Selecting a Generalizable Sample in Field Trials Using Cluster Analysis Stratified Sampling

Gleb Furman

QUALIFYING DOCUMENT
Presented to the Faculty of the Department of Educational Psychology in Partial Fulfillment of the Requirements for Advancement to Doctoral Candidacy

University of Texas - Austin
Fall 2017

Contents

Introduction

Field trials are experiments conducted to test the effect of a treatment in natural settings (Gerber and Green, 2012). This design is common in education research where treatments may take the form of new a curriculum or intensive behavioral interventions. Though internal validity, or the validity of inferring a causal relationship between the treatment and outcome (Shadish et al., 2001), should already have been confirmed, it may still be unclear how the treatment will perform in real world settings. This is because internal validity is usually tested in laboratory-like experiments where conditions are rigorously imposed to control for confounding variables. For instance, in order to test the causal relationship between a new reading intervention and end of year assessment scores, researchers may have a highly trained staff deliver the intervention under strict supervision to a handful of classrooms assigned to receive treatment. Yet the conditions of an average classroom are far from ideal and teachers with only a few days of professional development are much less effective at delivering a new and highly demanding curriculum. Conversely, field trials strive to be realistic and unobtrusive in order to have a better estimate of how well the intervention truly works in the real world (Gerber and Green, 2012).

Field trials answer the question "how well does this treatment perform in real settings", but they don't always address how representative these settings are of the real world. This is a question of external validity, or the extent to which a causal relationship exists across variations in persons, settings, treatments and outcomes (Shadish et al., 2001). Answering this question is complicated, but necessary. Consider policy makers and other stakeholders who have an interest in interventions performing as advertised. A principal coming across a reading program that has been shown to improve reading outcomes on students in rural Wyoming, may ask if the same intervention would work on their students in downtown Brooklyn. For this very purpose, the What Works Clearinghouse (WWC), which provides detailed and systematic reviews of research supported educational interventions, has recently begun reporting demographic variables. Yet generalizability is more complicated than simply matching students on a handful of arbitrary factors, and must be designed or accounted for statistically.

Finding a balance between internal and external validity is not simple. Ideally, dual randomization would be employed where units are randomly sampled into the study, then randomly assigned to a treatment or control group. Random treatment assignment supports internal validity by allowing us to assume that any observed differences between groups are due to receiving the treatment. Random sampling supports external validity by selecting a sample that is representative of a population. We define a representative sample as a miniature of a population (Kruskal and Mosteller, 1979), such that any effect present in the sample would also be present in the population. However, employing dual randomization is very expensive on a large scale in natural settings and is rarely done (Shadish et al., 2001). Most field trials eschew random sampling for this reason and instead rely on convenience samples (Gerber and Green, 2012), which make generalizations difficult. In contrast, observational studies may use probability sampling to select a highly representative sample while leaving treatment assignment to occur naturally (Shadish et al., 2001), which makes causal inference difficult.

Fortunately there has been much recent interest in developing alternative methods, propelled, in no small part, by government funding opportunities. Since its foundation, the Institute of Education Sciences (IES) has funded 956 studies (Fellers, 2017) contributing to their mission of providing and sharing scientific evidence in education. The IES funding structure is divided into five "Goals": (I) Exploration, (II) Development, (III) Efficacy and Replication, (IV) Effectiveness, and (V) Measurement. Of particular relevance here is Goal IV, which supports effectiveness studies, often referred to as "scale-ups". The purpose of these studies is to determine if an intervention, which has already been validated in an at least one efficacy trial, continues to show an effect in a "natural" setting, often at a national level. Scale-up studies are tasked with maintaining external validity, focusing on generalizing the treatment effect to a larger inference population. Yet a recent Government Accountability Office (GAO) review called attention to the lack of guidance provided for researchers in developing a representative sample (Fellers, 2017; Office, 2013). As a result, researchers are placing a larger emphasis on study designs and analytic methods that can support causal inference at the population level.

Several methods have been developed to account for non-random sample selection in field trials, and other randomized control trials (RCTs). These methods evaluate the similarity between the selected sample and population (Stuart et al., 2011) and reweigh the sample to estimate a population effect (O'Muircheartaigh and Hedges, 2014). Both methods are "retrospective", however, and coverage errors may still arise from a lack of representation of certain subsets of the population in the initial sample. This results, almost paradoxically, in the inference population being "trimmed" to resemble the sample, reducing the level of generalizability. Tipton (2013a) addressed the dearth in experimental design literature for in the literature for "designing generalizability", and proposed several methods for structuring the recruitment process to more purposefully target schools which are representative of the population.

Propensity score (Tipton et al., 2014) and cluster analysis (Tipton, 2013b) stratified sampling rely on splitting the population into several subsets, then prioritizing the sampling of units that are most representative of those subsets. The two methods differ in the manner of generating the subsets of units. The focus of the present study is the more general method of the two: cluster analysis stratified sampling (CASS; Tipton, 2013b). CASS relies on an optimization algorithm which maximizes the similarity within clusters and difference between clusters. Though some selection bias is ultimately unavoidable, this method potentially reduces coverage errors, which generates a more representative sample and improves the utility of retrospective generalization methods. Importantly, it also provides a more transparent, structured and well documented recruitment process which better lends itself to scrutiny by organizations like IES which provide funding for such studies, as well as other stakeholders and independent researchers. While potentially beneficial, there has been little exploration into the methodological aspects of the method. Tipton (2013b) provides only a single case study, and as of yet there have been no studies reporting the application of this method on a large scale study.

The purpose of the current study is to examine the ability of the CASS to generate a representative sample within the context of a large-scale field trial. In the next section we discuss the underlying framework for understanding causal inference and generalizability.

Then, we give an overview of various research designs that can be used for making casual inferences to a larger population and their limitations, followed by a formal introduction to CASS. Finally, the current research is proposed, followed by a brief discussion.

Targets of Inference

Potential Outcomes. Rubin (1974) provided a potential outcomes framework for defining and estimating the average treatment effect in randomized and non-randomized studies. Consider a trial examining the impact of a reading intervention on literacy exam scores. We begin with S defined as a sample of students participating in the trial. Let n be the number of students in S indexed by i, where $i = 1, .., n$. Let T_i be the binary treatment indicator, where $T_i = 1$ if student i receives the reading intervention and $T_i = 0$ if they do not. Prior to treatment assignment, all students in the trial have two potential outcomes. Let Y_i^1 be the outcome which would have been observed if student i were assigned to treatment ($T_i = 1$), and Y_i^0 the outcome which would have been observed if student i were assigned to control ($T_i = 0$). We define D_i as the difference between potential outcomes $Y_i^1 - Y_i^0$, or the potential treatment effect, for the individual unit i. We can now define the average treatment effect, τ^S, as the expected value of D_i across all units in S or

$$\tau^S = E(D_i) = E(Y_i^1 - Y_i^0) \tag{1}$$

Note that τ^S is the sample average treatment effect (SATE), which is specific to S.

If both Y_i^1 and Y_i^0, and therefore D_i, were observable for all units, we can calculate τ^S directly as the average of D_i across all units:

$$\tau^S = \frac{\sum_{i=1}^{n} D_i}{n} \tag{2}$$

However, as expressed by Holland (1986), due to the fundamental problem of causal inference it is not possible to observe both Y_i^1 and Y_i^0 for the individual unit once treatment begins. Rather, the observed outcome Y_i takes on the value of Y_i^1 for treatment units and Y_i^0 for control units:

$$Y_i = Y_i^1 T_i + Y_i^0 (1 - T_i) \tag{3}$$

Table 1

Potential Outcomes

Student	Y_i^1	Y_i^0	T_i	$1 - T_i$	Y_i
1	86	83	0	1	83
2	67	64	0	1	64
3	74	74	1	0	74
4	68	68	1	0	68
5	58	54	1	0	58
6	53	48	0	1	48

This is further illustrated in Table 1

A fundamental assumption of the potential outcomes model is the stable unit treatment value assumption (SUTVA; Rubin, 1980; Rubin, 1986). SUTVA states that treatment assignment for one unit should not affect the outcome of any other units in the sample, and that there should be only one version of each treatment. SUTVA is often violated in multilevel settings such as classrooms where one student receiving an intervention may affect other students, but this can be avoided by using cluster-randomization. In order to estimate SATE from observed outcomes, we need to make an additional assumption about the relationship between treatment assignment and the potential outcomes.

Here we must distinguish between randomized and non-randomized trials. In randomized trials, treatment assignment is completely independent of the potential outcomes:

$$T \perp [Y^1, Y^0] \tag{4}$$

This is known as the independence assumption (Holland, 1986). If independence is satisfied, we can show that the expected value of the potential outcomes is equal to the expected value of of the observed outcomes across all units in the respective treatment

and control groups groups:

$$E(Y^1) = E(Y^1|T = 1) = E(Y|T = 1) \tag{5}$$

$$E(Y^0) = E(Y^0|T = 0) = E(Y|T = 0) \tag{6}$$

This means that the expected value of Y^1 across all units in the trial is equal to the expected value of Y across all units assigned to the treatment. The same holds true for the control group. Using the linearity of expectation property of expected values, we can rewrite Equation 1 as the difference in expected observed outcomes for each group:

$$\tau^S = E(Y^1) - E(Y^0) = E(Y|T = 1) - E(Y|T = 0) \tag{7}$$

We can now calculate an estimate of τ^S as the simple difference in group means:

$$\hat{\tau^S} = \frac{\sum T_i Y_i}{\sum T_i} - \frac{\sum (1 - T_i) Y_i}{\sum (1 - T_i)} \tag{8}$$

where $\sum T_i$ is the number of units assigned to treatment, and $\sum (1 - T_i)$ is the number of units assigned to the control.

