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

Gleb Furman[1], James E. Pustejovsky[2], & Elizabeth Tipton[3]

[1] University of Texas at Austin

[2] University of Wisconsin-Madison

[3] Northwestern University

Abstract

Educational research aimed at informing policy decisions should ideally be designed to support 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 systematically 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 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. However, implementing these methods requires extensive recruitment efforts, especially when the population average willingness to participate is low. Stratified convenience sampling represents a potential compromise. Further research into recruitment and participation behavior is needed to design and test sampling methods under realistic conditions.

Keywords: generalizability, sampling, MRT

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

Sampling for generalization

The multi-site randomized trial (MRT) has become a common design for evaluating the effects 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 level 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 one 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, running an experiment across multiple sites 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 (Imai, King, & Stuart, 2008).

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; Tipton, Spybrook, Fitzgerald, Wang, & Davidson, 2020). 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, lac of representation in MRTs can also be a question of equity. For instance, small under-served rural districts are under-represented in RCTs sponsored by Institute of Education Sciences (Fellers, 2017; Stuart et al., 2017; Tipton et al., 2020), and thus may be less likely to benefit from federally funded education research.

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). Estimation entails re-weighing the sample (or otherwise adjusting) to represent a population of interest. These methods rely on the availability of extensive data clearly defining the inference population in order to justify a claim of generalizabilty. Specifically, data must enumerate all units in the population and include information on covariates that predict treatment effect heterogeneity. These requirements make retrospective methods particularly salient in the context of educational trials due to the availability of data on districts and schools (e.g., the Common Core of Data, the Stanford Education Data Archive, etc.).

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 benefits of 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.

One method in particular, Stratified Balanced Sampling (SBS), has received attention from intervention effectiveness researchers due to its accessibility.

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 approach to sample schools may even use a website (www.thegeneralizer.org) that guides them through the process using a sampling frame 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 et al., 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 should 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. Readily available extant data consist primarily of demographics and may not contain information on covariates that are more proximally related to variation in treatment effects, which can result in omitted variable bias

(Tipton, 2013b). Additionally, implementing SBS requires more resources than a simple convenience sample because 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 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.

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 have a more conscribed goal, seeking generalizability with respect to measured variables in the sampling frame, such as school demographics (e.g. ethnic composition, urbanicity) or general academic achievement information (e.g. Math, ELA). We will follow the latter conceptualization of generalizability.

Formally, we 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

$$SMD_p = \frac{\bar{x}_p - \mu_p}{\sigma_p} \tag{1}$$

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. By this measure, obtaining a $SMD_p \leq .25$ for all covariates in P would indicate that the sample is well balanced and can be generalized to the population. Covariates for which $SMD_p > .25$ would likely require additional statistical adjustment in the final analysis.

Should we foreshadow the B-Index here?

When recruiting a sample, selection bias–or imbalance between the sample and the population—can arise from at least two distinct sources: 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 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.

Formally, let P_j be an indicator for whether school j participates in the study. Let Z_j be an indicator for whether school j is recruited, or invited to participate in the study. Let E_j be an indicator for whether school j agrees to participate, if recruited. In order for school j to participate, it must both be recruited and agree; it follows that $P_j = Z_j \times E_j$. For a unit with covariates $\mathbf{x}_j$, the probability of participating in a trial is therefore

$$\Pr\left(P_j = 1|\mathbf{x}_j\right) = \Pr\left(E_j = 1|Z_j = 1, \mathbf{x}_j\right) \times \Pr\left(Z_j = 1|\mathbf{x}_j\right). \tag{2}$$

Non-representative samples can result from bias in the recruitment process, which determines $\Pr\left(Z_j = 1|\mathbf{x}_j\right)$, from bias in the response process, which determines $\Pr\left(E_j = 1|Z_j = 1, \mathbf{x}_j\right)$, or from the interaction of the two. In the remainder of this section, we introduce formal models for each component process.

Modeling Response Bias

To describe unit participation decisions (i.e., self-selection), we propose a response propensity score model that involves several key assumptions. First, we assume that each individual school is responsible for deciding whether to participate, and that school decisions are independent. That is to say, districts, administrators, teachers, research coordinators, etc., as individual units do not influence decision. Second, we assume that schools' response decisions are related to their measured covariates, $\mathbf{x}_j$, but not to unmeasured characteristics. Third, we assume that response decisions are not influenced by the recruitment process itself, so that

Any further comments about this assumption?

$$\Pr\left(E_j = 1 | Z_j = 1, \mathbf{x}_j\right) = \Pr\left(E_j = 1 | \mathbf{x}_j\right). \tag{3}$$

Under this assumption, E_j can be interpreted as an indicator for *potential response*, or whether a school would participate in the study *if approached for recruitment*, for any school in the population. This simplifying assumption could fail to hold if the likelihood that a school responds depends on the manner of recruitment.[1]

Fourth and finally, we assume that the probability of responding can be approximated with a linear transformation of demographic covariates ($\mathbf{x}_j$). Letting $\pi_j = \Pr\left(E_j = 1 | \mathbf{x}_j\right)$ represent the response propensity score, or the probability that unit j would agree to participate in a study if approached, we assume that:

$$\log\left(\frac{\pi_j}{1 - \pi_j}\right) = \beta_0 + \mathbf{x}_j \boldsymbol{\beta} \tag{4}$$

where $\mathbf{x}_j$ are 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:

$$E_j \sim B(\pi_j). \tag{5}$$

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

These assumptions are clearly a stylized, and perhaps over-simplified, representation of a complicated and multifaceted decision-making process. We know, for instance, that in many cases districts act as gatekeepers by determining whether or not to approve research proposals. However, a more nuanced model may not necessarily be

Cite?

[1] For example, a school might be more likely to respond to a recruitment attempt made through a face-to-face meeting than one from an email blast.

more accurate. Our understanding of school participation is limited to anecdotal evidence and studies that report sample characteristics. In the case of the latter, these characteristics themselves are biased as they are subject to the recruitment strategy employed by the researcher. For instance, if a study employed some form of non-probability sampling, differences between that sample and its' intended inference population could be attributed to an uknowable conflation of the recruitment method and non-response. The purpose of this model is therefore utility rather than accuracy.

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

added

Hrmm, any way to improve the rhetoric here?

Sampling Methods

We now 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 that researchers may face. 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. Thus, the recruitment indicators $Z_1, ..., Z_N$ are a function of the scores, $Z_j = Z_j(S)$ for $j = 1, ..., N$. 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,

$$Z_j(S) = I\left[n \geq \sum_{i=1}^{N} E_i I\left(S_i \leq S_j\right)\right] \tag{6}$$

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

$$R(S) = \sum_{j=1}^{N} I\left(S_j \leq S^{max}\right), \tag{7}$$

where $S^{max} = \max\{S_1 Z_1, S_2 Z_2, ..., S_N Z_N\}$.

Several of the sampling methods that we shall consider involve use of stratification. Stratified sampling gives researchers more control over the characteristics of their sample. Consider the case of random sampling, which only produces a representative sample on average. To reduce the likelihood of an "unlucky" sample, a population can be split into strata based on characteristics specified by the researcher. Random sampling from within each of the strata will then ensure that each of those characteristics are represented in the final sample. Stratification is routinely used with probability or non-probability sampling methods in survey research.

To model stratified sampling methods, we will assume that the population is divided into a set of K strata. Proportional allocation will be used to ensure the relative size of the strata in the sample reflects that of the population (Tipton, 2013b). 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

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

and

$$R_k(S) = \sum_{j=1}^{N} I\left(S_j \leq S_k^{max}, G_j = k\right), \tag{9}$$

with $S_k^{max} = \max\{S_1 Z_1 I(G_1 = k), ..., S_N Z_N I(G_N = k)\}$ and $R(S) = \sum_{k=1}^{K} R_k(S)$.

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, using probability sampling eliminates recruitment bias, so that any remaining selection bias is due entirely to schools' participation decisions (i.e., non-response bias). We can simulate unstratified random sampling (URS) by ranking each school at random, so that their order of recruitment S 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 defined by a cluster analysis and using a proportional allocation. For URS, $\Pr(Z_j = 1|\mathbf{x}_j)$ is constant for all $j = 1, ..., N$; for SRS, $\Pr(Z_j = 1|\mathbf{x}_j)$ is constant within each stratum but can vary between strata due to differences in average response rates.

