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

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

[1] University of Texas at Austin

Abstract

In order for educational research to be informative to policy makers, studies must be designed to make robust estimates of causal effects at the population level. Large scale multi-site randomized trials (MRT) often rely on vague convenience sampling methodology when recruiting districts and schools, resulting in relatively homogeneous samples that may differ greatly from the intended population of interest. Retrospective methods that quantify and statistically adjust for those differences are prosmising but have limited effect when the sample differs greatly from the population. Designing sampling methods that focus on generalizability may be a more ffective but costly solution, but limited methodological research has been performed to examine their effectiveness in the eduational context. This paper examines one promising method, stratified balanced sampling (SBS), in the context of recruiting a representative sample of schools for a large scale MRT. Using simulations based on real data, we compare SBS to stratified and unstratified versions of convenince sampling and probability sampling. Several models for generating school participation and emulating convenience sampling are proposed. Results indicate that SBS and stratified random sampling (SRS) result in highly generalizable samples. These methods are extremely costly to implement, however, especially when the population average willingness to participate is low. Stratified convenience sampling (SCS) is a potential compromise.

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

Introduction

In education research, multisite randomized trials (MRTs) are commonly used to inform policy decisions by examining intervention impacts at a large scale. Randomized treatment assignment, a gold standard for demonstrating causality, ensures such studies maintain a high level of internal validity. However, policy makers are often interested in generalizing treatment effects to populations that may not always be represented by the units being sampled. Treatment effect estimates in a sample may not generalize to a population if treatment effects vary across population units and the sampled units are not representative of the population units. Since treatment effect heterogeneity is common in the education research literature, it is necessary to design sampling methods that support generalizability if the goal of such research is to inform policy decisions. In practice, however, researchers often rely on samples of convenience potentially biasing treatment effect estimates in the population. As a result, MRTs have come under scrutiny for their limited external validity (Stuart, Cole, Bradshaw, & Leaf, 2011).

Probability sampling is commonly used in the survey literature to make unbiased estimates of population characteristics. When probability sampling is used along with randomized treatment assignment, also known as dual randomization, average treatment effect (ATE) estimates are unbiased in both the sample (SATE) and the population (PATE). However, probability sampling is rarely used in educational MRTs (Olsen, Orr, Bell, & Stuart, 2013; Shadish, Cook, & Campbell, 2002). Instead, recruitment of schools is often driven by convenience and cost effectiveness rather than representation, potentially resulting in biased estimates of PATE from unrepresentative samples. Though several methods have been proposed to statistically adjust for biased estimates, they invoke strong assumptions

and are subject to coverage errors (Tipton, 2014).

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 uses cluster analysis to split the population into smaller homogeneous strata and provides guidelines on how to sample from each strata 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) which guides them through this process using data from the Common Core of Data.

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

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 unit is to be approached by researchers, and participation refers to how likely the sampling unit 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 it's ability to select a generalizable sample, and the ease with which a full sample can be recruited.

Stratified Balanced Sampling

Stratified Balanced Sampling (SBS; Tipton, 2013b) applies cluster analysis in order to implement bias robust balanced sampling. The goal of SBS is to select a sample that is representative of a population along a set of covariates related to treatment heterogeneity. This process requires the availability of a rich data set of observed covariate for each unit in the population. The population is first divided into heterogeneous strata comprised of homogeneous units given the set of observed covariates. This is done using k-means clustering which assigns units to strata such that similarity within strata is maximized. Units most representative of each strata are prioritized by recruiters. This encourages the selection of units from subsections of the population which may otherwise not have been included in the sample.

Distance Metric. The first step is to select a distance metric, which summarizes the distance between two units on a set of covariates. This distance metric is used to maximize the similarity of units within clusters. Determining which metric to use largely depends on the type of covariates included in X and their relative importance. If all covariates are continuous, the weighted Euclidean distance can be used. Let $d_{ii'}^e$ be the Euclidean distance between unit i and unit i' where $i \neq i'$. Let X_{ih} and $X_{i'h}$ be the observed value of covariate $h = 1, ..., p$ for units i and i' respectively. Finally, let w_h be the weight associated with covariates X_{ih} and $X_{i'h}$. We calculate $d_{ii'}^e$ as:

$$d_{ii'}^e = \sqrt{\sum_{h=1}^{p} w_h (X_{ih} - X_{i'h})^2} \tag{1}$$

Setting the weights to $w_h = 1$ gives the most weight to covariate with the largest variances. Setting the weights to $w_h = 1/V(X_h)$ allows each covariate to contribute equally to calculating the distance measure. This latter approach is useful when the importance of

predictors of treatment effect heterogeneity is unknown.

