Estimating External Validity Bias in RCTs: A Simulation Study

Middlebury College

Trisha Singh

May 15, 2018

Abstract

Many, if not most, randomized controlled trials are conducted on sites that are chosen in a non-random manner. This can bias the impact estimate if the RCT sample is not representative of the population of interest for which researchers make recommendations. If estimates vary across sites and the inclusion of sites in the sample is correlated with site-level treatment impacts, the estimate obtained on the sample will not be externally valid. Using a dataset from an already existing RCT, I model different situations in which researchers oversample sites that are convenient to implement the treatment in and assess the external validity bias that arises from it. I compare this bias to a mathematical expression derived in the literature. I find a very high external validity bias in samples that are conveniently selected. Additionally, I test a reweighting procedure that utilizes the inclusion probability of each site and find that it generally reduces, but does not eliminate, external validity bias.

Acknowledgement

A big thank you to Professor Robbett for advising me through this process (and catching my countless typos!). I am also grateful to Professor Maluccio, Professor Gong, Professor Abel, and Professor Gregg for their extremely helpful feedback. I would also like to thank my peer reviewer and Javier del Cid.

Contents

List of Tables

List of Figures

1 Introduction

Randomized controlled trials (RCTs) are being increasingly used for evaluating the impact of policies in economics. Public policy and economics research organizations primarily use randomized controlled trials to help policymakers make evidence-based decisions. The Abdul Jameel Poverty Action Lab (J-PAL) has conducted 902 randomized evaluations in 79 countries and Innovations for Poverty Action (IPA), another well-known research organization, has conducted 650 evaluations in 51 countries. RCTs are considered to be the gold standard for impact evaluations, with the American Economic Association having 1718 RCTs listed in its registry.

The pervasiveness of RCTs in economics research can be explained by their ability to estimate causal effects when observational data are not available. If the RCT is implemented as planned, its estimate points to the average causal effect of the intervention of interest, unconfounded by unobserved variables. It also allows researchers to tailor their experiment to a specific intervention, while in observational data analysis, the individual components of the intervention may be indistinguishable from other inputs (Banerjee and Duflo, 2009).

Having an internally valid estimate of the policy intervention allows researchers to compare different treatments on the basis of their effectiveness and costs of implementation. For example, in the case of reducing school absenteeism, the conventional knowledge is that educational interventions such as free study materials, meals at school, teacher incentives, etc. will be the most effective. However, the study by Miguel and Kremer (2004) (mentioned later in the paper) provides evidence that deworming treatment for children is effective in reducing primary school absenteeism by 25%, an effect size that is much higher than those of some educational incentives (Banerjee and Duflo, 2009).

However, this example also highlights one of the main drawbacks of RCTs, which is the difficulty of generalizing their findings. The sample on which an RCT is performed is often non-randomly chosen due to budget constraints and drop-out of sites during the selection process. Thus, the sample of sites is representative of a very specific population [1]. For instance, in the Miguel and Kremer (2004) study carried out by J-PAL, the deworming treatment is implemented in two divisions in the Busia district of Kenya. Based on the result that deworming treatment improves attendance by 25%, J-PAL has recommended that the treatment be extended to the entire countries of Ethiopia, India, Kenya, and Vietnam with improved educational outcomes as a benefit of implementing the program. Even though the sample of this study is representative of a certain population, its findings may not be generalizable to larger populations. When deworming treatment was carried out in the slums of New Delhi in India, school attendance was shown to increase by 8%, which amounts to only about a third of the increase in Busia district (Bobonis et al., 2006). While an increase in school attendance is favorable nonetheless, this example showcases the difficulty of deciding the generalizability of an RCT and the importance of defining a population of interest for the RCT sample.

If the RCT is not representative of the population of interest (the population that the findings will be generalized to), it leads to external validity bias just as non-random selection at the individual level leads to a biased estimate. The most common way in which this occurs is the non-random selection of sites in multi-site evaluations. Several researchers have raised concerns about the external validity and scope of RCTs, citing the uncertainty of the impact estimate on the non-randomly selected sites in the sample generalizing to the population of interest (Allcott, 2015; Heckman, 1991; Pritchett, 2002; Rodrik, 2008). Olsen et al. (2013) develop a mathematical expression to show that the impact estimate obtained on non-randomly selected sites can be biased if treatment impacts vary across sites and are correlated with the inclusion of sites in the sample.

The purpose of this paper is to estimate the external validity bias and also adapt the theoretical expression derived by Olsen et al. (2013) to a real dataset using approximations.

[1] In particular, their sample is broadly representative of or externally valid for the population of co-educational primary schools in the Busia district of Kenya. As explained later, I use this study to test external validity bias and assume that the impact estimate obtained on this dataset is the 'gold standard' for the population of interest, which for my purposes I assume is the sample itself.

This will help to determine how heavily the impact estimate can be biased in non-random sampling of sites, and also whether Olsen et al.'s (2013) expression can be applied to policy evaluation RCTs, especially when randomization of treatment is at the site level. The approach involves simulating different site selection mechanisms and estimating the bias for each type of non-random RCT site selection. Additionally, I develop a reweighting solution based on Olsen et al.'s (2013) concept of inclusion probabilities that can help reduce external validity bias.

The paper begins by reviewing research that has been conducted on estimating external validity bias in RCTs in biostatistics and economics in Section 2. In Section 3, I explain the mathematical model developed by Olsen et al. (2013) and outline ways to estimate this bias on a real dataset. In Section 4, I describe the data I use to run my simulations, which are obtained from Miguel and Kremer's (2004) study of the impact of deworming on health and education outcomes. The sites mentioned in this paper are the primary schools in Busia district where Miguel and Kremer implemented their treatment. I present my results in Section 5, followed by a discussion of limitations and future work in Section 6.

To my knowledge, this is one of the few papers to estimate external validity bias, and the only one to do so by conducting simulations of site selection mechanisms. Additionally, this paper is the first to model a range of site selection mechanisms and test a reweighting procedure that attempts to reduce this bias. My findings suggest that external validity bias can be a large concern for RCTs with non-randomly selected sites whose findings are generalized to a larger population. In particular, I find that if most of the sites in my sample are selection based on convenience (reducing logistical costs), it can greatly bias the policy impact estimate. With regards to adapting the Olsen et al. (2013) expression, I find that the expression does tend to predict the direction of external validity bias, but its accuracy varies. Additionally, the reweighting procedure I test, which is inspired by the framework in Olsen et al.'s (2013) paper, generally succeeds in reducing this external validity bias.

2 Literature Review

As RCTs grow in popularity, there is an increasing focus in the literature on the external validity or generalizability of treatment effect estimates from these trials. Generalizability has often been discussed in the context of multi-site clinical trials. For instance, Rothwell (2005) explores external validity bias from a clinical perspective in which the treatment is often a health intervention. However, it provides useful insight on examining external validity of RCTs in general. Rothwell makes a checklist for clinicians containing factors contributing to external validity bias, such as choosing sites that will be easier to implement the trial in (usually those that have a good track record), sites with more resources, geographically convenient sites, etc. He does not estimate this bias, but discusses the importance of making the selection of sites transparent, since site selection can be one of the major determinants of bias.

