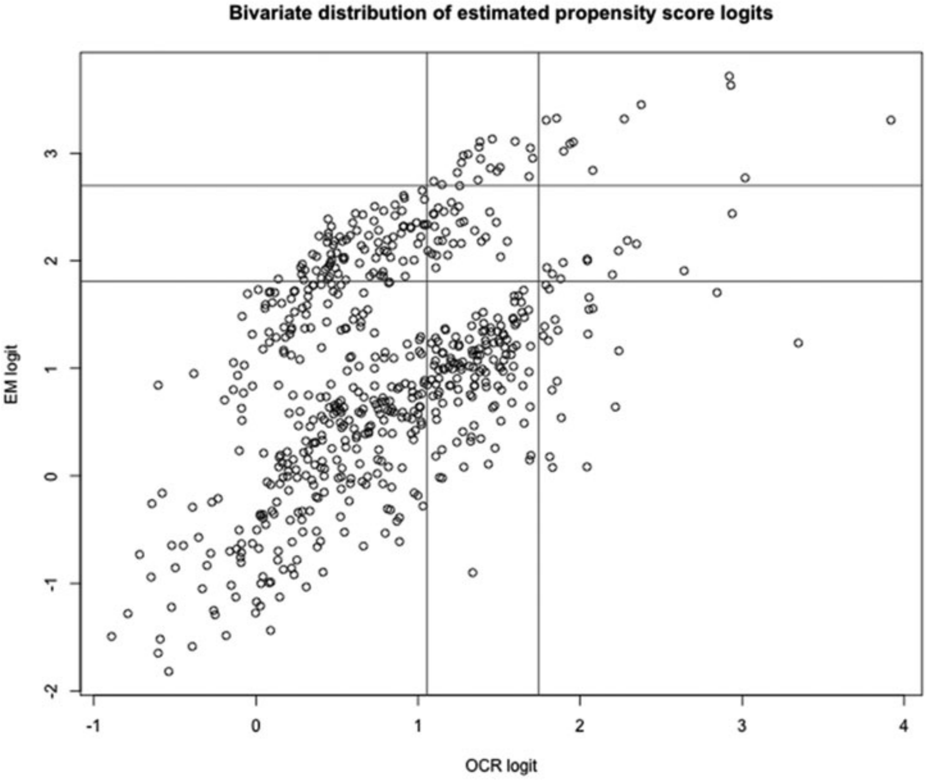

Figure 1. Bivariate distribution and nine strata for Open Court Reading (OCR) and Everyday Math (EM) study.

for districts in the P_{OCR} stratum that contains stratum *j*, and $L_{EM\bullet \mathrm{j}}$ is defined similarly. For example, in Table 3, stratum $j = 2$ is in the third marginal P_{OCR} stratum and the second marginal P_{EM} stratum. Note that as a result of the design of this experiment, exact balance cannot be achieved for either the P_{OCR} or P_{EM} populations because the same sample is used for both. Therefore, we have chosen to focus on approximate balance and to split the difference between the two; as such, we weight both distances, d_{OCRij} and d_{EMij}, equally. In Table 4 we illustrate what a table of results for these rankings might look like for a stratum.

Finally, note that in other situations, it may instead be desirable to let $d_{ij} = \lambda d_{\mathrm{OCRij}} + (1 - \lambda) d_{\mathrm{EMij}}$, for some $0 \leq \lambda \leq 1$. For example, when $\lambda > 1/2$, the primary goal is to achieve the best balance for P_{OCR}, and the secondary goal is to achieve approximate balance for P_{EM}. Here we use $\lambda = 1/2$.

Based on this analysis the recruitment team was given three instructions. First, sample from the most difficult strata first and give a greater proportion of resources to recruitment efforts in these strata. Second, within each stratum, use the distance ranks to determine the order in which districts are contacted or considered for recruitment. Third, when the last stratum (here Stratum 8) is reached, recalculate distances to offset any residual imbalances that arise from sub-optimal recruitment in the other strata.