If X contains both continuous and categorical variables, the general similarity measure may be used (Gower, 1971). This method relies on different calculations of distance depending on the type of covariates. Let $d_{ii'h}$ be the distance between observed values of covariate X_h for unit i and unit i' where $i \neq i'$. For categorical or dummy coded variables, $d_{ii'h} = 1$ if $X_{ih} = X_{i'h}$ and $d_{ii'h} = 0$ otherwise. For continuous covariates, we use the following formula:

$$d_{ii'h} = 1 - \frac{|X_{ih} - X_{i'h}|}{R_h} \tag{2}$$

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

$$d_{ii'}^g = \frac{\sum_{h=1}^{p} w_{ii'h} d_{ii'h}}{\sum_{h=1}^{p} w_{ii'h}} \tag{3}$$

where $w_{ii'h} = 0$ if X_h is missing for either unit and $w_{ii'h} = 1$ otherwise.

Selecting k Strata.   Before applying the clustering method, the number of strata to be generated must be determined. Generating more strata results in more homogeneity within strata, however in practical applications this may be more difficult to manage. For instance, if refusal and non-response rates are fairly high, having fewer 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.

One solution is to perform the analysis several times generating different numbers of strata, and comparing the proportion of variability between strata. Let k be the number of strata generated where $k = 1, 2, ..., q$ for some maximum allowable number of q strata. Let σ^2_{wk} be the total variability within each strata, and σ^2_{bk} be the total variability between each strata, for all covaraites 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:

$$p_k = \sigma^2_{bk}/(\sigma^2_{wk} + \sigma^2_{bk}) \tag{4}$$

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. Plotting p_k against k allows visual comparison of the results. The k for which the rate of change p_k slows is then selected.

Sample Selection. Finally, the sample must be selected and evaluated. Tipton (2013b) recommends selecting a balanced sample of order 1 within each stratum, then testing for robustness to model failure by comparing the selected sample on higher orders. First the number of units to be sampled from each strata must be identified using proportional sample allocation. Each strata contains N_j units where $N_1 + N_2 ... + N_k = N$. From each stratum j, n_j units must be sampled such that $n_j = [(N_j/N)n]$, where $[.]$ indicates that each value must be rounded to the nearest integer.

Next a method must be chosen for selecting units within each strata such that a balanced sample is achieved. Ideally this means that the expected value of covariate X_h across units in stratum j is equal to the expected value of covariate X_h across all units sampled from stratum j:

$$E(X_h|Z = 1, j) = E(X_h|j) \tag{5}$$

One method is to perform a simple random sample of n_j units. This is ideal as it results in a balanced sample on both observed and unobserved covariates over repeated samples in stratum j. This method is limited when n_j is small which may not result in balance for any particular sample. Another approach is to rank each unit within a strata using a distance measure, with units closer to the "center" of the strata ranked higher. Once again, the weighted Euclidean distance can be used, this time measuring the distance to the mean of each covariate:

$$d_{ij} = \sqrt{\sum_{h=1}^{p} w_h[X_{ijh} - E(X_h|j)]^2} \tag{6}$$

where w_h is the weight assigned to covariate X_h, $E(X_h|j)$ is the mean of the hth covariate in stratum j, and X_{ijh} is the value of the hth covariate for unit 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. The ranked list can then be used to prioritize units for recruitment, beginning with the highest ranked units available to the recruiter. If a unit is unavailable or refuses to participate the recruiter moves on to the next highest ranked unit until n_j units agree to participate.

