

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

Gleb Furman<sup>1</sup> & James E. Pustejovsky<sup>1</sup>

<sup>1</sup> University of Texas at Austin

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

### Introduction

The multi-site randomized trial (MRT) has become a common design in educational research, particularly in evaluations of intervention effectiveness. In essence, an MRT is a randomized control trial (RCT) that takes place accross multiple distinct sites (e.g. geographic locations), with random assignment taking place either at the site or unit level. In education such a design may consist of multiple schools being recruited from one or more districts. Once a sample of sites is recruited, students, teachers, classes or whole schools can be randomly assigned either to receive the intervention (treatment group) or to continue business as usual (control group).

The advantage of an MRT design over a single-site RCT is that it maintains the high level of internal validity via randomized treatment assignment, while also supporting a greater degree of generalizability by introducing cross-site variability. Ostensibly, this allows effect estimates to generalize to a larger population than estimates from a single-site design (Raudenbush & Liu, 2000). However, the ambiguity of how this larger population is defined, and the overall generalizability of large-scale MRTs, has increasingly come under scrutiny.

Generalizability is a multi-faceted causal inference problem that falls under the purview of external validity. For the purpose of the present study, we narrowly define generalizability as the extent to which a sample represents a well-specified population of interest. In the presense of treatment effect heterogeniety, or variability in response to intervention across persons or sites, estimating sample specific average effects may not be relevant beyond the units in the study. This makes intervention evaluations less informative to policy-makers who seek to make evidence-based decisions in populations that are unrepresented by the study sample.

In this context, inadequate sampling strategies have been shown to be a major threat to generalizability of MRTs. Purposive or non random samples may lead to substantially biased estimates of population level effects (Olsen, Orr, Bell, & Stuart, 2013; Shadish, Cook, & Campbell, 2002). Several studies have found substantial differences between schools and districts that participate in large scale randomized trials (Fellers, 2017; Stuart, Bell, Ebnesajjad, Olsen, & Orr, 2017). Beyond accurate effects estimates, generalizability can also be a question of equity. For instance, these studies have shown that small underserved rural districts are underrepresented in RCTs sponsored by Institute of Education Sciences, and therefore are less likely to benefit from federally funded research.

RCTs provide a high level of internal validity as they demonstrate the causality of the impact being estimated. In education research, multi-site RCTs, or multi-site randomized trials (MRTs), are used to inform policy decisions by evaluating intervention impacts on a larger scale. Recruitment and randomization occurs at the organizational level (e.g. schools), ostensibly increasing the diversity of settings in which the study takes place. In practice, however, the generalizability of MRTs is greatly limited by the sampling method being implemented. This is a major drawback for policymakers who are interested in generalizing treatment effects beyond the sample.

RCTs are commonly used in impact evaluations across many fields. Randomized treatment assignment ensures that estimated effects are causal, thus affording a higher level of internal validity. However, these effect estimates may be sample specific. If treatment effects are heterogeneous, then the impact of an intervention cannot be extrapolated beyond the study sample. This is a limiting feature of RCTs since policymakers seeking to implement interventions will have no way of gauging effectiveness in their population of interest.

Random sampling overcomes this limitation by selecting sites from a well-defined population with some known probability. When used in conjunction with randomized treatment assignment, this is known as dual randomization. Assuming no refusals during

recruitment, and full compliance without attrition after assignment, this design enables unbiased estimation of the sample average treatment effect (SATE). Using the known sampling probabilities, the SATE can then be used to estimate the population average treatment effect (PATE).

In absence of random sampling, PATE can still be estimated by using propensity score techniques, provided there is sufficient overlap between the sample and the population (Kern, Stuart, Hill, & Green, 2016; O’Muircheartaigh & Hedges, 2014; Tipton, 2013a). This is especially useful in impact evaluation as probability sampling is rarely used in this context (Olsen et al., 2013; Shadish et al., 2002). Instead, researchers often opt for convenience or purposive samples. These methods much less expensive to implement, but are not usually designed for representative sampling. This makes them susceptible to substantial differences between the sample and target population, referred to as under-coverage (Groves, 2004).

Under-coverage can be assessed using several techniques (Stuart, Cole, Bradshaw, & Leaf, 2011; Tipton, 2014) which identify how well a sample would generalize to a specific population. When under-coverage is too great, the full PATE cannot be recovered with the given sample, and the population needs to be trimmed by removing subsets of sites that are not represented. This can greatly diminish the relevance of study results and undermine the substantial investment into large-scale MRTs.

A series of recent papers instead advocate planning for generalizability at the recruitment stage (Tipton, 2013a, 2013b). These methods 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 attracted attention from researchers due to its accessibility. The method involves using cluster analysis to split the population into smaller homogeneous strata and ranking sites within each stratum to prioritize for recruitment in order to achieve a representative sample. Researchers who are interested in using this to sample schools may even use a website

([www.thegeneralizer.org](http://www.thegeneralizer.org)) which guides them through this process using data from the Common Core of Data.

Potential advantages of SBS include reducing under-coverage and greater recruitment transparency. However, little methodological work has examined this method’s effectiveness. Schools and districts with certain characteristics are unlikely to participate in large-scale MRTs (Fellers, 2017; Stuart et al., 2017; Tipton et al., 2016). If one or more strata are comprised of difficult schools, researchers may resort to convenience sampling within those strata. Furthermore, the additional resources required to recruit from all strata create concerns regarding practicality.

The goal of the current paper is two-part. First, we propose several methods for modeling two major sources of sampling bias: recruitment and participation. Recruitment refers to how likely a sampling site is to be approached by researchers, and participation refers to how likely the sampling site is to participate if recruited. This step will lay the groundwork for testing sampling methods in the educational context. Second, using the models proposed in the first step, we will compare SBS to several other recruitment models on the method’s ability to select a generalizable sample, and the ease with which a full sample can be recruited.

### **Stratified Balanced Sampling**

In this section we demonstrate SBS within the context of selecting schools for a large-scale MRT. This application will also serve as the basis for the present simulation study which will be further outlined in a later section. SBS is a flexible method that is adaptable to various data sets and, depending on study circumstances and design, may require different strategies or methods than presented here. For a more complete consideration of SBS, see Tipton (2013b).

SBS applies cluster analysis as a dimension reduction technique to drive bias-robust

balanced sampling based on a large set of covariates. The ultimate goal is to select a sample that is representative of a population along a set of covariates related to treatment heterogeneity and site participation. This process requires the availability of a rich data set of observed covariates for each site in the population. We therefore begin by discussing the data which will serve as the sampling frame and the covariates we have selected.

Following recommendations proposed by Tipton (2013b), we then implement k-means clustering to divide the population into heterogeneous strata comprised of homogeneous sites. K-means clustering assigns sites to strata such that similarity within each stratum is maximized. This step requires us to specify 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 determining the strata, we then use a second distance metric to rank sites in order of how representative each site is of its stratum. These ranks are later used to drive “balanced sampling” by prioritizing sites for recruitment. In practice, these rankings should encourage the selection of sites from subsets of the population which may otherwise not have been included given typical sampling strategies.

Finally, we discuss the advantages and limitations of SBS given our implementation, and the potential difficulties of using this to sample schools in the greater context of educational research.

## Sample Frame

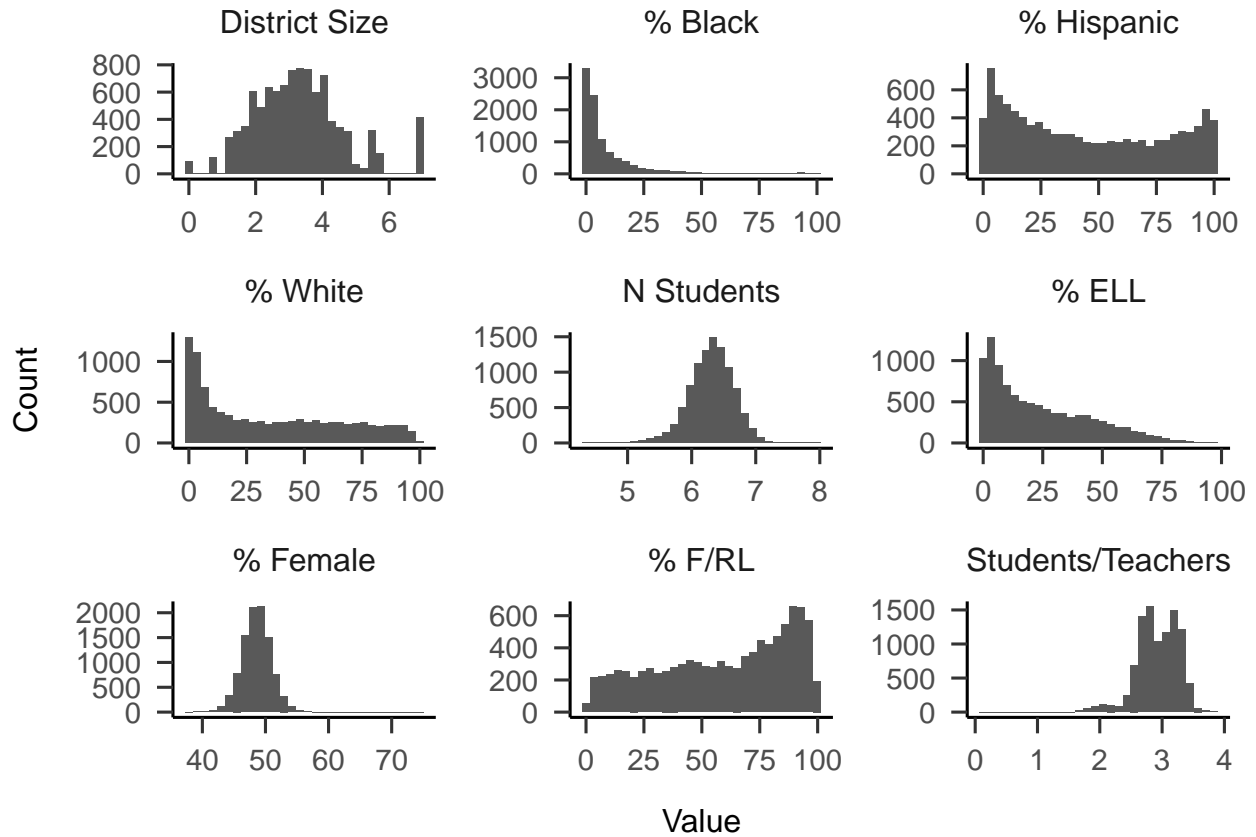
We first identify a population of schools from which we want to select a representative sample. Six States served as the geographic boundaries for this sample frame: California, Oregon, Pensilvania, South Carolina, Texas, and Wyoming. These six states were selected for arbitrary reasons, *blah blah blah*.