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 and stratified convenience sampling. To operationalize unstratified convenience sampling (UCS), 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 the ranks of participation propensity scores. 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. Let η_j be the integer rank of π_j within the distribution of participation propensity scores $\pi_1, ..., \pi_N$. Then, for UCS,

$$\Pr\left(Z_j = 1|\mathbf{x}_j\right) = c\eta_j \tag{10}$$

for a scaling constant c. Under this model of convenience sampling, the school with the highest response propensity is twice as likely to be sampled as the school with the median response propensity; the school at the 80th percentile is twice as likely to be sampled as the school at the 40th percentile; etc. Because they depend only on the ranks of the response propensities, the sampling probabilities will remain the same for location shifts and scale shifts in the distribution of response propensities. This property allows us to manipulate the parameters of the response propensity model (Equation (4)) without altering the behavior of the convenience sampling process.

We operationalize stratified convenience sampling (SCS) using the same process as for UCS, but with ranks determined independently within each stratum, and using a proportional allocation across strata.

Balanced Sampling

SBS entails first stratifying the population, then selecting units so that the sample from each stratum matches the corresponding sub-population on a set of covariates. Let μ_{pk} denote the mean of covariate p in population stratum k, for $p = 1, ..., P$ and $k = 1, ..., K$. Given a set of strata, SBS is designed to select units so that

$$\frac{1}{n_k}\sum_{j=1}^{N} X_{pj} \times I\left(Z_j = 1, G_j = k\right) - \mu_{pk}, \tag{11}$$

is small in absolute value for all $k = 1, ..., K$ and $p = 1, ..., P$.

Following Tipton (2013b), we operationalize balanced sampling by prioritizing the recruitment of sites based on their similarity to the “average” site in each stratum. Specifically, we ranked each site within a stratum using a weighted Euclidean distance measure, with sites closer to the center of the strata ranked higher. The scores for SBS are thus:

$$S_j = \sqrt{\sum_{p=1}^{P} w_p (X_{pj} - \mu_{pG_j})^2} \tag{12}$$

where w_p is a weight assigned to covariate X_p. Different weights could be used such that distances depend more heavily on covariates thought to be more related to treatment effect heterogeneity. In the absence of information about predictors of treatment effect heterogeneity, we set the weight of each covariate equal to the inverse of its population variance as $w_p = 1/\sigma_p^2$.

Is this consistent with the Generalizer/Tipton 2013 approach?

SBS is unique in that rankings are directly related to school characteristics and are not stochastic (i.e., they do not change across replications of the process). Scores within strata are based on equation (12), where schools closer to the "center" of the stratum are more representative 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

Using the formal models for unit participation and sample recruitment that we have described, it becomes possible to simulate the process of recruiting schools for an MRT. Monte Carlo simulations, such as the one we now describe, are a flexible method for assessing the performance of different sample recruitment methods under controlled conditions. Here, we use simulation to assess the generalizability of samples recruited through SBS relative to probability sampling or convenience sampling methods. We also examine the feasibility and equity with which the sampling methods can be employed.

The validity of simulation study findings rests on whether the models and assumptions employed are reasonable approximations to the processes involved in collecting real data. In particular, our simulation findings hinge on the assumptions of the participation propensity model and recruitment models. By relying on previous work and real data to inform our modeling assumption, we hope to reasonably represent trends in the performance of different recruitment methods. That said, given the speculative nature of some aspects of the simulation, our conclusions will necessarily be tentative.

Sampling frame

We study the performance of different recruitment methods for a study aimed at generalizing to the population of schools across 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. 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 the sample to include schools not normally found in large-scale evaluations of interventions (cf. Stuart et al., 2017; Fellers, 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

Leaving for now

Trim or condense this sentence?

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

Descriptive statistics for population of $N = 9,792$ *schools and log odds of participation associated with each covariate (Fellers, 2017)*

Variables	M	N	Log-odds of participation
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

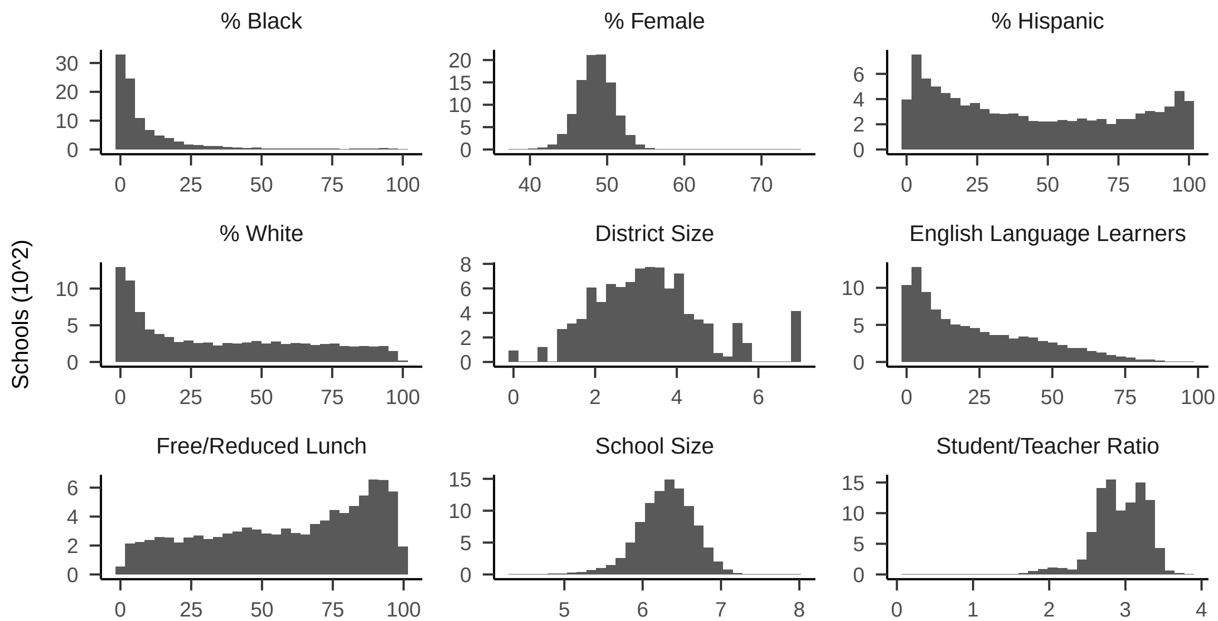

Figure 1. Distributions of continuous covariates. District Size, N Students, and Student/Teachers were transformed by taking natural logs.

Stratification

For recruitment methods involving stratification, we use strata generated through a cluster analysis of the school and student characteristics listed in Table 1. Our cluster analysis generally followed the approach and recommendations described by Tipton (2013b). Specifically, we implemented k-medoids clustering to divide the population into heterogeneous strata comprised of homogeneous sites. We used Gower's general similarity measure, which is appropriate for covariates that include both continuous and categorical variables (Gower, 1971; Tipton, 2013b). All analyses were performed in R (R Core Team, 2017) using the *cluster* package (Mächler, Rousseeuw, Struyf, Hubert, & Hornik, 2012). Section XX of the online appendix includes further details about the cluster analysis and stratification process.

Based on empirical criteria as well as subjective appraisals of what is feasible to implement, we chose a cluster solution with five strata. The strata ranged in size from

1,222 to 2,863 schools and explained 70.43% of the variation across all covariates. The proportion of variance explained by the strata differed by covariate, with five covariates having between 0 and 11% variance explained by the strata and eight covariates having from 45% to 87% variance explained. This discrepancy suggests that stratification will not be equally useful for balancing all covariates.

Further take-aways? Fore-shadow simulation findings?

Participation Propensity Score

In order to simulate the full process of recruiting a sample for an MRT, we need to choose parameter values for the participation propensity score model (Equation (4)). To do so, we drew 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 on the same variables used in our sampling frame. Note that this comparison is between schools that participated in trials (i.e., $P_j = 1$) and the full population of schools, and so differences may be the result of *both* non-response bias and researcher recruitment bias. In contrast, we are ideally interested in $\Pr(E_j = 1|\mathbf{x}_j)$, the *potential* participation probabilities across the full population. However, there is little data currently available to directly inform this participation model, and so we relied on the findings of Fellers (2017) to make tentative assumptions about the magnitude of the participation propensity model coefficients.

Fellers (2017) reported absolute standardized mean differences between schools that participated in IES-funded randomized trials and the population. We standardized the covariates and used reported SMDs as coefficients in Equation (4) to generated π_j values. The coefficients are reported in Table 1. These values lead to large variation in the response propensity scores. In order to examine the extent to which our results are driven by this assumption, we considered two further conditions, in which the propensity model coefficients were divided by a scaling factor of two or four. These conditions lead to successively weaker relationships between the covariates and the response propensity.