Table 4. Sample allocation plan for the combined OCR & EM studies

Stratum	OCR Stratum	EM Stratum	Number of Eligible Districts (M_j)	Number of Districts Needed in the Sample (n_j)	Sampling fraction (f_j)	Proportion of Resources (R_j/R)
1	3	3	12	3	0.250	42%
2	3	2	18	2	0.111	18%
3	2	3	12	1	0.083	14%
4	2	2	29	2	0.069	11%
5	1	3	21	1	0.048	8%
6	2	1	83	2	0.024	4%
7	1	2	132	1	0.008	1%
8	1	1	312	3	0.010	2%
9	3	1	0	0	0.000	0%

DISCUSSION

When using this sample selection method in practice, several important questions and issues arise. Because this approach is new, in this section we highlight some of the benefits of the approach and address particular concerns that are likely to arise.

Defining the Inference Population(s) and Eligibility Criteria

One of the strengths of this method is that it requires researchers to carefully define the inference population before the sample selection begins. For example, in the OCR and EM cases, we defined two inference populations: the population of school districts currently using OCR, and the population of school districts currently using EM. These inference populations were determined based on the assumption that future users of the programs are likely to be similar to those currently using the programs, and based on the desire to obtain evaluation results that may generalize to the current population of districts using OCR and EM. This assumption is based on the fact that both OCR and EM are widely known programs and are more likely to meet the needs of and appeal to particular types of school districts. It is important to note that other inference populations could have easily been chosen. For example, we could have focused on the entire population of school districts in the nation, or on only the subset of schools receiving Title I funds. Although agreement may not always be reached on what the best population is, the benefit of this approach is that the results of the experiment will generalize to *at least one* well-defined population, which is often not true of current practice.

Second, the method requires that eligibility criteria are carefully delineated and explained. In both the OCR and EM studies, which were combined for eligibility and recruitment, there were two requirements. The first came out of a concern for the internal validity of the experiment—that districts in the study could not currently be using either the OCR or EM curricula. The second requirement came out of the power analysis—that only districts with at least four elementary schools, each with at least 44 students in each of grades K–5, could be in the study. Our experience suggests that these two types of criteria—internal validity and power concerns—are likely to occur in most scale-up studies. For these reasons, we argue that the eligibles E and the inference population P will rarely coincide completely, making the method proposed here widely useful.

Unconfounded Sample Selection and Eligibility

The method we propose requires two important assumptions. The first, A1, requires that the sample selection is unconfounded; this means that our covariate vector $\mathbf{X}$ used for defining and estimating the propensity score $g(\mathbf{X})$ includes all the covariates that explain variation in potential treatment effects at the level of sample selection. For example, in the OCR and EM studies, $\mathbf{X}$ must include all covariates that explain variation in *district average treatment effects*.

Clearly an important concern is that in practice, this assumption may not always be met. Our ability to stratify on important covariates is limited by the population frames currently available at the national or state levels. On the plus side, certainly more data are available about the national population of school districts and schools than are available at the student level, making this approach particularly useful for selecting sites in cluster randomized or multisite studies. On the minus side, the type of covariates available tend to be demographic in nature. It would certainly be more ideal if this degree of data were available on proximal measures like student achievement, student motivation, or subject specific measures appropriate to the curriculum or program under study. However, we argue that although the list of covariates available for $\mathbf{X}$ likely does not include every variable of interest, by balancing the sample and population on 11 covariates, we are certainly better off than not balancing them at all (or on only one or two variables).