In non-randomized trials, treatment and control groups can differ systematically on a set of characteristics, which violates the independence assumption. However, this assumption can be relaxed to the strong ignorability assumption (Rosenbaum and Rubin, 1983), which has two components. First, treatment assignment is conditionally independent of potential outcomes given a set of pre-treatment covariates:

$$T \perp\!\!\!\perp [Y^1, Y^0]|X \tag{9}$$

where X is a vector of p covariates measured for each unit prior to exposure to treatment or control. Second, given X the probability of being assigned to treatment is nonzero:

$$0 < P(T = 1|X) < 1 \tag{10}$$

where $P(x)$ denotes the probability of x.

The first component of strong ignorability, sometimes referred to as unconfoundedess, is vulnerable to unobserved covariate bias. To the extent that any of the observed covariates

in the set X is correlated to any unobserved covariates related to treatment assignment, this becomes less of an issue. However, excluding any unobserved covariates which are unrelated to those that are observed would result in a violation of unconfoundedness. Though unconfoundedness can never be tested directly, Stuart (2010) summarizes several methods of assessing plausibility of violation and possible effects on study outcomes.

Population Average Treatment Effect. It is important to distinguish between sample average treatment effect (SATE) and population average treatment effect (PATE). Consider the study from the previous section examining the impact of a reading intervention on literacy exam scores. Suppose that the researchers, based in Austin, only selected schools in their general area to participate in the study. The SATE estimate would only be valid for this sample of schools. Researchers, policy makers, and other stakeholders who want to implement this intervention in their areas could argue, to the extent that their schools are "similar" to schools in Austin, that they would find the same effects. However, "similarity" is a vague concept here, and justifying this argument would require that the study was designed to support external validity, or generalizability.

Tipton (2013a) extends the SUTVA assumption from the potential outcomes framework to help define generalizability. Note that SUTVA as previously described relates treatment assignment with potential outcomes and must also be met. Here SUTVA has two components. First, the treatment effect of any unit in the sample is not affected by the treatment status of any other unit including in the population. This component can be violated when average treatment effects are sensitive to the proportion of the population being treated. The second component requires that the treatment effect of an individual unit should remain the same whether the unit is included in the sample. This second component can be violated when being included in an experiment has an impact on potential outcomes relative to being in similar circumstances but not a part of the experiment. One method of generalizing the finding of a study to units not include in the sample is to estimate the average treatment effect for the population that includes those units. For instance, if we estimate the population average treatment effect (PATE) for all schools in Texas, then superintendents in Houston can use these findings to justify

implementing the intervention in their schools.

In order to estimate PATE, we must first operationally define and enumerate the population. Let P be a well defined population of N units. Let each unit in the population be indexed by $i = 1, ..., N$. We define PATE, τ^P, as the average of the treatment effect across a population of units:

$$\tau^p = \frac{\sum_i^N D_i}{N} \tag{11}$$

To see why we can't estimate PATE directly, consider that PATE can be decomposed into the average treatment effect of units in the sample (SATE) and not in the sample (NATE; Imai et al., 2008). Let S be a sample which is a subset of P of size n. Let S' be the subset of P that is not in S of size $N - n$. Let τ^S and τ^N be the SATE and NATE respectively. We see that PATE is the weighted sum of SATE and NATE

$$\tau^P = \frac{n}{N}\tau^S + \frac{N-n}{N}\tau^N \tag{12}$$

Using Equation 8 we can readily estimate τ^S directly from the observed units in S, however doing so for τ^N can be more difficult. Probability sampling can be used to estimate τ^N by accounting for the selection probabilities for sample units and is common in survey sampling and other non-experimental designs (Groves et al., 2009). Due to limited resources of most large scale studies and the heavy logistical burden of implementing probability sampling followed by random assignment to treatment conditions, it is rarely implemented in field experiments (Shadish et al., 2001). However, probability sampling is frequently employed in observational studies where treatment assignment occurs naturally (Lohr, 2009; Shadish et al., 2001).

Observation Studies With Propensity Score Analysis

The primary characteristic of observational studies is that there is no treatment assignment mechanism that is controlled by the researchers Units either self select into treatment or control groups, or groups are simply observed in natural settings. Observational studies are employed when random assignment is not possible either to ethical or

practical concerns. Observational studies may consist of large samples, for instance, which would make randomizing unit into treatment and control conditions logistically difficult and costly. On the other hand, the large sample size makes it easier to collect a sample that is representative of the population

In the context of making causal inferences, Imai et al. (2008) identify three general categories of observational studies: (1) studies with random or known probability samples representative of the population, (2) studies with nonrandom samples but with enough information to correct for unrepresentativeness, and (3) studies based on convenience samples with no known relation to the population. Of particular interest to us is the first type. This design is commonly used in large scale nationally representative surveys like the Early Childhood Longitudinal Program (ECLS). Such studies rely on random or probability sampling to make population level inferences, making generalizations fairly straightforward. The trade-off is that this design is vulnerable to selection bias, where participants who receive treatment may systematically differ from those who do not on a range of preexisting characteristics related to treatment effect heterogeneity (Stuart and Rubin, 2008). We will therefore first discuss methods for estimating average treatment effects using this design, then discuss how these designs are used to make generalizations.

Propensity Score Analysis for estimating ATE. Observational studies may violate the independence assumption in that a set of covariates which predict treatment assignment also predict variation in potential outcomes. This assumption can be relaxed to the strong ignorability assumption when these covariates are observed and accounted for statistically. Various regression and propensity score techniques can be used to estimate causal effects while controlling for these differences (Schafer and Kang, 2008). Propensity score techniques are typically preferred over regression as they are robust to model mispecification and large differences between groups, provide more parsimonious outcome analyses, and encourage researcher honesty by isolating the propensity estimation from the outcome analysis (Ho et al., 2007; Stuart and Rubin, 2008; Williamson et al., 2012).

Propensity scores are a scalar summary of covariates which predict treatment assignment (Stuart and Rubin, 2008). Let $\pi_i(X)$, or π_i for short, be defined as the probability

that unit i is assigned to treatment ($T_i = 1$) given a set of covariates X_i:

$$\pi_i(X) = P(T_i = 1|X_i) \tag{13}$$

Rosenbaum and Rubin (1983) provided two key theorems enabling the use of propensity scores for estimating ATE. First, the distribution of X is the same in treatment and control groups for any given value of the propensity score. Thus, within a small range of the propensity score, the treatment assignment mechanism is approximately random with respect to the set of covaraites. Second, if treatment assignment is unconfounded given X, then treatment assignment is unconfounded given π_i:

$$T \perp\!\!\!\perp [Y^1, Y^0]|\pi_i \tag{14}$$

In practice, π_i is unknown outside of experimental designs and must be estimated. Typically this is done using logistic regression (Menard, 2002), however many other methods have recently been proposed for this application including machine learning, decision trees, meta-classifiers, discriminant analysis and others (Lee et al., 2010; McCaffrey et al., 2004; Westreich et al., 2010). Logistic regression is the simplest as it intuitively estimates π as a set of fitted values obtained by regressing the treatment indicator on a covariates related to treatment. This requires that researchers have some theoretical understanding of the treatment assignment mechanism, and were able to measure the covariates prior to treatment exposure. Let $\boldsymbol{X_i}$ be a 1 x p vector of covariates for each unit i measured prior to exposure to treatment, and let $\hat{\boldsymbol{\beta}}$ be a 1 x p vector of estimated logistic regression coefficients associated with $\boldsymbol{X}$. We estimate π as

$$\hat{\pi}_i = \frac{e^{\hat{\boldsymbol{\beta}}\boldsymbol{X}_i'}}{1 + e^{\hat{\boldsymbol{\beta}}\boldsymbol{X}_i'}} \tag{15}$$

Note that this example assumes a linear relationship between the covariates and the propensity score. This does not always have to be the case, as the model can include interactions and curvilinear relationships.

Once $\hat{\pi}$ is calculated, several methods can be used to estimate ATE including matching, stratification and weighing (Austin, 2011; Kang and Schafer, 2007; Schafer and Kang,

2008; Stuart, 2010). The goal of these methods is to leverage the balancing aspect of the propensity score to compare groups or individuals that differ on treatment assignment but are otherwise balanced on all other characteristics, thereby approximating a randomized experiment. When estimating propensity scores, model specification and diagnostics are not concerned with accurately estimating parameters. Instead, the primary goal is to create covariate balance and accurately estimate the propensity scores (Stuart and Rubin, 2008).

There are several important limitations to consider when performing the propensity score analysis (PSA). First, PSA requires balance on all covariates that predict both treatment assignment and potential outcomes. If there are additional unobserved covariates that are uncorrelated with observed covariates that are included in the model, any ATEs estimated are likely to be biased. Though it cannot be confirmed, there are several methods to test for sensitivity to violations of this assumption in the context of propensity scores (Imbens, 2004; Stuart, 2010), and to minimize the risk of violation (Schafer and Kang, 2008). Second, the PSA is ineffective if there is not a large enough area of common support (Stuart, 2010). This occurs when distributions of the propensity scores between treatment and control groups overlap very little. In this case, there aren't enough subjects in one group that are "like" the other, and additional methods are required to account for selection bias.