Once the sample has been recruited, it is necessary to evaluate the extent to which the sample is representative of the population by assessing the degree of balance of Order R achieved on the covariates. Let X_h^r be the observed value of the h'th covariate to the r'th power, where $h = 1, ..., p$ and $r = 1, ..., R$. The sample is a balanced sampled of the order R when the expected value of X_h^r in stratum j across all units in the sample is equal to the expected value of X_h^r across all units in stratum j:

$$E(X_h^r|j) - E(X_h^r|Z = 1, j) \tag{7}$$

As this equation approaches 0 for all covariates, the sample become more balanced and robust against model failure. Comparisons can be made using standardized mean differences or t-tests, though substantive criteria should also be considered.

If there is a high degree of homogeneity within strata, this sampling procedure should be fairly robust to moderate non-participation rate. However a very large non-participation rate may result in a lack of balance, especially in higher orders (Tipton, 2013b). If large differences are detected, post hoc methods described previously can be used to make further adjustments.

Outcome Analysis.   If proportional allocation is performed during the stratified sampling process, then the proportion of the sample in each strata is equal to the proportion of the population that is in each strata, $\frac{n_j}{n} = \frac{N_j}{N}$. Furthermore, if a bias-robust balanced sample is selected within each strata, then the expected value of X in the sample will equal to the expected value of X in the population within each stratum, $E(X|Z = 1, j) = E(X|j)$. When both of these conditions are satisfied, no additional adjustments need to be made to the outcome analysis. This is because the resulting sample will be self-weighing:

$$\begin{aligned} E(X|Z = 1) &= \sum (n_j/n)E(X|Z = 1, j) \\ &= \sum (N_j/N)E(X|j) \\ &= E(X) \end{aligned} \tag{8}$$

Given this, whatever random effects model which would ordinarily be applied may be use to estimate treatment effect. Furthermore, because the standard estimator is used, any

power analysis used to estimate the original sample size is unaffected. If proportional allocation is not achieved, the the sample needs to be reweighed to resemble the population, in which case the previously described retrospective methods can be applied to estimate treatment effects and make generalizations.

Advantages and Limitations. SBS has only recently 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 investigation 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 the 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 grants to the recruitment process. Clear documentation and reasoning for targeting units for recruitment is provided. 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, often a large source of bias. 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 other retrospective methods, SBS depends on the existence of a rich set of observed covariates related to treatment heterogeneity and sample selection for each unit 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). Such data is typically in aggregate form, as is the case with school/district censuses, which does not allow stratification at the individual level. Additionally, SBS requires more resources to implement than a simple convenience sample. Implementation of stratified sampling, both with propensity scores and cluster analysis, can be challenging. Depending on the context, developing a sample frame isn't always straightforward and must be thought out carefully (Tipton et al., 2016). Furthermore, recruiting ranked units 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 methodologists implementing this method.

Finally, although the findings in Tipton (2013b) were positive, 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 practice, 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 units in each strata 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. What if other states report fewer characteristics? How robust is this method to omitted variables?

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 population frame. Selection bias was generated to account for both self selection as well as researcher bias.

Data

In order to model realistic sampling scenarios, extant school data were used to develop the population frame. School characteristics were then used to generate a "participation propensity score" which indicated the probability of a school agreeing to participate if asked.

Because the overall participation rate for the population is unknown, this score was generated under severall assumed participation rates.

Population Frame.   The population frame is composed of data from three sources: (1) the Common Core of Data (CCD), (2) publicly available accountability data, and (3) the U.S. Census. The CCD is a comprehensive database housing annually collected national statistics of all public schools and districts. Accountability data was used to calculate the proportion of students within each school performing at or above proficiency in Math and ELA. Finally, local median income was obtained from the U.S. Census and was matched to each school by zip code.

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). 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, however, that some of these characteristics might also make it more likely that researchers would recruit these districts and schools in the first place. Anecdotal evidence from several research teams also suggests that schools are less willing to participate in experimental interventions for subjects in which their students are already excelling, therefore math and ELA achievement covariates were also included.