The second assumption, A2, is that eligibility is unconfounded. This means that the covariates that determine eligibility cannot also explain variation in potential treatment effects, conditional on the other covariates $\mathbf{X}$ included in the model. In the OCR and EM examples, this means that we must assume that holding constant the other 11 covariates in $\mathbf{X}$, district-average treatment effects do not differ for school districts that already use the respective program versus those that do not. Similarly, we must assume that district-average treatment effects do not vary in relation to district size (measured by both the number of elementary schools and the number of students per grade in these schools), holding constant the other 11 variables in $\mathbf{X}$. If it is believed that these assumptions cannot hold (e.g., if treatment effects were believed to be different for small districts), then inference could not be made using this method. However, inference would be possible for different inference populations—the inference population P could be redefined to exclude small districts, for example, or the power analysis recalculated to take into account a wider range of district and school sizes.

Strata Definitions, Ranking, and Within-Stratum Selection

The benefits of stratified random sampling, as compared to simple random sampling, are widely known and appreciated in the survey sampling context (Lohr, 1999). In large surveys—where units are randomly selected using either approach—the stratified random sampling approach is generally preferable because it leads to smaller sampling variance. Concerns about bias are not of importance in the survey sampling context, as by virtue of probability sampling, both methods lead to unbiased estimators.

It is important to highlight that the stratified selection approach introduced here is developed in a different context and serves a different purpose. First, the approach we are comparing this to is not that of a simple random sample, but instead of a convenience sample selected without relation to a well-defined inference population. If it is true that the potential treatment effects vary in relation to the covariates in **X**, then to the extent that the convenience sample and a particular inference population are unbalanced on these covariates, bias is induced. Here we use the stratified selection approach not to reduce variance but instead to reduce bias. By carefully defining an inference population, the covariates in **X**, the *k* strata, and the distance measures within strata, the goal of the method is to push toward a realized sample that is balanced on all covariates in **X**.

Because the goal of the method is to achieve balance, the more strata that are used, the more robust the method is toward the functional form of the relationship between the covariates in **X** and the potential treatment effects. For example, with only one stratum, balance may be achieved between the sample *S* and population *P* when the relationship between **X** and the potential effects is *linear*. However, this is a strong functional form assumption and is easily violated if the relationship is nonlinear. However, by using three or four or five (or more) strata, we can achieve greater balance between the sample *S* and inference population *P* on not just **X** but also on other moments (e.g., squares). This means that the use of additional strata makes the method more robust to model misspecification, which is desirable.

In addition, as the number of strata *k* increases, the sample also becomes more heterogeneous (on the covariates in **X**). This has two effects. First, if imbalances remain between the sample *S* and the population *P*, these can more easily be adjusted using a poststratification estimator with minimal costs in terms of increased sampling variance (Tipton, 2013). Second, because the variability in **X** is larger, the likelihood of a coverage error will be reduced, so that generalizations using the poststratification estimator to other populations will be better (Tipton, in press). This is important, as in many studies more than one inference population is of interest.

Finally, although the method we've focused on here is stratified selection using distance rankings within strata, it is certainly true that stratified sampling with random selection within strata could be used. Clearly the benefit of this approach is that the random selection of units would mean that the sample would be balanced not just on the covariates in **X** but on unobserved covariates as well. Although random selection is clearly preferable to any nonrandom selection process (ours included), it is important to highlight that both selection procedures require dealing with nonresponse. When units are selected randomly, the nonresponse process is typically accounted and adjusted for during post hoc analyses using methods that include propensity score matching on a set of observed variables. To the extent that the set of covariates used to meet A1 with our distance matching routine also include variables related to nonresponse, one benefit of our approach is that it requires these covariates to be accounted for *prior* to selection instead of primarily through post hoc adjustments (though clearly this could also be accounted for using random selection of units as well).