Probability Sampling. In order to generalize the ATE, non-experimental studies typically rely on probability sampling to select a sample that is representative of the population of interest. Simple random sampling (SRS) is, as the name implies, the simplest type of probability sampling, where the probability of being sampled is equal for all possible subsets of the population (Lohr, 2009; Salkind, 2010). When sample sizes are large enough, SRS is the ideal sampling method as it guarantees, within sampling error, that the ATE estimated from one sample is equal to the ATE estimated across units from the population that were not sampled (Shadish et al., 2001). That is to say, the estimated

SATE of a random sample is a good estimate of both the NATE and the PATE:

$$\hat{\tau}^P = \hat{\tau}^S = \hat{\tau}^N \tag{16}$$

Therefore, no additional adjustments are necessary to estimate PATE.

In order to implement random sampling, all units in the population must be enumerated in a list (Salkind, 2010). If the population of interest is all public schools in Texas, then there must be a list of all schools from which random selections can be made. In certain cases, there may not be a list of all units, however there may be a list of all clusters that those units are members of. For instance, it would be difficult to get a listing of all public school students in Texas if that were the population of interest. However, since all Texas public school students are members of Texas public schools, then we can instead take an SRS of schools. We can then decide to sample all students in the sampled schools, or get a list of students within each school and take an SRS of students within schools. This is known a cluster sampling (Lohr, 2009). Stratified random samples can also be used in cases where there are small subsets of the population which may be missed in SRS. In this case, strata based on these subsets can be created, and an SRS of students within strata can be taken. For instance, when trying to take a random sample of all students in a university, some departments may be so small that their students have a low probability of being sampled. Taking a random sample from each department ensures that students from all departments would be represented in the survey. Large scale surveys that use a combination of all or some of these sampling methods, such as the ECLS, are known as complex surveys.

Estimating PATE using PSA and these additional probability sampling designs is non-trivial. Recent work has compared methods using sample weights as covariates in logistic regression or as weights in weighted logistic regression to estimate propensity scores, in combination with matching, stratification, and weighting in the outcome analysis (Austin et al., 2016; DuGoff et al., 2014; Ridgeway et al., 2015). There is also limited research in applications of PSA in multilevel settings (Leite et al., 2015; Li et al., 2013) which occurs in cluster sampling. Another serious issue is missingness. In practice, not all units

that are selected for sampling will agree to participate, and not all units who agree to participate will be measured on all variables. This is known as unit non-response and item non-response, respectively (Lohr, 2009). In the case of item non-response, using PSA to estimate ATE is complicated when there is missingness along covariates used to estimate the propensity score. Several solutions include multiple imputation, missingness patterns, machine learning and others (Cham and West, 2016; Crowe et al., 2010; D'Agostino and Rubin, 2000; Mitra and Reiter, 2011; Qu and Lipkovich, 2009). In the case of unit non-response, samples are no longer representative of the population to the extent that non-responding units differ systematically from responding units on characteristics related to treatment effects, violating the unconfounded sample selection assumption. Notably there are currently no recommendations for implementing propensity score analysis for multilevel complex survey data in the presence of missingness.

Multi-Site Field Trials: Randomized Trials on Convenience Samples

Field Trials refer to randomized controlled trials (RCTs) conducted in "natural" settings where units are randomly assigned to treatment or control (Gerber and Green, 2012). RCTs are much preferred over observational studies for making causal inferences, and are considered to be the gold standard for doing so (Rubin, 2008). In educational intervention research, these studies are commonly employed as multi-site trials, where units are randomized within a site across multiple sites. For instance, several districts may be selected to participate, then schools within the districts are assigned to treatment or control. An ideal multi-site field trial study which seeks to generalize causal inferences would implement dual randomization. First, a sample S of n sites is drawn randomly from a well-defined population P of N sites. Next, units within each site are randomly assigned to treatment or control. This satisfies both SUTVA and Independence assumptions for potential outcomes and treatment effects.

Unfortunately, random samples are seldom used. In natural settings, treatment implementation and attrition are major roadblocks to an RCT's ability to maintain a probability sample while randomizing treatment assignment (Shadish et al., 2001). Units who

are selected for treatment don't always agree to participate (non-response, refusal), units who agree to participate don't always comply with their assignments (compliance), and units who agree and comply over time drop out (attrition). Working to address, control for, and prevent these issues can quickly use up a study's already limited resources, and so researchers often opt for convenience samples instead.

Convenience samples comprise of units that are easy to recruit or that are most likely to respond (Lohr, 2009). For example, recruiters can save resources by reaching out to districts where they already have relationships or to schools that have participated in previous research. Researchers sampling from multiple states can save money by only sampling from large urban centers. Instead of sending materials and intervention coaches to locations randomly dispersed across the states, they can concentrate on one central location in each state. Of course, this rarely results in a representative sample of initial target population (Lohr, 2009).

Convenience samples are subject to selection bias either on the part of the researcher or on the unit being sampled. For instance, in an attempt to deliberately select a representative sample, researchers may recruit units that they consider to be "average" representations of the population. This selection technique is also known as a judgement sample and is subject to the researchers' biases. Alternatively, if the researchers were to randomly sample units, but only those who agree to participate are selected into the sample, then the units have "self-selected" into the sample. In either case, the sample selected may no longer be representative of the population. If there are systematic differences between the sample and the population on characteristics that are related to differences in treatment effects, then the SATE estimated for that sample is no longer an unbiased estimate of PATE, reducing the external validity of the study. This difference must be then be accounted for statistically.

In multi-site research, the simplest method of generalizing beyond the sample of sites is using multilevel random effects models (MREM), also known as hierarchical linear modeling (HLM; Bryk & Raudenbush, 1992). Site level random effects with an assumed distribution are included in the model's estimate of the treatment effect, generalizing

the estimated effects to a population of sites that are "similar" to those in the sample. The major caveat here is the vagueness of this similarity. The assumed distribution of the random effects does not give more information beyond that findings generalize to a population of schools that are like those found in the study. If the researcher is interested in generalizing to a population not quite captured by the study's sample, a different approach is necessary.

Several retrospective methods have been developed to measure the sample's generalizability, or to reweigh it such that it resembles a specific population of interest. These methods both employ the propensity score and rely on populations being well defined. Furthermore, researchers must have access to extent data containing a rich set of variables describing characteristics of population units which may relate to sample selection and potential outcomes, such as the Common Core of Data (CCD).

Quantifying Generalizability. Stuart et al. (2011) proposed to use the propensity score as a metric to quantify the similarity of a sample to a target population, referring to it as a distance measure between the participants in the sample and the target population. This method relies on the assumption that sample selection is unconfounded, which is satisfied when unit treatment effects are independent of the sampling mechanism given a set of covariates:

$$D_i = [Y_i - Y_i] \perp\!\!\!\perp Z|X \tag{17}$$

This requires that X includes all covariates which explain variation in treatment effects and differ in distribution between the sample and population. Let π_i^S be defined as defined as the propensity for a unit in the inference population to be sampled given a the set of covariates that satisfy unconfounded sample selection:

$$\pi_i^S = P(Z_i = 1|X_i) \tag{18}$$

The propensity to be sampled can now be used to satisfy the unconfounded sample selection assumption by replacing the vector of covariates X_i:

$$Z_i \perp\!\!\!\perp [Y_i^1, Y_i^0]|\pi_i^S \tag{19}$$

This propensity score is also unobserved, and must be estimated using Equation 15. If estimated accurately, π^S summarizes the set of covariates for each unit such that units matched on $\hat{\pi}_i^S$ have equal probability of being selected into the sample.

Because $\hat{\pi^S}_i$ is a scalar summary of the set of covariates most related to the probability of being sampled, differences along $\hat{\pi^S}_i$ provide us with a measure of similarity between the units in the sample and units in the population. The differences between the sample and population can be summarized as D_p which is the difference between the mean of $\hat{\pi}_i^S$ in the sample and the population:

$$D_p = \frac{\sum Z_i \hat{\pi_i^S}}{\sum Z_i} - \frac{\sum (1 - Z_i)\hat{\pi_i^S}}{\sum 1 - Z_i} \tag{20}$$

If the sample is representative of the population, then the difference between $E(\hat{\pi}_i^S | Z_i = 1)$ and $E(\hat{\pi}_i^S | Z_i = 0)$ should be small and $E(D_p)$ should approach zero. As the systematic differences between the sample and population increase, so should $E(D_p)$. Stuart et al. (2011) provide no clear cut-point of an acceptable difference though previous literature on observational studies has suggested .25 or even .1 standard deviations to indicate a very large amount of extrapolation (Cochran and Rubin, 1973; Rubin, 1973).

Stratification for Reducing Bias. O'Muircheartaigh and Hedges (2014) proposed stratification based on propensity score matching as a method of reducing bias in causal effects estimates due to nonrandom sampling. Strata can be created by grouping units in the sample and population using cutoff scores across the range of the estimated propensity scores. Stratifying on the propensity score results in groups of sample and population units that are homogeneous with respect to the set of covariates and have roughly equal probability of being selected into the sample. Let sample S be defined as a subset of population P. Let J be the total number of strata created, with each strata being indexed as $j = 1, ..., J$. Let S_j be defined as a subset of S units that were grouped into stratum j, and let P_j be a subset of P units that were grouped into stratum j. Finally, let τ_j^S be the SATE of units grouped into stratum j, and let τ_j^P be the PATE of units grouped into stratum j. Within each stratum j, S_j should be representative of P_j, and therefore

τ_j^S would be an unbiased estimate of τ_j^P. Weighing by the proportion of population units within each strata we can estimate the overall τ^P as such:

$$\hat{\tau}^p = \frac{\sum_{j=1}^{J} N_j \hat{\tau}_j^S}{\sum_{j=1}^{J} N_j} \tag{21}$$

where N_j is the number of population units in stratum j. Rosenbaum and Rubin (1984) have shown that dividing into 5 strata can remove at least 90% of bias within each strata.

