# Sequential Rerandomization in the Context of Small Samples

Jiaxi Yang

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
under the Executive Committee
of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2021

**Abstract**

Sequential Rerandomization in the Context of Small Samples

Jiaxi Yang

Rerandomization (Morgan & Rubin, 2012) is designed for the elimination of covariate imbalance at the design stage of causal inference studies. By improving the covariate balance, rerandomization helps provide more precise and trustworthy estimates (i.e., lower variance) of the average treatment effect (ATE). However, there are only a limited number of studies considering rerandomization strategies or discussing the covariate balance criteria that are observed before conducting the rerandomization procedure. In addition, researchers may find more difficulty in ensuring covariate balance across groups with small-sized samples. Furthermore, researchers conducting experimental design studies in psychology and education fields may not be able to gather data from all subjects simultaneously. Subjects may not arrive at the same time and experiments can hardly wait until the recruitment of all subjects.

As a result, we have presented the following research questions:

1) How does the rerandomization procedure perform when the sample size is small?
2) Are there any other balancing criteria that may work better than the Mahalanobis distance in the context of small samples?
3) How well does the balancing criterion work in a sequential rerandomization design?

Based on the Early Childhood Longitudinal Study, Kindergarten Class, a Monte-Carlo simulation study is presented for finding a better covariate balance criterion with respect to small samples. In this study, the neural network predicting model is used to calculate missing

counterfactuals. Then, to ensure covariate balance in the context of small samples, the rerandomization procedure uses various criteria measuring covariate balance to find the specific criterion for the most precise estimate of sample average treatment effect. Lastly, a relatively good covariate balance criterion is adapted to Zhou et al.'s (2018) sequential rerandomization procedure and we examined its performance.

In this dissertation, we aim to identify the best covariate balance criterion using the rerandomization procedure to determine the most appropriate randomized assignment with respect to small samples. On the use of Bayesian logistic regression with Cauchy prior as the covariate balance criterion, there is a 19% decrease in the root mean square error (RMSE) of the estimated sample average treatment effect compared to pure randomization procedures. Additionally, it is proved to work effectively in sequential rerandomization, thus making a meaningful contribution to the studies of psychology and education. It further enhances the power of hypothesis testing in randomized experimental designs.

## Table of Contents

**List of Tables**

**List of Figures**

**Acknowledgments**

It is never easy to achieve a Ph.D. degree, let alone the 18 months accompany with COVID-19. I could never be grateful enough to those who have always been supporting and encouraging me. It is definitely not easy when you see your former classmates and friends are making progress in their career and life while you are still a *student*. But I would like to choose the same path if given the chance to live once again. It is a lifelong honor and approbation in my learning career.

To my advisor, Professor Bryan Keller, I cannot thank you enough for all your help and advice. I could not have completed my dissertation without your guidance. You taught me how to perfect my thought and framework. You read through my proposal sentence by sentence. You took the trouble to meet with me weekly during the pandemic. You always believe in me for my thoughts and presentation. As your second student, I could see that you have spent lots of effort in being a better advisor. Your criticism, your praise, your support and your kind words gave me the impetus to move forward. He who teaches me for one day is my teacher for life. I hope I could learn more from you in the future and continue our research.

I would also like to thank my committee: Professor James Corter, Professor Peter Bergman, Professor Caleb Miles and Professor Charles Lang. Thank you for your time and advice to my dissertation. Your feedback and comments are extremely helpful. Professor Corter allowed me to join his seminar in the first year of my study, which helped me to step in the field of research swiftly. Professor Lang referred a part-time data scientist job for me and I have been working for the company for 3 years. It really helped me to cumulate

working experience and trained me to program well. I also want to thank Professor Lee, Professor Tipton and Professor Decarlo from Teachers College, thanks for your help and guidance when I was studying in the department.

I would like to thank my first advisor Dr. Matthew Johnson and Mrs. Xiao Yang. It is you who arouse my interest in pursuing a Ph.D. degree and make up my mind. My best friends and comrades, Dr. Rui Lu and Dr. Jiaqing Zhang. I will never forget the days that we have hotpots together whenever we are struggling to graduation. I am so glad that we all come to a satisfying result in the first quarter of our lives. Rui, you are just like my second advisor. Jiaqing, I think you should share the credit of coming up with this dissertation topic.

Most of all, I would like to thank my family, my parents, my grandparents, uncles, aunts and cousins. Your love and support are always there with me. Mom and Dad, thank you for letting me move forward without any hesitation, it is lucky to be your child and I hope I did not let you down. I am trying my best to get you a grandson or granddaughter as soon as possible. You will always be my model of life. My dear cousin Hanbin Li, thank you so much for lending me your server. If it were not for this powerful computer, I would never finish my study smoothly.

Finally, I would like to thank my beloved girlfriend, Xuying. It is lucky to know you in such a hard period of time. Although you were at the other side of the earth, your company lightened my life and supported me through the journey. I cannot imagine the quarantined time without your words and kind concerns. We have walked through a hard but memorable time. I cannot wait to start our life together!