The cluster analysis depends on a complete enumeration of all population units and access to covariates that mediate treatment heterogeneity. Since we are only interested in selecting a representative sample, we instead identified covariates that predict school participation in RCTs. The goal is for our final sample to include schools not normally found in large-scale evaluations of interventions. To this end, selection of covariates was driven by prior research on district and school participation behavior in RCTs (Stuart et al., 2017; Tipton et al., 2016a; Fellers, 2017). 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.

It is important to note that some of these characteristics might also make it more likely that researchers would recruit these districts and schools in the first place. That is to say, school characteristics may drive selection bias by both impacting the types of schools that agree to participate, as well as by the types of schools that researchers recruit. For instance, the overrepresentation of larger schools in RCTs can mean that schools with more students are: (a) more likely to agree to participate, (b) more likely to be recruited by researchers, or (c) some combination of both (a) and (b). Therefore, the participation behavior outlined here is likely not representative of all schools in the US, rather only schools that are likely to be recruited by current practices.

Characteristics of the sample frame were sourced 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. Log-transformation was used on school size (number of students), district size (number of schools) and the student to teacher ratio. This is done to allow proportional comparisons at the extremes of the distributions (Hennig & Liao, 2013). For instance, the difference between two schools with 4000 and 3000 students was weighed as much as the difference between two schools with 400

and 300 students when generating clusters. Figure 1 displays the distribution of the continuous variables used. Two categorical covariates were used as well: urbanicity (urban, suburban, or town/rural) and a binary indicator of school wide title 1 eligibility. In all, the sample frame consists of 6 states, 2,016 districts and 9,792 schools.



*Figure 1.* Distributions of continuous covariates. Three covariates were transformed to logs: District Size, N Students, and Student/Teachers

## Stratification

We implemented stratification prior to simulation because the population is constant across iterations. Per Tipton (2013b)’s original recommendation, we use k-means clustering to partition the population into strata. This requires selecting a distance metric, choosing the number of strata, and generating the strata. All analyses were performed in R (R Core Team, 2018).



**Cluster Analysis.** We performed the cluster analysis using the *cluster* package (Maechler et al., 2017). First, the *daisy* function is used to compute an  $n$  by  $n$  pairwise distance matrix across all observations. This function requires two parameters: (1) the data matrix, and (2) the distance metric. The data matrix includes the full set of school level covariates used to compare schools for clustering. The distance metric summarises the difference between a pair of sites on a set of covariates in order to maximize the similarity of all sites within a cluster. As such, the appropriate metric varies on the type of data in the matrix.

In educational research, as is the case here, data are likely to contain both continuous and categorical variables. For mixed data such as this it is appropriate to use the general similarity measure (Gower, 1971; Tipton, 2013b). This measure relies on different calculations of distance depending on the type of covariates. Let  $X_{hi}$  and  $X_{hi'}$  be the observed value of covariate  $h = 1, \dots, P$  for sites  $i$  and  $i'$  respectively, where  $i \neq i'$ . Let  $d_{ii'h}$  be the distance between observed values of covariate  $X_h$  for site  $i$  and site  $i'$ . For categorical or dummy coded variables,  $d_{ii'h} = 1$  if  $X_{hi} \neq X_{hi'}$  and  $d_{ii'h} = 0$  otherwise. For continuous covariates, we use the following formula:

$$d_{ii'h} = 1 - \frac{|X_{hi} - X_{hi'}|}{R_h} \quad (1)$$

where  $|\cdot|$  indicates absolute value, and  $R_h$  is the range of observations for covariate  $X_h$ . This equation restricts the range of  $d_{ii'h}$  to  $[0, 1]$ . Finally, we calculate the general similarity between each site pair by taking the weighted average of the distances between all covariates. Let  $d_{ii'}^g$  be the general similarity between site  $i$  and site  $i'$ .

$$d_{ii'}^g = \frac{\sum_{h=1}^P w_{ii'h} d_{ii'h}}{\sum_{h=1}^P w_{ii'h}} \quad (2)$$

where  $w_{ii'h} = 0$  if  $X_h$  is missing for either site and  $w_{ii'h} = 1$  otherwise. Setting the distance metric to “gower” in the *daisy* function performs these calculations.

**Number of Strata.** Next we used the *kmeans* function to generate clusters, which employs an optimization algorithm to classify sites into  $k$  clusters by minimizing the total within cluster variance. For each  $k$ , it is recommended to run *kmeans* at least 10 times, and select the clustering that results in the smallest total within-cluster sum of squares. This function also requires two parameters: (1) the distance matrix from the previous step, and (2) the number of clusters to generate ( $k$ ).

Selecting an appropriate value for  $k$  is one of the most difficult problems in cluster analysis (Steinley, 2006). Tipton (2013b) states that both empirical and practical criteria should be used in selecting  $k$ . A larger set of strata would result in greater homogeneity within each stratum, however 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.

Hennig and Liao (2013) also argue 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 statistical estimation. With these considerations in mind, we examined three criteria: (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 analysis several times for each value of  $k$  and comparing 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 ???) package. Figure 2a displays the CH index for each  $k$  clusters generated. In this

case, generating 2 clusters maximises the CH-index.

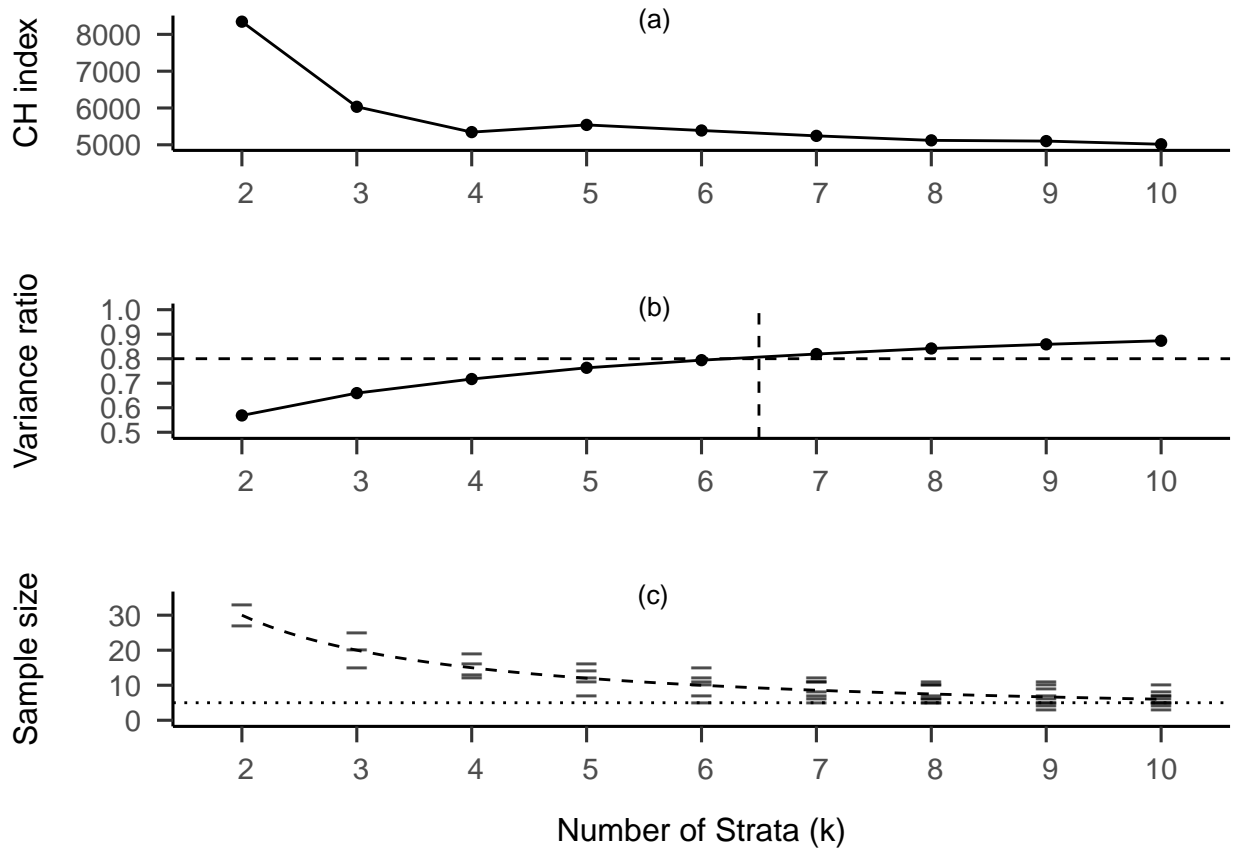
The proportion of between-cluster variance was calculated manually. Let  $k$  be the number of strata generated where  $k = 1, 2, \dots, q$  for some maximum allowable number of  $q$  strata. Let  $\sigma_{wk}^2$  be the total variability within each stratum, and  $\sigma_{bk}^2$  be the total variability between each stratum, for all covariates in  $X$  and for each set of  $k$  strata generated. Let  $p_k$  be the proportion of variability that is between strata for each set of  $k$  strata generated be defined as follows:

$$p_k = \sigma_{bk}^2 / (\sigma_{wk}^2 + \sigma_{bk}^2) \quad (3)$$

As  $p_k$  approaches 1, most of the variation is between strata, indicating homogeneity within strata. This increases the possibility of selecting a more balanced sample. Figure 2b plots  $p_k$  against  $km$  allowing visual comparison of the results. 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. Given this criteria it seems that at least 7 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 as we increase the number of clusters, the benefit of doing so decreases, while the difficulty of sampling from each cluster increases. In that case after 5 or 6 clusters the difficulty of sampling may not be worth such small increases in homogeneity within clusters.