Multi-Site Field Trials: Purposive Sampling

Tipton (2013) showed that retrospective methods such as those previously described are subject to coverage errors when used with convenience samples. Coverage errors occur when the sampling frame is sufficiently different from the population, that there are not enough units in the sample that are "like" those in the population. Furthermore, any attempts to reweigh the sample to fit the inference population will result in much larger standard errors greatly reducing the study's power. One way to consider this is to think of the sampling frame as a sub-population of units that are eligible to be selected as part of a convenience sample. This eligible population is defined by a set of eligibility criteria which can depend on power analysis specifications, budgetary limitations, and other researcher specified constraints. As study resources become limited, these eligibility criteria become more restrictive, and the eligibility population becomes less like the inference population. Furthermore, the sample selection process becomes iterative, informal and poorly documented (Tipton et al., 2014).

To address this, Tipton et al. (2014) and Tipton (2013b) suggest two strategies for purposive sampling from an eligible population: propensity score stratified sampling and cluster analysis stratified sampling. Both methods are based on selecting a bias-robust balanced sample which will be defined first. Each method guides the researcher through a process that is intuitive and easy to document and justify.

Bias-Robust Balanced Sampling. The goal of the balanced-sampling framework is to purposively select a sample such that it is a miniature of some well-defined population. The first step is to identify the inference population P of size N for which exists a measured

set of covariates. Next a sample S of size n must be selected such that τ^S is an unbiased estimate of τ^P. Bias in this case occurs when units in S differ in relation to units in P on a set of covariates which explain variation in treatment effects (Stuart et al., 2011; Tipton, 2013a). Since we do not know what these covariates may be prior to conducting the experiment, theory or previous research may be used to model a possible relationship. For instance, we may posit that the effect of a school based reading-intervention is linearly related to the prior year's reading test scores, which is observed for all units in P. Let D be the treatment effect of the intervention, and let X be the observed pre-treatment covariate (reading scores). We can now propose the following treatment heterogeneity model:

$$D = \beta_0 + \beta_1 X \tag{22}$$

If this model is accurate, then selecting a sample such that the expected value of X in the sample is equal to the expected value of X in the population, E(X|Z=1) = E(X), would result in a balanced sample, and $\hat{\tau}^S$ would be an unbiased estimator of τ^P.

Despite our best efforts, this model may still be incorrect. It is therefore necessary to develop method that is robust to model failure. Suppose in actuality the relationship between the treatment and covariate is curvilinear, and treatment is also related to another covariate, say school socioeconomic status (SES) which is also observed for all schools in the population prior to treatment. Let X_1 and X_2 be the observed pre-treatment reading scores and SES. The true model is now more complex:

$$D = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \beta_3 X_2 \tag{23}$$

If we are to rely on the simpler model we originally proposed, we end up with a biased estimate of τ^P. We can calculate that bias as follows:

$$\tau^S - \tau^P = \beta_2[E(X_1^2|Z = 1) - E(X_1^2)] + \beta_3[E(X_2|Z = 1) - E(X_2)] \tag{24}$$

To guard against selecting the wrong model, or model failure, we must develop a bias-robust balanced sampling strategy. Tipton (2013b) suggests selecting a balanced sample

by using a model of order R, where $r = 1, ..., R$ is the set of moments on which the sample is balanced on. Let $X_h = X_1, ..., X_p$ be a set of pre-treatment covariates which predict D. A balanced sample of order R is one where the expected value of each covariate for each power in the sample is equal to its expected value in the population:

$$E(X_h^r | Z = 1) = E(X_h^r) \tag{25}$$

A model of the form

$$D = \beta_0 + \sum_{h=1}^{p} \sum_{r=1}^{R} \beta_{hr} X_h^r \tag{26}$$

is robust against model failure as long as the true model is nested within it. This is because a sample balanced on higher moments of X leads to an unbiased estimate if the true model is based on equal or lower moments (Valliant et al., 2000). We could further guard against omitted variable bias by including more covariates which may correlate to any unmeasured variables related to differences in potential treatment effects (Brewer, 1999; Royall and Herson, 1973; Rubin and Thomas, 1996; Smith and Sugden, 1988; Stuart, 2010).

Stratified selection, discussed briefly as a probability sampling method in the form of stratified random samples, can be used to select a balanced sample. If the goal is to balance on a few categorical covariates, the strata are created naturally. Let X_1 and X_2 be two categorical variables with q_1 and q_2 levels respectively. To draw a sample balanced on X_1 and X_2 we can define a stratum for each possible combination of the variable values for a total of $k = q_1$ x q_2 strata. Let N_j be the number of P units in stratum $j = 1, ..., k$, and let $N = \sum N_j$ be the total number of units in P. Finally, let $\hat{\tau}_j^S$ be the estimated sample average treatment effect in stratum j. We can estimate the population average treatment effect from units selected into the strata by taking the weighted average of $\hat{\tau}_j^S$:

$$\hat{\tau}^P = \sum_{j=1}^{k} w_j \hat{\tau}_j^S \tag{27}$$

where $w_j = N_j/N$ is the proportion of P units in stratum j. As X becomes larger and begins to include continuous variables, defining these strata becomes much more difficult,

because a larger number of strata are required to account for all combinations of variables. The following sections describe two methods proposed by Tipton et al. (2014) and Tipton (2013b). Both methods are in essence dimension reduction techniques that enable multivariate stratified selection on a large and complex set of covariates.

Propensity Score Stratified Sampling. As with the retrospective methods, propensity score stratified sampling relies on the existence of a population frame, where all units in the population are enumerated and various characteristics are reported. We begin by identifying the following terms. Let P be the inference population of size N, let E be the population of eligibles of size M, and let S be the sample selected from E of size n. Both P and E must be well defined. As an example, P may consist of all public schools in Texas that offer 8th grade. A power analysis dictates that there need to be at least 40 students per grade, and financial restrictions mean the schools need to be from urban areas. Now E consist of all urban schools in P that have at least 40 students. The sample frame which represents E must contain all of the same information as the population in order to be able to sample S from E. Finally, n must be defined a priori based on a power analysis.

Next the set of covariates upon which to balance must be identified. This is guided by two assumptions: unconfounded sample selection (Stuart et al., 2011) and unconfounded eligibility (Tipton et al., 2014). Let $X = X_1, ..., X_p$ be defined as the set of p covariates that explain variation in treatment effects, and $G = G_1, ..., G_m$ be defined as the set of m covariates that define the eligible population (grade size, urbanicity). To satisfy unconfounded sample selection, S and P must be balanced on all covariates in X. Unconfounded eligibility states that if $E \neq P$ then X must not contain any covariates present in G. This is to say that eligibility must not be defined by covariates that explain variability in treatment effects. Violating this leads to biased estimates, and it is recommended to redefine P in order to satisfy the assumption.

Once all of these items are identified, the sample selection process begins. Strata must now be generated based on all of the covariates in X. Tipton et al. (2014) recommends using an eligibility propensity score for this purpose. Let Q be defined as an eligibility

indicator, where $Q = 1$ if a unit is in E and $Q = 0$ if a unit is not in E, and δ as the eligibility propensity score. We define δ as the probability of a unit begin included in G given X:

$$\delta = P(Q = 1|X) \tag{28}$$

where for every unit in P there must exist a similar unit in E:

$$0 < \delta < 1 \tag{29}$$

If there are units with $\delta = 0$ or $\delta = 1$, then this is a violation of the unconfounded eligibility assumption and will lead to coverage errors (Tipton, 2013a). In this case, it is recommended to redefine the population. As before, the true propensity score is unknown and must be estimated using the methods previously described.

The next step is to generate strata based on the estimated eligibility propensity score. Five strata is often recommended in the context of observational research (Cochran, 1968). Generating more strata would be beneficial, however depending on sample size requirements it may be difficult to adequately sample from too many strata. There are a number of methods for generating strata based on $\hat{\delta}$. One can generate strata such that $N_1 = N_2 = ... = N_j$ where each strata contains equal proportions of the population, or the range of $\hat{\delta}$ can be divided equally among k strata. Finally a simple random sample can be taken from the eligible units within each strata. Alternatively, units within strata can be ranked in order of how well they represent the strata using a distance measure. Units with higher ranks are then prioritized by recruiters. In either case, proportional sample allocation should be used such that the sample selected from each stratum is proportional to the number of units from the population that are in that strata. Let n_j be the number of units selected into the sample such that $n = \sum n_j$. We calculate n_j as:

$$n_j = n\frac{N_j}{N} \tag{30}$$

There are several advantages of applying this method. First, the inference population is explicitly defined by researchers, and the sample selection process is well documented.

Second, This method can readily be extended, with additional assumptions, to situations where eligible population and inference population don't overlap. Third, it does not preclude the use of the retrospective methods and in fact can reduce coverage errors when employing them (Tipton et al., 2014).

There are several limitations as well. The eligible population isn't always stable, and may change as the study progresses. For instance, researchers may want to sample 60 schools from 3 states, but only have enough resources to support treatment implementation in one state at a time. Several clusters of districts from each state are selected. Schools in these sets of districts now represent the entire sample frame, and the strata are created based on this information. After the first year, one state drops out and the strata are no longer valid. Second this method can't be applied when all or most of the units in the population are eligible. A more general method, cluster analysis stratified sampling (Tipton, 2013b), addresses these issues.

