Assessing sampling methods for generalization from RCTs: Modeling recruitment and participation

Gleb Furman[1] & James E. Pustejovsky[2]

[1] University of Texas at Austin

[2] University of Wisconsin-Madison

Abstract

Educational research aimed at informing policy decisions should ideally be designed to inform causal inferences at the population level. Large scale, multi-site randomized trials often rely on vague convenience sampling methodology when recruiting districts and schools, resulting in relatively homogeneous samples that may differ greatly from the intended population of interest. Retrospective methods that quantify and statistically adjust for those differences are promising but may have difficulty overcoming substantial selection bias. Designing sampling methods that focus on generalizability may be a more effective approach. However, there is a lack of methodological research examining the effectiveness of such strategies in education research contexts. We propose a framework for conducting such research based on formal models for study recruitment and participation. Using this framework, we then examine one promising method, stratified balanced sampling (SBS), in the context of recruiting a representative sample of schools for a large trial. Using simulations based on real sampling frames, we compare SBS to stratified and unstratified versions of convenience sampling and probability sampling. Under our modeling assumptions, we find that SBS and stratified random sampling result in highly generalizable samples. These methods are extremely costly to implement, however, especially when the population average willingness to participate is low. Stratified convenience sampling represents a potential compromise.

One more sentence about future research needs.

Keywords: generalizability, sampling, MRT

Assessing sampling methods for generalization from RCTs: Modeling recruitment and participation

Introduction

The multi-site randomized trial (MRT) has become a common design for evaluating the effectiveness of educational interventions. An MRT is a randomized control trial that takes place across multiple distinct sites, such as school districts, medical clinics, or geographic areas, with random assignment taking place either at the site or unit level. In education research, an MRT might entail recruiting multiple schools in each of several districts. Once a sample of schools is recruited, students, classes, teachers, or whole schools are randomly assigned either to receive an intervention (treatment group) or to an alternative condition, such as continuing business-as-usual (control group).

Well-executed MRTs with low attrition support a high level of internal validity, in that we can be confident that observed differences in outcomes between the treatment and control groups represent causal effects of treatment assignment. One might also expect that running a trial across multiple sites would support a high level of external validity, or generalizability, because it provides a basis for detecting and investigating cross-site variation in treatment effects (Raudenbush & Bloom, 2015). Ostensibly, this features allows researchers to generalize findings to a larger population than estimates from a single-site design (Raudenbush & Liu, 2000). However, such claims rest on strong and rather ambiguous assumptions about the goal of generalization.

In the context of intervention studies, generalizability describes how well the effect of an intervention would hold for units outside of the study. For instance, results of a randomized trial are often interpreted as estimates of the sample average treatment effect (SATE), or the average effect of intervention for the set of units that actually participated in the study (Gerber & Green, 2012). If the study sample is representative of a larger population, a SATE estimate can also be interpreted as a reasonable estimate of the average effect of intervention within the larger population of units, or population average treatment effect (PATE), without any additional adjustments. However, if the sample is compositionally different from the population of interest, and if response to intervention varies across units or sites in the population, then the SATE no longer provides an unbiased estimate of the PATE.

Cites supporting this?

Several studies have found that schools and districts that participate in large scale randomized trials differ substantially from the national population and from policy-relevant sub-populations (Fellers, 2017; Stuart, Bell, Ebnesajjad, Olsen, & Orr, 2017). This suggests that current research practices yield non-representative samples. In turn, non-representative samples may lead to substantially biased estimates of population-level effects (Olsen, Orr, Bell, & Stuart, 2013; Shadish, Cook, & Campbell, 2002). Beyond accurate effects estimates, generalizability can also be a question of equity. For instance, small under-served rural districts are underrepresented in RCTs sponsored by Institute of Education Sciences (Fellers, 2017; Stuart et al., 2017), and thus may be less likely to benefit from federally funded education research.

Update with recent refs

One way to achieve strong generalizability is to select sites from a well-defined population with known probabilities of selection. Assuming random assignment with full compliance and low attrition, this design enables unbiased estimation of the SATE. Using the known sampling probabilities, the SATE can then be adjusted to estimate the PATE. Unfortunately, probability sampling is rarely used in large-scale impact evaluations (Olsen et al., 2013; Shadish et al., 2002). Instead, researchers often opt for convenience or purposive sampling. These methods are much less expensive to implement, but are not usually designed to achieve population representation.

Retrospective generalization methods

A growing body of methodological research has considered how to estimate PATEs from non-representative samples using retrospective propensity score analysis (Kern, Stuart, Hill, & Green, 2016; O'Muircheartaigh & Hedges, 2014; Stuart, Cole, Bradshaw, & Leaf, 2011; Tipton, 2013a). In order to justify a claim of generalizabilty, these methods rely on the availability of data for a clearly specified inference population. The data must include covariates that predict treatment effect heterogeneity. The methods then work by re-weighting the sample (or otherwise adjusting) to generate estimates of the PATE. A shortcoming of retrospective generalization methods is under-coverage (Groves, 2004) which occurs when a sample lacks units corresponding to a sub-group of the inference population. Under-coverage can be assessed using several techniques that identify how well a sample would generalize to a specific population (Stuart et al., 2011; Tipton, 2014). If under-coverage is great enough to prevent use of statistical adjustment for estimating the PATE, then the inference population must be redefined. However, re-defining the inference population diminishes the relevance of study results, undermining the substantial investment into large-scale MRTs.

Stratified balanced sampling

Rather than relying on retrospective adjustments, a series of recent papers instead advocate designing robust sampling methods that focus on generalizability at the recruitment stage (Tipton, 2013a, 2013b). Like retrospective adjustment, these methods also require a well-defined and enumerated population for which there is extant data, making them especially relevant in the educational context. One method in particular, Stratified Balanced Sampling (SBS), has received attention from intervention effectiveness researchers due to its accessibility.

Weird way to bring this point up. Move closer to top?

SBS involves using cluster analysis to split the population into smaller, more homogeneous strata. Sites within each stratum are then ranked according to how well they represent the stratum and highly representative sites are prioritized for recruitment. Researchers who are interested in using this to sample schools may even use a website (www.thegeneralizer.org) that guides them through this process using data from the Common Core of Data.

Several studies have described potential theoretical benefits of SBS (Tipton, 2013b; Tipton et al., 2016). First, SBS requires that researchers specify the inference population and justify targeting specific sites. Thus, it enforces a degree of clarity regarding the goal of generalization, which has often been lacking even in large-scale MRTs in education (Tipton, Spybrook, Fitzgerald, Wang, & Davidson, 2020). Second, SBS should theoretically improve sample representation and reduce under-coverage, thereby improving generalization to the specified target population. Further, SBS easily integrates with retrospective statistical adjustment techniques. Even if balance is only partially improved at the sampling stage, coverage errors will still be reduced, mitigating the need to redefine the inference population post-hoc. Finally, SBS requires researchers to carefully document the recruitment process, which supports transparency and enables more rigorous critique of the sampling design and study inferences, as well as enabling follow-up analysis on participation behavior such as systematic differences between participants and non-responders.

There are, of course, several limitations as well. SBS depends on the existence of a rich set of observed covariates related to treatment heterogeneity and sample selection for each site in the population. Most readily available extant data primarily consist of demographics and may not contain information on covariates that are more proximally related to variation in treatment effect, which can result in omitted variable bias (Tipton, 2013b). Additionally, SBS requires more resources to implement than a simple convenience sample. Recruiting ranked sites from multiple strata requires a coordinated effort between recruiters (Tipton, Hallberg, Hedges, & Chan, 2017). This means that recruiters cannot work independently and must rely on a partnership with researchers implementing this method.

At least one research group has implemented SBS in a large-scale multi-site educational intervention study and documented their experiences (Tipton & Matlen, 2019). The authors reported success in selecting a highly generalizable sample, but substantial efforts were required to develop the sample frame, generate optimal strata, and coordinate with recruiters. This was compounded by the unavailability of data necessary for identifying the sampling frame. Additionally, though the recruiters reported that working within the strata did not burden their efforts, certain strata were more difficult to recruit from than others. These findings raise an important and pragmatic question: do the advantages of SBS justify the additional resources necessary to implement it? And how do other methods of sample recruitment compare in terms of their ability to obtain representative samples and their cost to implement?

Aims

In order for researchers to make informed decisions about sampling strategies, there is a need to better understand both the performance and the cost of SBS, relative to other sample recruitment methods. The goal of the current paper is to describe a framework for making such assessments and to demonstrate how it can be applied. Central to our proposed framework is that sample representation is influenced by two distinct processes: the researcher's recruitment method and schools' participation decisions. The recruitment method influences whether a population unit is approached by researchers, and includes methods such as SBS, probability sampling, and convenience sampling. Because convenience sampling can take many forms in practice, we put forth a simple model, which prioritizes sites that are more likely to agree to participate. To model school district participation decisions, we use extant demographic data and reported characteristics of schools that have been recruited to large-scale trials to simulate a participation propensity score.

After describing formal models for recruitment and participation, we report a simulation study comparing the performance and feasibility of the sampling methods. We conceptualize generalizability in terms of whether obtained samples are similar to the target population on observable site characteristics. We conceptualize feasibility in terms of researcher recruitment effort to obtain a target sample size and how sample participation is distributed across the population. As with many simulation studies, findings from those reported below are tentative and limited due to some of the simplifying assumptions that we impose–particularly assumptions about the school participation model and the process of convenience sampling. Despite these assumptions, we argue that the simulations nonetheless provide insights into the relative performance of different recruitment methods. Moreover, the limitations of the simulations highlight the need for more and better empirical data about recruitment efforts, so that key assumptions of the framework can be further refined.