Figure 2c plots the sample size that needs to be selected from each cluster to fulfill the proportional allocation requirement such that the number of sites sampled from each cluster is proportional to the size of the cluster in the population. 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  $k > 8$  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.



*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 line indicates a cluster sample size requirement of 5 schools.

### 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  $j$ :

$$E(X_i h | Z_i = 1, j) = E(X_h | j) \quad (4)$$

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  $N_1 + N_2 \dots + N_k = N$ . From each stratum  $j$ , we calculated the number of sites to be sampled,  $n_j$ , such that  $n_j = [(N_j/N)n]$ , where  $[\cdot]$  indicates that each value is 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:

$$d_{ij} = \sqrt{\sum_{h=1}^p w_h [X_{hij} - E(X_h | j)]^2} \quad (5)$$

where  $w_h$  is the weight assigned to covariate  $X_h$ ,  $E(X_h | j)$  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. 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.

## Advantages and Limitations

SBS has only recently been proposed and there have not been any large scale implementations of this method reported. Also, beyond the original proposal article, there have not been any methodological investigations into the method. Tipton (2013b) illustrated SBS by comparing it to a previous study which did not use a formal sampling method (Roschelle et al., 2010). Data from the previous study was used to generate two hypothetical samples. The first sample was the ideal in which the highest ranked schools in each stratum agreed to participate. The second sample was a “worst case” non-response sample where the first 50 highest ranked schools in each stratum refused to participate, resulting in a non-response rate of at least 83% in each stratum. These two samples were then compared to the original sample on the first 3 moments of 26 covariates. Both samples achieved with SBS resulted in better balance than the original sample on at least 19 of the 26 covariates in the first two moments, and 14 out of 26 in the third moments. The samples were then compared on how well they would generalize using the method proposed by Stuart et al. (2011). This final test showed that the samples achieved with SBS resulted in less coverage errors than the original sample and would thus be easier to generalize using the retrospective methods.

Stratified sampling methods offer three key advantages: transparency, non-response analysis, and integration with and improvement of post hoc adjustments. Perhaps the most appealing advantage of this method is the transparency it requires of the recruitment process. Researchers are forced to provide clear documentation and reasoning for targeting sites for recruitment. This allows for a more careful critique of the study and a better justification of the recruitment process for funders and other stakeholders. It also allows for a better analysis of non-responders. By identifying a set of observed covariates which may predict treatment heterogeneity prior to conducting the study, non-responders or refusals can be tracked and later analyzed for any systematic differences from the inference population or study participants. Finally, implementing this method does not preclude the use of the

retrospective methods previously discussed. SBS may not alleviate all balancing issues, and additional statistical adjustments may need to be made. Even if balance is only partially improved at the sampling stage, coverage errors will still be reduced and less of the inference population will need to be discarded.

There are, of course, several limitations as well. As with the retrospective methods, 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 data sets primarily consist of demographics and may not contain all of the covariates 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.

Although the findings in Tipton (2013b) were promising, it is unclear how generalizable they are to other potential studies. The inference population consisted of 1,713 non-charter schools serving seventh graders in Texas. Of these schools, 73 (4.3%) were selected into the sample across nine strata. In other applications, the inference population may be much larger and more heterogeneous, requiring more strata to be created. Sampling from too many strata is difficult when sample sizes are restricted. Would this method be beneficial to researchers making inferences on a national scale? In order to create a non-response condition, Tipton (2013b) selected the first 50 sites in each stratum to be refusals, resulting in smaller strata having higher non response rates.

Furthermore, what if schools in larger strata had higher non-response rates? How many schools would recruiters have to contact before collecting a full sample? School recruitment is a time consuming and complicated process which requires approval at several levels. Researchers may not want to invest in recruiting schools from strata with particularly

high non response rates. Finally, the data came from a single state which happened to provide information on 26 covariates. If a national population was of interest, more states would need to be included, but data reporting is not uniform across all states.

## Methods and Models

In this section we develop the framework for designing a study that simulates recruitment in educational research.

### Framework for Generalizability

Following the framework developed by Tipton (2013b), we begin with a data set enumerating a population of  $N$  sites (schools), indexed as  $i = 1 \dots N$ . Each site also has a vector of observed characteristics  $X_i$  of length  $P$ . The covariates are a mixture of continuous, binary, and categorical data. The goal is to select a sample of  $n$  sites such that there is balance along  $X_i$  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.  $SMD$  is calculated as

$$SMD = \frac{\bar{X} - \mu}{\sigma} \quad (6)$$

where  $\bar{X}$  is a vector of covariate means in the sample,  $\mu$  is the vector of covariate means in the population, and  $\sigma$  is the vector of covariate standard deviations in the population.  $SMD$  values closer to zero indicate greater balance between the sample and the population.

### Modeling Selection Bias

Selection bias stems from two sources: (1) site participation behavior (self-selection) and (2) researcher recruitment behavior. To model the former we propose a simple



participation propensity score. Let  $\pi_i^P$  represent the participation propensity score, or the probability that a site agrees to participate in an RCT if approached for recruitment. Using a simple logistic transformation we model  $\pi_i^P$  as follows:

$$\log\left(\frac{\pi_i^P}{1 - \pi_i^P}\right) = \beta_0 + \mathbf{X}_i\boldsymbol{\beta} \quad (7)$$

where  $\mathbf{X}_i$  is an  $N$  by  $P$  matrix of covariates that predict sample selection for each site, and  $\boldsymbol{\beta}$  is a vector of coefficients associated with those covariates. Participation for site  $i$  will be determined by sampling from a Bernoulli distribution with probability equal to  $\pi_i^P$ .

Unlike participation bias which is almost always present, selection bias largely depends on the sampling method being used. Probability or random sampling avoids this by selecting units with known probabilities, which can then be accounted for in the analysis stage. However this is not often feasible in the case of large-scale MRTs. Subsets of schools are likely to be overlooked if samples are too small, or when participation rates are too low. Further, random sampling from a broad population (such as an entire state) is likely to produce samples that are geographically dispersed, creating logistic difficulties for data collection and treatment implementation. As a result, researchers often eschew probability sampling in favor of convenience sampling (Gerber & Green, 2012; Shadish et al., 2002).

selection bias can be avoided using empirical sampling techniques such as probability sampling. However, these methods are often difficult and costly.

To generate selection bias stemming from researcher recruitment, we propose a “low hanging fruit” approach to convenience sampling. We assume that researchers have some knowledge of how likely each site is to participate, and prioritize easy to recruit sites as a way of maximizing recruitment efficiency. While there are many other ways to define convenience sampling, we posit this as a first step. Sites are prioritized for recruitment by sequentially sampling from the population of sites using  $\pi_i^P$  as a sampling weight. Once a

site is added to the list in order of priority, the probability of choosing the next site is proportional to the weights of the remaining items. In this way, sites that have higher values for  $\pi_i^P$  are more likely to be recruited earlier on.

## Sampling Methods

Five sampling methods were compared: stratified and unstratified random sampling, stratified and unstratified convenience sampling, and stratified balanced sampling (SBS). We model the sample selection process in two stages. In the first stage, we generated indicators of *potential participation*—that is, whether a school would participate in the trial *if approached for recruitment*. Let  $N$  be the total number of schools in the population, and let schools be indexed by  $j = 1, \dots, N$ . We define  $E_j$  as a binary indicator that school  $j$  will agree to participate if contacted by recruiters, where  $E_j = 1$  if the school agrees, and  $E_j = 0$  if the school refuses. Each school was checked for approval by sampling from a Bernoulli distribution with probability equal to  $\pi_j^P$  for each school  $j$

$$E_j \sim B(\pi_j^P) \quad (8)$$

The potential participation indicators were generated at the beginning of each iteration and were therefore common across the five sampling methods examined.

We modeled 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  $S = S_1, \dots, S_N$ . Let  $Z_j(S)$  be an indicator whether school  $i$  is sampled based on the score  $S$ . For the unstratified sampling methods, we determined  $Z_1(S), \dots, Z_N(S)$  by sorting schools according to  $S$  and selecting the first 60 schools where  $E_j = 1$ . Specifically,

$$Z_j(S) = I \left[ 60 \geq \sum_{k=1}^N E_k I(S_k \leq S_j) \right] \quad (9)$$

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 calculated the number of schools contacted as

$$R(S) = \sum_{j=1}^N I(S_j \leq S_{max}), \quad (10)$$

where  $S_{max} = \max\{S_1 Z_1, S_2 Z_2, \dots, S_N Z_N\}$ .