Cluster Analysis Stratified Sampling

Cluster analysis stratified sampling (CASS; Tipton, 2013b) applies cluster analysis in order to implement bias robust balanced sampling. The goal of CASS 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 supports 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 $d^e_{ii'}$ 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 $d^e_{ii'}$ as:

$$d^e_{ii'} = \sqrt{\sum_{h=1}^{p} w_h (X_{ih} - X_{i'h})^2} \tag{31}$$

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:

$$d_{ii'h} = 1 - \frac{|X_{ih} - X_{i'h}|}{R_h} \tag{32}$$

where $|.|$ indicates absolute value, X_{ih} and $X_{i'h}$ are values of the h^{th} 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 $d^g_{ii'}$ be the general similarity between unit i and unit i' where $i \neq i'$.

$$d^g_{ii'} = \frac{\sum_{h=1}^{p} w_{ii'h} d_{ii'h}}{\sum_{h=1}^{p} w_{ii'h}} \tag{33}$$

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:

$$p_k = \sigma^2_{bk}/(\sigma^2_{wk} + \sigma^2_{bk}) \tag{34}$$

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:

$$E(X_h|Z = 1, j) = E(X_h|j) \tag{35}$$

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:

$$d_{ij} = \sqrt{\sum_{h=1}^{p} w_h[X_{ijh} - E(X_h|j)]^2} \tag{36}$$

where w_h is the weight assigned to covariate X_h, $E(X_h|j)$ is the mean of the h^{th} covariate in stratum j, and X_{ijh} is the value of the h^{th} 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:

$$E(X_h^r|j) - E(X_h^r|Z = 1, j) \tag{37}$$

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:

$$\begin{aligned} E(X|Z = 1) &= \sum (n_j/n)E(X|Z = 1, j) \\ &= \sum (N_j/N)E(X|j) \\ &= E(X) \end{aligned} \tag{38}$$

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. CASS 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 CASS 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 CASS 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 CASS resulted in less coverage errors than the original sample and would thus be easier to generalize using the retrospective methods.

Tipton (2013b); Tipton et al. (2014) showed that 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. CASS 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, CASS 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, CASS 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., 2016b). Furthermore, recruiting ranked units from multiple strata requires a coordinated effort between recruiters (Tipton et al., 2016a). 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?

Proposed Research

The purpose of this study is to test the conditions under which CASS is effective for selecting a sample representative of a targeted population. CASS is an untested time-

intensive method, and researchers must be aware of whether it is appropriate or even feasible given their resources and area of study prior to applying it. To this end, the following research questions are proposed:

RQ1. How does the generalizability of samples selected by CASS compare to that of samples selected by convenience sampling and random sampling?

RQ2. How much additional units need to be contacted in order to fully select a sample compared to convenience sampling and random sampling?

RQ3. How does omitting a variable affect performance of CASS when modeling strata?

RQ4. How does overall participation rate affect performance of CASS relative to convenience sampling and random sampling?

RQ5. How does homogeneity of units in the population affect the performance of CASS relative to convenience sampling and random sampling?

These questions will be addressed within the context of a hypothetical example where researchers are selecting a state or nationally representative sample for a large scale multi-site field trial. Researchers must recruit districts and schools to participate in a 5th grade reading intervention. Samples achieved by various methods will be compared on how balanced and generalizable they are, as well as the difficulty with which they were achieved.

Overview

All research questions will be addressed by comparing CASS to convenience sampling (CS) and random sampling (RS). CS represents sampling as it is currently conducted in large scale studies. If CASS outperforms CS, then it is a more effective method that researchers should consider. Conversely, random sampling (RS) will serve as the standard, thought it should be noted that RS is also subject to selection bias in the presence of non-response, as are CASS and CS. Though it is unfeasible in most large scale designs, comparing CASS to RS will provide insight into the shortcomings of the CASS method.

RQ1 will be addressed by comparing the performance of the three sampling methods. Performance refers to how generalizable and well balanced a sample selected by a given method is. We expect that overall CASS will out-perform CS, and that random sampling will out-perform both CASS and CS.

RQ2 will be addressed by tracking the number of rejections from districts and schools accrued by each sampling method before achieving a full sample. Recruitment is a limited resource which may easily be exhausted pursuing schools that are highly valuable for research purposes but unlikely to participate. It may be the case that under certain conditions, CASS is too restrictive to fully implement and some compromise or mixed method must be applied to loosen the restrictions. We expect that in highly heterogeneous populations or under very low participation rates, the strict application of CASS Will become less feasible.

RQ3 will be addressed by performing CASS both using the full set of covariates predicting sample selection (CASS-F) to inform strata generation, and by omitting a variable (CASS-OV) to inform strata generation. Researchers may not always know all of the reasons that individual units chose to participate in experimental studies, or may not always be able to access all of the necessary data prior to sampling. By comparing the performance of CASS-F and CASS-OV, we will be able to measure how robust CASS is to omitted variable bias. We expect that CASS-F will outperform CASS-OV, however the degree to which this will occur may depend on the covariance of the omitted variable and the other variables in the model.

RQ4 will be addressed by generating high, medium and low overall participation rates. CASS may not be applicable in a population with a very low participation rate, because the pool of possible participants is so small. Furthermore, as the participation rate increases, we expect that the number of rejections to increase. As such, a researcher may want to settle for a less complicated model if they expect a very low participation rate. We expect that as participation rates decrease, the range in performance of the three models also decreases.

RQ5 will be addressed by using different sets of target populations which vary in het-

erogeneity on a set of characteristics. The purpose of CASS is to create several strata from a population such that sampling units within each strata are similar and between strata are dissimilar on a set of variables. If the set of variables are already similar across the population, then the CASS will be less effective in creating distinct strata. We expect that in populations with less heterogeneity, the CASS will perform more similarly to CS.

In total, three conditions will be manipulated for each sampling method: sampling method, participation rate, and heterogeneity in the population. The pretense of an omitted variable will be a condition specific to the CASS method.

Method

The proposed study will use a Monte Carlo technique repeatedly sample from a single population. Repeated sample designs can sample from real or generated populations. Because the population parameters are known, samples selected by different methods can be compared based on how closely their characteristics approximate those of the population. Extant data collected and housed by federal and state agencies will serve as the population frame. Using real data rather than a hypothetical generated population improves the external validity of the study because methods compared under realistic conditions are more readily applicable to researchers.

To simulate schools and districts agreeing to participate, a response generation model (RGM) will be created to generate a "participation propensity" based on a set of school and district characteristics. The source of variability between replications is added by sampling each unit's participation from a Bernoulli distribution with this participation propensity as the probability. Within each replication, a sample will be selected using different sampling methods under various conditions. The balance of the samples, their generalizability to the population, and the number of rejections will be compared across sampling methods.

Response Generation Model. To realistically model responses, it must be understood why units choose to participate in experimental studies. Prior research has shown districts are more likely to agree to participate if they have more English language learner

(ELL) students, more economically disadvantaged (ED) students, fewer White students, more Black/African American students, families with lower educational achievement, more poverty, and are Urban (Stuart et al., 2017; Tipton et al., 2016a). Although there are currently no studies reporting indicators predicting school participation, anecdotaly prior research teams have found that schools with a high level of performance in academics related to the intervention would refuse to participate. The extant data used as the basis for the proposed simulation contains variables that represent these characteristics.

The purpose of the response generating model (RGN) is to model the propensity to participate of districts and schools, which is the true probability that a district or school will participate if approached. The process of recruiting schools to RCTs often begins at the district level since many districts serve as gatekeepers for their schools. First a proposal to do research is submitted, typically to a research and evaluation department at the district level. If a proposal is rejected, all schools within the district become inaccessible. If the proposal is accepted, then researchers may begin recruiting schools in the district, which in turn may accept or reject the offer to participate. To better represent the two step process of first approaching districts for approval, and then approaching schools, two RGMs will be developed. Let π^D and π^S be the propensity to participate for districts and schools respectively. The following base model will be used for the district RGM:

$$\log\left(\frac{\pi^D}{1-\pi^D)}\right) = \beta_0 + \beta_1 X_{Suburban} + \beta_2 X_{Town/Rural} + \beta_3 X_{pELL} + \beta_4 X_{pED} + \beta_5 X_{pMin} + \beta_6 X_{MedInc} \tag{39}$$

where $X_{Suburban}$ is the percentage of schools in the district which are suburban, $X_{Town/Rural}$ is the percentage of schools in the district which are in towns or are rural, X_{pELL} is the average percentage of ELL students in the district's schools, X_{pED} is the average percentage of ED students in the district's schools, X_{pMin} is the average percentage of minority students in the district's schools, and X_{MedInc} is the average median household income in the district's school communities.

The following base model was used for the school RGM:

$$\log\left(\frac{\pi^S}{1-\pi^S}\right) = \gamma_0 + \gamma_1 W_{Suburban} + \gamma_2 W_{Town/Rural} + \gamma_3 W_{pELL} + \gamma_4 W_{pED} \tag{40}$$
$$+ \gamma_5 W_{pMin} + \gamma_6 W_{MedInc} + \gamma_7 W_{ELA} + \gamma_8 W_{Math}$$

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

Sampling Methods. Four sampling methods will be compared: random sampling (RS), convenience sampling (CS), cluster analysis stratified sampling full (CASS-F), and cluster analysis stratified sampling with omitted variable (CASS-OV). For all four methods, the decision of whether or not a school agrees to participate will be the same and will be generated at the start of every iteration. Let J be the total number of districts in the population, and let districts be indexed by $j = 1, ..., J$. We define E_j as a binary indicator that district j will agree to participate if contacted by recruiters, where $E_j = 1$ if the district agrees, and $E_j = 0$ if the district refuses. Let n_j be the total number of schools in district j, and let schools be indexed by $i = 1, ..., n_j$. We define E_{ij} as a binary indicator that school i will agree to participate if contacted by recruiters, where $E_j = 1$ if the school agrees, and $E_j = 0$ if the school refuses.