The remainder of the paper is organized as follows. In the next section, we describe the framework for assessing recruitment methods by proposing a model for school participation decisions and several models for sample recruitment, including SBS and other sampling methods. In the following section, we provide an illustration of using cluster analysis to define population strata—the critical initial step in SBS, and one that may have broader utility as well. We then demonstrate how to implement our proposed framework in a simulation, in which we assess the relative performance of SBS and other recruitment methods. In the final two sections, we report the simulation results and discuss limitations and directions for future research in this area.

Stratification

One of the innovative aspects of SBS is use of cluster analysis to stratify the population prior to sampling. However, the potential benefits of stratification are not limited to use with SBS. In survey research, stratification is routinely used with probability or non-probability sampling. Thus, in the framework that we propose in the following section, we consider stratified versions of several sampling methods beyond just SBS. Before describing those methods, we first illustrate how a population can be stratified using a cluster analysis. Our illustration follows the approach originally proposed by Tipton (2013b), who provides a more complete discussion of its use in the context of SBS. The results of this exercise also serve as the basis for the simulation study described in a later section.

More about rationale and benefits of stratification.

We demonstrate stratification within the context of selecting schools for a large-scale MRT, where the goal is to select a sample that is representative of a population with respect to a set of covariates related to treatment heterogeneity and site participation. When used for survey sampling or blocking in experimental studies, strata may be defined based on one or two covariates that lend themselves to categorization (Lohr 1999). In contrast, our purpose here is to reduce bias in the estimate by controlling for potential moderators, we will require a larger number of covariates of varying complexity. Traditional stratification at this scale would result in too many strata to handle and an overly complicated recruitment process.

Add citation

We begin by explaining the data which will serve as the sampling frame and the covariates used in the cluster analysis process. We then implement k-medoids clustering to divide the population into heterogeneous strata comprised of homogeneous sites, generally following recommendations put forth by Tipton (2013b). K-medoids clustering assigns sites to strata such that similarity within each stratum is maximized. This step consists of: (1) selecting a distance metric to compute a dissimilarity matrix, and (2) choosing the number of strata that need to be generated. Here we rely on empirical criteria as well as subjective appraisals of what is feasible to implement. After arriving at a final clustering of schools, we illustrate how the strata explain variation in the underlying clusters.

Sample Frame

For purposes of illustration, we will consider a population of schools from a diverse set of six states: California, Oregon, Pennsylvania, South Carolina, Texas, and Wyoming. These six states were selected because they provided ready access to school- and district-level achievement data, which can be used to expand the current research. Also for illustrative purposes, we selected as covariates school characteristics that previous research has found to be associated with participation in RCTs. The goal is for our final sample to include schools not normally found in large-scale evaluations of interventions (Fellers, 2017; Stuart et al., 2017; Tipton et al., 2016). These studies found that districts and schools with higher proportions of students who are English language learners (ELL), economically disadvantaged (ED), non-White, and living in urban settings are more likely to participate, as are larger districts and schools.

Weird justification

We developed a sampling frame for this population using data from the Common Core of Data (CCD; https://nces.ed.gov/ccd/index.asp). The CCD is a comprehensive database housing annually collected census data of all public schools and districts. Table 1 displays population descriptives based on the selected covariates. Prior to stratification, we calculated log-transformations of school size (number of students), district size (number of schools) and the student-to-teacher ratio in order to allow proportional comparisons at the extremes of the distributions (Hennig & Liao, 2013). Figure 1 displays the distribution of the continuous variables used in the cluster analysis. In all, the population frame consisted of 6 states, 2,016 districts and 9,792 schools.

Table 1

Population descriptives and log odds of participation associated with covariate (Fellers, 2017)

Variables	M	SD	log_odds
School Data			
District Size	84.51	202.73	0.520
School Size	578.62	200.80	0.374
Student/Teacher Ratio	19.75	6.13	-0.101
Schoolwide Title I	0.63	0.48	0.019
Suburban	0.41	0.49	0.007
Town/Rural	0.19	0.39	-0.403
Urban	0.40	0.49	0.433
Student Data			
% Black	9.81	16.03	0.291
% Hispanic	44.67	32.78	0.395
% White	34.69	30.49	-0.538
% Female	48.56	2.35	-0.019
English Language Learners	23.32	20.78	0.412
Free/Reduced Lunch	59.24	28.72	0.081

Stratification with Cluster Analysis

Cluster analysis serves as a dimension reduction tool to condense the population into smaller set of homogeneous strata. Cluster analysis entails selecting a distance metric, specifying the number of strata, and generating the strata. We describe each of these considerations in turn. All analyses were performed in R (R Core Team, 2018) using the *cluster* package (Maechler et al., 2017).

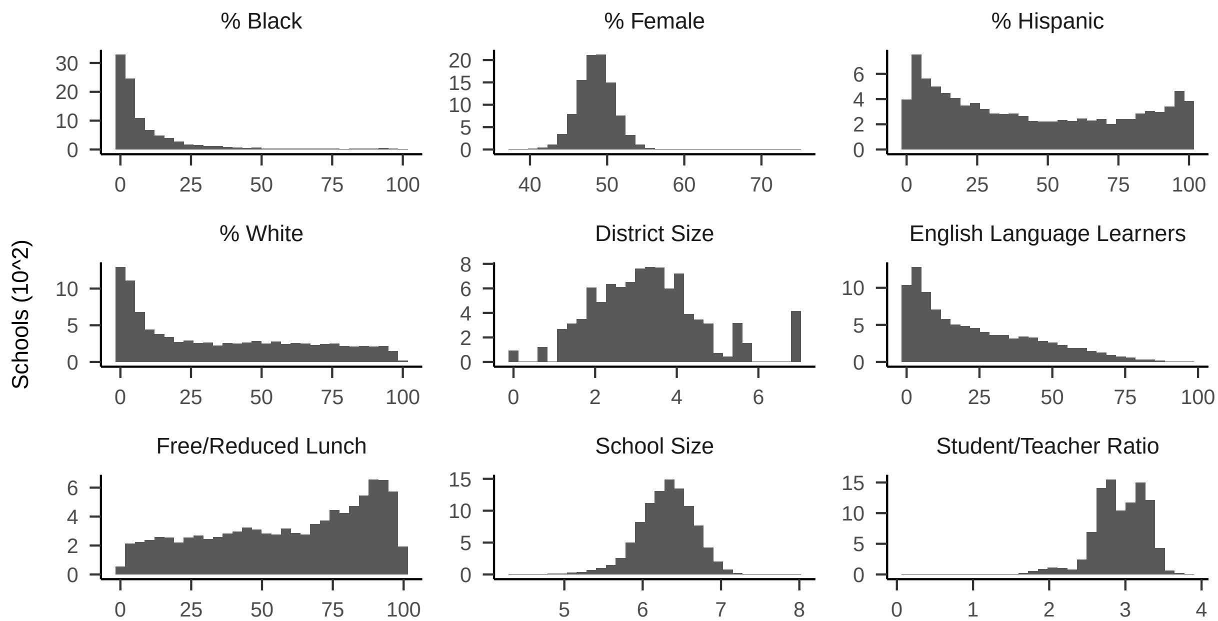

Figure 1. Distributions of continuous covariates. District Size, N Students, and Student/Teachers were transformed by taking natural logs.

Distance metric. In a cluster analysis, the distance metric is how dis-similarity between pairs of units on a set of covariates is quantified. As such, the appropriate metric varies depending on the type of data in the matrix. In educational research contexts, data are likely to contain both continuous and categorical variables, and such was the case here. For mixed data such as this, it is appropriate to use Gower's general similarity measure (Gower, 1971; Tipton, 2013b). This measure relies on different calculations of distance depending on the type of covariates. Let x_{pi} and x_{pj} be the observed value of covariate $p \in \{1, ..., P\}$ for units i and j respectively, where $i \neq j$. Let d_{pij} be the distance between observed values of covariate p for sites i, j. For categorical or dummy coded variables, $d_{pij} = 1$ if $x_{pi} = x_{pj}$ and $d_{pij} = 0$ otherwise. For continuous covariates, we use the following formula:

$$d_{pij} = 1 - \frac{|X_{pi} - X_{pj}|}{R_p} \quad (1)$$

where $|\cdot|$ indicates absolute value and R_p is the range of observations for covariate p. This equation restricts the range of d_{pij} to $[0,1]$. We calculated the general similarity between each site pair by taking the weighted average of the distances between all covariates. Let d^g_{ij} be the general similarity between site i and site j.

$$d^g_{ij} = \frac{\sum_{p=1}^{P} w_{pij} d_{pij}}{\sum_{p=1}^{P} w_{pij}} \tag{2}$$

where $w_{pij} = 0$ if x_p is missing for either site and $w_{pij} = 1$ otherwise. We computed a dissimilarity matrix for all $N \times N$ pairs of units based on the the full set of school-level covariates in Table 1.

Number of Strata. Selecting an appropriate number of clusters is one of the most difficult problems in cluster analysis (Steinley, 2006). Tipton (2013b) argued that both empirical and practical criteria should be used in selecting k. Hennig and Liao (2013) also argued that the method of selecting k should depend on the context of the clustering, framing the issue as one of obtaining an appropriate subject-matter-dependent definition rather than a purely statistical question.