For the stratified sampling methods, the above process was applied within each stratum. We determined target sample sizes for each stratum based on a proportional allocation. Letting  $N_k$  denote the total number of schools in stratum  $k$ , we set a target sample size of  $n_k = \lceil 60 \times N_k / N \rceil$  for stratum  $k$ , where  $\lceil x \rceil$  is the integer nearest to  $x$ .

**Random Sampling.** As previously mentioned, unstratified random sampling (URS) is typically highly impractical in the context of educational MRTs. However, URS is nonetheless interesting as a theoretically simple ideal against which to compare other sampling methods. We simulated URS sampling by ranking each school according to a random number from a uniform distribution, so that their order of recruitment was determined by the score  $S_j \sim U(0, 1)$ .

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 simulated a stratified random sample (SRS), with strata determined by the results of the cluster analysis and using a proportional allocation.

**Convenience Sampling.** We also simulated two convenience sampling methods: unstratified (UCS) and stratified convenience sampling (SCS). Here, we assumed that recruiters have some knowledge of each schools' probability of participating, if approached, and that they prioritize higher probability schools in order to minimize effort. To operationalize unstratified convenience sampling, we assumed that schools are approached for recruitment one at a time, with their order determined by sampling without replacement, with probability proportional to participation propensity scores  $\pi_1^P, \dots, \pi_N^P$ . Once a school was selected and assigned a rank, the next school was selected with a probability proportional to the weights of the remaining schools. Once all ranks were assigned, schools

were again approached until 60 schools agreed to be in the sample. We implemented 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 were based on equation (5) (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 were ordered randomly. Because Tipton (2013b) proposed balanced sampling in connection with stratification based on a cluster analysis, we only consider the stratified version, SBS.

## Simulation Study

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. Real data was used to inform the sample frame. Selection bias was generated to account for both self selection as well as researcher bias.

### Simulation design

#### Data Generation

**Participation Propensity Score.** A response generating model (RGM) was developed to simulate self selection. The RGM creates a propensity score for each school that indicates the probability of the school agreeing to participate if targeted by any of the sampling methods. This model assumes that schools can be recruited directly by researchers. Recall equation (7) where  $\pi_i^P$  is generated for each site  $i$  using a linear transformation of the observed covariates with a logit link function. The set of covariates,  $X_i$ , was composed of

variables described 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. Absolute SMD between the schools that participated and the population were reported. By standardizing  $X_i$  and using the reported SMDs as coefficients, we generated  $\pi_i^P$  values that should on average mimic the samples that Fellers studied.

Intercept values were manipulated to generate different levels of population participation rates. Since these rates are unknown, we selected parameters for 9 levels of participation rates. Table 1 reports the variables, variable codes, and coefficients used in the RGM. Covariate coefficients remained constant across response rates.

## Analysis

**Generalizability.** There are several methods to determine 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. Let  $X_j$  be the full set of covariates identified in Table 1 indexed by  $j = 1, \dots, 9$ . Let  $\bar{X}_j$  and  $M_j$  be the mean of covariate  $j$  in the sample and population respectively. Finally, let  $\sigma_j$  be the standard deviation of  $X_j$  in the population. We will calculate the *SMD* of each covariate as

$$SMD_j = \frac{\bar{X}_j - M_j}{\sigma_j} \quad (11)$$

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. First all sites in the population are divided into  $k$  bins. For bins  $j = 1, \dots, k$ , let  $w_{pj} = N_j/N$  be the proportion of the population

Table 1

*Ods ratio coefficients for  
Response Generating Model*

Variables	log_odds
District Size	0.52
% Black	0.29
% Hispanic	0.40
% White	-0.54
N Students	0.37
% ELL	0.41
% Female	-0.02
% F/RL	0.08
Students/Teachers	-0.10
Suburban	0.01
Schoolwide Title I	0.02
Town/Rural	-0.40
Urban	0.43

and  $w_{sj} = n_j/n$  be the proportion of the sample in each bin. We will calculate  $B$  as:

$$B = \sum_{j=1}^k \sqrt{w_{pj}w_{sj}} \quad (12)$$

Bins must be defined such that  $\sum w_{pj} = \sum w_{sj} = 1$ . Selecting the correct number of bins is important, as too many bins will underestimate the similarity between distributions, and too few will overestimate. Tipton (2014) recommends generating equal bins of size  $h$  calculated as

$$h = 1.06s(N + n)^{-1/5} \quad (13)$$

where  $s^2$  is the pooled variance across the sample and population:

$$s^2 = \frac{(n-1)s_s^2 + (N+1)s_p^2}{(N+n-2)} \quad (14)$$

**Feasibility.** In order 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.

## 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 critical value of  $B = .95$  would be needed to expect full generalizability, with lower values requiring additional statistical assistance. At population participation rates below 50%, only SBS consistently selected highly generalizable samples. 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 rates 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 counterintuitive, as survey literature suggests that stratified samples are more representative. Because the B-index is an overall measure of generalizability across many covariates, it is difficult to untangle why these methods performed as they did. The next section will examine the models' performance for each covariate.

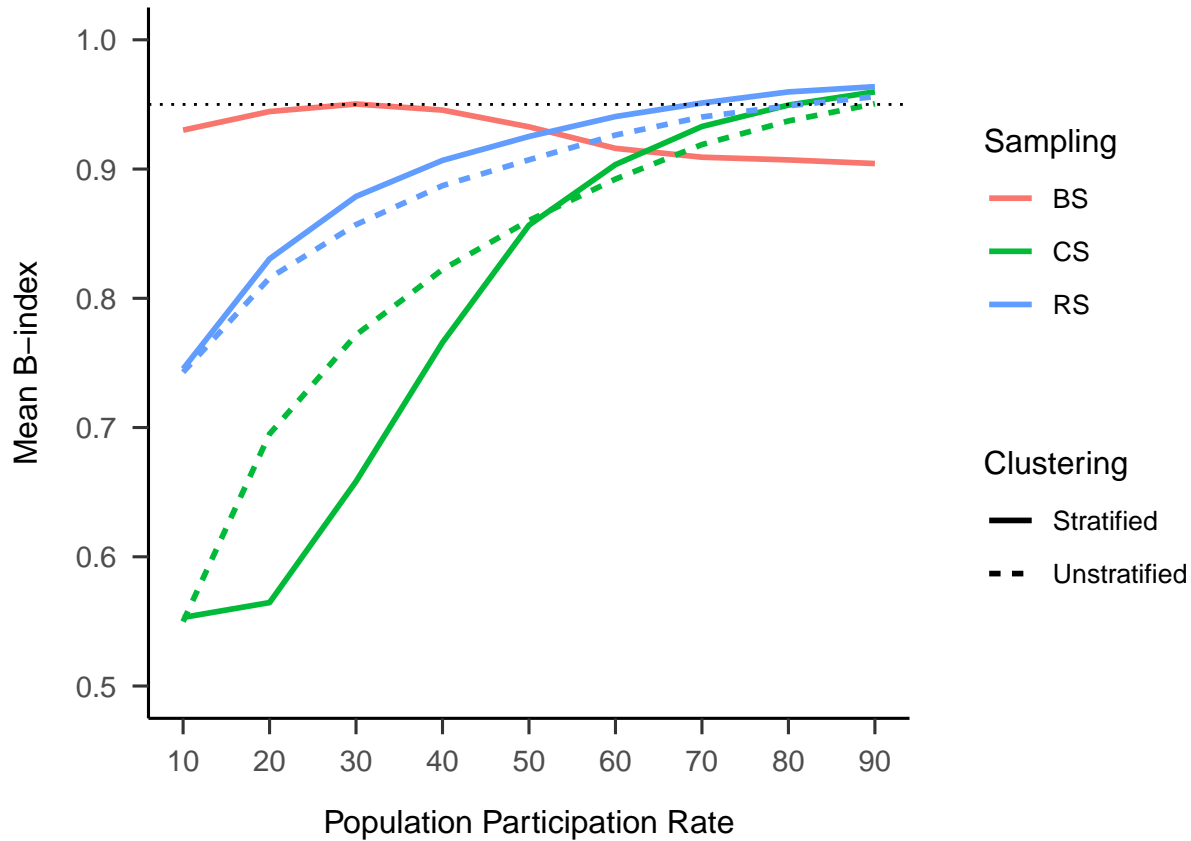


Figure 3. Average B-index for varying participation rates, by sampling method. Horizontal dotted line represented index of .95 indicating a high level of generalizability.

**Standardized Mean Differences.** One potential explanation for why stratification may underperform in certain circumstances is if the strata do not explain any variance in a covariate that is related to participation. To examine this we calculated the ICC for each covariate across strata and compared it to the log-odds coefficients associated with that covariate in the participation response model. Table 2 displays these values. Examining these along with SMDs uncovered several patterns which are summarised in figure 4.