Design and Analysis Issues

This article has focused on developing a stratified selection method for selecting a balanced sample. Once the sample has been selected, an important question is if and how the study design and analysis should take into account the sample selection process. In general, there are two different approaches to analyzing data under selection and assignment processes—randomization- versus model-based inferences. In survey sampling design-based inferences (based on the random selection probabilities) are most common, whereas in experimental design, model-based inferences (using analysis of variance models or their multilevel counterparts) prevail (Fienberg & Tanur, 1987). The model-based method commonly used in the analysis of effectiveness and efficacy trials uses a superpopulation perspective. In this framework, inferences regarding the population average treatment effect Δ are considered unbiased whenever the n units in the sample S and the N units in the population P can both be considered finite samples from the same theoretical superpopulation. So long as the model is correctly specified—where by model we here mean assumptions A1 and A2 have been met—this means it makes no difference if the data are drawn randomly or not (Duncan, 2008; Neyman, 1934). The benefit of this perspective is that the sample selection process does not need to be directly accounted for in the estimation of the treatment effect or standard errors, which is to say that the same estimators currently standard in the field can continue to be used.

An additional question is if and how the strata used for selection should be accounted for in the study design. In the OCR and EM cases studied here, which are multisite studies in which the treatment is assigned within each school district, the additional level of stratification amounts to blocking at a level higher than treatment assignment. When estimating the district-average treatment effect, here the precision gains to including the strata in the analysis are minimal.

In study designs where clusters (e.g., schools) are first selected via the stratified selection method and later randomized to treatment conditions, there are more clear benefits to including the strata in the experimental design. Here the question is if the clusters should be randomized to treatment conditions without respect to the strata, or if randomization to treatment should occur within each stratum. We note that there are two advantages to randomizing within each stratum. First, there are precision gains, because the strata act as blocks, reducing sampling variance. Second, separate stratum specific treatment effects can be estimated, allowing for a measure of treatment effect variability to be estimated. If this is desired, then the number of strata in the final sample will need to be defined before sample selection begins in order to correctly estimate the statistical power of the design for testing hypotheses about the population average treatment effect.

CONCLUSION

In this article we present a new method for sample selection for scale-up experiments. An important feature of this method is that it requires researchers to very explicitly define who they believe the inference population should be and how this population relates to the definition of eligibility for the experiment. The method we present here applies when the population of eligible units and the inference population do not fully overlap. In these cases, particular assumptions are required, which this method articulates. This is an important feature of the method—it makes explicit the assumptions required for generalization from an experiment.

The method developed here is a combination of a stratified selection approach and propensity score matching methods. The goal is to create a sample that is similar in composition to a well-defined inference population. The method we present is flexible and practical in the sense that it identifies units to be targeted for recruitment, and when they are not available, identifies similar units for replacement. In addition, this method helps researchers determine which areas of the population may be most difficult to recruit from, enabling recruitment resources to be deployed strategically throughout the recruitment period.

REFERENCES

Borman, G. D., Dowling, N. M., & Schneck, C. (2008). A multi-site cluster randomized field trial of Open Court Reading. *Educational Evaluation and Policy Analysis*, *30*, 389–407.

Borman, G. D., Slavin, R. E., Cheung, A., Chamberlain, A. M., Madden, N. A., & Chambers, B. (2005). Success for All: First-year results from the National Randomized Field Trial. *Educational Evaluation and Policy Analysis*, *27*, 1–22.

Boruch, R. F. (1996) *Randomized experiments for planning and evaluation: A practical guide* (Applied Social Research Methods Series, Vol. 44). Thousand Oaks, CA: Sage.

Cochran, W. G. (1968). The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics*, *24*, 295–313.

Cook, T. D. (1993) A quasi-sampling theory of the generalization of causal relationships. In L. Sechrest & A. G. Scott (Eds.), *Understanding causes and generalizing about them: New directions for program evaluation* (pp. 39–82). San Francisco, CA: Jossey-Bass.

Cornfield, J., & Tukey, J. W. (1956). Average values of mean squares in factorials. *The Annals of Mathematical Statistics*, *27*, 907–949.

Deming, W. E., & Stephan, F. F. (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *The Annals of Mathematical Statistics*, *11*, 427–444.

Duncan, G. J. (2008). When to promote, and when to avoid, a population perspective. *Demography*, *45*, 763–784.