Proportional allocation dictates that each stratum should contribute a number of units to the full sample that is proportional to the size of the strata. Having unequal sized strata means that recruiters will need to focus more on larger strata. Generating a larger set of strata would result in greater homogeneity within each stratum, but it may also be more difficult to manage for recruiters. For instance, if refusal and non-response rates are fairly high, having fewer sites spread across more strata may make it difficult to adequately recruit from all strata. Resource constraints (e.g. time, funding, recruiters) may also be a factor in the number of strata selected.

With these considerations in mind, we examined three criteria for choosing the number of strata: (1) a generalized form of the Calinski-Harabasz index (Caliński & Harabasz, 1974) proposed by Hennig and Liao (2013), (2) the proportion of between-cluster variance as recommended by Tipton (2013b), and (3) the practicality of

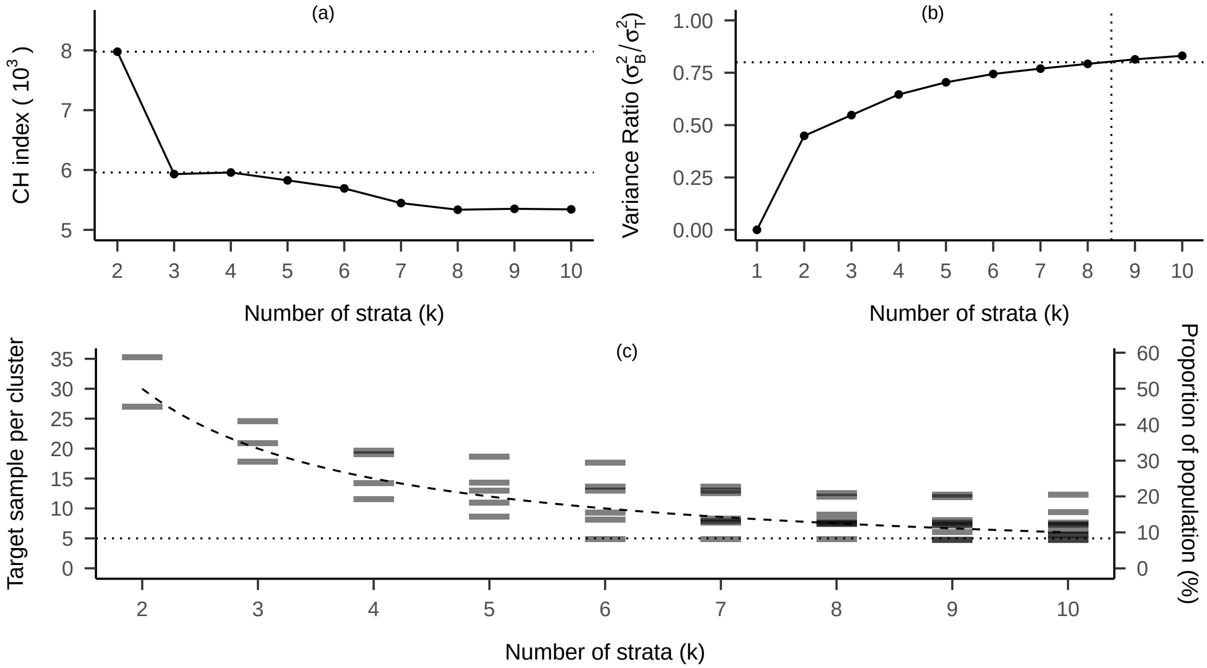

Figure 2. Plots used to determine value for k. (a) Calinski-Harabasz index; peaks indicate better fit. (b) Ratio of between cluster sum of squares to total cluster sum of squares; horizontal line indicates cutoff of .8, vertical line indicates minimum number of clusters needed to achieve cutoff. (c) Sampling requirements for each cluster given proportional allocation; horizontal dotted line indicates a minimum sample size requirement of 5 schools.

sampling from fewer clusters. Our strategy was to perform the cluster analysis several times for a specified number of clusters, then compare all performance criteria for each set of strata generated (Figure 2).

We first calculated the Calinski-Harabasz (CH) index using the *cluster.stats* function from the *fpc* (Hennig, 2019) package. Figure 2a displays the CH index for each k clusters generated. In this case, generating 2 clusters maximizes the CH-index. Another potential solution is at 4 clusters where there is also a local maxima.

We also considered the proportion of variance that lay between clusters. Let K be the number of strata generated where $K = 1, 2, ..., q$ for some maximum allowable number of q strata. For each set of K strata, we computed the between and within cluster variability for each covariate. We then calculated the proportion of variability that is between strata by taking the sum of the between-cluster variance across all covariates, divided by the sum of the between-cluster and within-cluster variance across all covariates.

Figure 2b plots p_k against k. The k for which the rate of change p_k slows is considered favorable. Tipton (2013b) also recommends selecting the number of clusters such that at least 80% of the variability is between clusters, indicated by the figure as a dashed line. In light of this criterion, it seems that at least 9 clusters should be generated. However, we also see that after a sharp initial increase, the slope of the graph begins to level out. This indicates that each additional cluster increases the sampling complexity while explaining less variability in covariates. In practice, the difficulty of sampling may not be worth the small increases in homogeneity within clusters obtained when using more than 4 or 5 clusters.

Figure 2c plots the sample size that needs to be selected from each cluster to fulfill the proportional allocation requirement. The dashed line indicates the ideal allocation if all clusters were of equal size. We see that the variability in cluster sizes decreases as more clusters are generated. A sensible cutoff may be determined by looking at the size of the smallest cluster. At $k > 5$ it seems that the smallest clusters would require less than 5 sites being sampled, which may be very difficult in a practical setting. We determined that this would be the most likely criteria to be considered in the field, and ultimately decided to generate 5 clusters for this analysis.

Balanced Sampling

The goal of balanced sampling is to recruit in such a way that the expected value of covariate X_h across sites in stratum j is equal to the expected value of covariate X_h

across all sites sampled from stratum j:

$$E(X_{hi}|Z_i = 1, j) = E(X_h|j) \tag{3}$$

where $Z_i = 1$ if site i is recruited into the sample and $Z_i = 0$ otherwise. Following Tipton (2013b), we implemented balanced sampling by prioritizing the recruitment of sites based on their similarity to the "average" site in each stratum. First we identified the number of sites to be sampled from each stratum using proportional sample allocation. Each stratum contains N_j sites where $N_1 + N_2 ... + N_k = N$. From each stratum j, we calculated the number of sites to be sampled, n_j, such that $n_j = [(N_j/N)n]$, where [.] indicates the value rounded to the nearest integer.

Next we ranked each site within a stratum using a distance measure, with sites closer to the "center" of the strata ranked higher. We calculated the weighted Euclidean distance to the mean of each covariate:

$$d_{ij} = \sqrt{\sum_{h=1}^{p} w_h (X_{hij} - \mu_{hj})^2} \tag{4}$$

where w_h is the weight assigned to covariate X_h, μ_{hj} is the population mean of covariate h in stratum j, and X_{hij} is the value of covariate h for site i in stratum j. As with generating the strata, different weights can be used such that distances depend more heavily on covariates thought to be more related to treatment effect heterogeneity. To weigh covariates equally, we set the weight weight of a covariate to the inverse of its population variance as $w_h = 1/\sigma_h^2$. We then used the ranked list to prioritize sites for recruitment, beginning with the highest ranked sites. If a site was unavailable or refused to participate, a recruitment attempt was made with the next highest ranked site until n_j sites agreed to participate.

Methods and Models

The goal of recruitment is to obtain a study sample that is compositionally similar to a population of interest, such that treatment effects detected in the study can be generalized to that population. Ideally, we would like to obtain a sample that is similar to the population on any and all characteristics that are related to treatment effect heterogeneity. However, both SBS and retrospective generalization methods aim toward a more limited goal of obtaining a sample that is similar to the population on a set of measured covariates. We follow this conceptualization of generalizability.

More discussion of what these covariates should be

Formally, we shall consider a population of N units (such as schools), indexed by $j = 1, ..., N$. We assume that we have a dataset containing P covariates, $\mathbf{x}_j = (x_{1j}, ..., x_{Pj})$, for every unit in the population. The covariates may be a mixture of continuous, binary, and categorical data. Following Tipton (2013b), we assume that the researcher's goal is to select a sample of n sites such that there is balance along $\mathbf{x}_j$ between the sample and the population, indicating that the population is compositionally represented by the sample. We measure balance using the standardized mean difference (SMD) between the sample and population for a given covariate. For covariate p, the SMD is calculated as

$$SMD_p = \frac{\bar{x}_p - \mu_p}{\sigma_p} \tag{5}$$

where $\bar{x}_p$ is the sample mean of covariate p, μ_p is the covariate mean in the population, and σ_p is the standard deviation of the covariate in the population. SMD values closer to zero indicate greater balance between the sample and the population.

More discussion of how to interpret balance across the set of P covariates.

When recruiting a sample, imbalance between the sample and the population—or selection bias—can arise from at least two distinct processes: recruitment methods and the decisions of individual units about whether to participate. Our goal is to assess the effectiveness and costs of different recruitment methods, such as SBS, probability sampling, and convenience sampling. However, any of the recruitment methods must contend with unit non-response, such as when school leaders decline to participate in the research study. Our proposed framework treats the two processes as distinct, allowing us to isolate the influence of different recruitment methods.