The factors mentioned by Rothwell (2005) influence what is formally known as 'convenience sampling,' that is, choosing sites that are more convenient to implement the trial in. This describes the common scenario in which researchers oversample sites with certain characteristics in order to minimize costs and make treatment implementation more feasible (Emerson, 2015). It is often found in RCTs with a small number of sites, especially in developing countries. Another determinant of external validity bias is the interaction of site characteristics with treatment effects, which I will further discuss.

Within economics, there is a small but growing body of literature on the causes and magnitude of external validity bias. Banerjee and Duflo (2009), in one of their seminal discussions on RCTs, mention the generalizability of these trials and how the observed treatment effect can depend on the environment. They state that this environment dependence can manifest itself in two ways: individual level heterogeneous treatment effects that are not captured in the model, and unobserved differences in the implementation of the intervention across sites. They propose replication of RCTs as a possible solution to assessing the generalizability of the treatment effect estimate. However, they admit that the pressure to publish novel findings reduces the incentive for multiple replications of the same study.

Vivalt (2015) presents a way to assess generalizability using the existing literature. She examines heterogeneous treatment effects within studies and across studies of the same treatment and outcome. The studies that show similar magnitudes of treatment effect estimates and low individual level heterogeneity are said to be more generalizable. However, a meta-analysis would require a large number of replications of the RCTs in different settings, which would require a long waiting period before a new causal relationship can be externally validated. She also suggests other ways to assess generalizability: specifying a causal chain to model the mechanism through which the treatment affects the outcome, reporting outcomes considered by other studies, or providing more detailed information about the implementation of the treatment.

Allcott (2015) estimates generalizability with data from within the RCT on the Opower energy conservation program. However, Allcott's analysis applies to a very specific context in which there are a number of smaller experiments as a part of a larger experiment to test for energy consumption and the sample grows over time. In this case, the first few sites are chosen according to convenience, and the other less convenient sites are included in the sample as it grows. Allcott finds that the first few experiments show higher treatment effects than the later experiments. Thus, he showed that in the Opower project, if only the more convenient sites are included in the sample, it biases the treatment effect for the population of interest, which is all the energy-consuming households in the United States.

Another way to consider external validity in multi-site RCTs, which I will further develop, is to model the site selection process and adjust the treatment impact estimate for it. In what is to my knowledge the only work attempting to correct for external validity bias, Cole and Stuart (2010) suggest reweighting as a way to standardize observed trial results to the population of interest. They use an HIV treatment trial which is known to be unrepresentative of the population of interest and has a biased treatment impact estimate. They conduct Monte Carlo simulations while applying a generalizability correction to the estimate and find that the new estimate is largely unbiased. However, it involves the strong assumption that the researchers have measured and correctly modeled the determinants of selection that reflect heterogeneity in the treatment effect.

Before conducting further work on correcting external validity bias, it is important to estimate the magnitude of this bias in relation to the treatment impact. Olsen et al. (2013) develop a theoretical expression for external validity bias based on a model of purposive site sampling for multi-site RCTs for policy impact evaluation. Purposive selection refers to nonrandom selection of sites subject to constraints such as cost and whether eligible sites choose to participate. The authors propose that the impact estimate pooled across such sites may not be an estimate for the population of interest. Thus, non-random sampling of sites can lead to external validity bias and the authors develop a theoretical expression for calculating this bias.

The expression for external validity bias depends on the following factors: variance of impacts across sites, the coefficient of variation in inclusion probabilities across sites in the population, and the correlation between site-level impacts and the site inclusion probabilities. Site inclusion probability is the probability that the site is included in the sample in purposive selection.

According to this expression, external validity bias only exists when there is variation in site-level impacts. There is a considerable body of research on cross-site variation of treatment impact estimates. Greenberg et al. (2003) attempt to explain cross-site variation in a specific type of policy impact evaluation, which are welfare-to-work programs. They attribute differences in outcomes across sites to variations in participant characteristics and program implementation. In Chapter 8 of Riccio et al. (1994) the cross-site variation in California's Greater Avenues for Independence Program (another welfare-to-work program) is examined. The existence of this variation deserves attention because if we assume that program implementation is uniform across all sites, the variation in participant characteristics can bias the average treatment effect if the sample is not representative of the population of interest.

The second component of this estimation of external validity bias is site selection modelling. In order to model different components of site selection, it is important to consider several RCTs and how they select sites. For instance, Gleason et al. (2010) estimate the effect of charter schools on students' educational outcomes in the US. It is a fairly large scale RCT, with 36 charter schools chosen across 15 states. They outline the process followed for the selection of these charter schools, namely, charter schools that had been open for at least two years, had relatively stable organization and procedures, and had saturated membership. Within these selected schools, some decided not to participate due to unobserved reasons. The authors state that these charter schools did vary from the population of interest, especially since the participating school served more advantaged students. Kraemer (2000) discusses the problems arising from lack of information or incorrect information about sites in a multi-site evaluation. If the researchers are selecting a small number of sites to be representative of the population of interest, it is possible that they choose sites in such a manner that the sample is skewed away from the population of interest due to misinformation.

Thus the existing literature explains some sources of external validity bias based on theory, intuition, and experience in the field. There has been some work on determining whether the results from a study are generalizable, either by comparing to other studies or examining variation across sites within the study. A more recent body of research is focused on estimating the magnitude of the bias that can arise from the sample not representing the population of interest, as seen in the works of Vivalt (2015) and Allcott (2015). I will extend this body of research and will utilize existing literature to model hypothetical situations of how sites are non-randomly selected, estimate the bias that can arise from this type of sampling, and test a reweighting solution to reduce this bias.

3 Data

Olsen et al. (2013) suggest modelling hypothetical site selection mechanisms in order to estimate the magnitude of external validity bias. I implement this idea by assuming that the sample is the population of interest, and hence the impact estimate is the true population statistic. Then, I choose subsets by modelling different site selection mechanisms that can lead to bias. The estimates obtained on each sample containing a subset of sites from the original dataset can be compared with the true estimate to obtain the external validity bias. This estimation can help to verify the expression provided by Olsen et al. (2013), which I adapt to use on real datasets, or propose methods such as reweighting sites in order to reduce this bias.

To estimate external validity bias using this method, it is important to choose a randomized controlled trial with a large number of sites and a strong impact estimate in order to take subsets that are suitable for comparison. I use the randomized controlled trial conducted by Miguel and Kremer (2004) on the impact of deworming treatment on children's health and education outcomes. This experiment is an evaluation of the Primary School Deworming Project (PSDP) conducted in the Busia district in Kenya from 1998 to 2002. The data are publicly available and were accessed through Harvard Dataverse.

The sample contains 75 rural primary schools in two divisions of the southern Busia district: Budalangi and Funyula. The treatment administered was medical treatment for intestinal helminths (worms) provided by Internationaal Christelijk Steunfonds Africa (ICS), a local non-governmental organization. Randomization was conducted at the school level. Due to limited resources, the 75 schools were divided into three groups of 25 schools that received the treatment at different times. Group 1 was given deworming treatment in 1998 and 1999, Group 2 in 1999, and Group 3 in 2001. Thus the treatment impact on the schools was compared in the manner described in Table 1. This also allows for the estimation of the effect of one or two years of deworming treatment.