The results of Fellers (2017) do not provide any information about the overall potential participation rate across the full population–that is, the fraction of the full population of schools that would agree to participate in a study, if the entire population were recruited. This is a key parameter of the simulation that has a strong influence on the performance of all recruitment methods. We therefore varied the overall participation propensity over a wide range, by manipulating the intercept term in Equation (4), given the other coefficients of the propensity score model and the scaling factor. We considered nine different levels for the average participation rate, ranging from 10% to 90%.

Simulation procedures

We designed the simulation to examine the performance of 5 recruitment methods across 9 levels of population response rate and 3 levels of the scaling factor in the response propensity model. The population of schools and their assigned strata were constant across all conditions and replications. The participation propensity score was constant across all sampling methods and replications within each condition for the response propensity model. Each replication involved first simulating potential participation indicators (E_j's) according to Equation (5), which we then treated as constant across each sampling method. That is, if a given school was selected in each of three sampling methods, the school would provide the same response for all three methods, either agreeing or refusing to participate. Next, we generated five sets of rankings (one for each recruitment 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. The order in which schools were approached varied by recruitment method and across replications. The one exception was the stratified balanced sampling method, where the order in which schools are approached was constant because it is a function of stratum assignment and school characteristics.

Performance measures

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 (1). 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 full generalizability, however, we should consider how the full distribution of the sample compares to that of the population.

A more comprehensive measure of generalizability is the B index (B; Tipton, 2014), which provides a single score summary of overlap between two groups. This index offers several advantages. First, it takes into account both the variance and distribution of the variables in each group by measuring the overlap of multiple density plots. Second, it quantifies what is typically a visual comparison and provides an intuitive interpretation. Third, it simultaneously compares two groups across a range of covariates and can easily be scaled to include various interactions and transformations.

The B index is defined in terms of a distribution of propensity scores, which are used to summarize into a single dimension the differences between an obtained sample and the population distribution of the covariates. For a given recruitment method, let ψ_j represent the sample selection propensity score, or the logistic transform of the probability that school j was included in an obtained sample:

$$\psi_j = \log\left(\frac{\Pr(P_j = 1|\mathbf{x}_j)}{1 - \Pr(P_j = 1|\mathbf{x}_j)}\right).$$

Note that this sample selection propensity score is distinct from the potential participation propensity model (Equation (4)), in that ψ_j is based on a sample actually obtained under a given recruitment method, allowing for units to self-select or decline to participate. For each recruitment method and each obtained sample, we estimated ψ_j using a basic logistic regression:

$$\psi_j = \gamma_0 + \mathbf{x}_j \boldsymbol{\gamma} \tag{13}$$

The estimated values of the sample selection propensity score, $\hat{\psi}_j$, serve as scalar summaries of differences across all covariates in $\mathbf{x}_j$.

The generalizability index compares the distribution of $\hat{\psi}_j$ in the sample to the corresponding distribution in the population. The 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. Given the estimated propensity scores $\hat{\psi}_1, ..., \hat{\psi}_N$, we calculated the B index as

$$B = \int_{\hat{\psi}_{min} - 3h_{max}}^{\hat{\psi}_{max} + 3h_{max}} \sqrt{\hat{f}_s(z)\hat{f}_p(z)} dz \tag{14}$$

where $\hat{\psi}_{min}$ and $\hat{\psi}_{max}$ are the minimum and maximum values of $\hat{\psi}$ respectively, $\hat{f}_s(z)$ is a Gaussian kernel density estimate of the distribution of propensity scores in the sample, calculated using bandwidth h_s, and $\hat{f}_p(z)$ is a Gaussian kernel density estimate of the distribution of propensity scores in the non-sampled population, calculated using bandwidth h_p. We calculated the bandwidth h for the population or sample as follows:

$$h = \sigma_\pi \left(\frac{4}{3n} \right)^{1/5} \tag{15}$$

where σ_π is the standard deviation of the propensity scores.

If need be, we could probably move these two paragraphs to on-line appendix to save space.

Feasibility. Recruiters expend a lot of resources contacting districts and schools, scheduling meetings and traveling between prospective schools. A project with limited resources may not be able to afford to go through a large list of potentially uninterested schools. To capture the feasibility of different recruitment methods, we tracked the total number of schools approached for each full sample obtained (Equations (6) and (8)). We then calculated the average number of refusals needed to achieve a full sample for each recruitment method. This measure allows us to compare the feasibility or difficulty of recruiting a full sample using each method.

Sampling Inequity. This simulation design also enables us to study sampling inequality. The distribution of opportunities to participate in research is an important issue when considering federally funded research. Non-random sampling for such studies create the possibility of systemic under-representation of some types of schools, which then do not share in the benefits of participating in research, or systemic over-representation of other schools, which more often bear the burden of participation in such research.

To assess sampling inequity, we tracked the frequency with which schools were recruited by each method. We then summarized these data by calculating a Gini coefficient for each recruitment method and each population participation rate. This index ranges from 0 to 1, with 0 indicating perfect equity, and 1 indicating total inequity. In economic contexts, the Gini coefficient is used to describe the degree of income inequality by showing the disproportionate distribution of wealth across levels of incomes. Similarly, we use it in the context of sampling to summarize the distribution of opportunities to participate in research. We calculated Gini coefficients using the *Gini* function from the *ineq* package (Zeileis, 2014).

Results

Generalizability

B-Index. Figure 2 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%, SBS consistently out-performed the other recruitment methods. Further, across population response rates, SBS never resulted 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.

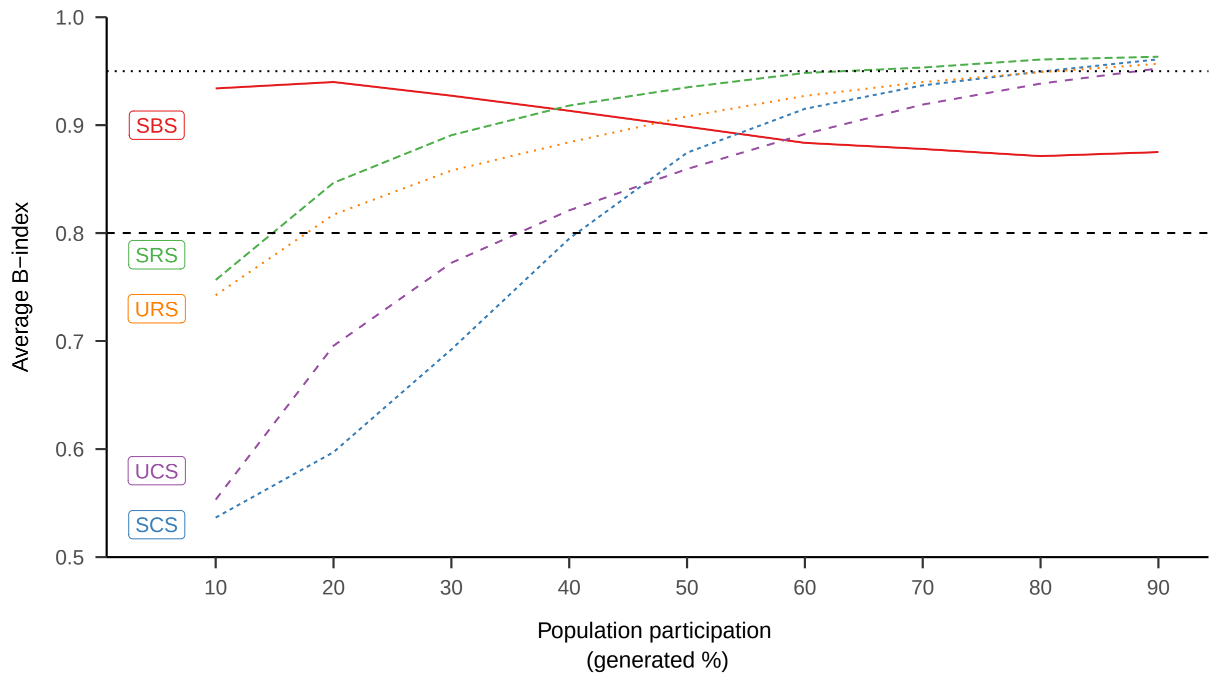

Figure 2. Average B-index for varying participation rates, by sampling method. Horizontal dotted and dashed lines represented indices of .80 and .95, respectively, indicating acceptable or high levels of generalizability.

We also found several trends that were unexpected. At 50% and beyond, SBS performance slowly degraded, while other methods maintained 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 achieved better balance than stratified convenience sampling. This seems counter-intuitive, as survey literature suggests that stratification produces more representative samples (Lohr, 2019). Because the B-index is an overall measure of generalizabilty across many covariates, it does not provide much indication of why these performance differences occur. To untangle why the methods performed as they did, we need to examine the balance obtained by each recruitment method on individual covariates.