Modeling Selection Bias

To model unit participation decisions (i.e., self-selection) we propose a simple participation propensity score. Let E_j be a binary indicator of *potential participation*—that is, whether a school would participate in the study *if approached for recruitment*, where $E_j = 1$ if the school would agree and $E_j = 0$ if the school would refuse. Let π_j represent the participation propensity score, or the probability that unit j would agree to participate in an RCT if approached. We will model π_j using a basic logistic regression, where:

$$\log\left(\frac{\pi_j}{1-\pi_j}\right) = \beta_0 + \mathbf{x}_j\boldsymbol{\beta} \tag{6}$$

where $\mathbf{x}_j$ is a $1 \times P$ vector of covariates that predict sample selection for each unit, and $\boldsymbol{\beta}$ is a vector of coefficients associated with those covariates. Participation for unit j is then determined by sampling from a Bernoulli distribution with probability equal to π_j:

$$E_j \sim B(\pi_j). \tag{7}$$

In this manner, π_j is a constant school characteristic, whereas E_j can vary across replications of the sampling process.

Discuss assumptions behind this model.

Discuss difficulty of estimating this model from extant data, due to non-representative samples. Also how will we develop assumptions about values of β?

Sampling Methods

In the remainder of this section, we propose formal models for several sampling procedures. The scope of procedures that we review here is not meant to be comprehensive. However, by formalizing them in this fashion we hope to make the procedures transparent, reproducible, and modifiable to easily fit other circumstances which may be of interest for study.

For each method, we model the sampling process as observing the potential participation indicator for schools in a ranked list, where the order in which schools are contacted is determined by a score $S = S_1, ..., S_N$. Different sampling procedures are defined by different methods of determining the scores. Let $Z_j(S)$ be an indicator whether school i is sampled based on the score S. For the un-stratified sampling methods, we determined $Z_1(S), ..., Z_N(S)$ by sorting schools according to S and selecting the first n schools with $E_j = 1$. Specifically,

$$Z_j(S) = I\left[n \geq \sum_{i=1}^{N} E_i I\left(S_i \leq S_j\right)\right] \tag{8}$$

where $I(C)$ denotes the indicator function, equal to 1 if C is true and otherwise equal to 0. Based on the sample selection indicators, we calculate the number of schools contacted as

$$R(S) = \max\{S_1 Z_1, S_2 Z_2, ..., S_N Z_N\}. \tag{9}$$

Several of the sampling methods that we shall consider involve use of stratification. For these methods, we will assume that the population is divided into a set of K strata and that the target sample size for each stratum is based on proportional allocation. Justify proportional allocation Letting N_k denote the total number of schools in stratum k, we set a target sample size of $n_k = [n \times N_k/N]$ for stratum $k = 1, ..., K$, where $[x]$ is the integer nearest to x. Let $G_j \in \{1, ..., K\}$ be the stratum assignment of unit j, for $j = 1, ..., N$. For the stratified sampling methods, the process for determining participation is applied separately within each stratum, so that

$$Z_j(S) = I\left[n^{G_j} \geq \sum_{i=1}^{N} E_i I\left(S_i \leq S_j, G_i = G_j\right)\right] \tag{10}$$

and

$$R_k(S) = \max\{S_1 Z_1 I(G_1 = k), S_2 Z_2 I(G_2 = k), ..., S_N Z_N I(G_N = k)\}, \tag{11}$$

with $R(S) = \sum_{k=1}^{K} R_k(S)$.

Random Sampling

As we previously noted, probability sampling is typically impractical in the context of educational MRTs. However, probability sampling methods are nonetheless interesting as a simple, theoretical ideal against which to compare other sampling methods. In particular, we can simulate unstratified random sampling (URS) sampling by ranking each school at random, so that their order of recruitment is determined by sampling without replacement from the integers $1, ..., N$. In practice, methods such as cluster sampling, stratified sampling, or a combination of both would likely offer advantages over unstratified random sampling. We therefore also consider stratified random sampling (SRS), with strata determined by the results of the cluster analysis and using a proportional allocation.

Convenience Sampling

Some form of convenience sampling is probably the most common approach to sample recruitment. However, researchers rarely operationalize or report their process for selecting a convenience sample, leaving open the questions of what drives recruitment bias and how to model it. As a first step, we posit that the purpose of convenience sampling is to minimize recruitment effort. If we further assume that recruiters have some prior knowledge (based on field experience) of how likely schools are to participate if approached, then we can model convenience sampling as prioritizing schools with a higher propensity to participate. We refer to this as the "low hanging fruit" approach to convenience sampling.

As with probability sampling, we will consider two forms of convenience sampling: unstratified (UCS) and stratified convenience sampling (SCS). To operationalize unstratified convenience sampling, we assume that schools are approached for recruitment one at a time, with their order determined by sampling without replacement, and with probability proportional to participation propensity scores $\pi_1, ..., \pi_N$. Once a school is selected and assigned a rank, the next school is selected with a probability proportional to the weights of the remaining schools. Once all ranks are assigned, schools are again approached until 60 schools agreed to be in the sample. We operationalize stratified convenience sampling using the same process, but with ranks determined independently within each stratum, and using a proportional allocation across strata.

Balanced Sampling

SBS is unique in that rankings are directly related to school characteristics and do not change across iterations. Scores within strata are based on equation (4) (i.e., $S_j = d_{jk}$), where schools closer to the "center" of the stratum are more representative of it and are therefore a higher priority. Though extremely unlikely, it is possible that several schools could be equidistant from the center of the stratum; in such cases, schools are ordered randomly. Because Tipton (2013b) proposed balanced sampling in connection with stratification based on a cluster analysis, we only consider the stratified version, SBS.

Simulating sample recruitment

In this section we describe the simulation study developed to assess the generalizability of the samples selected by SBS relative to several other sampling methods, and the feasibility with which the sampling methods can be employed. A simulation study is the most convenient method for comparing multiple sampling techniques in a controlled environment. The external validity of this study largely rests on our ability to accurately model the participation propensity score. By relying on previous work and real data to inform our model, we hope to maximize how well our results represent reality.

Data Generation

Strata Assignments. In the Stratification section, we detailed how the cluster analysis was performed and resulted in schools being assigned to one of five strata.

Strata are tied directly to school characteristics, and are independent of any of the simulation study's conditions. Strata assignments are therefore constant across iterations and conditions.

Participation Propensity Score. We modeled the participation propensity score (equation (6)) using observed covariate values from extant school data. The set of covariates, X_i, consisted of the same variables used in the cluster analysis described in Sample Frame section above. The corresponding coefficients, B, were based on work by Fellers (2017) who compared 571 elementary schools that participated in IES funded studies to the full population of U.S. elementary schools. The authors reported absolute SMD between the schools that participated and the population. We standardized X_i and used reported SMDs as coefficients in equation (6) to generated π_i^P values. The covariates along with the coefficient values are reported in table 1. We generate different levels of population participation rates by manipulating the intercept values. Since these rates are unknown, we selected parameters for 9 levels of participation rates. As such, participation propensity scores were constant for each school only within each participation rate condition.

Simulation procedures

We designed the simulation to examine the performance of 5 sampling methods across 9 levels of population participation. The population of schools and their assigned strata were constant across all conditions and iterations. The participation propensity score was constant across all sampling methods and iterations within each level of population participation. Within each iteration, all schools were checked for approval (equation (8)) which then remained constant across each sampling method. That is, if a school happened to be approached by three sampling methods within an iteration, the school would provide the same response to all three methods (agree or refuse to participate). The order in which schools were approached varied by method and across iterations. The one exception is the balanced sampling method, where the order in which schools are approached is constant since it is tied to strata assignment and school characteristics.

In total, 45,000 samples were generated by running 1000 iterations for each of 9 population participation levels, and using 5 sampling methods in each iteration. Each iteration proceeded as follows. First, we generated school responses (E_j) for all schools. Next, we generated 5 sets of rankings, one for each sampling method, for all schools. We checked each school in order of rank for recruitment until a sample of 60 schools was selected by each method. For each sample within an iteration, we tracked which schools agreed to participate, which schools refused to participate, and how many schools refused to participate. We also calculated the B-index and the SMD between each sample and the population. If the sampling method was a stratified version, B-indices and SMDs were calculated within each stratum as well.

Analysis

Generalizability. There are several methods to quantify how generalizable a sample is to a target population. One common method is to compare the sample to the population on a range of covariates by examining SMDs as shown in equation (5). This method is limited as it only provides us with a measure of how close the sample means are to the population means. To have true generalizability, sample variance must also be representative of the population variance. Therefore, in addition to SMDs we also estimated the generalizability index (B; Tipton, 2014).

The generalizability index is bounded between 0 and 1, with 0 indicating no overlap between the sample and the population, and 1 indicating the sample is representative of the population.

$$B = \int_{\pi_{min} - 3h_{max}}^{\pi_{max} - 3h_{max}} \sqrt{\hat{f}_s(z)\hat{f}_p(z)} dz \tag{12}$$

where $\hat{f}_s(z)$ is a Gaussian kernel density estimate of sample distribution of propensity scores using bandwidth h_s, and $\hat{f}_p(z)$ is a Gaussian kernel density estimate of population distribution of p-scores using bandwidth h_p. We calculate the bandwidth h for the population or sample as follows:

$$h(\pi_i, ..., \pi_n) = \sigma_\pi(\frac{4}{3n})^{1/5} \tag{13}$$

where σ_π is the variance of the propensity scores.