Table 2

*Odds ratio coefficients and strata ICC*

Variables	ICC	Absolute Log Odds	Quality
Schoolwide Title I	0.94	0.02	Good1
% F/RL	0.76	0.08	Good1
Urban	0.91	0.43	Good1
% White	0.67	0.54	Good1
% Hispanic	0.66	0.40	Good1
% ELL	0.54	0.41	Good1
Suburban	0.58	0.01	Good2
Town/Rural	0.27	0.40	Good2
District Size	0.12	0.52	Good2
% Female	0.00	0.02	Neutral
Students/Teachers	0.10	0.10	Neutral
N Students	0.06	0.37	Bad
% Black	0.06	0.29	Bad

First, figure 5 displays six covariates where both ICC values and RGM parameters were high. For these covariates, SBS performed as well as or better than all stratified methods, and all stratified methods performed better than their unstratified counterparts. Figure 6 displays two covariates where both ICC values and RGM parameters were low. SBS seemed to perform slightly better than all other methods, however all methods performed well across all population participation rates.

Figure 7 reports the first set of unexpected results. For three covariates, stratified random sampling and stratified convenience sampling outperform their unstratified counterparts. However, for district size, SBS seems to perform more poorly as the

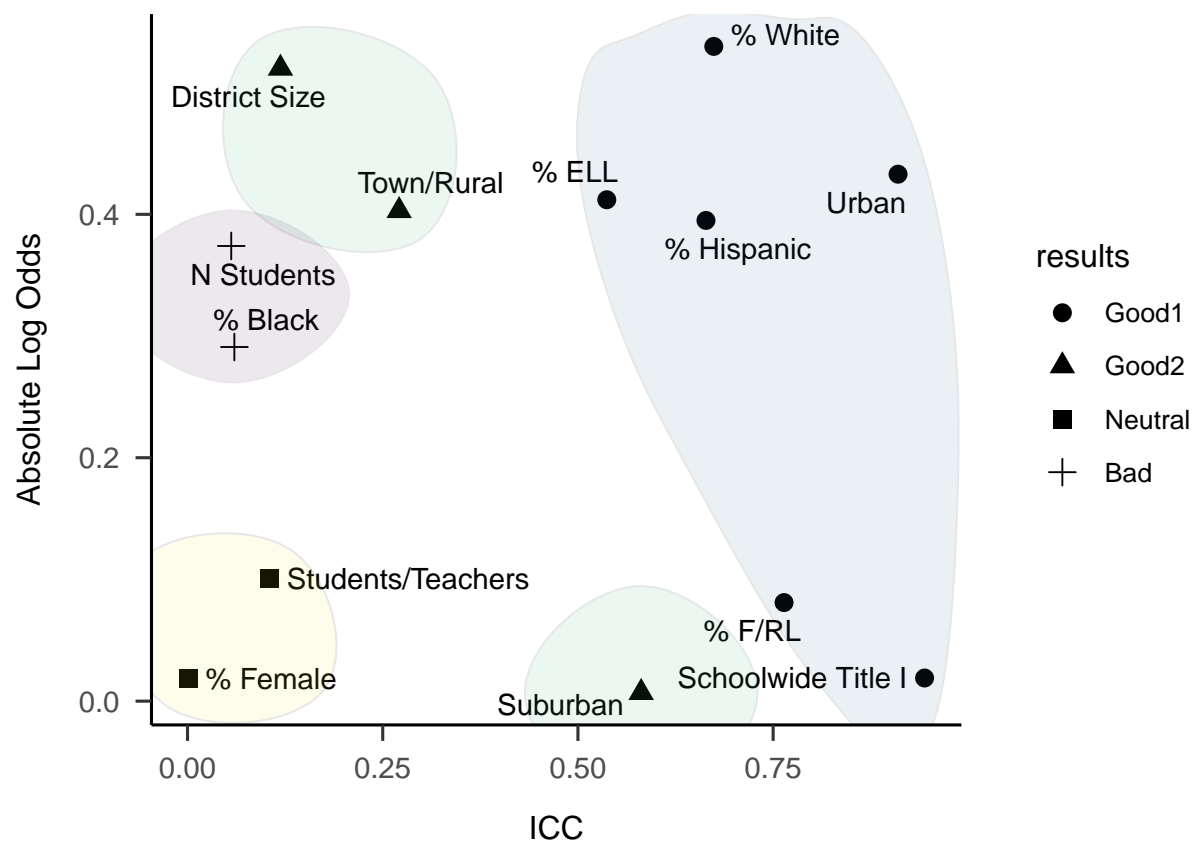


Figure 4. ICC Values vs absolute log odds. Shaded areas illustrate potential patterns in generalizability measured by SMDs.

population participation rate increases. For suburban and town/rural indicators, this is the only time where SBS performs poorly using the SMD criteria, but only in lower population participation rates. Examining table 2 suggests that this pattern emerges when ICC values are moderate and RGM parameters either high or low.

Finally, figure 8 reports the second set of unexpected results. For two covariates, SBS seems to perform well. However, stratified random sampling and stratified convenience sampling perform worse than their unstratified counterparts. For % black SBS performance also seems to slightly degrade as population participation rates increase. Examining table ?? suggests that this pattern emerges when ICC values close to zero and RGM parameters are moderate or higher

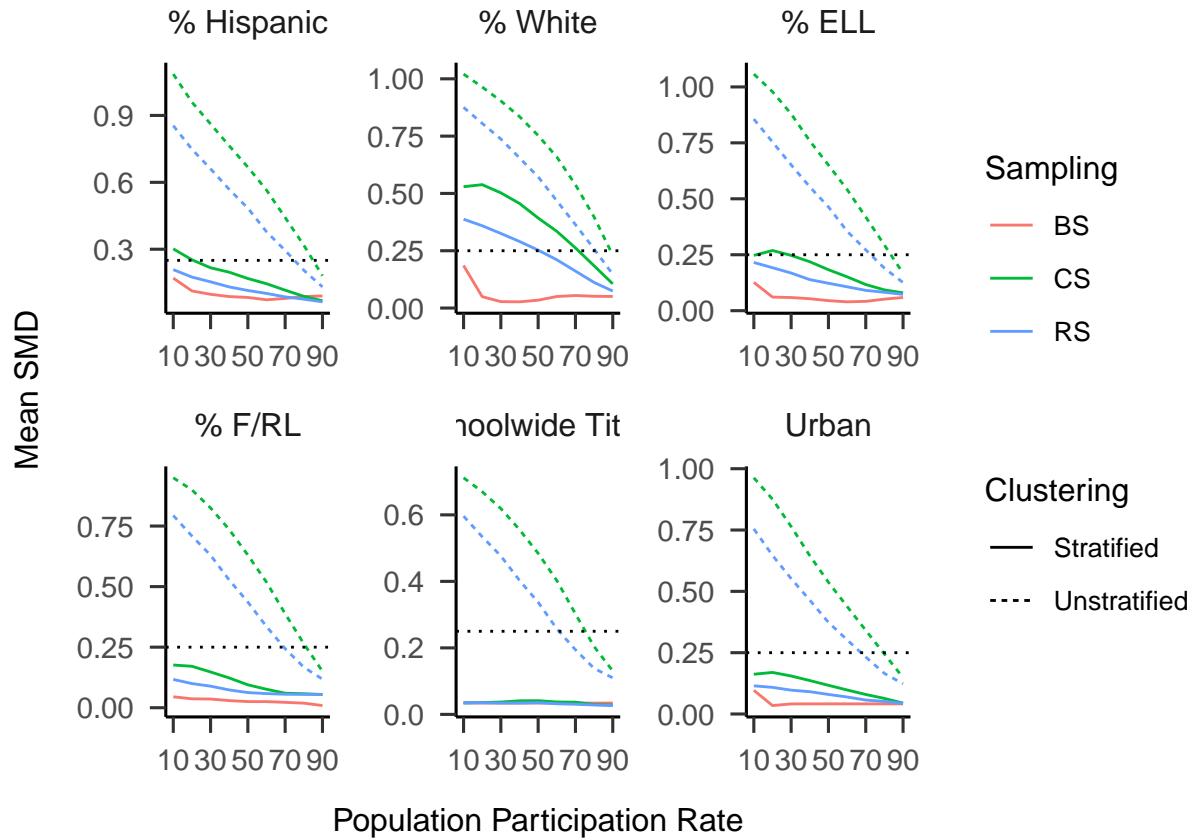


Figure 5. Good balance for stratified methods. Dotted horizontal line indicates a .25 cutoff.