Standardized Mean Differences. We began this analysis by plotting mean SMDs for each recruitment method on each individual covariate versus the population participation rate. We considered a method to perform well if it resulted in an average imbalance of no more than 0.25 standard deviations. The relative performance of sampling methods varied depending on the covariate. Several patterns emerged, based on which we identified three groups of covariates. Figure 3 displays an example of each pattern; results for specific covariates can be found in Online Appendix Figures 8, 9, and 10.

The first group of covariates included nine variables where stratified methods outperformed unstratified methods at every population participation rate. The second group included two covariates where all sampling methods resulted in good balance. The third group included three covariates where where at least one unstratified method performed better than the corresponding stratified method. Across all groups, stratified balanced sampling almost always resulted in acceptable balance and generally yielded better balance than 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. Figure 4 contains a plot of the proportion of variance explained by the strata for each covariate versus the log-odds coefficients associated with that covariate in the participation propensity model. It can be seen that Group 1 consisted of covariates where a substantial fraction of the variance was between clusters. Groups 2 and 3 both consisted of covariates that were poorly clustered, with no

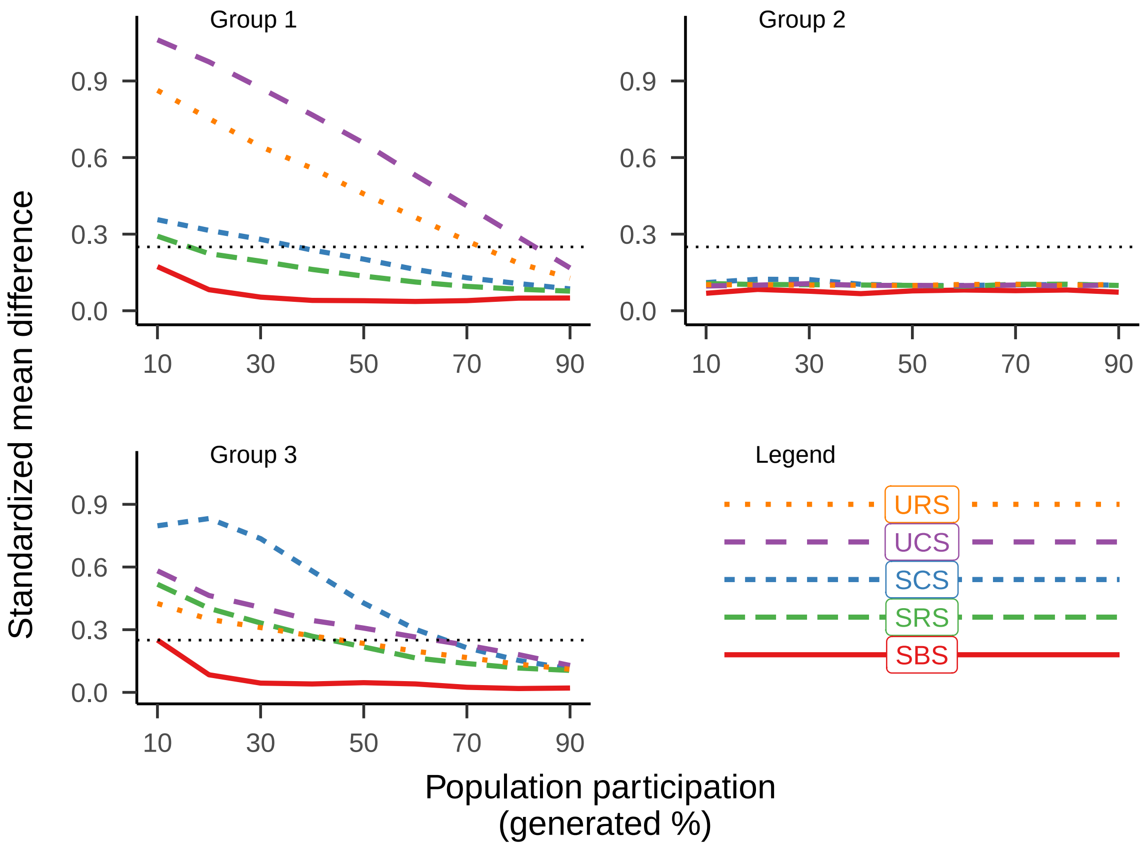

Figure 3. Patterns of relative performance based on standardized mean differences. The dotted horizontal line represents a threshold of 0.25 for acceptable balance.

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.

Further explain each group. How does the plot explain the performance of stratified methods for each set of covariates?

Feasibility

Recruitment Attempts. Figure 5a depicts the total number of schools contacted to recruit a full sample of 60 schools versus the population participation rate, for each recruitment method. Figure 5b depicts the response rate, or percentage of schools that agreed to participate among those where recruitment was attempted, versus the population participation rate, for each recruitment method. For each of these

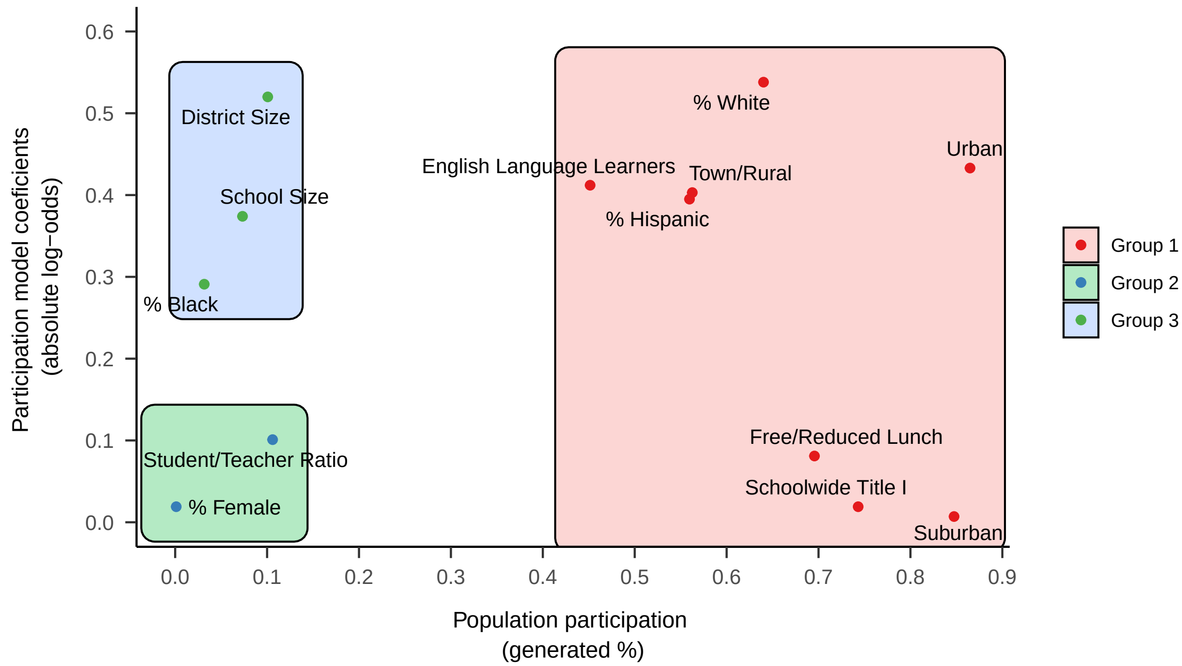

Figure 4. Variance explained by clustering vs absolute log odds. Shaded areas illustrate patterns in generalizability measured by SMDs.

metrics, there were substantial differences between methods at lower population participation rates. At higher participation rates, the differences were less pronounced, becoming negligible at population participation rates of 60% or more.

So what? What is the take-away?

It is important to note that, particularly in the lower population participation rates, the total schools contacted values are quite extreme. Due to the speculative nature of the simulation, a more meaningful interpretation would be to compare the relative performance of the recruitment methods. Overall, UCS required the smallest recruitment effort to obtain a full sample, followed by URS and SCS, SRS, and finally SBS. Online Appendix Figure 11 provides further detail about the relative feasibility of each recruitment method.