Do these equations look right? Should (11) be $\pi_{max} + 3h_{max}$?

Feasibility. To assess feasibility, the total number of schools approached to achieve a full sample was tracked. The average number of refusals each sample method resulted in prior to selecting the full sample was calculated across replications. Recruiters expend a lot of resources contacting districts and schools, scheduling meetings and traveling between interested locations. A project with limited resources may not be able to afford to go through a large list of potentially uninterested sites. This measure allows us to compare the difficulty with which a full sample is recruited using each method.

Sampling Inequality. To assess sampling inequality, we tracked the frequency with which schools were recruited by each sampling method across iterations and conditions. We then summarized this data by calculating the Gini coefficient for each method. This was done in R using the *Gini* function from the *ineq* package (Zeileis 2014). The Gini coefficient is typically used in economics to assess the degree of income inequality by showing the disproportionate distribution of wealth across levels of incomes. Similarly it could be used in the context of sampling by identifying how often under-represented schools are recruited to participate in studies. The Gini index ranges from 0 to 1, with 0 indicating perfect equality, and 1 indicating perfect inequality.

Results

Generalizability

B-Index. Figure 3 displays the average B-index for each method across participation rates. Acceptable values of B for generalizability vary depending on the size of the sample, the size of the population, and the number of covariates (Tipton, 2014). Given our design, a value of $B = .95$ would indicate very good overlap with no need to adjust estimates. Generally $B >= .80$ still supports generalizability with adjustment and likely no need to redefine the population. At population participation rates below 40%, only SBS consistently performs better than other models. Across population response rates it also never results in a B value below .85. This indicates that SBS is successful at sampling schools that are unlikely to participate and therefore tend to be underrepresented by the other sampling methods—particularly when overall participation propensities are low. Stratified random sampling consistently outperformed simple random sampling across all participation rates, though only slightly.

We also found several trends that were unexpected. At 50% and beyond, SBS performance slowly degrades, while other methods maintain a steady increase. We expected a constant positive relationship between the population participation rate and the performance of all methods. Furthermore, at low response rates unstratified convenience sampling seemed to perform better than stratified convenience sampling. This seems counter-intuitive, as survey literature suggests that stratified samples are more representative. Because the B-index is an overall measure of generalizabilty across many covariates, it is difficult to untangle why these methods preformed as they did. We next examined the performance of the sampling methods on each individual covariate.

Standardized Mean Differences. We began this analysis by plotting mean SMDs for each covariate across all sampling methods against the population participation rate. Sampling methods were considered to perform well if they resulted in

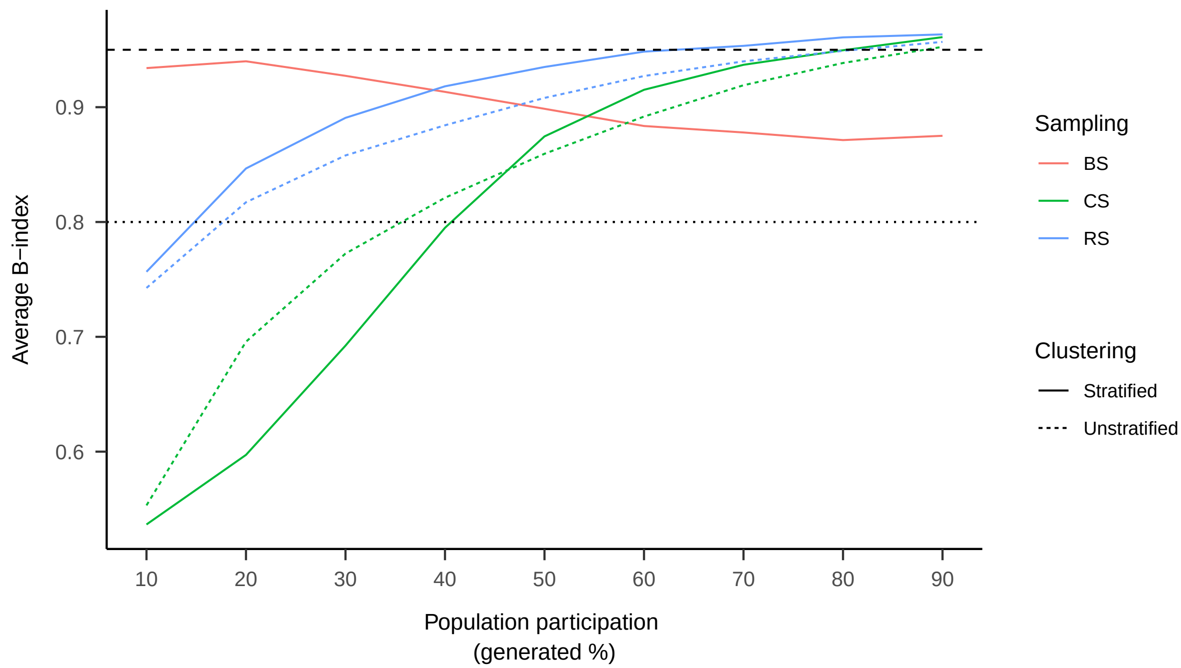

Figure 3. Average B-index for varying participation rates, by sampling method. Horizontal dotted line represented index of .95 indicating a high level of generalizability.

an average SMD value below .25 across iterations. The relative performance of sampling methods varied across covariates. Several patterns did emerge, however, which allowed us to designate three groups of covariates. Figure **??** displays an example of each pattern. Group 1 represents nine covariates where stratified methods outperformed unstratified methods across population participation rates. Group 2 represents two covariates where all sampling methods resulted good balance. Group 3 represents three covariates where stratified methods performed worse than their unstratified counterparts, or where at least one unstratified method performed better than a stratified method. Across all groups, stratified balanced sampling almost always resulted in good balance and generally outperformed all other sampling methods.

One potential explanation the patterns found in groups 2 and 3 is that the strata were poorly specified for these covariates. To examine this we calculated the the proportion of variance explained by the strata for each covariate and compared it to the

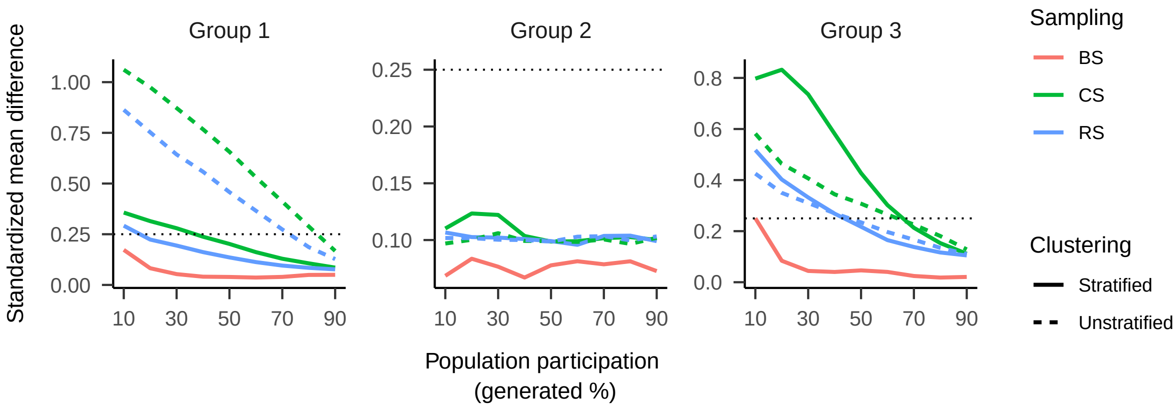

Figure 4. Patterns of relative performance based on standardized mean differences. The dotted horizontal line represents a cut off of .25 under which good balance is achieved.

log-odds coefficients associated with that covariate in the participation propensity model. We then plotted these values in figure 5. Several patterns emerged here as well. Group 1 consisted of covariates where at least 45% of the variance was between clusters. Groups 2 and 3 both consisted of covariates that were poorly clustered, with no more than 15% of variance between clusters. The major difference between groups 2 and 3 appears to be the relationship between the covariates and the likelihood of participation.