For my purposes of estimating external validity bias, I will restrict the sample to schools

Table 1: Comparing Schools across Groups

Year	Treatment schools	Control schools
1998	Group 1	Group 2, Group 3
1999	Group 1, Group 2	Group 3

in group 1 and group 3 and only the year 1998, thus including only 50 schools as sites. There are 21,754 students in the sample. I will focus on the impact of one year of deworming treatment on school participation outcomes. On this sample, Miguel and Kremer find that undergoing one year of deworming treatment increases school participation by about 4.7 percentage points. [2] I will consider this sample as the population of interest, and the impact estimate on this 'population' as the true statistic.

It is important to note that the results from this randomized controlled trial are themselves prone to external validity bias and may not be representative of the population of interest, namely rural schools in Kenya. There are only two divisions under consideration. Only 75 out of 95 schools in the Budalagi and Funyula districts were chosen to participate. Reasons for excluding schools included: socioeconomic disparity among schools (town schools that were more resourced than rural schools were discarded), transportation difficulties, single sex schools, schools that had already received deworming treatment, and newly opened schools. However, for the purposes of this study, I assume that the sample of 50 schools is itself the population of interest and I examine how impact estimates on subsets deviate from the 'true' population estimate.

The dataset also contains useful site characteristics for modelling site selection mechanisms and creating subsets of the existing sites. There is some geographic information about each site such as the zone (ward) in which the school is located and the total number of schools within varying distances of each school. There are 8 zones, with 5 zones in Funyala division (Namoboboto, Nambuku, Bwiri, Agenga/Nanguba, Funyula) and 3 zones in Bunyala division

[2] Participation rate for a child is calculated as the number of times the child was present at school as a percentage of the total number of times school attendance is observed.

Table 2: Distribution of Schools in Different Zones

Zone	Frequency
Agenga/Nanguba	10
Bunyala Central	7
Bunyala North	6
Bunyala South	3
Bwiri	6
Funyula	7
Namboboto	6
Nambuku	5

(Bunyala Central, Bunyala North, Bunyala South). The divisions of Funyula and Bunyala together make up the district of Busia in which the RCT is carried out. The distribution of sites in different zones is shown in Table 2.

The researchers also record whether the school was affiliated with another non-governmental organization project known as the School Assistance Project (SAP), which provides financial assistance for textbook purchase and classroom construction, and teacher performance incentives. There are 15 schools affiliated with the SAP program in this sample (7 in treatment, 8 in control).

Additionally, I use geographic data about the mentioned zones to model site characteristics that are unobserved in the researchers‘ analysis. In particular, I use data on the zone and district boundaries, primary and secondary road networks, and location of nearest NGO office that implemented the treatment. These data are obtained through the Kenya Open Data Initiative (KODI) website and online information about ICL, which is the non-governmental organization that implemented the treatment. [3] The usage of these variables is outlined in the following section.

[3] KODI website: `https://hub.arcgis.com/datasets/66cfcf6d3724405bb15b0099faa46142_0/data`, Zone boudaries obtained from: `https://github.com/mikelmaron/kenya-election-data/blob/master/output/counties.zip`, ICS office location: `https://www.icsafrica-sp.org/contact/`.

4 Methodology

The specification to find the impact estimate on school participation mentioned previously is as follows (Miguel and Kremer, 2004),

$$\text{Participation Rate}_{ijt} = \alpha + \beta_1 T_{1it} + \gamma \text{Ext} + \delta \boldsymbol{X}'_{ijt} + e_{ijt}$$

for student i in school j at time t. β_1 is the impact of undergoing deworming treatment one additional year, γ is the additional positive externality of being located near other treatment schools, and $\boldsymbol{X}'$ is a vector of controls for school and student characteristics.

The goal of this estimation is to model site selection mechanisms that can lead to external validity bias and estimate the magnitude of this potential bias. For each site selection mechanism, I assign inclusion probabilities to each school based on its characteristics. Then, I run Monte Carlo simulations with 100 iterations in order to form samples containing a subset of the sites. The sites chosen in each iteration vary due to the probabilistic nature of site selection. In each iteration, I calculate the external validity bias (EVB) in two ways:

1. Empirically, EVB $= \hat{\Delta} - \Delta$

2. Adapting the expression from Olsen et al., EVB $= \rho_{\Delta P}\ \sigma_{\Delta}\ \text{CV}_P$,

where Δ is the magnitude of the true impact, $\hat{\Delta}$ is the estimate of the impact on the current sample, $\rho_{\Delta P}$ is the correlation between site level impacts Δ_s (where s is a site)and site inclusion probabilities P, σ_{Δ} is the variance in Δ, and CV_P is the coefficient of variation of P for all the sites in the population.

An important contribution of Olsen et al. (2013) is that they conceptualize site selection as a sampling process from a well defined population, in which each site site has an inclusion probability. The inclusion probability is the proportion of times a site s is included in the sample if the sample selection process is repeated infinitely. This is a useful way to model site selection and I build upon this concept to assign each site an inclusion probability based on its characteristics for each scenario. In general, sites that are more convenient based on certain criteria are assigned a higher inclusion probability and sites that are less convenient have a lower inclusion probability. I outline these particular scenarios and how inclusion probabilities are assigned in section 4.1. In addition to modelling site selection, I use the concept of inclusion probabilities to test a reweighting solution for external validity bias, which is described in section 4.3.

4.1 Empirically Estimating External Validity Bias

As discussed in the literature review, a common source of external validity bias is the non-random selection of sites based on convenience. Researchers may select their sites based on location or the ease and cost-effectiveness of carrying out an RCT. In order to model this site selection, I assign all the sites in the population inclusion probabilities based on their characteristics and sample 20 sites from the population of 50 based on these probabilities using Monte Carlo simulations for each site selection scenario. In each of the scenarios outlined below (selecting sites based on affiliation with non-governmental organizations, proximity to other sites, connectivity by roads, and distance to the research center), I assign a threshold such that a site is classified as convenient or not convenient based on each characteristic. This threshold is designed such that there are 20-22 sites above the convenience threshold with an equal number of treatment and control sites.[4] This is done so that there is enough variation in convenient sites in the sample for each iteration and the number of treatment and control sites in the sample is balanced.

Additionally, the proportion of convenient sites in the sample, represented as q_c, is varied for each scenario in order to examine the effect on external validity bias as site selection becomes more convenient or less non-random. In general, q_c starts at .9 and is reduced by steps of .1 as sampling become more random and finally approaches the true proportion of

[4] The only exception is when I model affiliation with non-governmental organizations (NGOs), in which case there are 15 sites over the convenience threshold, since there are only a total of 15 sites affiliated with NGOs in this dataset.

convenient sites in the population (that is, the proportion of sites in the population that lie above the convenience threshold in each scenario). At this point, the site selection process is random since inclusion probabilities of sites above and below the convenience threshold are the same.

Based on the number of sites above the convenience threshold and the proportion of convenient sites in the sample, I assign inclusion probabilities defined in the following manner,

$$p_1 = q_c \, \frac{20}{n_1}$$
$$p_0 = \frac{20(1 - q_c)}{50 - n_1}$$

where p_1 is the inclusion probability for sites above the convenience threshold, p_0 is the inclusion probability for sites below the convenience threshold, q_c is the proportion of convenient sites in the sample, and n_1 is the number of convenient sites in the population. The inclusion probabilities are calculated such that given the above constraints, there are 10 treatment and control sites each that make up the sample of 20 in each iteration.