## Feasibility

**Recruitment Attempts.** Figure 9 reports the sampling statistics for each method across population participation rates. Differences between sampling methods were substantial at lower participation rates. As participation rates increased, the differences decreased exponentially and became negligible. It's important to note that the magnitudes of these results are quite extreme. This is likely due to some mispecification on the part of the simulation either in the response generating model parameters or the population 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. Figure 10 plots proportional relative difficult of the models to

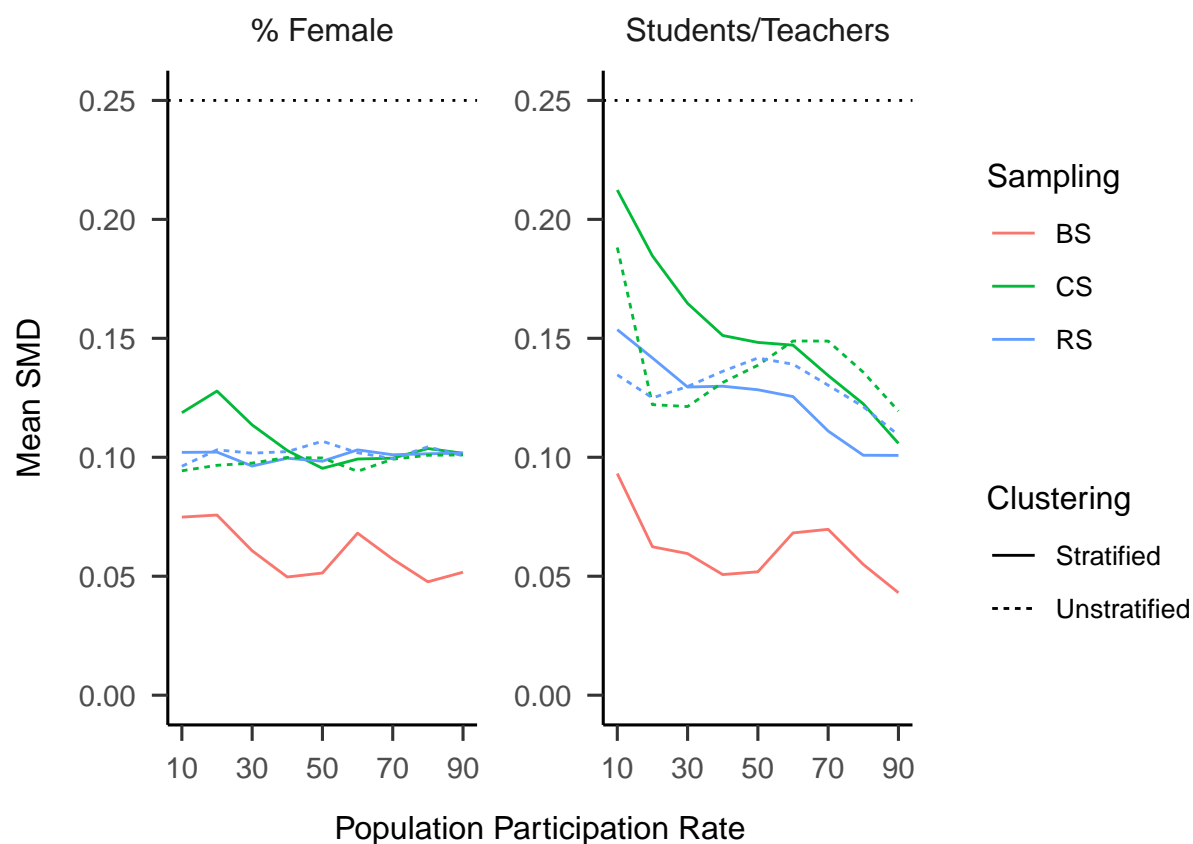


Figure 6. All methods performed well. Dotted horizontal line indicates a .25 cutoff.

each other. At a population response rate of 10%, SBS compared to UCS required approximately 22 times the number of schools to be contacted. Relative to URS, SBS required about 7 times as many recruitment attempts. The 10% response rate is likely a severe underestimate of the true population response rate and should therefore be read interpreted as an absolute worst case.

**Gini Plot.** We calculated the Gini coefficient for each sampling method to examine the equality of sampling probabilities across methods. Figure Gini coefficient across participation response rates for each sampling method. A coefficient of 1 indicates major inequality in sampling probability. displays these for each sampling method across population participation rates. A coefficient of 1 indicates substantial sampling inequality. In the context of the simulation, this indicates that across iterations, only a small subset of the

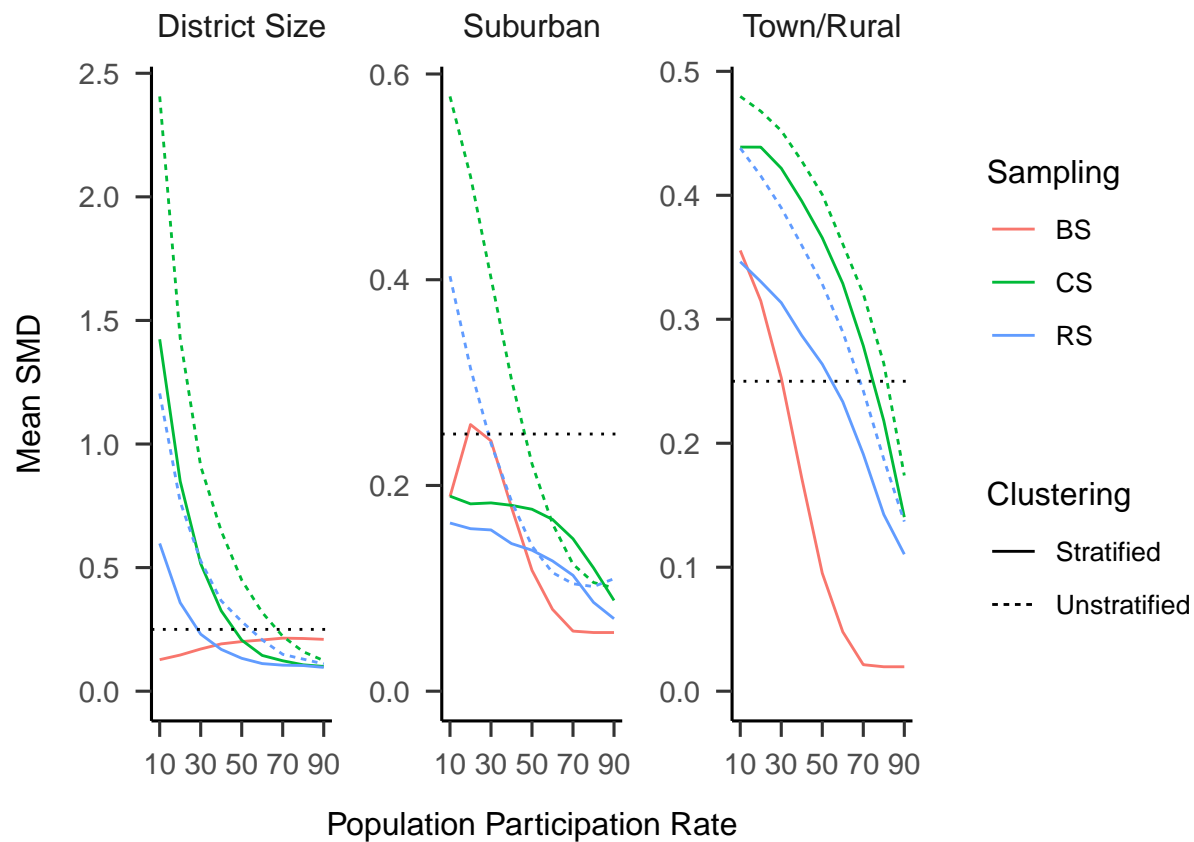


Figure 7. Good balance for stratified methods. Dotted horizontal line indicates a .25 cutoff.

population was ever actually sampled. Several trends emerged in this analysis. As population participation rates increased, inequality increased when using SBS, and 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 inequality, followed by URS, SCS and UCS. One exception is when the population response rate was set to 90%, where the unstratified and stratified versions of RS and CS seemed to perform equally. This finding suggests that stratifying the population results in a larger potential sampling pool. However, the method for selecting units from within each stratum may ultimately nullify this difference.

### 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

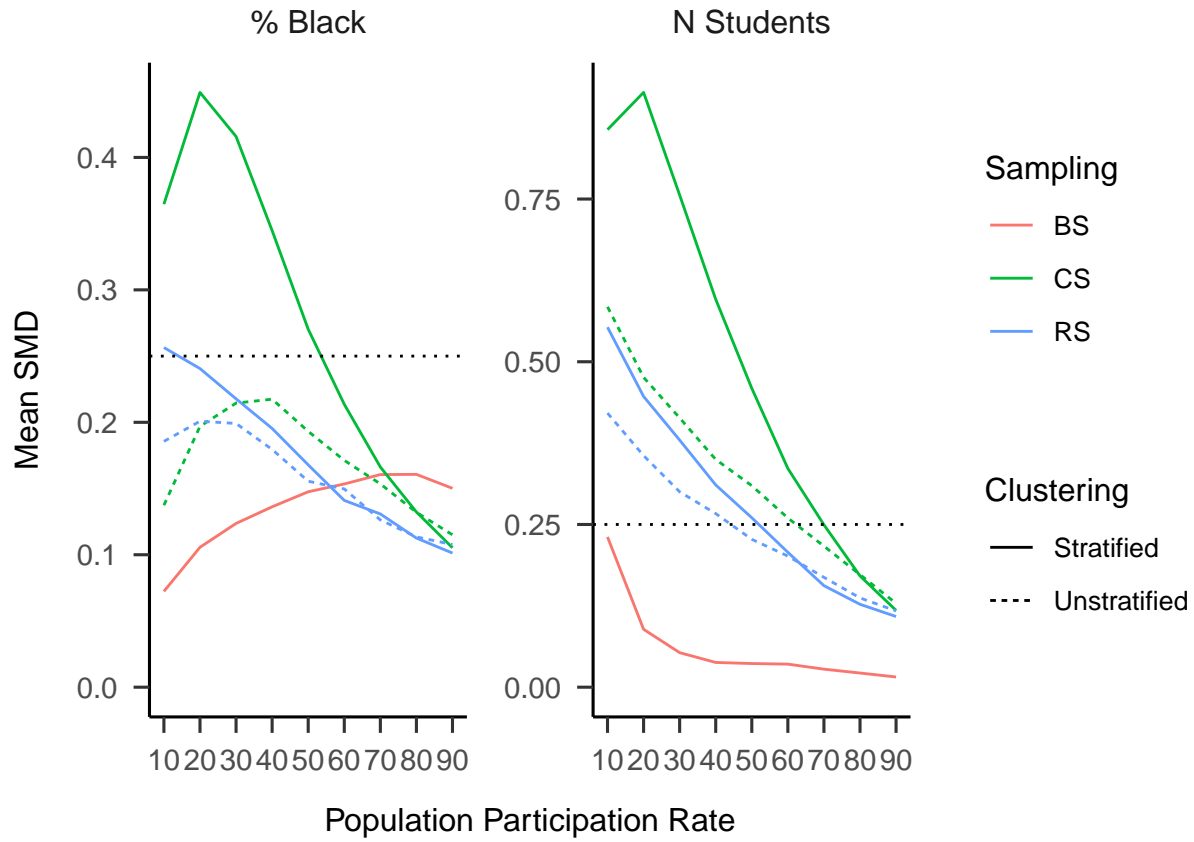


Figure 8. Poor balance for stratified methods. Dotted horizontal line indicates a .25 cutoff.

recruited, SBS is unlikely to be fully implemented in the ideal form. Instead, SCS may be a reasonable compromise. Our findings indicate that any sampling method is greatly improved by first stratifying the population. We show that, in certain cases, performing a convenience sample within strata (SCS) is comparable to simple random sampling (URS) both in terms of generalizability and feasibility.