Feasibility

Recruitment Attempts. Figure 6a reports the total number of units that needed to be contacted to recruit a full sample of 60 schools across population participation rates for all sampling methods. Figure 6b reports the percent of schools that agreed to participate across population participation rates for all sampling methods. Differences between sampling methods along these two measures were substantial at lower participation rates. As participation rates increased, the differences decreased exponentially and became negligible. It is important to note that the magnitudes of these results are quite extreme. This is likely due to some misspecification on the part of the simulation either in the response generating model parameters or the population

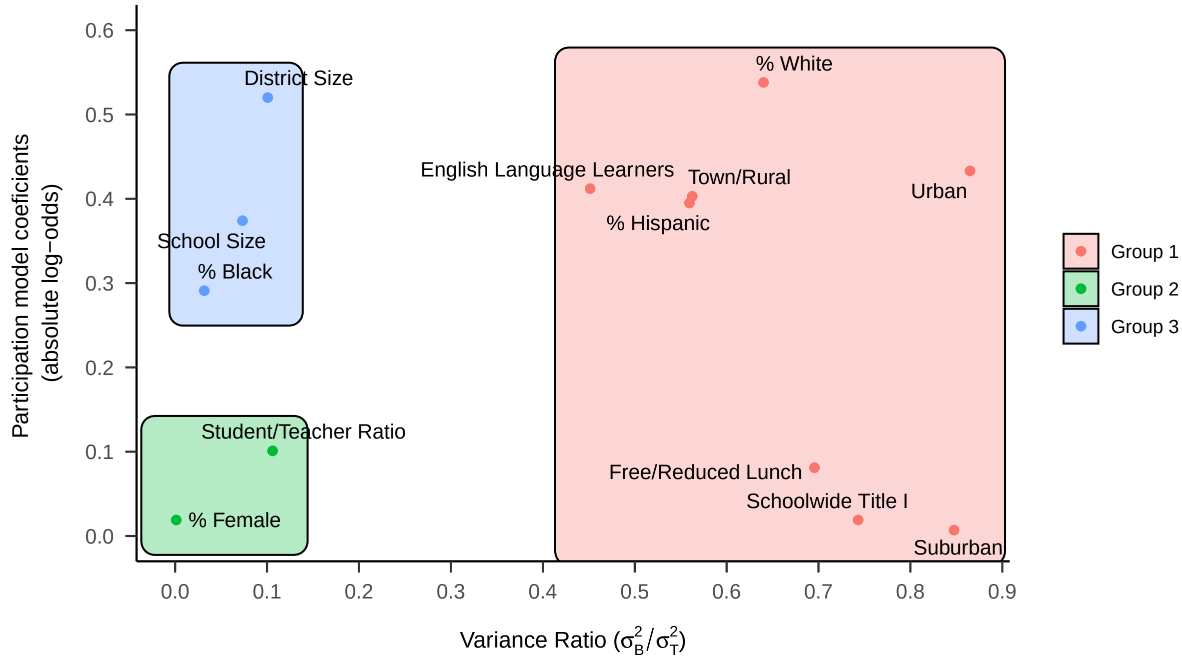

Figure 5. Variance explained by clustering vs absolute log odds. Shaded areas illustrate patterns in generalizability measured by SMDs.

participation rates. Rather than looking at the raw values, a more meaningful interpretation would be to compare the performance of the models relative to each other. Overall, UCS required the least "effort" to recruit a full sample, followed by URS and SCS, SRS, and finally SBS.

Sampling Inequality. We calculated the Gini coefficient for each sampling method to examine the equality of sampling probabilities across methods. Figure 7 displays these for each sampling method across population participation rates. A coefficient of 1 indicates substantial sampling inequality. In the context of the simulation, a high Gini coefficient indicates that across iterations, only a small subset of the population was ever actually recruited. Several trends emerged in this analysis. As population participation rates increased, inequality increased when using SBS, but decreased when using any other method. These four other methods also performed consistently relative to one another, with SRS resulting in the least amount of sampling

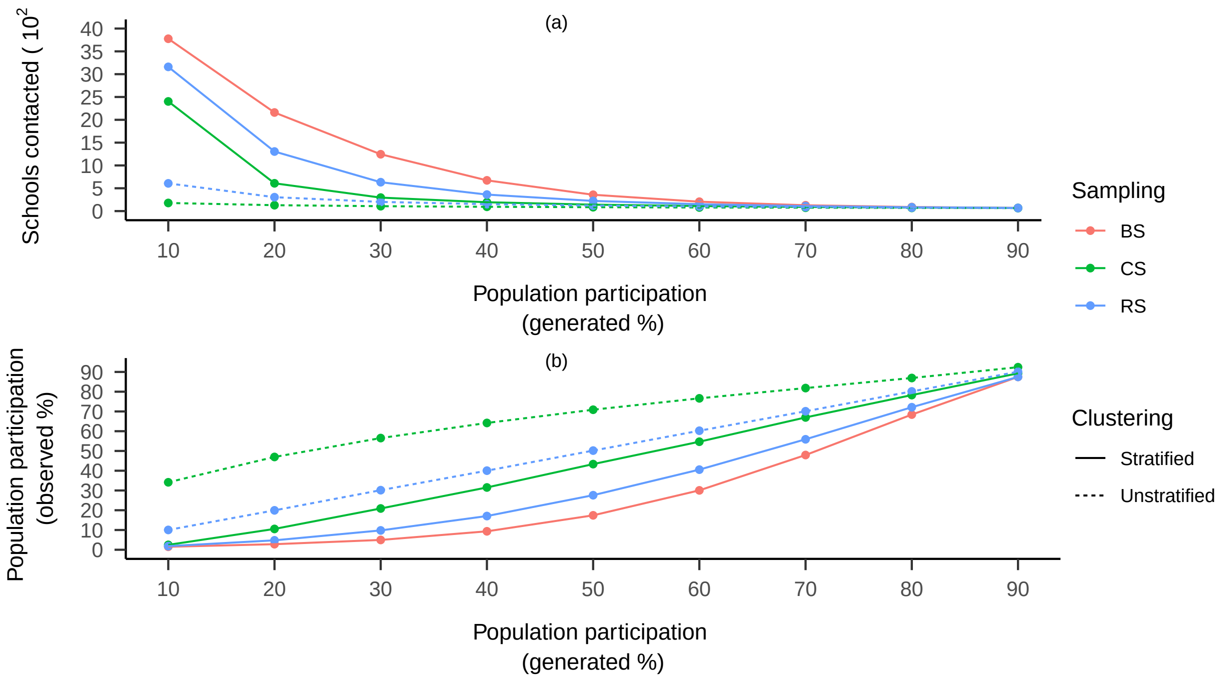

Figure 6. Sample recruitment statistics and response rates. Plot (a) shows the total number of units contacted to achieve a full sample of 60 schools. Plot (b) shows the percent of schools that agreed to participate when recruited.

inequality, followed by URS, SCS and UCS. As population rates increased, differences between stratified and unstratified versions of sampling methods seemed to decrease. This finding suggests that stratifying the population results in a larger potential sampling pool when the overall population response rate is low.

Summary of Trends

In terms of selecting a generalizable sample, SBS resulted in a considerable improvement compared to UCS. However, given the difficulty with which those samples are recruited, SBS is unlikely to be fully implemented in the ideal form. Instead, SCS may be a reasonable compromise. Our findings indicate that convenience and probability sampling methods are often improved by first stratifying the population. In certain cases, convenience sampling within strata (SCS) is comparable to simple random

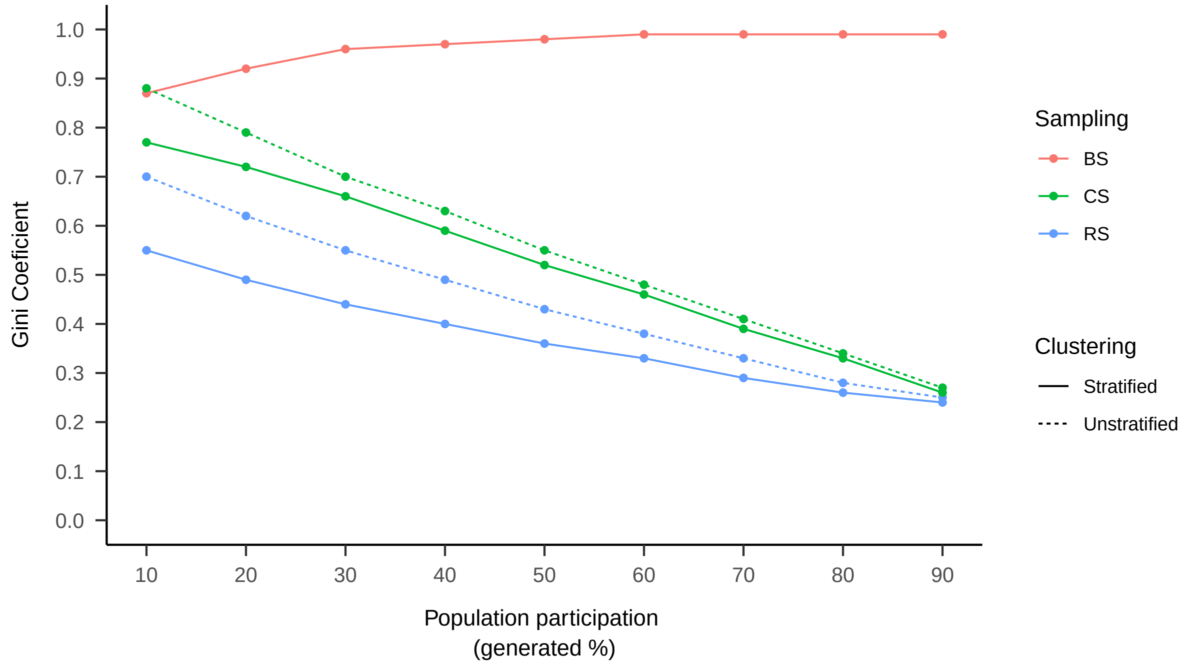

Figure 7. Gini coefficient across participation response rates for each sampling method. A coefficient of 1 indicates major inequality in sampling probability.

sampling (URS) both in terms of generalizability and feasibility.

An important observation which has not yet been addressed in the literature relates the specification of the clustering method to the response model. For instance, in Figure **??** we see that all methods of sampling result in balance on percentage of Black students, except for stratified convenience sampling. This is likely a result of the covariate having a strong relationship with participation, but not being weighed enough in the cluster analysis used in determining the strata. Figure 5 displays the calculated intraclass correlation coefficient (ICC) for each covariate along the strata, and the coefficient for each covariate in the response generating model. For covariates where one value (ICC or coefficient) is high, while the other is low, stratified sampling techniques resulted in poorer balance.