In the following sections, I outline some convenience sampling mechanisms that I model on the basis of site characteristics.

4.1.1 Affiliation with Non-Governmental Organiations (NGOs)

As shown by most RCTs on the J-PAL and IPA websites, a major source of external validity bias is when researchers choose sites that are already affiliated with NGOs. This makes it easier to implement the RCT in that site and increases the chances of the implementation of the treatment proceeding as intended.

Miguel and Kremer (2004) recorded whether each site is affiliated with the SAP program which provides financial assistance to improve the quality of education. There are 8 control schools and 7 treatment schools affiliated with the SAP project. Although this variable was recorded in order to control for contamination of the treatment effect during the timing of

Table 3: Distribution of Primary Students within 3 km of RCT Site

	Minimum	Median	Mean	Maximum	Standard Deviation
Number of Students	0	1071	1212	3054	722.18

the RCT, I use it as an indicator that the school is affiliated with an NGO.

Since there are only 15 such sites in the sample, that is, $n_1 = 15$, the proportion of convenient sites in the sample q_c begins at .7. Additionally, since there are an unequal number of treatment and control sites above the convenience threshold, the inclusion probabilities are modified as follows,

$$p_1 = q_c \frac{20}{n_1} \quad \text{(remains same)}$$

$$p_0 = \begin{cases} \frac{10 - q_c * n_{1,\text{treatment}}}{25 - n_{1,\text{treatment}}} & \text{if treatment site} \\ \frac{10 - q_c * n_{1,\text{control}}}{25 - n_{1,\text{control}}} & \text{if control site} \end{cases}$$

4.1.2 Proximity to Other Sites

Geographic characteristics of some sites may limit the ability of researchers from accessing them. Additionally, conducting randomized controlled trials in sites that are not easily accessible can be costly. Thus, researchers may choose sites that are closer to each other in order to minimize transportation costs. The existing recorded site characteristics provide the number of primary schools (and consequently the number of primary schools students) close to each site within 3km or 6km distances. The distribution of primary student density is shown in Table 3. Using this distribution, I assign a convenience threshold of 1358. Thus, schools having 1358 or more students within a 3km radius are said to be convenient and assigned a higher inclusion probability, and vice versa for schools having a lower proximity to other sites.

Table 4: Distribution of Main Roads in Zones

Zones	Number of Main Roads
Agenga/Nanguba	25
Bunyala Central	6
Bunyala North	16
Bunyala South	0
Bwiri	9
Funyula	3
Namboboto	3
Nambuku	3

4.1.3 Road Networks

Another way in which researchers are constrained is the availability of transport. The 50 schools in the deworming RCT come from eight wards or zones in the Busia division in Kenya. Using GIS data from OpenStreetMap, I find the connectivity of these wards in terms of the number of main roads in each ward. I assign the convenience threshold to be 9. Thus, zones with 9 or more main roads are said to be convenient and more easily accessible sites and the sites within these zones are assigned a higher inclusion probability and vice versa for less convenient sites. The distribution of main roads in each zone is shown in Table 4. The presence of tertiary roads is highly correlated with the the presence of primary roads such that the convenience threshold defined by taking both primary and tertiary roads into account leads to the same zones being above the convenience threshold. Thus, I restrict my analysis to primary roads.

4.1.4 Distance to Research Center

Another important factor related to transport costs and ease of implementation is the location of the research center of either the organization or an affiliated NGO. In the case of this dataset, I obtained information about the location of the affiliated NGO's office, named ICL, in Busia district and measured the distance of the centroid of each zone to the office.

Table 5: Distance to ICL Office of Zones

Zones	Distance to ICL Office (km)
Agenga/Nanguba	225
Bunyala Central	382
Bunyala North	325
Bunyala South	405
Bwiri	279
Funyula	169
Namboboto	129
Nambuku	129

The convenience threshold is set at about 170 km, thus sites that are in zones whose centroid is 170km or closer to the ICL office are considered more convenient and assigned a higher inclusion probability and vice versa for less convenient sites. The distribution of the distance to the ICL office for each zone is shown in Table 5.

4.2 Predicting External Validity Bias

Calculating the Olsen et al. (2013) expression requires finding the variation in all sites in the population rather than just the sample. Since the estimation of the statistics such as the site-level treatment effect is not possible for sites in the entire population, I approximate them by applying this formula to the available sample. It is also important to note that the model proposed by Olsen et al. (2013) is developed for RCTs in which randomization is at the individual level. However, due to ethical concerns, an increasing number of RCTs are randomized at the site level. Additionally, taking the example of the dataset being used, randomization at the site level allows for estimation of positive externalities of the treatment. Thus, I adapt the estimation strategy so that the expression of external validity bias is applicable. The main problem posed is the inability to know the treatment impact for sites in the control group needed in order to calculate $\rho_{\hat{\Delta}P}$ and $\sigma_{\hat{\Delta}}$. However, propensity score matching can be used to assign a hypothetical outcome after treatment to each of these sites. I use nearest neighbor matching to link control sites to treatment sites on the basis of all observed baseline characteristics, and use the difference between the baseline school participation and the hypothetical school participation rate after receiving treatment as the treatment impact estimate for that site. Due to these factors, the Olsen et al. (2013) expression is heavily approximated and I do not expect its magnitude to closely match the empirically estimated bias, but it would be expected that the expression is able to predict the direction of bias.

4.3 Reducing External Validity Bias

I also test whether the bias that may arise in the above situations can be reduced through a reweighting procedure. I weight each observation by the inverse of inclusion probability of the site that the student belongs to. This paper is the first to use such a method in the context of external validity bias, although the concept of weighting by the inverse probability is a standard procedure to normalize the sample to fit another population. This method is often used in handling missing data and in using propensity scores to estimate treatment effects (Seaman and White, 2013; Mansournia and Altman, 2016).

5 Results

When conducting these simulations, I measured the empirical bias, the predicted bias obtained by adapting the expression from Olsen et al. (2013), and the reduction in bias due to reweighting. For the empirical bias and reduction in bias due to reweighting, I also generated bootstrapped standard errors and bootstrap-t confidence intervals, which are the recommended statistics in this kind of study (Carpenter and Bithell, 2000).

In terms of notation, the proportion of convenient sites in the sample is represented as q_c, sample estimate as $\hat{\Delta}$, and empirical bias as $\hat{\Delta} - \Delta$.

5.1 Empirical and Predicted Bias

First, I calculate empirical bias in order to estimate how high bias values can be due to non-random sampling from the population of interest. Then, I compare the adapted version of Olsen et al.'s expression to see if it can be used as a predictor for bias when it is not possible to directly measure it.

5.1.1 Affiliation with NGO

In this scenario, the true proportion of convenient sites in the population is 0.3. As shown in Table 6 and Figure 1, at high levels of convenience sampling $q_c = 0.5$ onwards, I see that there is external validity bias with a 95% confidence interval of about $-.02$ to $-.01$, which is considerably large given that the treatment impact on the population is about .04. Additionally, most of the confidence intervals (other than at $q_c = .4$) are below 0, showing non-random sampling leads to underestimation of the result. It is important to note that although applying the Olsen et al. (2013) expression shows the correct direction of bias, it does not accurately predict the magnitude.