Log-transformation was used on school size (number of students) and median income. 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 should be weighed as much as the difference between two schools with 400 and 300 students when generating clusters. Figure 1 displays a comparison of the distribution of these variables and their logs. Figure 2 displays the distribution of the remaining continuous variables.
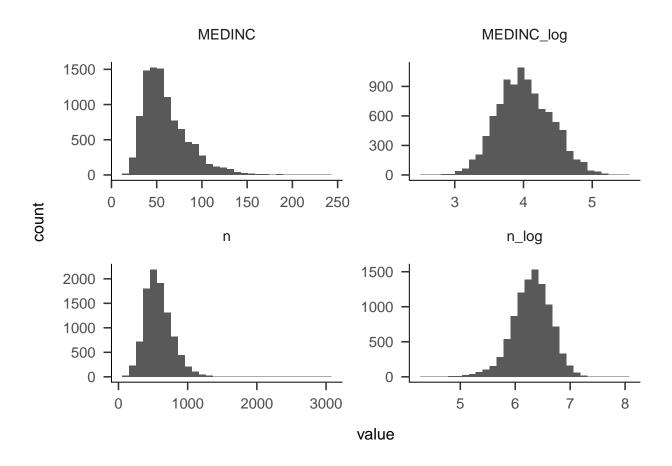
Figure 1. Comparison of covariate distributions and their log transformations.

Participation Propenity 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 method. This model assumes that schools can be recruited directly by researchers.

Let and π^S be the participation propensity score. The following base model will be used for the RGM:

$$\log\left(\frac{\pi^S}{1-\pi^S}\right) = \gamma_0 + \gamma_1 X_{Suburban} + \gamma_2 X_{Town/Rural} + \gamma_3 X_{pELL} + \gamma_4 X_{pED} \tag{9}$$

$$+ \gamma_5 X_{pMin} + \gamma_6 X_{MedInc} + \gamma_7 X_{ELA} + \gamma_8 X_{Math}$$
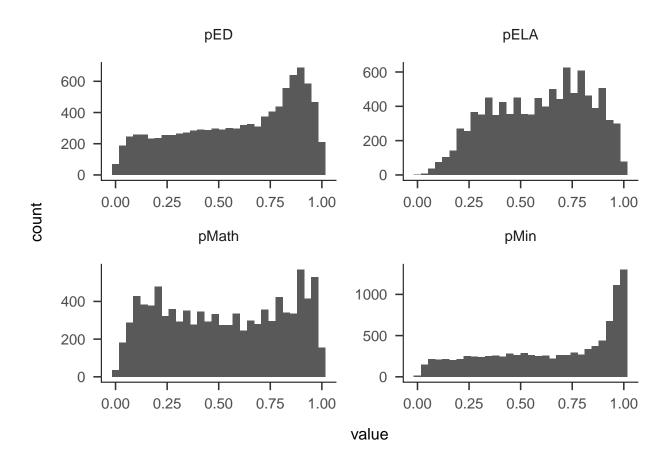
Figure 2. Distributions of the remaining continuous covariates.

Where $X_{Suburban}$ indicates if the school is suburban, $X_{Town/Rural}$ indicates if the school is in a town or is rural, X_{pELL} is the percentage of ELL students in the school, X_{pED} is the percentage of ED students in the school, X_{pMin} is the percentage of minority students in the school, and X_{MedInc} is the average median household income in the school community, X_{ELA} is the percentage of students in a school scoring at or above proficiency on English language arts exams, and X_{Math} is the percentage of students in a school scoring at or above proficiency on math exams.

Parameters for the RGM were selected by iterating across values until the model produced samples with desired covariate characteristics.

Table 1

Var	RR = 10%	RR = 20%	RR = 30%	RR = 40%	RR = 50%	RR = 60%	RR = 70%
Intercept	-2.99	-1.83	-1.09	-0.27	0.77	1.52	4.61
n	-0.17	-0.16	-0.20	-0.20	-0.21	-0.24	-0.21
pED	0.74	0.91	0.89	0.75	1.09	1.05	1.28
pELA	0.17	0.30	0.41	-0.70	0.25	-0.86	0.11
pELL	-0.17	-0.15	-0.20	-0.19	-0.20	-0.21	0.10
pMath	0.15	0.06	-0.06	0.96	0.16	1.19	0.19
pMin	-0.95	-1.24	-1.16	-1.35	-1.75	-2.39	-5.57
Suburban	-0.18	-0.07	-0.28	-0.64	-0.50	-0.19	-0.51
ToRu	0.55	0.34	0.73	0.81	1.64	1.49	2.00
Urban	-0.33	-0.22	-0.44	-0.84	-0.67	-0.43	-0.62

Stratified Balanced Sampling

Stratification was performed prior to simulation because the population is constant across iterations. Per Tipton's (2013) 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.