Discussion

The main goal of this study was to develop a framework for exploring the performance of sampling methods within the educational context. We put forth several models for algorithmically representing how researchers may select convenience sample of schools. Using prior work we also attempted to model how schools may decide whether or not to participate in a study if approached by recruiters. The methods we proposed for modeling these behaviors can, in principle, be extended to more complex and realistic specifications and adapted to other population frames.

A secondary goal was to use this framework in a demonstration of several sampling methods and their relative performance in terms of generalizability and feasibility. From this work we have drawn several conclusions. Stratified balanced sampling as proposed by Tipton (2013b) has the potential to greatly increase the generalizability of samples selected for MRTs. However, this method is not without limitations. Within our simulation, strict adherence came at a great cost in terms of sheer number of schools that needed to be contacted. Thus, implementing this method in practice may require allocating many more study resources to sample recruitment.

We found that ignoring the response model when specifying covariate weights during the cluster analysis stage may attenuate the generalizability of the resulting sample. Covariates that were not prioritized when generating the strata wound up being poorly represented in the final samples if they strongly predicted participation. Ignoring these relationships has the potential to undermine the investment of resources into SBS. Finally, while the balanced sampling approach does result in greater generalizability, it also appears to limit the actual pool of potential participants. Particularly in larger population response rates, the same subset of schools were likely to be recruited across iterations.

One potential compromise between current practice and SBS is to generate strata, but then implement convenience sampling within strata. As demonstrated in the simulations, stratified convenience sampling often resulted in better balance on individual covariates than simple random sampling. This may also elicit greater buy in from recruiters by placing less restrictions on what units they must sample.

Beyond generalizability, stratifying in this manner requires that researchers make sampling decisions in the study design phase, and to track changes in the sampling plan as recruitment progresses. Documenting and reporting this process would in turn support further research into developing more efficient and effective sampling methods.

Limitations

The models that we have studied make several key assumptions which represent limitations of the findings from the simulation study. First, in modeling convenience sampling, we assumed that recruiters always prioritize schools that are most likely to participate. In reality, other factors play a role as well, such as proximity of sample sites to the researcher and to each other, existing relationships between the recruiters and the sample sites, and other researcher assumptions about the sample site's characteristics. This limits how well our results reflect the performance of models in reality. Addressing this in future work can lead to better guidelines for future sampling designs.

Another implication of this assumption is that recruiters have approximate knowledge of how likely a sampled site is to participate. Though researchers may speculate about sites that are more willing to participate (such as schools in larger urban districts) and prioritize recruiting such sites, it is not likely that their estimation of "willingness" would be as close to the truth as we have estimated. Given this, it is possible that the "feasibility" of the convenience methods is over-stated, and that the degree of generalizability for some covariates is under-stated.

It is worth refining and exploring additional methods for modeling convenience sampling. The algorithms used in these methods can easily be tuned to include additional factors that might influence school recruitment priorities. For instance, location data is readily available and can therefore be incorporated into the model for how researchers prioritize schools in convenience sampling. Further work here could lead to more realistic and practical assessments of feasibility and generalizabilty, potentially providing researchers with a tool for evaluation sampling methods given their unique circumstances during the study design phase.

The second set of assumptions deals with the speculative nature of our participation model. In practice, the decision of whether a school participates in such a study often involves multiple stages. Generally, districts serve as gatekeepers, requiring submission and approval of research requests prior to recruitment. If the request is denied, no schools within the district may be recruited. If approved, researchers may work with a district-wide school coordinator, or may have to contact schools individually. In either cases, the ultimate decision may then rest with administrators, school research coordinators, or the teachers themselves.

A further limitation of the simulations is that the parameters in the response generating model are based on values from a study that examined the difference between schools participating in large-scale RCTs and the overall population of schools. However, these RCTs themselves typically rely on some form of convenience sampling. Consequently, our parameters reflect participation rates of schools that are likely to participate in RCTs, rather than the full population of schools. There simply isn't a solid understanding of what drives school participation in the current literature.

We believe that our findings reasonably represent the relative performance of the various sampling methods we tested in the context of educational research. However, a more exhaustive examination at what drives school participation in the population could address the above limitations and give us a better sense of how sampling methods would perform in reality. Disentangling participation bias from sampling bias requires researchers to implement probability based sampling or to be more transparent about their sampling practices. Doing so would also provide a deeper insight into school behavior and representation in research. If we can identify schools that are consistently and systematically under-represented in funded research, we can develop strategies to target such schools and increase the inclusivity of studies that strive for truly representative population level inferences.

Future Directions

This study has laid the groundwork for several avenues of research that are worth exploring further. First, additional work is needed on how best to optimize the cluster analysis. We have shown that the extent to which balance is achieved on a given covariate is related to how much that covariate drives strata generation and school participation. If some covariates are known to have greater influence on school participation, they should be weighted more heavily in generating the strata. Further work is also needed to understand the relationship between number of clusters generated, generalizability, and feasibility. It is expected that more clusters would increase generalizability, but also make recruitment more difficult. A better understanding of these relationships would help drive decision making during the design phase, which would make SBS much more accessible and quicker to initiate.

Further work also needs to examine the impact of these sampling methods on the bias and accuracy of population average treatment effect estimates. In this study, our goal was only to select a generalizable sample, where generalizability was operationalized as balance between the sample and population on a set of covariates. To the extent that the same covariates that dictate selection are also predictive of variance in treatment effects, we could extrapolate that a sample that is balanced on these covariates can be used to estimate an unbiased PATE. Reality is likely more complicated, however, and it would be worth examining the relationship between sampling and estimating unbiased treatment effects further.

Earlier we stated that if treatment effects are constant across units in a population, nonrandom samples of the population should still lead to unbiased estimates of average treatment effects. However, if only a narrow slice of the population is studied, there may not be enough variability in potential moderators to detect heterogeneity. Adding variation by selecting a more diverse sample may be useful if the presence of heterogeneity is unknown. This further complicates the specification of the cluster analysis. How should covariates be weighed relative to each other depending on whether they predict participation, differences in treatment effects, or some combination of both? To address this our work must be extended to study the relationship between sampling methods and bias in treatment effect estimation.

Large scale MRTs are expensive to implement, and resource allocation for such studies presents many difficult trade-offs. Researchers who wish to invest in robust recruitment strategies to amplify the impact and relevance of of their work should be better equipped to anticipate the costs and benefits of various sampling strategies. We hope that by showing the relative performance of these sampling methods, and by demonstrating the implementation of stratified sampling, we have contributed to future researchers being better informed in making these decisions.

References

Caliński, T., & Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics*, *3*(1), 1–27. https://doi.org/10.1080/03610927408827101

Fellers, L. (2017). *Developing an approach to determine generalizability: A review of efficacy and effectiveness trials funded by the Institute of Education Sciences* (Ph.D.). Columbia University, United States – New York. Retrieved from https://search.proquest.com/docview/1865595768/abstract/40FD82F4A0C24535PQ/1

Gerber, A. S., & Green, D. P. (2012). *Field experiments: Design, analysis, and interpretation* (1st ed). New York: W. W. Norton.

Gower, J. C. (1971). A General Coefficient of Similarity and Some of Its Properties. *Biometrics*, *27*(4), 857–871. https://doi.org/10.2307/2528823

Groves, R. M. (Ed.). (2004). *Survey methodology*. Hoboken, N.J: Wiley-Interscience.

Hennig, C. (2019). *Fpc: Flexible procedures for clustering*. Retrieved from https://CRAN.R-project.org/package=fpc

Hennig, C., & Liao, T. F. (2013). How to find an appropriate clustering for mixed-type variables with application to socio-economic stratification: How to Find an Appropriate Clustering. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *62*(3), 309–369. https://doi.org/10.1111/j.1467-9876.2012.01066.x

Kern, H. L., Stuart, E. A., Hill, J., & Green, D. P. (2016). Assessing Methods for Generalizing Experimental Impact Estimates to Target Populations. *Journal of Research on Educational Effectiveness*, *9*(1), 103–127. https://doi.org/10.1080/19345747.2015.1060282

Olsen, R. B., Orr, L. L., Bell, S. H., & Stuart, E. A. (2013). External Validity in Policy Evaluations That Choose Sites Purposively. *Journal of Policy Analysis and*

Management, *32*(1), 107–121. https://doi.org/10.1002/pam.21660

O'Muircheartaigh, C., & Hedges, L. V. (2014). Generalizing from unrepresentative experiments: A stratified propensity score approach. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *63*(2), 195–210. https://doi.org/10.1111/rssc.12037

Raudenbush, S. W., & Bloom, H. S. (2015). Learning about and from a distribution of program impacts using multisite trials. *American Journal of Evaluation*, *36*(4), 475–499.