Sampling Inequity. We calculated the Gini coefficient for each sampling method to examine the equity implications of different recruitment methods. Figure 6 displays these for each sampling method across population participation rates. A

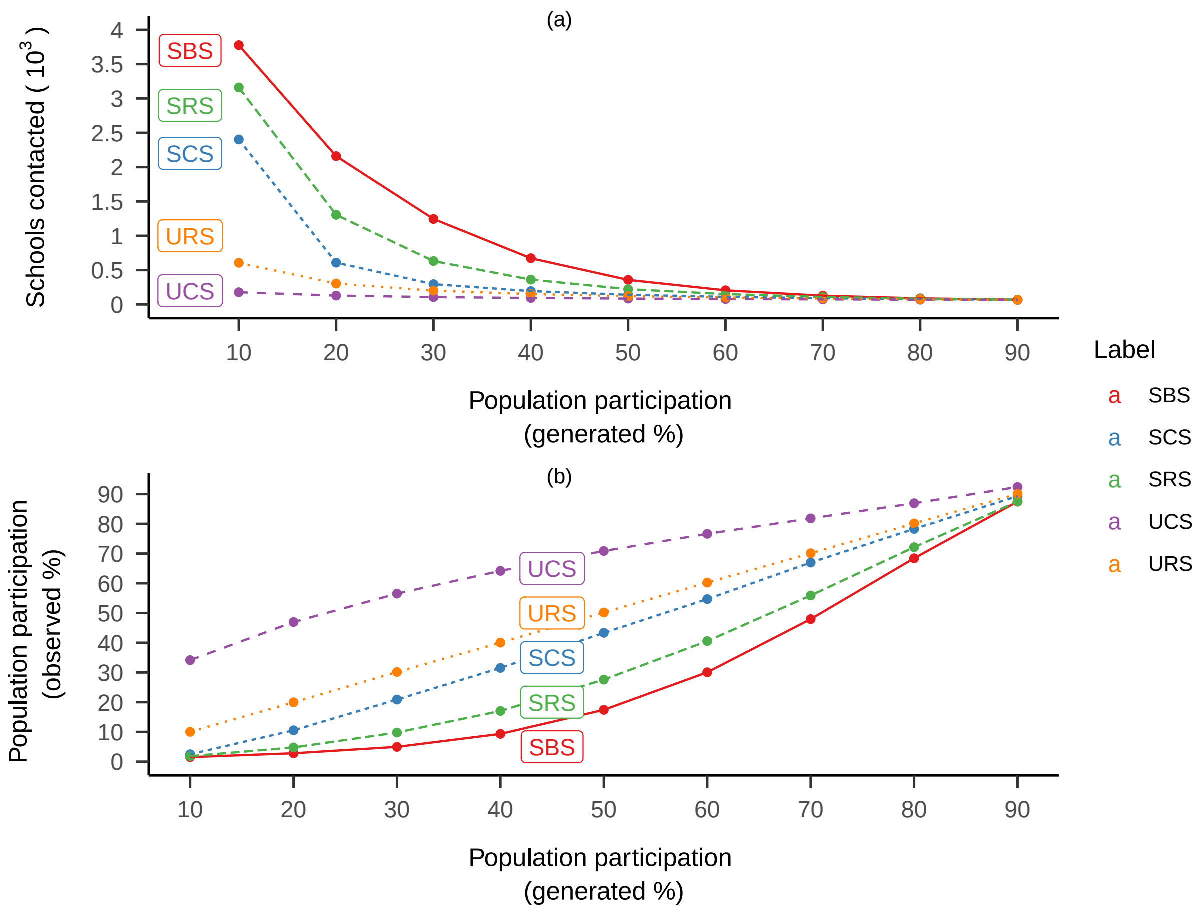

Figure 5. Sample recruitment statistics and response rates. Plot (a) shows the total number of units contacted to achieve a full sample of 60 schools. Plot (b) shows the percent of schools that agreed to participate when recruited.

coefficient of 1 indicates absolute sampling inequity. In the context of the simulation, a high Gini coefficient indicates that across replications, only a small subset of the population actually participated. Indovidual Gini curves and coeficiencts for each method accross response rates can be found in Online Appendix Figure 12.

Recruited or participated?

Participated

Then the methods section needs to clarify that too.

Several trends emerged in this analysis. Overall, SBS was the most inequitable sampling process. This is consistent with the balanced sample approach because order of recruitment is tied to school characteristics and therefore the population frame. SRS was the most equitable sampling process, followed by URS, SCS, and UCS. As population participation increased, inequity increased when using SBS, but decreased considerably when using the other methods, and differences between these methods diminished. This finding suggests that stratifying the population results in a larger potential sampling pool when the overall population response rate is low. However, stratification alone is still limited by the recruitment method within strata, and any method that does not involve probability sampling inherently results in increased sampling inequality.

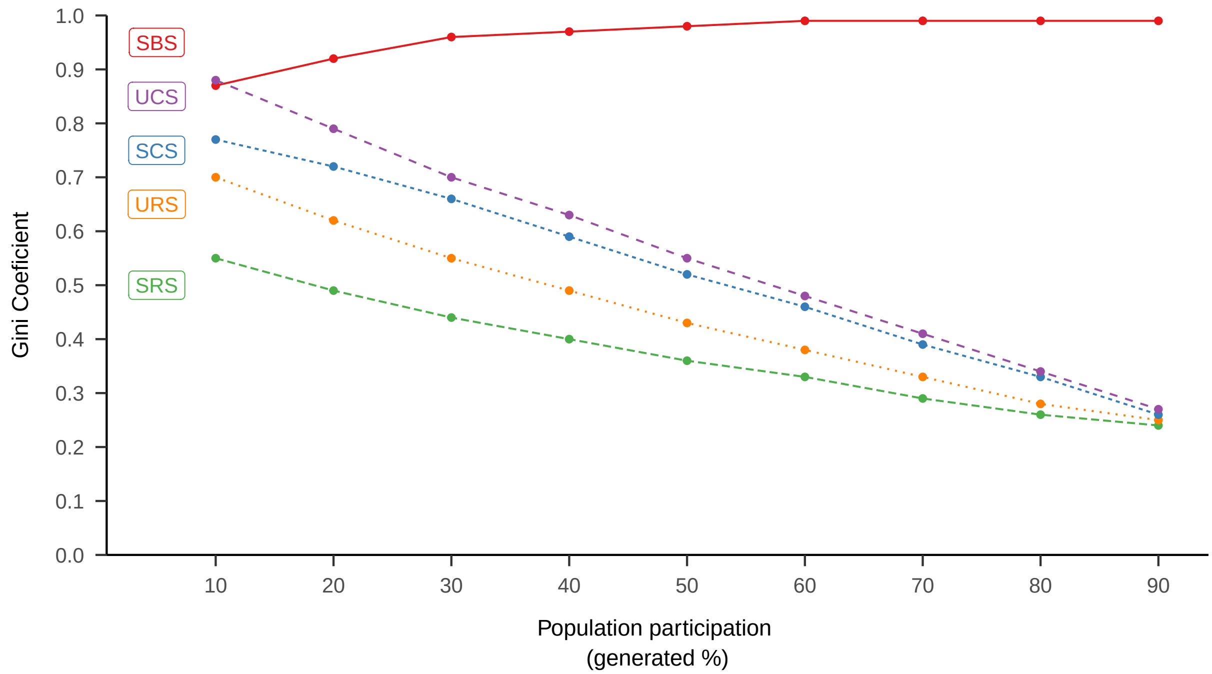

Figure 6. Gini coefficient across participation response rates for each sampling method. A coefficient of 1 indicates major inequality in sampling probability.

Summary of Trends

Under the assumptions of our simulation models, recruiting samples using SBS yielded highly generalizable samples. In particular, SBS resulted in a considerable improvement compared to UCS—the method that we suspect most closely approximates current practice. However, given the large number of schools contacted to obtain a full sample, it may be infeasible to fully implement SBS in the ideal form. Another consistent pattern was that convenience and probability sampling methods were often improved by first stratifying the population. In fact, in cases where the cluster analysis performed well for a given covariate, or where popoulation participation rates were high, convenience sampling within strata (SCS) performed as well as or better than simple random sampling (URS) on measures of both generalizability and feasibility.

A further pattern in these simulation results relates the specification of the clustering method to the response model. Covariates that are highly related to determining participation must be prioritized in generating the strata. If such covariates are poorly allocated across strata (i.e., there is little heterogeneity between strata) then the purpose of stratification is undercut and it will not result in good balance on the individual covariate. It is therefore important to not only examine overall variance explained when performing the cluster analysis, but to use individual covariate variability as a criterion as well.

Wonder if this isn't more of a discussion point (because it's drawing implications, rather than just summarizing.)