Fienberg, S. E., & Tanur, J. M. (1987). Experimental and sampling structures: Parallels diverging and meeting. *International Statistical Review*, *55*, 75–96.

Groves, R., Dillman, D., Eltinge, J., & Little, R. J. A. (2002). *Survey nonresponse*. New York, NY: Wiley.

Groves, R. M., Fowler, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2009). *Survey methodology* (2nd ed.). Hoboken, NJ: Wiley.

Hedges, L. V., & O'Muircheartaigh, C. A. (2011). *Improving generalizations from designed experiments* (Working paper). Northwestern University, Evanston, IL.

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

Institute for Education Sciences. (2011). Funding opportunities: Search funded research and contracts. Retrieved from http://ies.ed.gov/funding/grantsearch

Kalton, G. (1968). Standardization: A technique to control for extraneous variables. *Applied Statistics*, *17*, 118–136.

Kish, L. (1987). *Statistical design for research* (Wiley Series in Probability and Mathematical Statistics). New York, NY: Wiley.

Kitagawa, E. M. (1964). Standardized comparisons in population research. *Demography*, *1*, 296–315.

Kruskal, W. H., & Mosteller, F. (1979). Representative sampling. III. The current statistical literature. *International Statistical Review*, *47*, 245–265.

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (Wiley Series in Probability and Mathematical Statistics). New York, NY: Wiley.

Lohr, S. (1999) *Sampling: Design and analysis*. New York, NY: Duxbury.

National Center for Education Statistics. (2011). *Common core of data*. Retrieved from http://nces.ed.gov/ccd

National Reading Panel. (2000). *Report of the National Reading Panel: Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its* *implications for reading instruction*. Washington, DC: National Institute of Child Health and Human Development.

Neyman, J. (1934). On the two different aspects of the representative method: The method of stratified sampling and of purposive selection. *Journal of the Royal Statistical Society*, *97*, 558–625.

Oaxaca, R. (1973). Male–female wage differentials in urban labor markets, *International Economic Review*, *14*, 693–709.

Olsen, R. B., Orr, L. L., Bell, S. H., & Stuart, E. A. (2012). External validity in policy evaluations that choose sites purposively. *Journal of Policy Analysis and Management*. Advance online publication. doi:10.1002/pam.21660

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models* (2nd ed.). Thousand Oaks, CA: Sage.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*, 41–55. doi:10.1093/biomet/70.1.41

Rosenbaum, P. R., & Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association*, *79*, 516–524. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/20199225

Rosenberg, P. (1962). Test factor standardization as a method of interpretation. *Social Forces*, *41*, 53–61.

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

Schneider, B., & McDonald, S. K. (2007). *Scale-up in education: Ideas in principle* (Vol. 1). Lanham, MD: Rowman & Littlefield.

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

Slavin, R. E., & Lake, C. (2007). *Effective programs in elementary mathematics: A best- evidence synthesis*. Baltimore, MD: Johns Hopkins University.

Spybrook, J., & Raudenbush, S. W. (2009). An examination of the precision and technical adequacy of the first wave of group randomized trials funded by the Institute of Education Sciences. *Educational Evaluation and Policy Analysis*, *31*, 298–318.

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. *Journal of the Royal Statistical Society, Series A*, Part *2*, 369–386.

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

Tipton, E. (in press). *Stratified sampling using cluster analysis: A balanced-sampling strategy for improved generalizations from experiments*. Evaluation review.

Valliant, R., Dorfman, A. H., & Royall, R. M. (2000). *Finite population sampling and inference: A prediction approach* (Wiley Series in Probability and Statistics). New York, NY: Wiley.

What Works Clearinghouse. (2007). *Elementary school math*. Retrieved from http://ies.ed.gov/ncee/wwc/reports/elementary_math/topic/references.asp

What Works Clearinghouse. (2010). *Intervention: Everyday mathematics*. Retrieved from http://ies.ed.gov/ncee/wwc/reports/elementary_math/eday_math/