First, each district will be checked for approval by sampling from a Bernoulli distribution with probability equal to the districts' participation propensity.

$$E_j \sim B(\pi_j^D) \tag{41}$$

If the district declines, all schools in that district are excluded from the sample. Otherwise, the school is then checked for approval by again sampling from a Bernoulli distribution, this time with the probability equal to the school's participation propensity:

$$E_{ij} \sim \begin{cases} B(\pi^S_{ij}) & \text{for } Z_j = 1 \\ 0 & \text{otherwise} \end{cases} \tag{42}$$

Thus at the beginning of each iteration, whether or not a school is willing to participate will already be determined.

In random sampling, a simple random sample will be taken of all the schools in the population. In real world settings a simple random sample is not practical. With smaller samples some subsets of schools may be overlooked, and schools that were sampled would be randomly scattered throughout the state making data collection and treatment implementation logistically difficult. More likely methods would be clustered randomization, stratified random sampling, or some combination of both. However, the goal is to have a standard as a comparison across multiple iterations, which in this case is simple random sampling. To achieve this, the order in which schools are approached will be indexed by rank, r, where $r = 1$ if a school is approached first. Rank will be randomized such that each school has an equal probability of being approached. Once schools are ranked, each school will be approached until 60 schools agree to be in the sample:

$$\sum_{r=1}^{R} Z^S_r = 60 \tag{43}$$

where R is the total number of schools approached. This method will allow tracking of the number of schools that declined to participate, $R - 60$.

In convenience sampling we will assume that recruiters have some knowledge the schools' and districts' likelihoods of participating, and will prioritize recruitment based on that knowledge in order to minimize effort. To achieve this, districts will be ranked in order of π^D overall such that $r^D = 1$ for the district with the highest participation propensity (most likely to participate). Schools will then be ranked within districts in order of of π^S such that $r^S = 1$ for the school with the highest participation propensity within the district with the highest participation propensity. In this way, ranking will be nested within districts, and all schools in higher ranked districts will be approached prior to schools in a lower

ranked district, regardless of the schools' participation propensities. Once again districts and schools will be approached until 60 schools agree to be in the sample:

$$\sum_{r^S=1}^{R} Z_{r^S}^{S} = 60 \tag{44}$$

Sampling using CASS-F and CASS-OV will follow the procedure outlined in the section on CASS. The two methods will vary in the model used for generating strata, and ranking schools within strata, with CASS-F using the full RGM, and CASS-OV omitting a variable. This may result in differences in the number of strata generated, school membership of the strata, and school rankings within strata. Schools will be ranked within strata such that $r_k = 1$ for the school that is most representative of strata k. The percent of the total sample recruited from each strata should be proportional to the percentage the population of schools that are in the strata (proportional sample allocation). Therefore, schools will be approached independently within each strata in order of rank until the proportional sample allocation requirements are met:

$$\sum_{r_k=1}^{R_k} Z_{r_k} = [\frac{n_k}{N} 60] \tag{45}$$

where n_k is the total number of schools in the strata, N is the total number of schools in the population, and R_k is the total number of schools approached in strata k with brackets indicating rounding to the nearest whole number.

Though extremely unlikely, in the case of the CS and both CASS conditions, several schools and/or districts may have the same rank. In such cases, schools with equal rank will be ordered randomly.

Population Frame. Three sources of data were used to inform the population frame: the Common Core of Data (CCD), publicly available State DOE Data, and data from the US Census. The CCD is a comprehensive database housing annually collected national statistics of all public schools and districts. The CCD maintains general descriptive information on schools and districts (name, address, phone number), student and staff demographics, and fiscal data (revenues, cost expenditures). The most recent complete data set is from the 2013-2014 academic school year.

Beginning with the No Child Left Behind (NCLB) act in 2001, all state education agencies are required to provide reasonable public access to achievement data at the school level. This data typically includes performance on English and Language Arts (ELA) and Math exams mandated by the state, often broken down by categories of demographics (gender, ethnicity, economic disadvantage status), as well as grade. The data is made available via government run state education websites in various degrees of specificity. In this way it is possible to access school level achievement and compositions of English language learners (ELLs), students with disabilities, and students who are economically disadvantaged.

Data from the 2015 American Community Survey was accessed via the American FactFinder tool made available on the US Census government website. This data provided us with median 12-month household incomes within each zip code. This data was linked with the CCD data which provided school zip codes providing us with a school community measure of socioeconomic status.

The following district and school level variables are available from these three data sources for response generation and sample modeling: urbanicity, percentage of ELL students, percentage of minority students, percentage of economically disadvantaged (ED) students, average school performance on scaled English and language arts (ELA) and math assessments, and median household income.

Data from six states have been selected for the simulation study: California, Oregon, Pennsylvanian, South Carolina, Texas and Wisconsin. Each state separately will serve as individual inference populations, with all six states together serving as a "national" inference population. Table 2 displays the total number of districts, schools and students within each population. The six states were selected because they all report vertically scaled assessment scores and represent different levels of heterogeneity on the previously identified covariates of interest. Table 3 displays the standardized mean differences between each state population and the overall national population with magnitudes greater than .5 in bold. Table 4 displays the means and standard deviations for each population along the covariates. Generally speaking, public funding as provided by IES is focuses

on research in public education, therefore only public schools, as indicated in the CCD, will be included in the population. The study will also assume that the hypothetical intervention does not span multiple grades, therefore schools that offer 5th grade and have at least 50 students enrolled at that grade will be included in the population.

Table 2

Total units by population

Population	Districts	Schools	Students
ALL	1,991	8,928	4,235,304
CA	523	3,421	1,132,221
OR	94	404	188,855
PA	459	1,006	578,301
SC	79	462	153,012
TX	592	3,103	1,930,750
WI	244	532	252,165

Table 3

Standardized mean difference between states and total population

Var	CA	OR	PA	SC	TX	WI
Median Income ($)	0.31	-0.25	-0.09	**-0.54**	-0.18	-0.13
ED (%)	-0.03	0.14	-0.36	0.16	0.21	**-0.56**
ELA (%)	-0.47	-0.39	-0.03	**-0.95**	**0.84**	**-0.71**
ELL (%)	**0.63**	-0.44	**-0.84**	**-0.69**	-0.14	**-0.71**
Math (%)	**-0.70**	-0.38	-0.3	-0.44	**1.04**	-0.29
Minority (%)	0.34	**-0.81**	**-1.00**	-0.47	0.32	**-1.14**
Suburban (%)	0.17	-0.22	0.21	-0.09	-0.18	-0.15
Town/Rural (%)	-0.26	0.28	0.18	**0.62**	0.03	0.46
Urban (%)	0.05	-0.01	-0.36	-0.41	0.16	-0.22

Table 4

Descriptive statistics for each target population

Var	ALL	CA	OR	PA	SC	TX	WI
Median Income ($\$10^4$)	6.04 (2.52)	6.83 (2.76)	5.41 (1.47)	5.81 (2.18)	4.69 (1.42)	5.59 (2.47)	5.71 (1.54)
ED (%)	0.58 (0.29)	0.57 (0.3)	0.62 (0.3)	0.47 (0.25)	0.62 (0.2)	0.64 (0.28)	0.42 (0.23)
ELA (%)	0.61 (0.23)	0.51 (0.21)	0.52 (0.16)	0.61 (0.21)	0.4 (0.17)	0.8 (0.12)	0.45 (0.16)
ELL (%)	0.31 (0.33)	0.52 (0.38)	0.17 (0.16)	0.03 (0.06)	0.08 (0.1)	0.26 (0.22)	0.08 (0.11)
Math (%)	0.56 (0.29)	0.35 (0.22)	0.45 (0.17)	0.47 (0.22)	0.43 (0.17)	0.86 (0.1)	0.47 (0.19)
Minority (%)	0.63 (0.31)	0.74 (0.24)	0.38 (0.2)	0.32 (0.32)	0.49 (0.25)	0.73 (0.26)	0.28 (0.26)
Suburban (%)	0.4 (0.49)	0.48 (0.5)	0.29 (0.46)	0.5 (0.5)	0.35 (0.48)	0.31 (0.46)	0.33 (0.47)
Town/Rural (%)	0.2 (0.4)	0.09 (0.29)	0.31 (0.46)	0.27 (0.44)	0.45 (0.5)	0.21 (0.41)	0.38 (0.49)
Urban (%)	0.4 (0.49)	0.42 (0.49)	0.4 (0.49)	0.22 (0.42)	0.2 (0.4)	0.48 (0.5)	0.29 (0.45)

RGM Parameters. Because the RGM identifies which individual units will agree to participate if approached, it also dictates the overall participation rate of the population. Therefore, three sets of school and district RGMs are needed to generate low (10%), medium (25%) and high (50%) participation rates. Table 5 will display the coefficients used for each participation rate condition:

Table 5

Response model coeficients for school and district RGMs

	District			School		
β	10%	25%	50%	10%	25%	50%
Intercept	-	-	-	-	-	-
Suburban	-	-	-	-	-	-
Town/Rural	-	-	-	-	-	-
pELL	-	-	-	-	-	-
pED	-	-	-	-	-	-
pMin	-	-	-	-	-	-
MedInc	-	-	-	-	-	-
ELA				-	-	-
Math				-	-	-

The parameter values reported in Table 5 will be selected by iteratively modifying the RGMs until they produced samples with desired covariate characteristics. Tipton et al. (2016a) reported standardized mean differences (SMDs) between a sample of districts participating in an RCT and the population of districts along the set of covariates used in the RGMs, which served as the target characteristics. SMDs were calculated as

$$SMD = \frac{\bar{X}_j^S - \bar{X}_j^P}{\sigma_j} \tag{46}$$

where $\bar{X}_j^P$ is the population mean of variable j and σ_j is the population standard deviation

of variable j. $\bar{X}_j^S$ is the expected sample's mean of the variable j and is calculated as:

$$\bar{X}_j^S = \frac{\sum \frac{1}{1-P(S_i=1)} X_{ij}}{\sum \frac{1}{1-P(S_i=1)}} \tag{47}$$

where $P(S_i = 1)$ is the probability of the unit i being sampled as calculated by the appropriate RGM. X_{ij} is the value of variable j for unit i. Table 6 will display the standardized mean differences between the sample and population as reported by (Tipton et al., 2016a) and as produced by the district RGMs for each participation rate.