A possible reason for the underestimation of the treatment impact is that the schools that had already received the benefits of the SAP program may not be as receptive of the treatment. The attendance outcomes in these schools may have reached a plateau due to the success of the previous treatment and may not show as high an effect as other sites in the population would. This is also an important aspect to consider as more and more RCTs are carried out through partnerships with NGOs and often in the same sites. Banerjee and Duflo (2009) also voice this concern about how the same individuals being offered different treatments over time may bias the impact estimate.

It is difficult to determine why the Olsen et al. expression does not yield accurate magnitudes in this scenario other than recognizing the limiting factor that applying the expression to this dataset required a lot of assumptions, including assuming that the variation in the sample can approximate the variation in the population.

5.1.2 Proximity to Other Sites

The true proportion of convenient sites in the population is about 0.42. From Table 7 and Figure 2, we can see that at high levels of convenience sampling, the external validity bias is significantly higher than that at levels close to random sampling. When the proportion of convenient sites in the sample is high, $q_c = 0.5$ onwards, the 95% confidence interval for external validity bias ranges from .02 to .05, which is even larger than the treatment impact on the population. This indicates that sampling sites based on this type of convenience can lead to a very large overestimation of the result. In this case, the Olsen et al. (2013) expression performs well in terms of predicting the magnitude and predicts the direction of bias accurately.

A possible explanation for the overestimation on the sample is that schools that are located close to other schools may have better resources. Thus it is possible that the sites in the sample show a higher treatment impact because they are better able to implement the treatment. This overestimation is not confounded with externality effects since the regression speficiation used to generate this estimate controls for the treatment impacts on other neighboring sites. This is an important result to consider, especially when generalizing to a higher scale in which a significant fraction of the sites sites may be isolated and may not have adequate resources to implement the treatment as planned.

5.1.3 Road Networks

In this scenario, the true population proportion of convenient sites is about .44. From Table 8 and Figure 3, we see again that high levels of convenience sampling lead to high external validity bias with the 95% confidence interval ranging from .02 to .03 at high levels of $q_c = .8$ and $q_c = .9$. The expression of from Olsen et al. (2013) once again performs well in terms of predicting the magnitude and direction of the bias.

The intuition for the overestimation of the treatment impact observed here could be similar to the one provided for the previous scenario in which sites that are more accessible by road may have more resources to implement the treatment properly and see more gains than less accessible sites that may not have as many resources.

5.1.4 Distance to Office

The true population proportion of convenient sites in this scenario is 0.44. Table 9 and Figure 4 show almost a clear trend of increased convenience sampling leading to higher external validity bias. At the highest levels of convenience, $q_c = .8$ and $q_c = .9$, the 95% confidence interval of external validity bias ranges from $-.045$ to $-.035$. The Olsen et al. (2013) expression again predicts the magnitude and direction of bias fairly accurately.

The intuition behind the large underestimation of the treatment effect may be very similar for the one in the scenario of affiliation with NGOs. The sites that are close to the research center may have already been subject to a number of RCTs and may not show a large effect for this new intervention.

Table 6: External Validity Bias when Selecting on the Basis of NGO Affiliation

p_c	$\hat{\Delta}$	Empirical bias			Predicted bias
		$\hat{\Delta} - \Delta$	$\text{SE}_{\hat{\Delta}-\Delta}$	$\text{CI}_{\hat{\Delta}-\Delta}$	
0.7	0.0301*	-0.0166	0.0024	(-0.0213, -0.0119)	-0.0615
0.6	0.0315	-0.0151	0.0023	(-0.0197, -0.0106)	-0.0467
0.5	0.0320*	-0.0147	0.0028	(-0.0201, -0.0093)	-0.0313
0.4	0.0420**	-0.0047	0.0028	(-0.0102, 0.0008)	-0.0153
0.3	0.0396*	-0.0070	0.0035	(-0.0139, -0.0002)	-0.0016

Table 7: External Validity Bias when Selecting on the Basis of Proximity to Other Sites

p_c	$\hat{\Delta}$	Empirical bias			Predicted bias
		$\hat{\Delta} - \Delta$	$\text{SE}_{\hat{\Delta}-\Delta}$	$\text{CI}_{\hat{\Delta}-\Delta}$	
0.9	0.0820**	0.0353	0.0081	(0.0194, 0.0513)	0.0356
0.8	0.0821**	0.0354	0.0065	(0.0227, 0.0481)	0.0279
0.7	0.0804**	0.0337	0.0077	(0.0187, 0.0488)	0.0204
0.6	0.0563*	0.0096	0.0053	(-0.0008, 0.0201)	0.0130
0.5	0.0536**	0.0070	0.0043	(-0.0015, 0.0154)	0.0057
0.4	0.0501*	0.0035	0.0063	(-0.0088, 0.0158)	-0.0014

Table 8: External Validity Bias when Selecting on the Basis of Accessibility by Road

p_c	$\hat{\Delta}$	Empirical bias			Predicted bias
		$\hat{\Delta} - \Delta$	$\mathrm{SE}_{\hat{\Delta}-\Delta}$	$\mathrm{CI}_{\hat{\Delta}-\Delta}$	
0.9	0.0706^{***}	0.0239	0.0026	(0.0189, 0.0289)	-0.0347
0.8	0.0702^{***}	0.0235	0.0027	(0.0182, 0.0288)	-0.0272
0.7	0.0629^{**}	0.0162	0.0031	(0.0100, 0.0224)	-0.0196
0.6	0.0544^{**}	0.0077	0.0038	(0.0003, 0.0151)	-0.0121
0.5	0.0513^{**}	0.0047	0.0029	(-0.0010, 0.0103)	-0.0045
0.4	0.0428^{*}	-0.0039	0.0036	(-0.0108, 0.0031)	0.0030

Table 9: External Validity Bias when Selecting on the Basis of Distance to Office

p_c	$\hat{\Delta}$	Empirical bias			Predicted bias
		$\hat{\Delta} - \Delta$	$\mathrm{SE}_{\hat{\Delta}-\Delta}$	$\mathrm{CI}_{\hat{\Delta}-\Delta}$	
0.9	0.0054	-0.0412	0.0024	(-0.0459, -0.0365)	-0.0417
0.8	0.0077	-0.0390	0.0028	(-0.0445, -0.0334)	-0.0326
0.7	0.0173	-0.0293	0.0028	(-0.0348, -0.0238)	-0.0235
0.6	0.0281^{*}	-0.0186	0.0042	(-0.0269, -0.0103)	-0.0145
0.5	0.0417^{*}	-0.0050	0.0032	(-0.0114, 0.0014)	-0.0054
0.4	0.0482^{**}	0.0015	0.0038	(-0.0060, 0.0090)	0.0362

Figure 1: External Validity Bias when Selecting on the Basis of Affiliation with NGO