Distance Metric. The set of covariates include continuous variables as well as binary indicators for urbanicity (urban, suburban, and town/rural). Within this context it is generally recommended to use Gower's (1971) general dissimilarity distance (Everitt, 2011; Tipton, 2013). This method relies on different calculations of distance depending on the type of covariates. Let $d_{ii'h}$ be the distance between observed values of covariate X_h for unit i and unit i' where $i \neq i'$. For categorical or dummy coded variables, $d_{ii'h} = 1$ if $X_{ih} = X_{i'h}$ and

$d_{ii'h} = 0$ otherwise. For continuous covariates, we use the following formula:

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

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

$$d_{ii'}^g = \frac{\sum_{h=1}^{p} w_{ii'h} d_{ii'h}}{\sum_{h=1}^{p} w_{ii'h}} \tag{11}$$

where $w_{ii'h} = 0$ if X_h is missing for either unit and $w_{ii'h} = 1$ otherwise. This produces an n by n distance (or dissimilarity) matrix.

Number of Clusters.   Selecting the number of clusters, k, is one of the most difficult problems in cluster analysis (Steinley, 2006). To date, the most extensive investigation of methods for determining k was conducted by Milligan and Cooper (1985) who analyzed 30 methods. However, aside from the limited generalizability of this study, many methods are also inappropriate in the context of non-hierarchical clustering and thus do not support k-means clustering. Tipton (2013) states that both statistical and practical criteria should be used in selecting the number of clusters. For instance, a large number of clusters would result in more homogeneous strata and, in turn, a more robust sample. However as strata become smaller, they also become more difficult to adequately sample from. Hennig and Liao (2013) argue that the method of selecting k should depend on the context of the clustering and frame the issue as one of obtaining an appropriate subject-matter-dependent definition of rather than a statistical estimation. Ultimately three considerations were used to select the number of clusters: the ratio of variability between clusters to the sum of within and between cluster variability as recommended by Tipton

(2013), a generalized form of the Calinski-Harabasz index (Calinski and Harabasz, 1974) proposed by Hennig and Liao (2013), and the practicality of sampling from fewer clusters.

- Everitt (2011), p126
  - clusterSim
  - Continuous data?
    * Calinski and Harabasz (1974)
    * Duda and Hart (1973)
  - Steinley, D. (2006) K-means clustering: a half-century synthesis. British Journal of Mathematical & Statistical Psychology, 59, 1–34.
- Milligan and Cooper (1984)
  - list 30
- Sugar and James (2003) via Hennig & Liao 2013 p 314
  - Modern look

Cluster Analysis. Cluster analysis was performed using the *cluster* package (Maechler et. al. 2017) in R. 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 included the full set of school level covariates. The metric was set to "gower". Next the *kmeans* function was used to generate clusters. This method uses an optimization algorithm to classify units into k clusters by minimizing the total within cluster sum of squares. This function also requires two parameters: (1) the distance matrix, and (2) the number of clusters to generate (k). 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. [Get Citation]

Figure 3 displays the Calinski-Harabasz (CH) index for each k clusters generated for both the SBS-F and SBS-OV. The value of k that maximizes the CH index should be selected. However we see several local maxima: $k = [2, 5, 8, 11]$ for SBS-F, and $k = 2, 6, 10$

for SBS-OV.

Taking the ratio of between-cluster SS to within-cluster SS and plotting it against number of clusters creates a chart similar to an upside down elbow graph. Figure 4 displays this for both SBS-F and SBS-OV. Tipton (2013) 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 10 clusters should be generated for SBS-F, and 8 for SBS-OV. 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 6 or 7 clusters the difficulty of sampling may not be worth such small increases in homogeneity within clusters.