Table 6

Standardized mean differences generated using school and district RGMs

Variables	Tipton (2016a)	10%	25%	50%
Urban	0.43	-	-	-
Suburban	-0.60	-	-	-
Town/Rural	0.22	-	-	-
pELL	0.95	-	-	-
pED	0.67	-	-	-
pMin	0.56	-	-	-
MEDINC	-0.66	-	-	-

Because school characteristics have not been reported in prior works, values used as target SMDs for schools were the same as those used for districts. Math and ELA SMDs were selected to be -.25 to reflect anecdotal evidence suggesting schools refuse to participate in intervention research if students already perform highly on measures related to the intervention.

Simulation Steps

The following sections outline the proposed simulation steps. The cluster analysis stage of CASS, which generates sample strata, is constant across replications for each target population, and will therefore be completed first. This will be followed by response generation and sample selection, which will be repeated for each replication.

Cluster Analysis. Strata need to be generated such that each represents unique subgroups of the target population from which to begin sampling. This step will follow the description of CASS outlined previously. K-means clustering will be used to generate strata. The set of covariates used will depend on the CASS condition. For CASS-F, the full set of covariates used in the school RGMs will be used: urbanicity, percent ELL, percent ED, percent minority, median income, ELA performance and math performance. For CASS-OV, the median income variable will be omitted. This variable was selected for omission because all other variables can generally be found in education related extant data. The median income was found using US census data which may often be overlooked. Since both versions contain urbanicity, a categorical variable, the general similarity measure (Gower, 1971) will be used as the distance measure to generate strata.

Next, for each target population, the number of strata (k) to be generated will be determined. This will be done by repeating the clustering method to generate all sets of strata between $k = 2$ and $k = 20$. The proportion of variability that is between strata will be calculated using Equation 34 and plotted against each k strata. The k strata at which the rate of change for the proportion of variability slows will be selected for each target population.

Finally, schools within strata will be ranked in order of representativeness of that strata using the weighted Euclidean distance calculated with Equation 35. This will result in 7 sets of ranks (one for each target population) for each CASS method (CASS-F and CASS-OS). Once again, these ranks are independent across all other conditions and will remain constant across replications. They will only be used to specify the order in which schools and districts are contacted for participation.

Response Generation and Sample Selection. Within each replication, participation willingness will be determined for districts (E_j) and schools (E_{ij}) using Equations 41 and 42. This will be done separately for each of the three participation rate conditions using the appropriate RGMs. Next schools will be selected into the sample using RS, CS and the two CASS methods until Equations 43, 44, and 45 are satisfied. This will result in 48 sets of samples: 7 populations by 3 response rates by 4 models.

Outcomes of Interest

Two primary measures of performance are of interest: the extent to which the sampling method used achieves a representative sample of the population (sample balance and generalizability), and the difficulty of applying the sampling method (feasibility).

Sample Balance. The balance of each sample will be determined by calculating the standardized mean difference (SMD) of the covariates between the sample and the population. Let X_j^r be the full set of 9 covariates identified in Table 4 indexed by $j = 1, ..., 9$ to the order of r where $r = 1, .., 3$ if X_j is continuous and $r = 1$ if X_j is binary. Let $\bar{X}_j^r$ and M_j^r be the mean of covariate X_j^r in the sample and population respectively. Finally, let σ_j^r be the standard deviation of X_j^r in the population. We will calculate the SMD of each covariate as

$$SMD_j^r = \frac{\bar{X}_j^r - M_j^r}{\sigma_j} \tag{48}$$

To get a sense of the overall balance of the sample for each order, the mean of the SMDs for all variables in the sample will also be calculated:

$$\overline{SMD^r} = \frac{\sum SMD_j^r}{p^r} \tag{49}$$

Where p^r is the number of covariates of order r. SMD_j^r and $\overline{SMD^r}$ will be calculated done for each set of 48 samples.

Generalizability. Sample balance is only 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, the second method of determining generalizability will be to estimate the generalizability index (B; Tipton, 2014a). This index takes into account the distribution similarity of the estimated sampling propensity score, (Equation 18), between the sample and population. 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 units in the population are divided into k bins. For bins $j = 1, ..., k$, let $w_{pj} = N_j/N$ be the proportion of the population and $w_{sj} = n_j/n$ be the proportion of the sample in each bin.

We will calculate B as:

$$B = \sum_{j=1}^{k} \sqrt{w_{pj} w_{sj}} \tag{50}$$

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

$$h = 1.06 s (N + n)^{-1/5} \tag{51}$$

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

$$s^2 = \frac{(n-1)s_s^2 + (N+1)s_p^2}{(N+n-2)} \tag{52}$$

Feasibility. Finally, in order to assess feasibility, the total number of schools approached before a sample of 60 was collected will be kept track of. The average number of refusals each sample method resulted in prior to selecting the full sample will be 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 units. This measure will allow us to compare the difficulty with which a full sample is recruited using each method.

Analysis. The replication results of three sets of outcomes will be plotted and inspected visually. Box-plots will be used to simultaneously judge the mean and variability of the outcomes across replications. Separate plots will be generated for each population and, in the case of the SMD's, each variable and order. Results will be plotted on the Y axis against participation rate conditions on the X axis. Methods used to obtain the results will be color coded. Methods that result in a SMD_j^r closer to zero for each variable and order will have selected a sample that is more balanced on that variable and order. Methods that result in a $\overline{SMD^r}$ closer to zero will have selected a sample that is more balanced on that order overall. Methods that results in a B index on average closer

to zero will have selected a sample that is more generalizable. Such methods will also have reduced coverage errors when retrospective generalization methods are employed for estimating PATE. Methods that on average come across fewer refusals will be less difficult to implement and therefore more feasible.

Discussion

This paper proposes a quantitative examination of CASS, a model which enables purposive sampling for generalizing treatment effects to a larger, well-defined population. Currently, there is only one case study comparing CASS to convenience sampling (Tipton, 2013b) which lacks generalizability. The CASS has several potential advantages over how sample selection is currently performed in large-scale RCTs. First, by carefully selecting a sample rather than relying on post-hoc statistical methods alone, possible coverage errors are reduced, and when necessary, the effectiveness of the post-hoc methods are increased. Second, it provides a structure for the selecting of the sample and forces the researcher to document the process enabling a better justification of the sampling process relative to convenience sampling.

There are, of course, several limitations to the CASS method. This method requires that the target population is well defined and that data exists which reports characteristics of this population. Necessarily, this data must contain variables that predict treatment effect heterogeneity as well as the probability of being sampled. The availability of such data varies across fields, however a general trend towards quality data collection and sharing suggests that this limitation will become less restrictive over time. This method is also fairly resource intensive. It requires recruiters to contact units such as districts and schools that they otherwise not have bothered with (at times rightly so) and to work with methodologists where they would otherwise be fairly independent. If not manged well this can be frustrating for the research team.

Though there are potential benefits, a large goal of the present study is to determine if this method is worth the effort. CASS may outperform convenience sampling across all conditions, and be more feasible than random sampling. However, if it requires 5 times

the phone calls and site visits to recruit unwilling units, it may be difficult to justify from a practical perspective. Knowing the limitations of the method allows researchers to make a more informed decision before allocating valuable resources.

As is the case with any study of a novel methodological application, the present study is only a first step and contains several limitations. The states (populations) were selected for the study primarily because of the availability and completeness of the data they provided. As such, the set of populations is not representative of all possible populations across research fields. Further research may look at finding or generating more extreme examples of heterogeneity across covariates. Missingness may also be examined as many state data sets were incomplete or suppressed information due to avoid sharing identifiable data.

Another caveat would the be impact of omitting a variable from the CASS method. Median income was selected as the omitted variable because in practice it needed to be extracted from a separate database and merged, a task some researchers may avoid. However if this variable is highly correlated with other variables included in the model, then the impact of its omission would be lessened. It would be necessary to experiment with omitting variables with varying degrees of correlation both with other covariates as well as treatment effect and sample selection. It may also be interesting to include a nuisance variable, as researchers often include additional variables that may be unnecessary, or detrimental.

The school and district RGMs are relatively simple linear models with likely unrealistic parameters. It may be worthwhile to explore more complex RGMs which include curvilinear relationships, interactions, multilevel structures, various distributions and more variables. Ultimately, better models for school and district participation are needed to inform this line of research, which, ironically, can be developed via application of CASS. By virtue of this selection framework, well defined units are contacted for participation in studies. If units do not respond, or refuse to participate, this information can be kept and used to perform a non-response/refusal analysis. This helps both the researchers of the individual and other members of the research community better understand differences

between units that participate in research and units that refuse.

The application of the CASS method in this study is fairly limited in scope. Tipton (2013b) gives many recommendations but is careful to also provide alternatives at each step where a methodological decision needs to be made. Given the recency of this method there is limited methodological research comparing these alternatives to the recommendations, or stating when these alternatives may be more appropriate. K-means clustering is recommended for generating strata, however many other alternatives exist including dimension reduction methods (principle components), unsupervised learning, and hierarchical cluster analysis.