Raudenbush, S. W., & Liu, X. (2000). Statistical power and optimal design for multisite randomized trials. *Psychological Methods*, *5*(2), 199–213. https://doi.org/10.1037//1082-989X.5.2.199

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

Steinley, D. (2006). K-means clustering: A half-century synthesis. *British Journal of Mathematical and Statistical Psychology*, *59*(1), 1–34. https://doi.org/10.1348/000711005X48266

Stuart, E. A., Bell, S. H., Ebnesajjad, C., Olsen, R. B., & Orr, L. L. (2017). Characteristics of School Districts That Participate in Rigorous National Educational Evaluations. *Journal of Research on Educational Effectiveness*, *10*(1), 168–206. https://doi.org/10.1080/19345747.2016.1205160

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: Use of Propensity Scores to Assess Generalizability. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *174*(2), 369–386. https://doi.org/10.1111/j.1467-985X.2010.00673.x

Tipton, E. (2013a). Improving Generalizations From Experiments Using Propensity Score Subclassification: Assumptions, Properties, and Contexts. *Journal of Educational and Behavioral Statistics*, *38*(3), 239–266. Retrieved from https://www.jstor.org/stable/41999424

Tipton, E. (2013b). Stratified Sampling Using Cluster Analysis: A Sample Selection Strategy for Improved Generalizations From Experiments. *Evaluation Review*, *37*(2), 109–139. https://doi.org/10.1177/0193841X13516324

Tipton, E. (2014). How Generalizable Is Your Experiment? An Index for Comparing Experimental Samples and Populations. *Journal of Educational and Behavioral Statistics*, *39*(6), 478–501.

Tipton, E., Fellers, L., Caverly, S., Vaden-Kiernan, M., Borman, G., Sullivan, K., & de Castilla, V. R. (2016). Site Selection in Experiments: An Assessment of Site Recruitment and Generalizability in Two Scale-up Studies. *Journal of Research on Educational Effectiveness*, *9*(sup1), 209–228. https://doi.org/10.1080/19345747.2015.1105895

Tipton, E., Hallberg, K., Hedges, L. V., & Chan, W. (2017). Implications of Small Samples for Generalization: Adjustments and Rules of Thumb. *Evaluation Review*, *41*(5), 472–505. https://doi.org/10.1177/0193841X16655665

Tipton, E., & Matlen, B. J. (2019). Improved Generalizability Through Improved Recruitment: Lessons Learned From a Large-Scale Randomized Trial. *American Journal of Evaluation*, *40*(3), 414–430. https://doi.org/10.1177/1098214018810519

Tipton, E., Spybrook, J., Fitzgerald, K. G., Wang, Q., & Davidson, C. (2020). Toward a system of evidence for all: Current practices and future opportunities in 37 randomized trials. *Educational Researcher*, 0013189X20960686.

Online Appendix

Sampling Feasibility. Figure 8 compares each sampling method to a reference method by plotting the factor of increased difficulty, calculated as the number of schools contacted by comparison method divided by the number of schools contacted by reference method. This gives us another perspective on the relative difficulty of each method. The straight horizontal line represents the reference method.

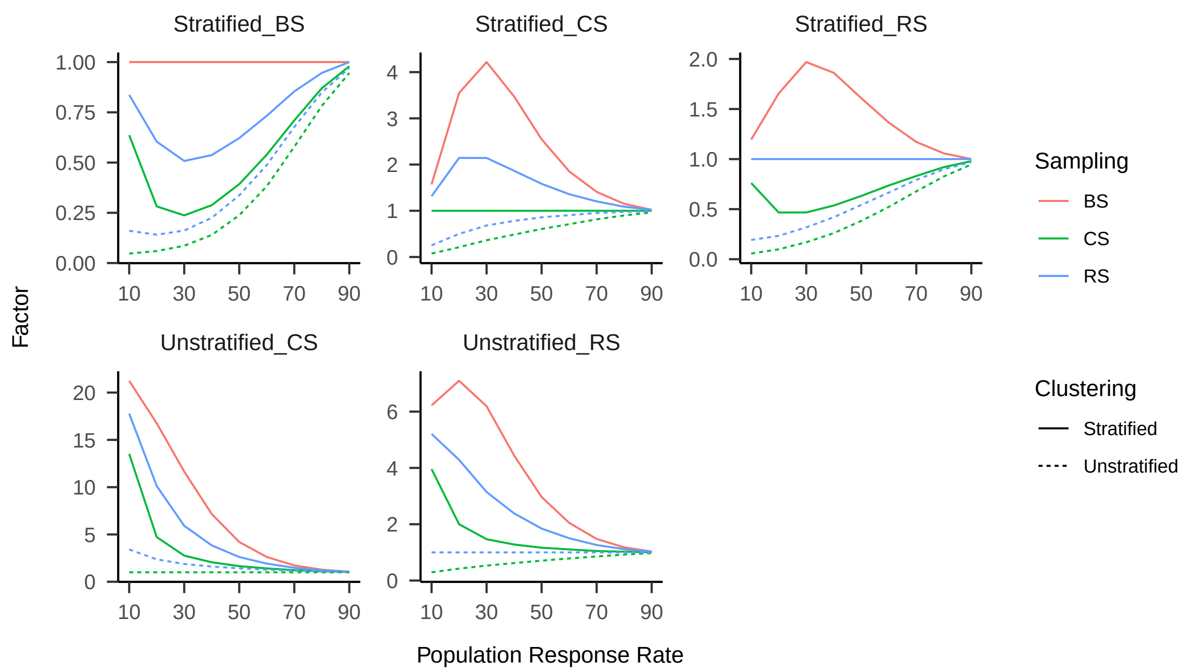

Figure 8. Relative sampling difficulty of each method compared to other methods. The straight horizontal line indicates the reference method being compared to.

Sampling Inequality. Figure 9 displays the Gini curve and coefficient for all sampling methods across participation rates. The index is calculated by computing the area between the diagonal line and the curve. Coefficients of 0 indicate uniform equality across all sampling units, i.e. all schools have an equal opportunity to be sampled. Coefficients of 1 indicate complete inequality, i.e. very few schools are constantly being sampled across iterations. Overall, stratification results in lower inequality. However, since balanced sampling prioritizes schools according to set characteristics, the same schools are likely to be sampled each time.

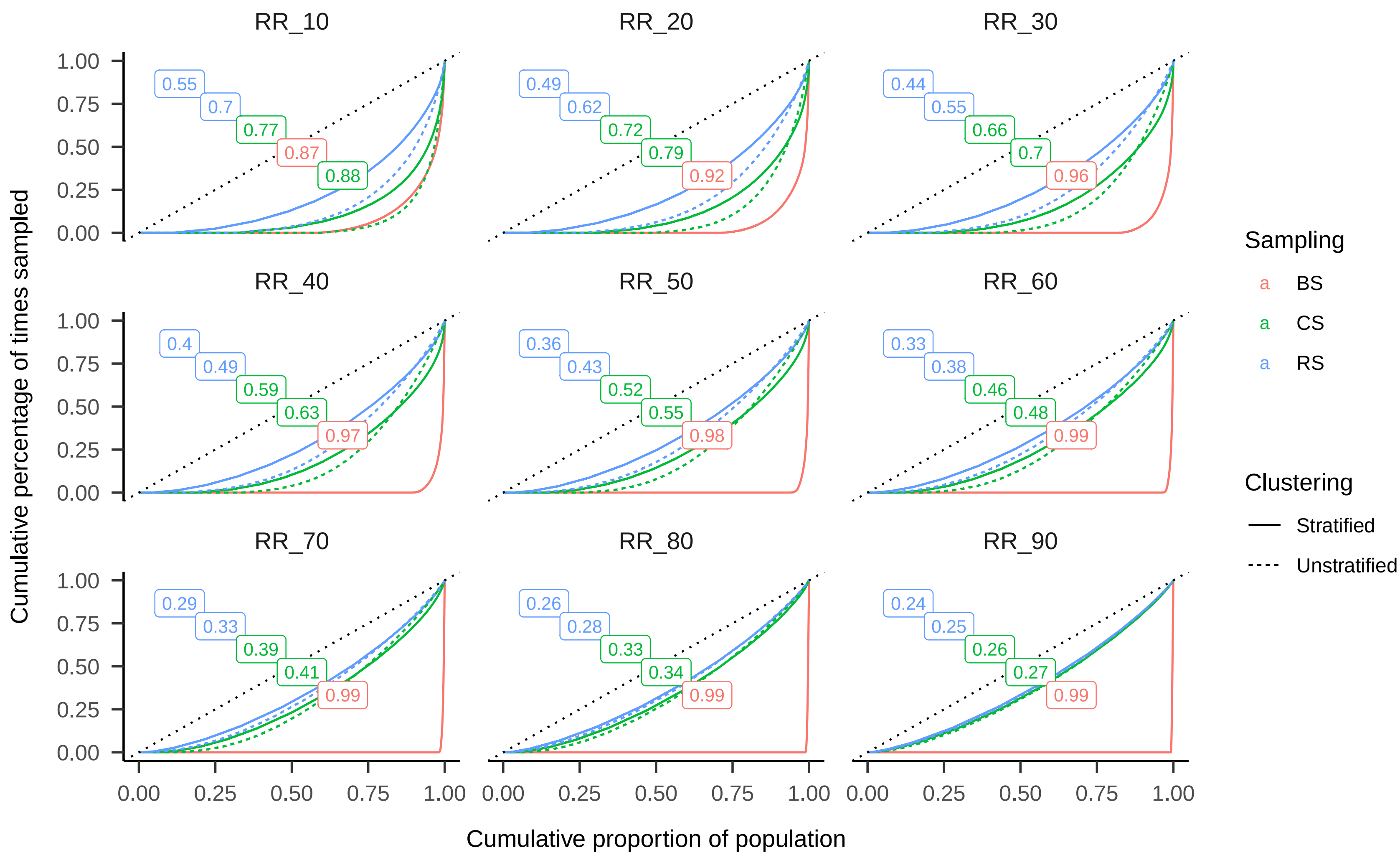

Figure 9. Cumulative probability plot and Gini coefficients representing the inequality of school sampling across sampling methods and population response rates.