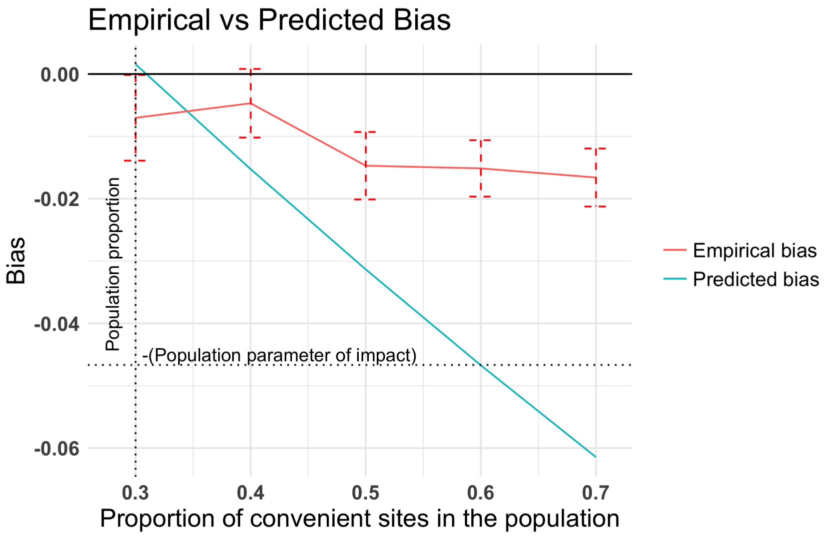


Figure 2: External Validity Bias when Selecting on the Basis of Proximity to Other Sites

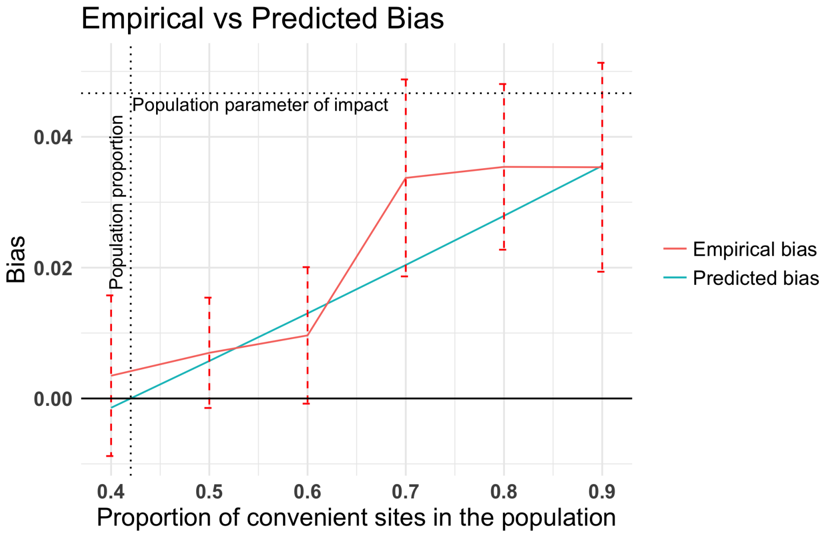


Figure 3: External Validity Bias when Selecting on the Basis of Accessibility by Road

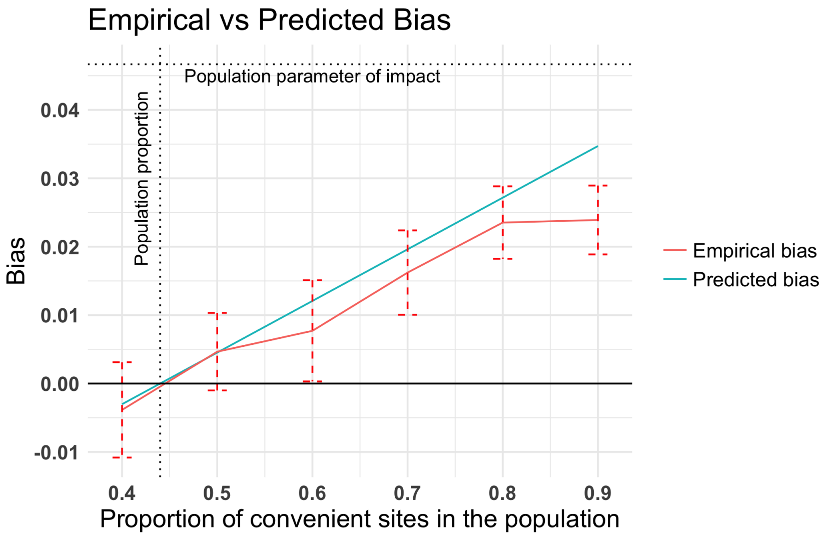
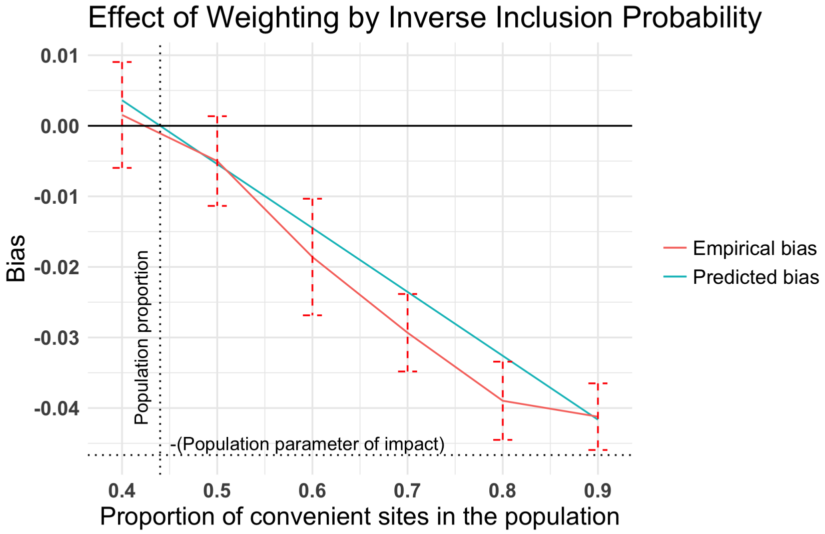


Figure 4: External Validity Bias when Selecting on the Basis of Distance to Office



Thus, the results indicate that if site selection is largely non-random, it is possible to obtain a treatment impact estimate that is heavily biased away from the population statistic. Additionally, the expression proposed by Olsen et al. (2013) is useful for predicting the direction of external validity bias, even though the magnitude of the predicted bias may not be of the same scale as the actual bias as seen in scenario 1.

5.2 Reweighting Solution

From Tables 10-13 and Figures 5-8, we can see the results from applying the reweighting solution to all four scenarios. In general, weighting each observation by the inverse of its site's inclusion probability reduces external validity bias. However, it is difficult to explain the variation in the performance of this reweighting solution.

For instance, in scenarios 2 and 3 (proximity to other sites and accessibility by roads) as seen from Figures 6 and 7, we see that the trend of reduction in bias is very similar. The solution becomes better at reducing bias as site selection becomes more non-random, but its performance drops at $q_c = 0.9$. It is possible to explain the improvement in performance with the increased variation in inclusion probabilities, but it is difficult to explain the behavior at $q_c = 0.9$.

In scenarios 1 and 4 (NGO affiliation and distance to office) as seen from Figures 5 and 8, the performance is almost flipped along the x axis and by taking the confidence intervals into account, we can see that the performance of the reweighting solution remains uniform as sampling becomes more non-random. Without conducting more simulations and comparative statics, it is difficult to determine the reasons for this performance.

However, it is encouraging to see that in all four scenarios and all different values of q_c, the reweighting solution does succeed in reducing the bias and does not ever increase it. Thus, this may be a feasible solution for researchers who are unable to conduct non-random sampling and are able to identify the population of interest and assign inclusion probabilities to their sites.