An exception to this seems to be driven largely by the accuracy of the clustering method with respect to the response model. For instance, in Figure 8 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 during stratification. Figure 4 displays the calculated intraclass correlation coefficient (ICC) for each covariate

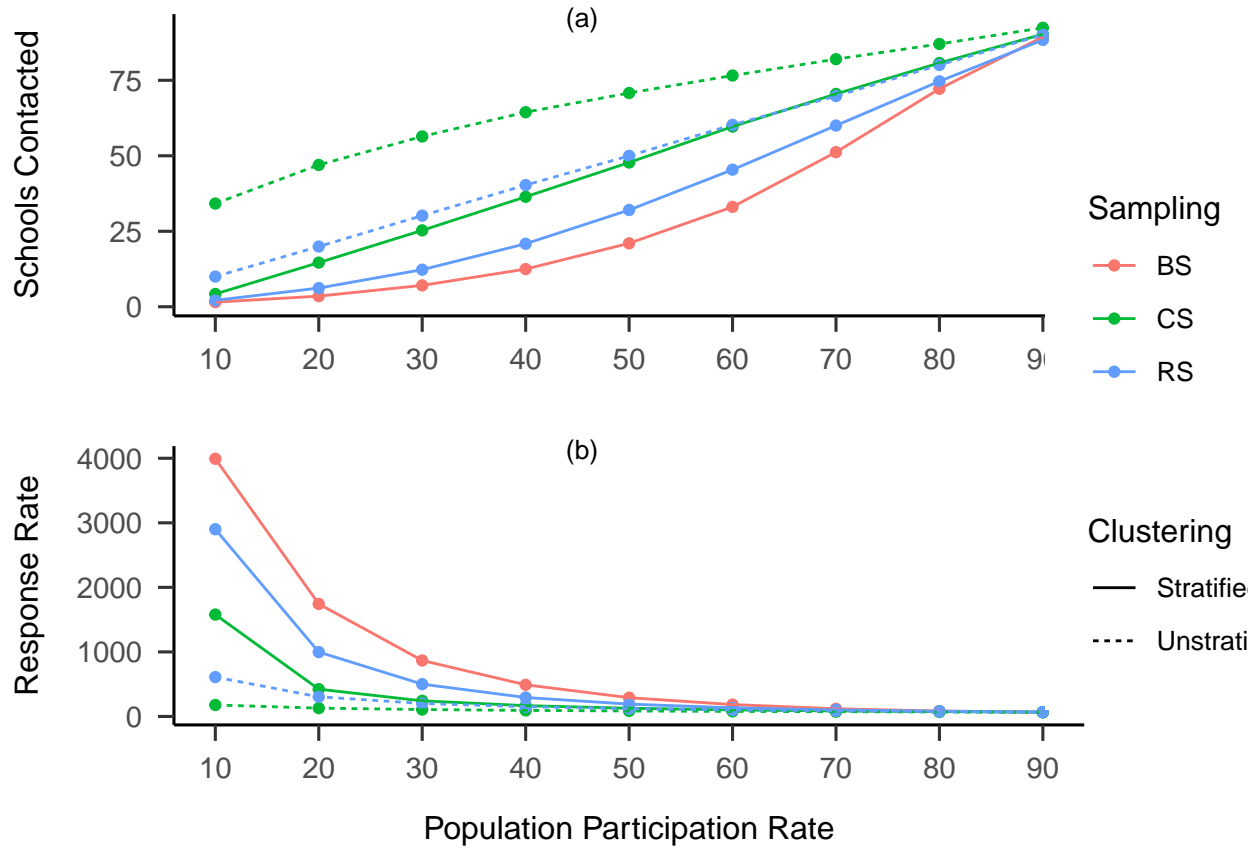


Figure 9. Sampling statistics. The top plot shows the total number of units contacted to achieve a full sample of 60 schools. The bottom plot shows the percent of schools that agreed to participate when recruited.

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 poor balance.

## Discussion

The main goal of this study was to develop a framework for exploring the performance of sampling methods within the educational context. Prior to our work there has been no precedent for developing an algorithm that mimics researchers performing a convenience sample of schools. Similarly we found no prior work attempting to model whether or not schools would participate in a study if recruited. The methods we proposed for modeling

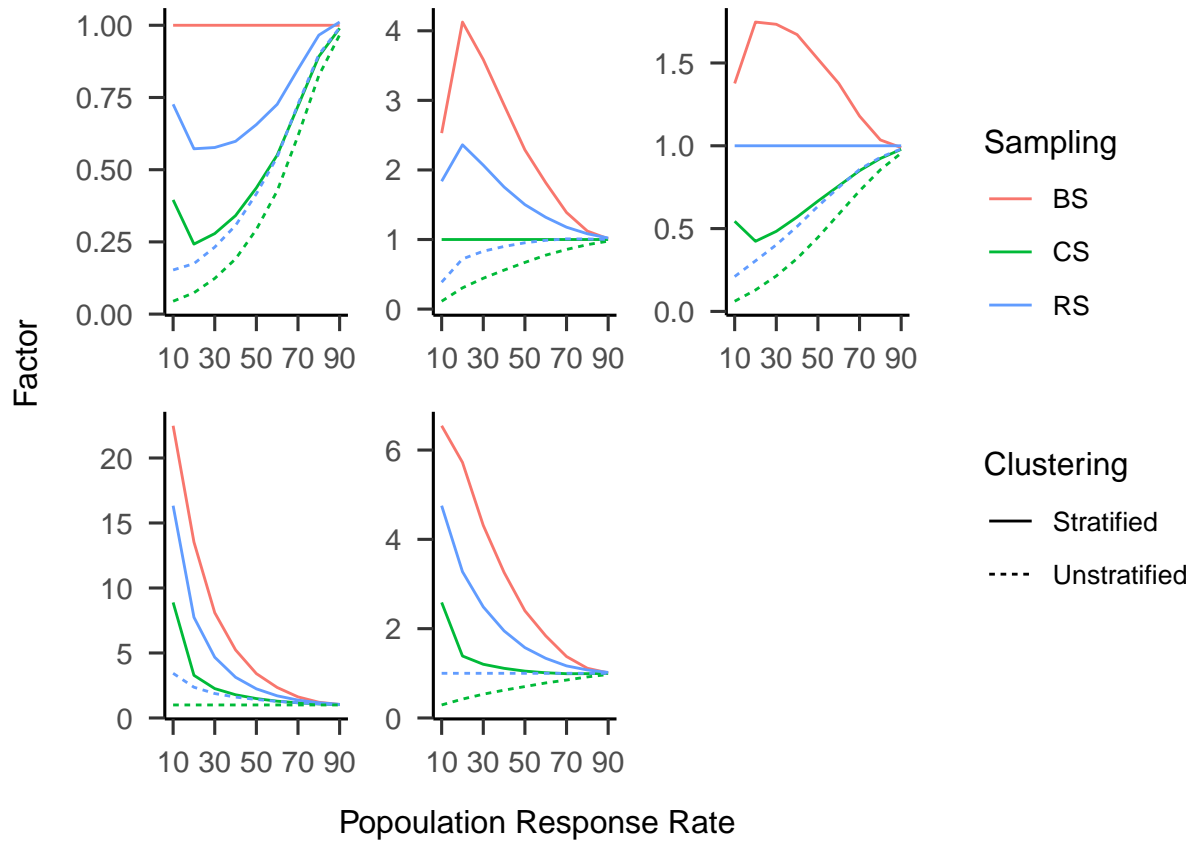


Figure 10. Proportionate comparisons of difficulties between models. The straight horizontal line is the focus of comparison in each panel.

these behaviors can easily be extended to more complex and realistic specifications and adapted to any relevant population frame.

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 of greatly increasing the generalizability of samples selected for MRTs. However, there are also several limitations. Within our simulation, strict adherence came at a great cost in terms of sheer number of schools that needed to be contacted. In practice this could mean that many more resources must be allocated to recruitment if this method is to be implemented.