Discussion

The main goal of this study was to develop a framework for exploring the performance of different recruitment methods within the context of multi-site trials in education research. We have described several models for algorithmically representing how researchers might select samples of schools. Drawing on prior empirical 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 second goal was to use this framework to investigate the relative perfomance of different several sampling methods in terms of generalizability and feasibility. Based on these simulation, we draw several tentative 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 application of SBS came at a great cost in terms of the 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 also found that the properties of the strata generated by cluster analysis have implications for the generalizability of samples recruited through stratified methods. If a covariate was predictive of participation propensity, then it was important for the stratification to explain a substantial fraction of its total variation. If a covariate was predictive of participation, but not prioritized when generating the strata, it had poor representation in samples collected through stratified methods. Ignoring the the quality of the cluster analysis solution has the potential to undermine the performance of SBS. Finally, while the balanced sampling approach does result in strong generalizability, it might also limit the pool of potential participants. Particularly for larger population response rates, the same subset of schools were likely to be recruited across replications of the simulation.

One potential compromise between current practice (such as UCS) and SBS is to combine stratification with some form of convenience sampling. As demonstrated in the simulations, stratified convenience sampling often resulted in better balance on individual covariates than simple random sampling, and required lower recruitment effort compared to balanced sampling. This may also elicit greater buy-in from recruiters by placing less restrictions on which 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 on 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 may 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 the researchers' own personal biases. Due to this limitation, caution is warranted in interpreting the simulation findings regarding absolute performance and absolute feasibility of convenience sampling methods.

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, their estimation of "willingness" may be further from the truth than what we have assumed. If so, it is possible that the feasibility of the convenience methods is over-stated and that their degree of generalizability is under-stated.

It would be worth refining and exploring additional methods for modeling convenience sampling. The algorithms used in these methods could be tuned to include additional factors that might influence school recruitment priorities. For instance, location data is readily available and could 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 of recruitment methods given the unique circumstances that they face during the study design phase.

A second limitation arises from the simplified and speculative nature of our participation model. In practice, the decision of whether a school participates in such a study often involves multiple stages. District offices often serve as gatekeepers, requiring submission and approval of research requests prior to starting recruitment of individual schools. 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. Our participation model assumed that only school level characteristics play a roll in this decision, when in reality a multilevel model accounting for district and individual administrative characteristics would be a more realistic representation of the process. However, we are not aware of data sources that would allow us to develop even very rough estimates of such a model's parameters.

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.

This limitation points towards the need for more extensive research regarding factors that influence school participation in RCTs. In particular, research is needed that can disentangle recruitment bias from selection bias. Such information could be generated through careful record-keeping during the recruitment stages of MRTs, particularly if the recruitment methods are operationally defined and if the study aims for large-scale population representation. Such efforts could provide 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

In this study, we have sought to lay the groundwork for several avenues of further research. First, additional work is needed on how best to implement cluster analysis in the context of stratified sampling methods. We have shown that the extent to which balance is achieved on a given covariate is related to how strongly the strata explain variation in the covariate and how strongly the covariate is related to school participation. If some covariates are known to have greater influence on school participation, it may be useful to weight them more heavily in generating the strata. Further work is also needed to understand the relationship between the number of strata, 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, potentially making SBS more accessible and quicker to implement.

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 may extrapolate that a sample that is balanced on these covariates can be used to estimate an unbiased PATE. Still, we think it would be worth investigating the full process–from study design to parameter estimation–to understand how sampling methods might influence the bias in impact estimates for a target population.

If treatment effects are constant across units in a population, non-representative 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 sufficient variability in potential moderators to detect heterogeneity. Adding variation by selecting a more diverse sample may be useful if the extent 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.

I don't quite follow how this fits in. Let's discuss this point.

We didn't discuss this, should I drop for now?

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. Through further refinement of the models and simulations demonstrated here, we hope to help researchers make more informed decisions about sample recruitment and representation.

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

Imai, K., King, G., & Stuart, E. A. (2008). Misunderstandings between experimentalists and observationalists about causal inference. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *171*(2), 481–502. https://doi.org/10.1111/j.1467-985X.2007.00527.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

Lohr, S. L. (2019). *Sampling: Design and Analysis* (2nd Edition). Chapman and Hall/CRC.

Mächler, M., Rousseeuw, P., Struyf, A., Hubert, M., & Hornik, K. (2012). *Cluster: Cluster Analysis Basics and Extensions* (Vol. 1).

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

R Core Team. (2017). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

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. https://doi.org/10.3102/0013189X20960686

Zeileis, A. (2014). *Ineq: Measuring inequality, concentration, and poverty.* Retrieved from https://CRAN.R-project.org/package=ineq

Online Appendix

Stratification with Cluster Analysis

This section provides further details about how we used cluster analysis to determine the population strata used in the simulation study. Cluster analysis serves as a dimension reduction tool to condense a high-dimensional distribution of population characteristics into a small set of homogeneous strata. Cluster analysis requires 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, 2017) using the *cluster* package (Mächler et al., 2012).

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:

$$d_{pij} = 1 - \frac{|X_{pi} - X_{pj}|}{R_p} \tag{16}$$

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_{ij}^g be the general similarity between site i and site j.

$$d_{ij}^g = \frac{\sum_{p=1}^{P} w_{pij} d_{pij}}{\sum_{p=1}^{P} w_{pij}} \tag{17}$$

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 unequally 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 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 7).

We first calculated the Calinski-Harabasz (CH) index using the *cluster.stats* function from the *fpc* (Hennig, 2019) package. Figure 7a 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. For each set of $K = 1, .., 10$ total 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 7b plots p_k against K. The K for which the rate of change p_k slows is considered favorable. Tipton (2013b) also recommended selecting the number of clusters such that at least 80% of the variability is between clusters; this threshold is indicated in the figure by 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 7c plots the required sample size 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. When $K > 5$, the smallest clusters would require less than 5 sites being sampled, which may be very difficult in practice. Considering both the statistical criteria and the pragmatic constraints, we generated 5 clusters for purposes of stratification.