Figure @ref(fig:k-size-full plots the sample size that needs to be selected from each cluster to fulfill the proportional allocation requirement such that the number of units 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 for SBS-F when 8 or less clusters are generated, they are more equally sized, with the exception of 3 and 6 clusters where one is much larger than the others. For SBS-OV this is less apparent. Instead a sensible cutoff may be determined by looking at the size of the smallest cluster. At $k > 7$ it seems that the smallest clusters would require less than 5 units being sampled, which may be very difficult in a practical setting.

In order to maintain comparability between methods, it was determined that 6 clusters would be generated for both models, though in practice 7 clusters for the full model may be more prudent.
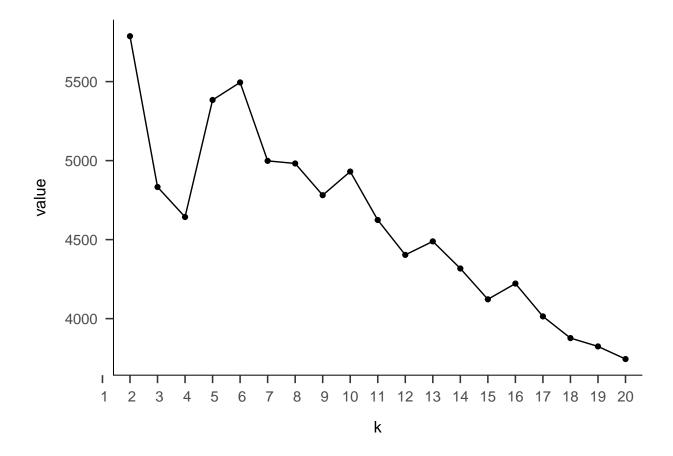
Figure 3. Generalizd Calinski-Harabasz index

References

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

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

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*
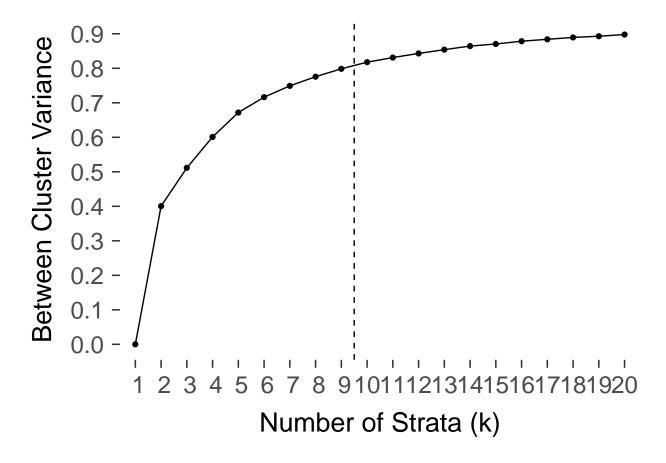
Figure 4. Ratio of between cluster sum of squares to total cluster sum of squares

Statistics), *62*(3), 309–369. doi:10.1111/j.1467-9876.2012.01066.x

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

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:
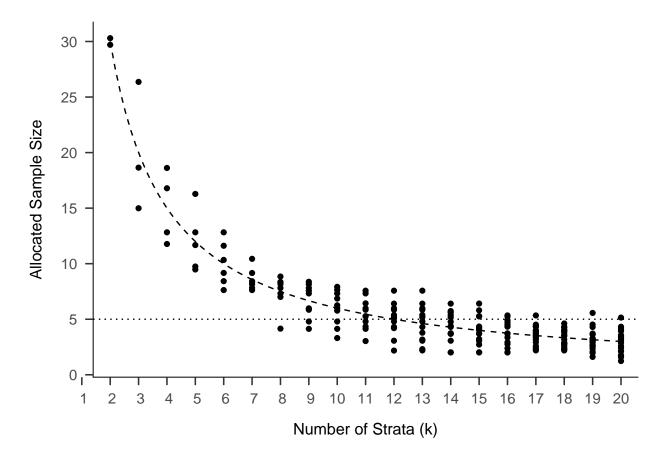
Figure 5. Sampling requirements for each cluster

Houghton, Mifflin and Company.

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