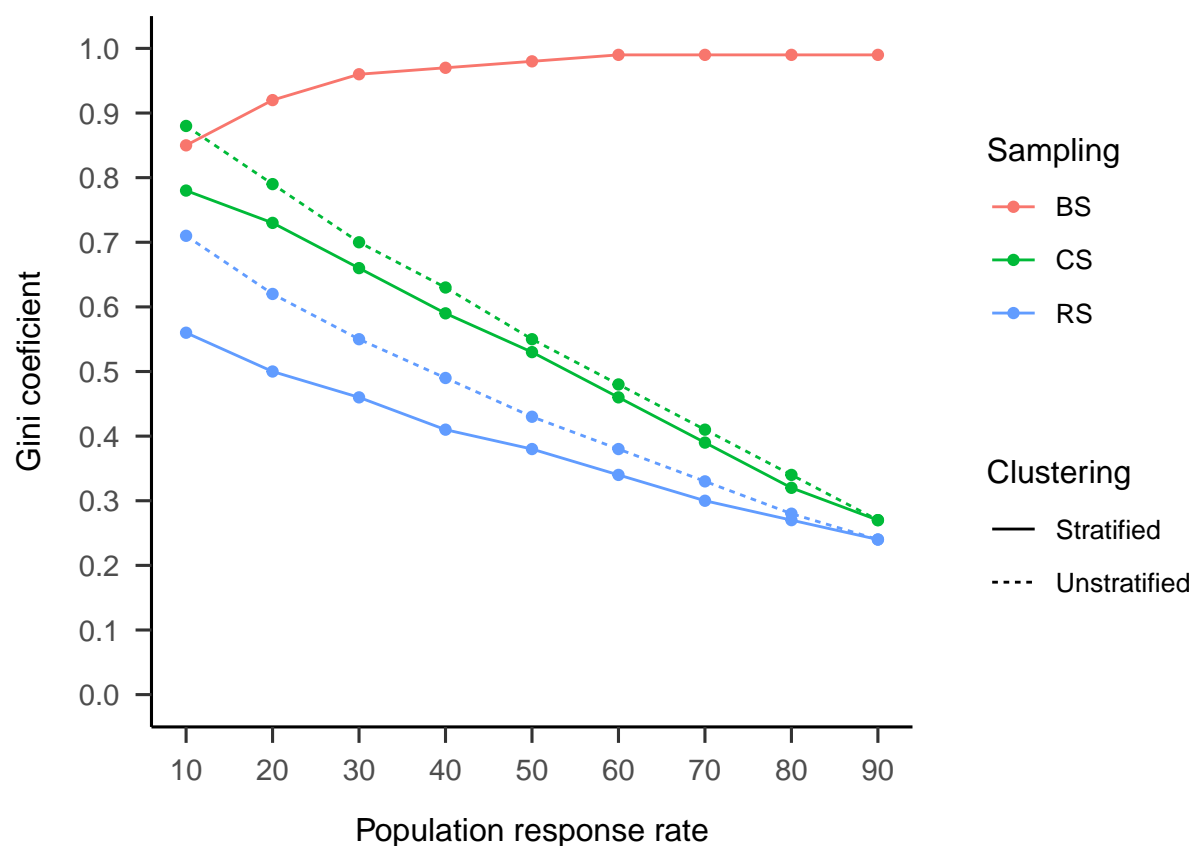


Figure 11. Gini coefficient across participation response rates for each sampling method. A coefficient of 1 indicates major inequality in sampling probability.

We found that improper specification during the cluster analysis stage may also attenuate the generalizability of the sample. Covariates that were improperly accounted when generating the strata wound up being poorly represented in the final samples. This can potentially nullify 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 set of schools were likely to be recruited across iterations.

One potential compromise is to perform a convenience sample within each stratum after performing the cluster analysis, as demonstrated by SCS. This often resulted in better balance on individual covariates than simple random sampling, which is touted as the

generalizability gold standard. Beyond generalizability, stratifying in this manner also offers the additional advantage of transparency by forcing the researcher to make sampling decisions in the study design phase, and to keep track of sampling decisions as recruitment is implemented.

The sampling methods we developed made several 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 truth, many other factors play a role 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.

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 (schools in larger urban districts) and prioritize recruiting such sites, it is not likely that they would estimate willingness as well as we have simulated. 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.

The second set of assumptions deals with the speculative nature of our participation model. The parameters in our response generating model are based on values from a study that examined the difference between schools participating in large-scale randomized control trials (RCT) and the overall population of schools. However, these RCTs themselves typically rely on some form of convenience sampling. In that sense our parameters reflect participation rates of schools that are likely to participate in RCTs, rather than the full population of schools.

Furthermore, the decision of whether a school participates in such a study is

multi-leveled. Generally, districts serve as gatekeepers, requiring research requests to be submitted and approved before recruitment can begin. If the request is denied, no schools within that district may be recruited. If approved, schools may be contacted individually. The ultimate decision may then rest with administrators, or they may be passed on to teachers most affected by the study. Despite these limitations, we believe that our findings reasonably represent the relative performance of the various sampling methods we tested in the context of educational research.

Finally, we hope to make clear the disadvantages of convenience sampling in this context. Large scale MRTs are expensive to implement, however by not investing in robust recruitment strategies researchers severely limit the impact and relevance of their work. We believe that the increased cost of a sampling method designed for generalizability is greatly outweighed by the benefit of an intervention whose impacts we can estimate more accurately and for a wider population.

## References

- Caliński, T., & Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics*, 3(1), 1–27. doi: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. doi:10.2307/2528823

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

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. doi: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. doi: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. doi: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. doi:10.1111/rssc.12037

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

Roschelle, J., Shechtman, N., Tatar, D., Hegedus, S., Hopkins, B., Empson, S., . . . Gallagher, L. P. (2010). Integration of Technology, Curriculum, and Professional Development for Advancing Middle School Mathematics: Three Large-Scale Studies. *American Educational Research Journal*, 47(4), 833–878.

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. doi: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. doi: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. doi: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. doi: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. doi: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. doi:10.1177/0193841X16655665

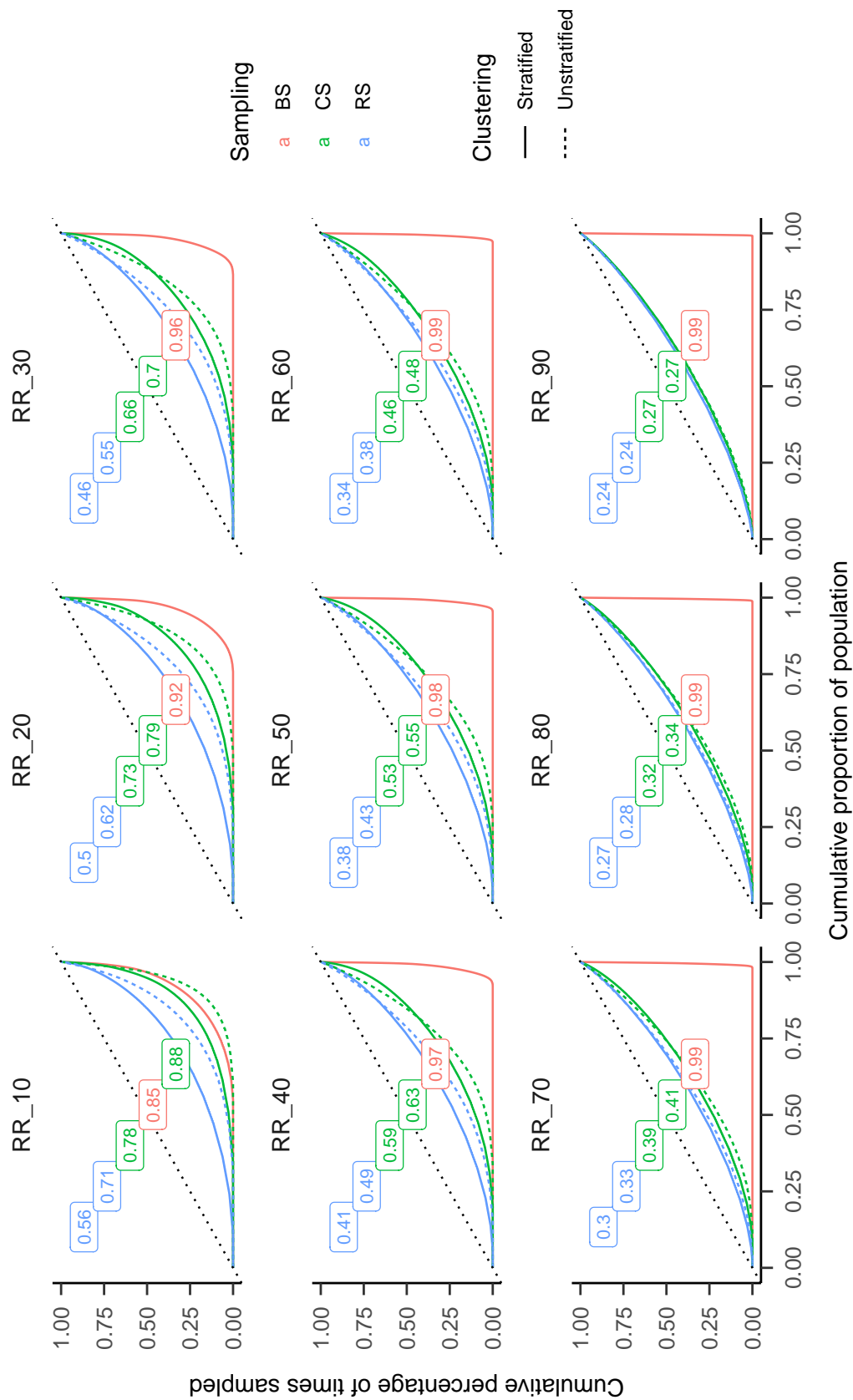


Figure 12. Gini plot and coefficient representing the inequality of school sampling across sampling methods and population response rates.