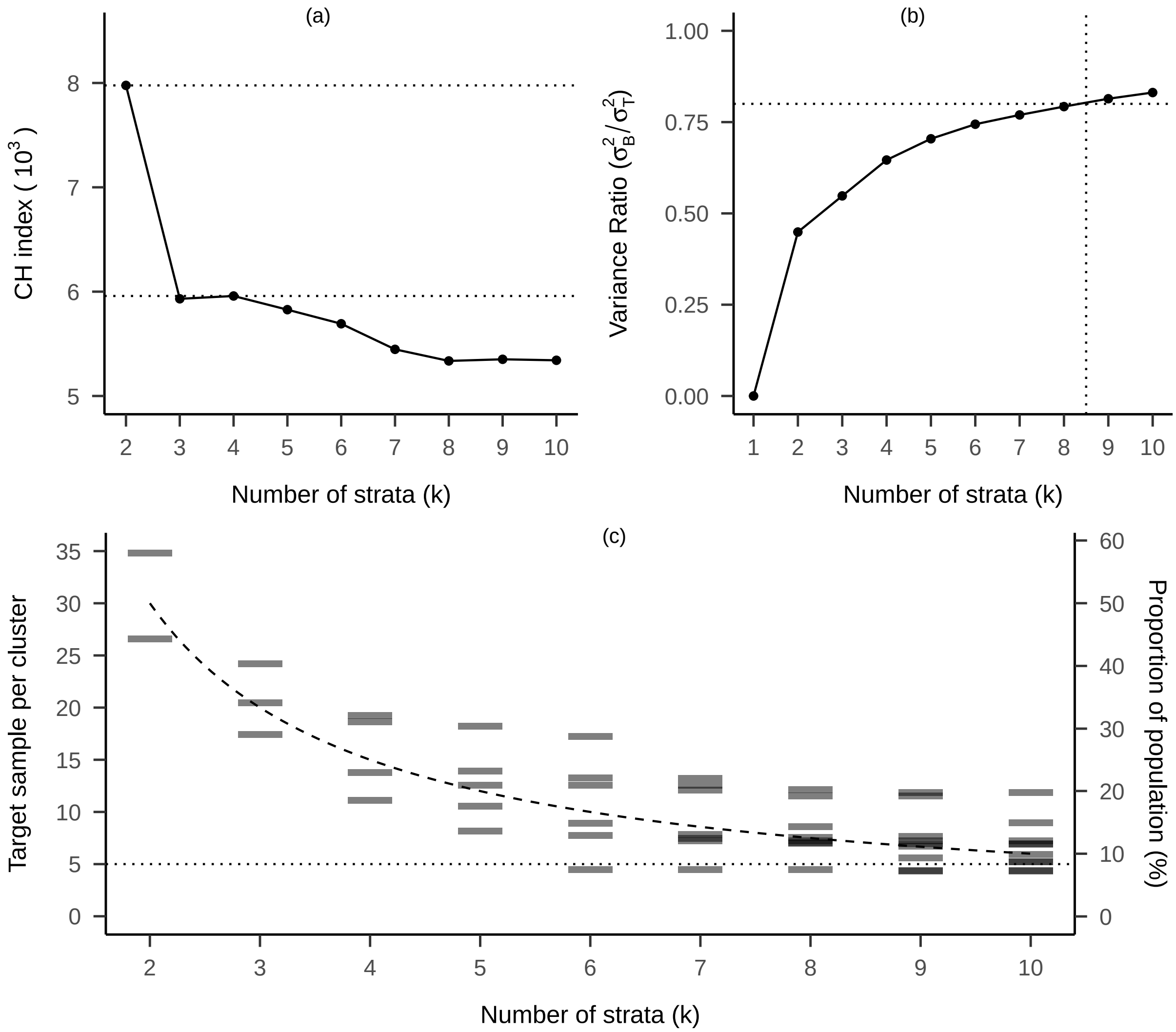

Figure 7. Plots used to determine the number of strata (K). (a) Calinski-Harabasz index; peaks indicate better fit. (b) Ratio of between cluster sum of squares to total cluster sum of squares; horizontal line indicates cutoff of .8, vertical line indicates minimum number of clusters needed to achieve cutoff. (c) Sampling requirements for each cluster given proportional allocation; horizontal dotted line indicates a minimum sample size requirement of 5 schools.

Variation explained by the strata

The strata resulting from the cluster analysis range in size from 1,222 to 2,863 schools. The strata explain 70.43% of the variation across all covariates (Figure 7b), with a CH-index of 5,827.31 (Figure 7a). We also computed variance explained for individual covariates. The proportion of variance explained by the strata differed greatly by covariate, with 5 covariates having between 0 and 11% variance explained by the strata and 8 variables having from 45% to 87% variance explained.

Standardized Mean Differences

Figures 8, 9, and 10 display individual covaraite balance as measured by standardized mean differences. Each group of plots correspond to the three patters on performance described in the paper.

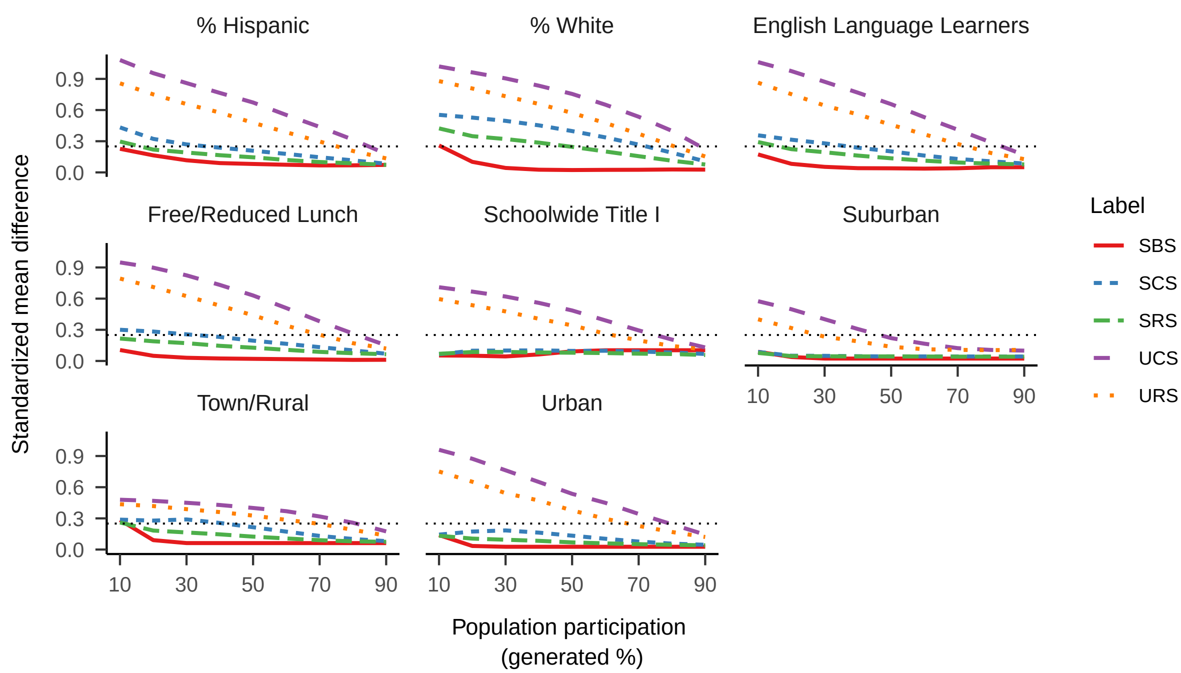

Figure 8. Group 1: Sampling methods performed as expected. All stratified methods outperformed non-stratified methods. Better balance was attained by balanced sampling followed by random sampling and convenience sampling.

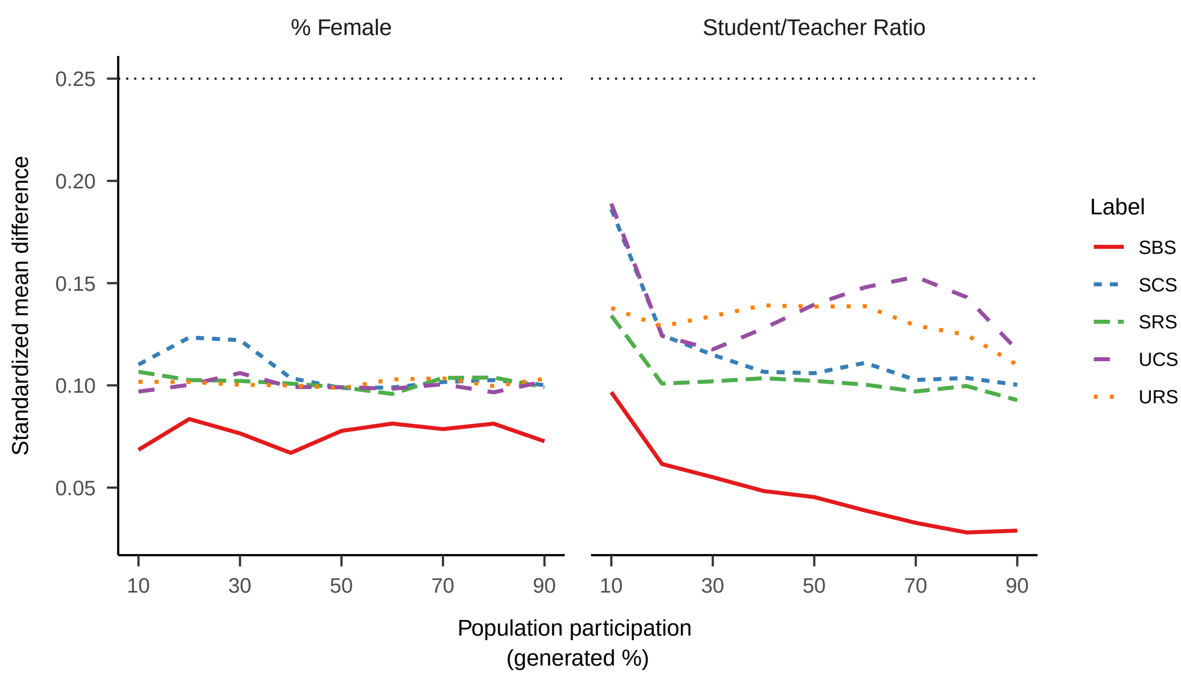

Figure 9. Group 2: Sampling methods equally well resultiing in good balance.