Table 10: Affiliation with NGO: Effect of Weighting by Inverse Site Inclusion Probability

p_c	$\hat{\Delta}$	Bias after weighting		
		$\hat{\Delta} - \Delta$	$\text{SE}_{\hat{\Delta}-\Delta}$	$\text{CI}_{\hat{\Delta}-\Delta}$
0.7	0.0387*	0.0159	0.0013	(0.0133, 0.0184)
0.6	0.0386*	0.0156	0.0013	(0.0130, 0.0181)
0.5	0.0416*	0.0139	0.0011	(0.0119, 0.0160)
0.4	0.0459*	0.0154	0.0014	(0.0127, 0.0181)
0.3	0.0434*	0.0110	0.0011	(0.0089, 0.0132)

Table 11: Proximity to Other Sites: Effect of Weighting by Inverse Site Inclusion Probability

p_c	$\hat{\Delta}$	Bias after weighting		
		$\hat{\Delta} - \Delta$	$\text{SE}_{\hat{\Delta}-\Delta}$	$\text{CI}_{\hat{\Delta}-\Delta}$
0.9	0.0814**	0.0125	0.0013	(0.0099, 0.0151)
0.8	0.0721*	0.0204	0.0030	(0.0146, 0.0263)
0.7	0.0747**	0.0148	0.0011	(0.0126, 0.0171)
0.6	0.0518***	0.0092	0.0008	(0.0075, 0.0108)
0.5	0.0522**	0.0040	0.0003	(0.0034, 0.0046)
0.4	0.0508*	0.0010	0.0001	(0.0008, 0.0012)

Table 12: Road Accessibility: Effect of Weighting by Inverse Site Inclusion Probability

p_c	$\hat{\Delta}$	Bias after weighting		
		$\hat{\Delta} - \Delta$	$\text{SE}_{\hat{\Delta}-\Delta}$	$\text{CI}_{\hat{\Delta}-\Delta}$
0.9	0.0685**	0.0102	0.0012	(0.0079, 0.0124)
0.8	0.0584*	0.0154	0.0015	(0.0125, 0.0182)
0.7	0.0545**	0.0108	0.0009	(0.0090, 0.0127)
0.6	0.0479**	0.0084	0.0007	(0.0070, 0.0098)
0.5	0.0491**	0.0032	0.0002	(0.0028, 0.0036)
0.4	0.0448*	0.0024	0.0002	(0.0020, 0.0028)

Table 13: Distance to Office: Effect of Weighting by Inverse Site Inclusion Probability

p_c	$\hat{\Delta}$	Bias after weighting		
		$\hat{\Delta} - \Delta$	$\text{SE}_{\hat{\Delta}-\Delta}$	$\text{CI}_{\hat{\Delta}-\Delta}$
0.9	0.0211	0.0214	0.0016	(0.0182, 0.0246)
0.8	0.0261	0.0264	0.0020	(0.0226, 0.0303)
0.7	0.0331*	0.0231	0.0022	(0.0188, 0.0274)
0.6	0.0380*	0.0233	0.0025	(0.0184, 0.0283)
0.5	0.0465**	0.0191	0.0018	(0.0155, 0.0226)
0.4	0.0457**	0.0393	0.0102	(0.0194, 0.0592)

Figure 5: NGO Affiliation: Effect of Weighting by Inverse Site Inclusion Probability

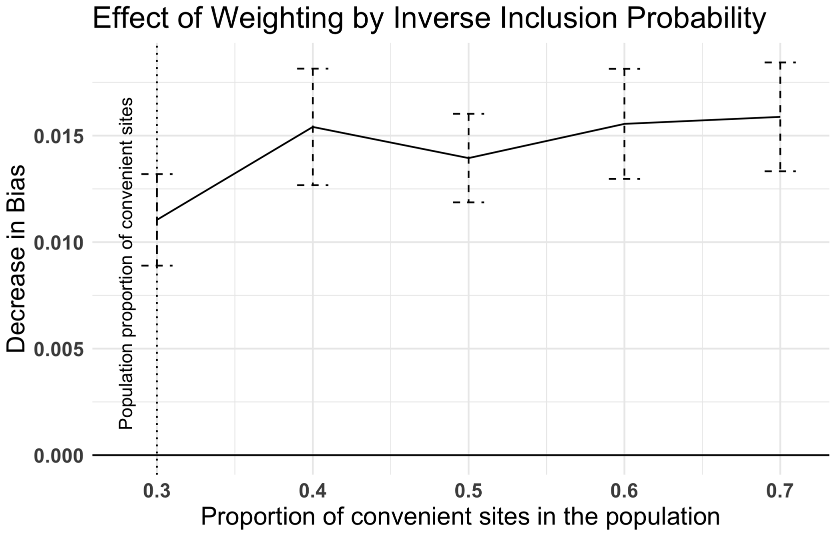


Figure 6: Proximity to Other Sites: Effect of Weighting by Inverse Site Inclusion Probability

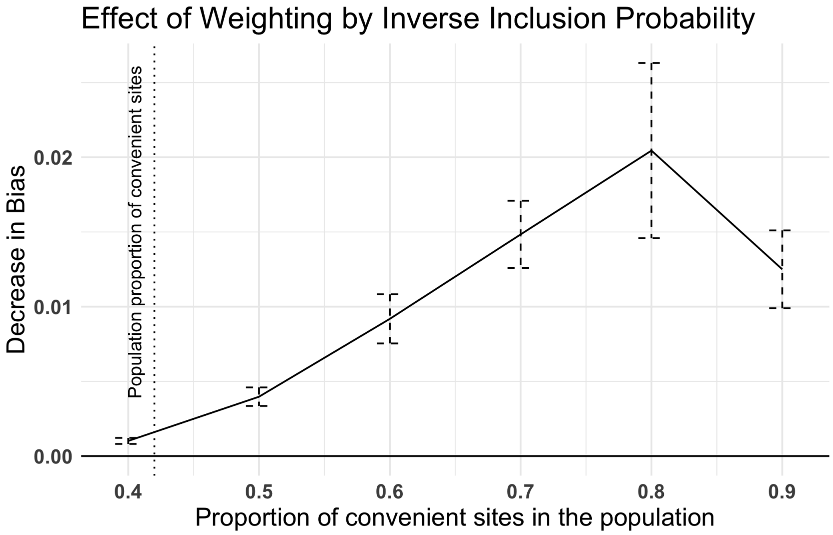


Figure 7: Road Accessibility: Effect of Weighting by Inverse Site Inclusion Probability