Even within k-means clustering there are several options not explored. There is a variety of alternative distance measures which can be used for generating strata and ranking units which may be more appropriate under different conditions. Various weighing methods also exist for calculating the distance measure. It may be helpful to weigh variables differently given their relationship to sample selection. Finally alternative criteria for determining the number of strata to generate may also be explored.

References

Austin, P. C. (2011). An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. *Multivariate Behavioral Research*, 46(3):399–424.

Austin, P. C., Jembere, N., and Chiu, M. (2016). Propensity score matching and complex surveys. *Statistical Methods in Medical Research*.

Brewer, K. R. W. (1999). Design-Based or Prediction-Based Inference? Stratified Random vs Stratified Balanced Sampling. *International Statistical Review / Revue Internationale de Statistique*, 67(1):35–47.

Cham, H. and West, S. G. (2016). Propensity score analysis with missing data. *Psychological Methods*, 21(3):427–445.

Cochran, W. G. (1968). The Effectiveness of Adjustment by Subclassification in Removing Bias in Observational Studies. *Biometrics*, 24(2):295–313.

Cochran, W. G. and Rubin, D. B. (1973). Controlling Bias in Observational Studies: A Review. *SankhyÄĄ: The Indian Journal of Statistics, Series A (1961-2002)*, 35(4):417–446.

Crowe, B. J., Lipkovich, I. A., and Wang, O. (2010). Comparison of several imputation methods for missing baseline data in propensity scores analysis of binary outcome. *Pharmaceutical Statistics*, 9(4):269–279.

D'Agostino, R. B. and Rubin, D. B. (2000). Estimating and Using Propensity Scores with Partially Missing Data. *Journal of the American Statistical Association*, 95(451):749–759.

DuGoff, E. H., Schuler, M., and Stuart, E. A. (2014). Generalizing Observational Study Results: Applying Propensity Score Methods to Complex Surveys. *Health Services Research*, 49(1):284–303.

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.

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

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

Groves, R. M., Fowler, F. J. J., Couper, M. P., Lepkowski, J. M., Singer, E., and Tpiramgeai, R. (2009). *Survey Methodology*, volume 561 of *Wiley Series in Survey Methodology Ser.* John Wiley & Sons, Incorporated, 2 edition.

Ho, D. E., Imai, K., King, G., and Stuart, E. A. (2007). Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference. *Political Analysis*, 15(3):199–236.

Holland, P. W. (1986). Statistics and Causal Inference. *Journal of the American Statistical Association*, 81(396):945–960.

Imai, K., King, G., and Stuart, E. A. (2008). Misunderstandings between experimentalists and observationalists about causal inference. *Journal of the Royal Statistical Society:*

Series A (Statistics in Society), 171(2):481–502.

Imbens, G. W. (2004). Nonparametric Estimation of Average Treatment Effects under Exogeneity: A Review. *The Review of Economics and Statistics*, 86(1):4–29.

Kang, J. D. Y. and Schafer, J. L. (2007). Demystifying Double Robustness: A Comparison of Alternative Strategies for Estimating a Population Mean from Incomplete Data. *Statistical Science*, 22(4):523–539.

Kruskal, W. and Mosteller, F. (1979). Representative Sampling, III: The Current Statistical Literature. *International Statistical Review / Revue Internationale de Statistique*, 47(3):245–265.

Lee, B. K., Lessler, J., and Stuart, E. A. (2010). Improving propensity score weighting using machine learning. *Statistics in Medicine*, 29(3):337–346.

Leite, W. L., Jimenez, F., Kaya, Y., Stapleton, L. M., MacInnes, J. W., and Sandbach, R. (2015). An Evaluation of Weighting Methods Based on Propensity Scores to Reduce Selection Bias in Multilevel Observational Studies. *Multivariate Behavioral Research*, 50(3):265–284.

Li, F., Zaslavsky, A. M., and Landrum, M. B. (2013). Propensity score weighting with multilevel data. *Statistics in Medicine*, 32(19):3373–3387.

Lohr, S. L. (2009). *Sampling: Design and Analysis*. Duxbury Press, Boston, Mass, 2 edition edition.

McCaffrey, D. F., Ridgeway, G., and Morral, A. R. (2004). Propensity Score Estimation With Boosted Regression for Evaluating Causal Effects in Observational Studies. *Psychological Methods*, 9(4):403–425.

Menard, S. W. (2002). *Applied logistic regression analysis*. Number no. 07-106 in Sage university papers. Quantitative applications in the social sciences. Sage Publications, Thousand Oaks, Calif, 2nd ed edition.

Mitra, R. and Reiter, J. P. (2011). Estimating propensity scores with missing covariate data using general location mixture models. *Statistics in Medicine*, 30(6):627–641.

Office, G. A. (2013). Education research: further improvements needed to ensure relevance and assess dissemination efforts. report to the Committee on Education and

the Workforce, House of Representatives, United States Government Accountability Office, Washington, D.C. Report.

O'Muircheartaigh, C. and Hedges, L. V. (2014). Generalizing from unrepresentative experiments: aÂăstratified propensity score approach. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 63(2):195–210.

Qu, Y. and Lipkovich, I. (2009). Propensity score estimation with missing values using a multiple imputation missingness pattern (MIMP) approach. *Statistics in Medicine*, 28(9):1402–1414.

Ridgeway, G., Kovalchik, S. A., Griffin, B. A., and Kabeto, M. U. (2015). Propensity Score Analysis with Survey Weighted Data. *Journal of Causal Inference*, 3(2):237–249.

Roschelle, J., Shechtman, N., Tatar, D., Hegedus, S., Hopkins, B., Empson, S., Knudsen, J., and 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.

Rosenbaum, P. R. and Rubin, D. B. (1983). The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika*, 70(1):41–55.

Rosenbaum, P. R. and Rubin, D. B. (1984). Reducing Bias in Observational Studies Using Subclassification on the Propensity Score. *Journal of the American Statistical Association*, 79(387):516–524.

Royall, R. M. and Herson, J. (1973). Robust Estimation in Finite Populations I. *Journal of the American Statistical Association*, 68(344):880–889.

Rubin, D. B. (1973). The Use of Matched Sampling and Regression Adjustment to Remove Bias in Observational Studies. *Biometrics*, 29(1):185–203.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701.

Rubin, D. B. (1980). Randomization Analysis of Experimental Data: The Fisher Randomization Test Comment. *Journal of the American Statistical Association*, 75(371):591–593.

Rubin, D. B. (1986). Statistics and Causal Inference: Comment: Which Ifs Have Causal

Answers. *Journal of the American Statistical Association*, 81(396):961–962.

Rubin, D. B. and Thomas, N. (1996). Matching Using Estimated Propensity Scores: Relating Theory to Practice. *Biometrics*, 52(1):249–264.

Salkind, N. J., editor (2010). *Encyclopedia of research design*. SAGE Publications, Thousand Oaks, Calif. OCLC: 436031218.

Schafer, J. L. and Kang, J. (2008). Average causal effects from nonrandomized studies: A practical guide and simulated example. *Psychological Methods*, 13(4):279–313.

Shadish, W. R., Cook, T. D., and Campbell, D. T. (2001). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton Mifflin, Boston. OCLC: 49823181.

Smith, T. M. F. and Sugden, R. A. (1988). Sampling and Assignment Mechanisms in Experiments, Surveys and Observational Studies, Correspondent Paper. *International Statistical Review / Revue Internationale de Statistique*, 56(2):165–180.

Stuart, E. A. (2010). Matching Methods for Causal Inference: A Review and a Look Forward. *Statistical Science*, 25(1):1–21.

Stuart, E. A., Bell, S. H., Ebnesajjad, C., Olsen, R. B., and 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.

Stuart, E. A., Cole, S. R., Bradshaw, C. P., and Leaf, P. J. (2011). The use of propensity scores to assess the generalizability of results from randomized trials. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 174(2):369–386.

Stuart, E. A. and Rubin, D. B. (2008). Matching Methods for Causal Inference. In Osborne, J., editor, *Best Practices in QuasiExperimental Designs*, pages 155–176. Sage Publications.

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.

Tipton, E. (2013b). Stratified Sampling Using Cluster Analysis: A Sample Selection Strategy for Improved Generalizations From Experiments. *Evaluation Review*, 37(2):109–

139.

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., and Castilla, V. R. d. (2016a). Site Selection in Experiments: An Assessment of Site Recruitment and Generalizability in Two Scale-up Studies. *Journal of Research on Educational Effectiveness*, 9(1):209–228.

Tipton, E., Hallberg, K., Hedges, L. V., and Chan, W. (2016b). Implications of Small Samples for Generalization: Adjustments and Rules of Thumb. *Evaluation Review*, page 0193841X16655665.

Tipton, E., Hedges, L., Vaden-Kiernan, M., Borman, G., Sullivan, K., and Caverly, S. (2014). Sample Selection in Randomized Experiments: A New Method Using Propensity Score Stratified Sampling. *Journal of Research on Educational Effectiveness*, 7(1):114–135.

Valliant, R., Dorfman, A. H., and Royall, R. M. (2000). *Finite population sampling and inference: a prediction approach*. Wiley series in probability and statistics. Survey methodology section. John Wiley, New York.

Westreich, D., Lessler, J., and Funk, M. J. (2010). Propensity score estimation: neural networks, support vector machines, decision trees (CART), and meta-classifiers as alternatives to logistic regression. *Journal of Clinical Epidemiology*, 63(8):826–833.

Williamson, E., Morley, R., Lucas, A., and Carpenter, J. (2012). Propensity scores: From naive enthusiasm to intuitive understanding. *Statistical Methods in Medical Research*, 21(3):273–293.