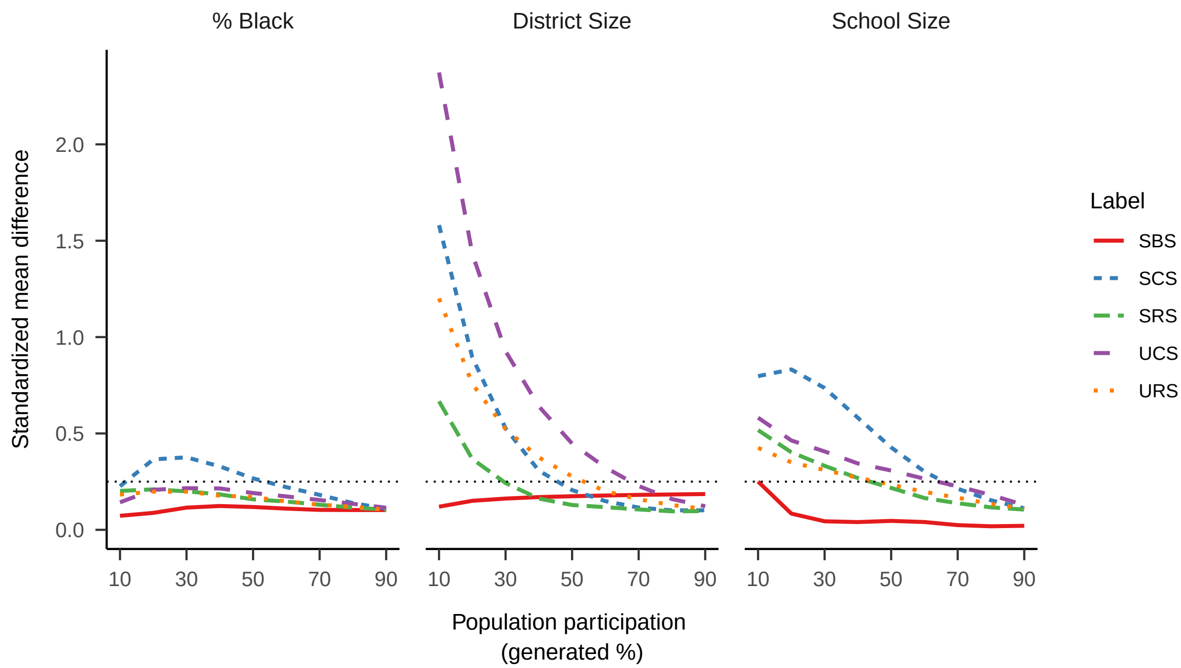

Figure 10. Group 3: Stratification sometimes resulted in poorer balance.

Sampling Feasibility

Figure 11 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.

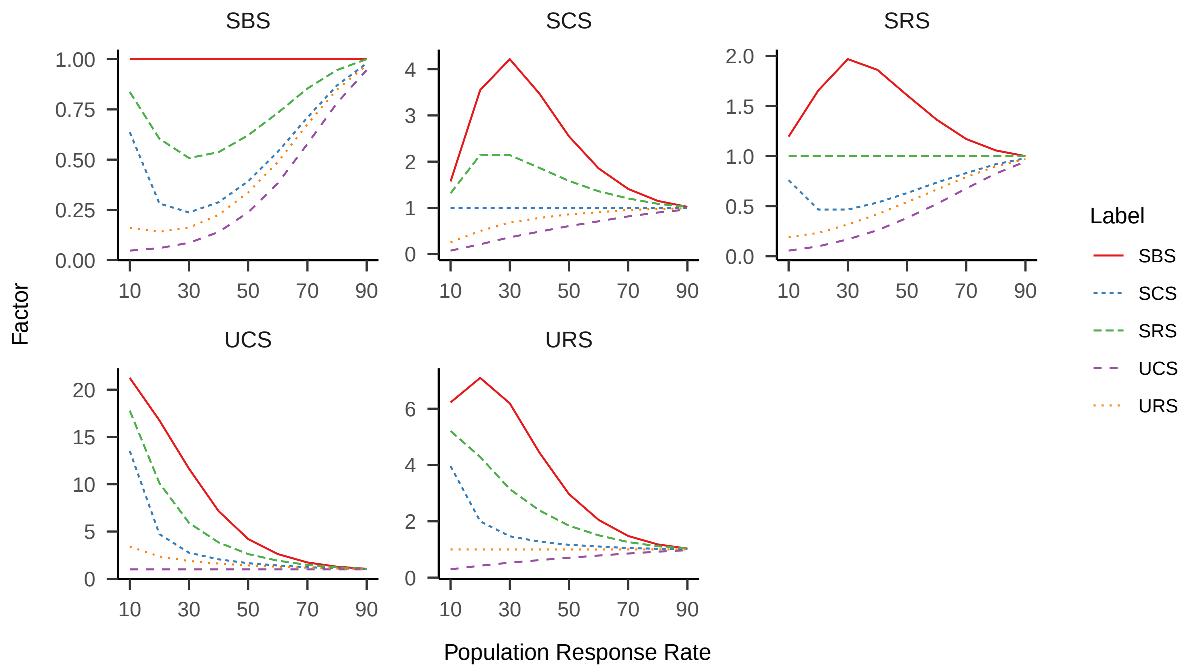

Figure 11. Relative sampling difficulty of each method compared to other methods. The straight horizontal line indicates the reference method being compared to.

Sampling Inequality

Figure 12 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.

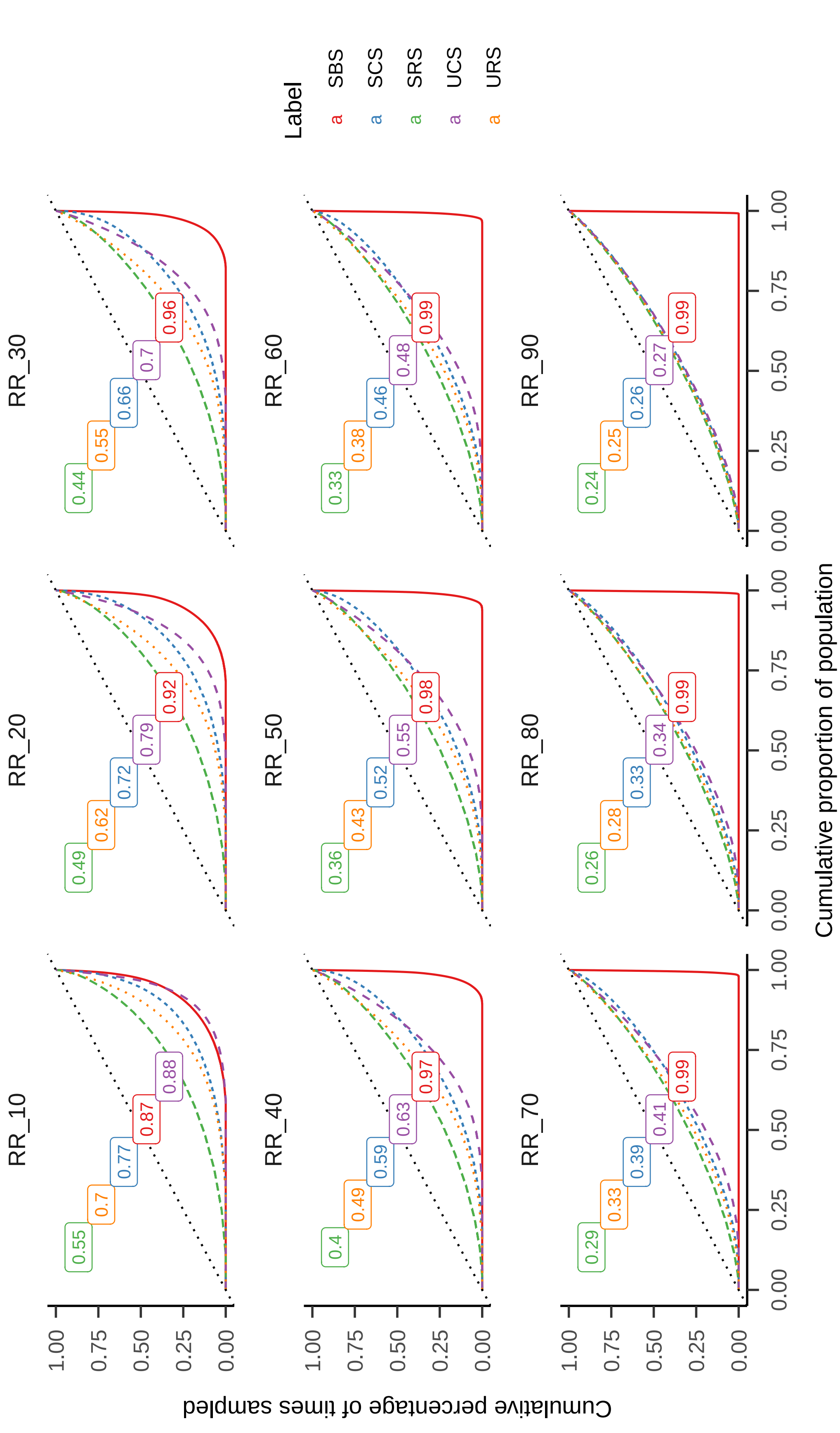

Figure 12. Cumulative probability plot and Gini coefficients representing the inequality of school sampling across sampling methods and population response rates.