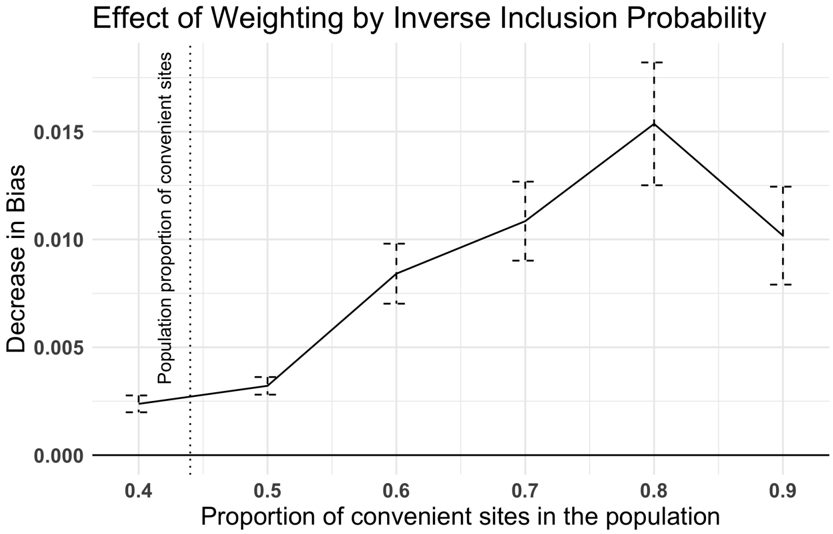
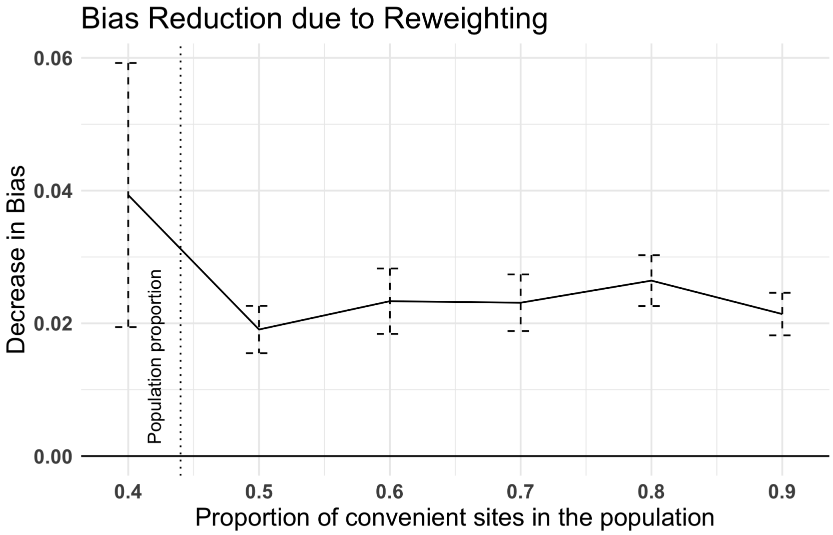


Figure 8: Distance to Office: Effect of Weighting by Inverse Site Inclusion Probability



6 Discussion

My findings suggest that external validity bias in RCTs is a problem that deserves more attention than it is currently paid in the literature. This bias has the potential to be very large and this misinformation can limit the performance of policies that are scaled up on the basis of RCTs. My research contributes to answering the question about the magnitude of external validity bias in different site selection mechanisms, which largely lacks empirical evidence in the current literature. However, it is important to keep in mind that in order to increase the credibility of the findings from this paper, the same analysis should be run on different RCT datasets to see if they yield similar results. It is also important to test the external validity of economic theory rather than just estimates on different samples. Lucas (2003) examines interesting questions about the validity of theoretical constructs, which could help set the framework for further considerations of external validity bias.

A limitation of this paper is in modelling geographic convenience sampling. The lack of GPS coordinates for individual sites makes it difficult to use more granular data, such as the distance of each site to roads, the population density in that area, economic indicators, etc. Researchers should also be encouraged to record the geographic coordinates of each site in order to model their own site selection mechanisms and assign inclusion probabilities to sites. In this manner, they can increase transparency about their population of interest and the way in which samples are selected.

Additional future work on this topic could focus on refining the reweighting procedure. The results from this study show that it is a possible way to reduce external validity bias. In order to better understand this solution, it is important to run different simulations and document its performance. It is possible to try other nonlinear functions of the inclusion probability rather than its inverse to check if it shows a better performance than the method I use.

References

Allcott, H. (2015, August). Site Selection Bias in Program Evaluation. *The Quarterly Journal of Economics 130*(3), 1117–1165.

Banerjee, A. V. and E. Duflo (2009). The Experimental Approach to Development Economics. *Annual Review of Economics 1*(1), 151–178.

Bobonis, G. J., E. Miguel, and C. Puri-Sharma (2006, October). Anemia and School Participation. *Journal of Human Resources XLI*(4), 692–721.

Carpenter, J. and J. Bithell (2000, May). Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians. *Statistics in Medicine 19*(9), 1141–1164.

Cole, S. R. and E. A. Stuart (2010, July). Generalizing Evidence From Randomized Clinical Trials to Target Populations: The ACTG 320 Trial. *American Journal of Epidemiology 172*(1), 107–115.

Emerson, R. W. (2015, April). Convenience Sampling, Random Sampling, and Snowball Sampling: How Does Sampling Affect the Validity of Research? *Journal of Visual Impairment & Blindness (Online); Huntington 109*(2), 164.

Gleason, P., M. Clark, C. C. Tuttle, and E. Dwoyer (2010, June). The Evaluation of Charter School Impacts. Mathematica Policy Research Reports, Mathematica Policy Research.

Greenberg, D., R. Meyer, C. Michalopoulos, and M. Wiseman (2003, August). Explaining variation in the effects of welfare-to-work programs. *Evaluation Review 27*(4), 359–394.

Heckman, J. J. (1991, July). Randomization and Social Policy Evaluation. Working Paper 107, National Bureau of Economic Research.

Kraemer, H. C. (2000). Pitfalls of multisite randomized clinical trials of efficacy and effectiveness. *Schizophrenia Bulletin 3*(26), 533–541.

Lucas, J. W. (2003, September). Theory-Testing, Generalization, and the Problem of External Validity. *Sociological Theory 21*(3), 236–253.

Mansournia, M. A. and D. G. Altman (2016, January). Inverse probability weighting. *BMJ 352*, i189.

Miguel, E. and M. Kremer (2004, December). Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities. *Econometrica 72*(1), 159–217.

Olsen, R. B., L. L. Orr, S. H. Bell, and E. A. Stuart (2013). External Validity in Policy Evaluations that Choose Sites Purposively. *Journal of policy analysis and management : [the journal of the Association for Public Policy Analysis and Management] 32*(1), 107–121.

Pritchett, L. (2002, December). It pays to be ignorant: A simple political economy of rigorous program evaluation. *The Journal of Policy Reform 5*(4), 251–269.

Riccio, J., D. Freedlander, and S. Freedman (1994, September). *GAIN: Benefits, Costs, and Three-Year Impacts of a Welfare-to-Work Program. California's Greater Avenues for Independence Program.* Manpower Demonstration Research Corporation, Three Park Avenue, New York, NY 10016.

Rodrik, D. (2008, October). The New Development Economics: We Shall Experiment, but How Shall We Learn? SSRN Scholarly Paper ID 1296115, Social Science Research Network, Rochester, NY.

Rothwell, P. M. (2005, January). External validity of randomised controlled trials: "To whom do the results of this trial apply?". *The Lancet 365*(9453), 82–93.

Seaman, S. R. and I. R. White (2013, June). Review of inverse probability weighting for dealing with missing data. *Statistical Methods in Medical Research 22*(3), 278–295.

Vivalt, E. (2015, May). Heterogeneous Treatment Effects in Impact Evaluation. *American Economic Review 105*(5), 467–470.