## Introduction

Randomized experimental designs help estimate causal effects, as they generally balance all potential confounding factors, including both observed and unobserved (Krause & Howard, 2003). However, if there is a substantial imbalance on key covariates before conducting the experiment, it could yield bad allocations. Therefore, researchers must rerandomize the subjects instead of carrying out the experiment. Morgan and Rubin (2012) presented theoretical results and concrete structures that support rerandomization. Their rerandomization framework is considered practical, as all subjects and covariates are randomized and balanced simultaneously. They also suggest using the Mahalanobis Distance for measuring covariate balance when deciding whether the assignment is balanced or not on the account of two advantages: (1) Mahalanobis Distance is symmetric in the treatment assignment, resulting in the unbiased estimation of the average treatment effect (ATE) under rerandomization; (2) Mahalanobis Distance is an equal-percent variance that reduces when the covariates are ellipsoidally symmetric, indicating that it reduces the variance of the mean differences of all covariates by the same percentage (Morgan & Rubin, 2012). It is theoretically appealing to balance all covariates in the design stage. However, there are several concerns in real case studies. For instance, researchers may wish to balance the covariates that are considered more important before conducting the experiment and ignore the covariates that are not considered important. Morgan and Rubin (2015) created a rerandomization procedure for treatment-versus-control experiments, consisting of tiers of covariates with varying importance. Researchers conducting experimental design studies in psychology and education fields may not be able to experiment on all subjects

simultaneously. Subjects may not arrive at the same time and experiments can hardly wait until the recruitment of all subjects. To solve this problem, Zhou, Ernst, Morgan, Rubin and Zhang (2018) created a rerandomization scheme for sequential designs.

In rerandomization studies, there are several theoretical contributions such as the use of $2^K$ factorial designs for rerandomization (Branson, Dasgupta & Rubin, 2016), asymptotic results for rerandomization (Li, Ding & Rubin, 2018), and asymptotic results for the combination of regression adjustment and rerandomization (Li & Ding, 2020). However, there are only a limited number of studies adapting rerandomization designs. Additionally, there is not much research on the covariate balance criteria which are defined prior to the rerandomization procedure. Branson and Shao (2018) presented ridge rerandomization that utilizes ridge Mahalanobis distance as the balance criterion to manage the condition that covariates are not ellipsoidally symmetric. Furthermore, discussion and investigation on the performance of rerandomization with small samples are limited. According to Morgan and Rubin (2012), researchers should ensure that the number of acceptable rerandomization is not too small when the sample size is small and the acceptable rerandomized assignment plan could be limited, but more research need to be done.

The main focus of this paper is on the design stage of causal inference studies. It aims to find better covariate balance criteria that can enhance the rerandomization procedure performance when the sample size is small. Also, if a better balancing criterion is found, we will attempt to adapt it to sequential rerandomization to evaluate the performance. The main research questions are:

1) How does the rerandomization procedure perform when the sample size is small?

2) Are there any other balancing criteria that work better than the Mahalanobis distance with respect to small samples?

3) How well does the balancing criterion work in a sequential rerandomization design?

The rest of this dissertation will proceed as follows. Chapter 2 covers a brief introduction of the history of causal inference and the Rubin Causal Model framework with the potential outcome notation. In this chapter, we also decompose the estimation error of the average causal effect and introduce the development of rerandomization along with its procedure and theoretical background and sequential rerandomization. This dissertation primarily focuses on the way to reduce the estimation error due to imbalance treatment via rerandomization. Chapter 3 introduces multiple covariate balance criteria that have previously been shown to effectively function when the sample size is small. Also, we propose a procedure that allows performance comparison of various covariate balance criteria in complete rerandomization as well as sequential rerandomization. This chapter further introduces the simulation study setting. Chapter 4 presents a Monte-Carlo simulation study based on the Early Childhood Longitudinal Study, Kindergarten Class to identify and determine a relatively good covariate balance criterion with respect to small samples. In the simulation study, we use the neural network predicting model to estimate missing counterfactuals. Furthermore, after adapting the selected covariate balance criterion to the sequential rerandomization procedure, we examine the performance. This chapter also covers the study results. Additionally, an analysis is performed to investigate whether rerandomization enhances the power of hypothesis testing in the experiment. Chapter 5 includes the conclusion and limitations of the dissertation. Lastly, Chapter 6 includes the discussion of possible research that might be

extended from this topic. In this chapter, we also discuss some interesting insights that arises from the simulation study and discuss more general cases of experimental design, including multi-treatment or unequal-sized experimental design.

## Literature Review

### 2.1 Notation

In causal inference and research in observational studies and randomized experiments, there is rich statistical literature. These types of studies and experiments particularly require a proper design. In terms of design, researchers can think that any study aimed at estimating the effect of certain interventions has two important stages – design and outcome analysis (Stuart, 2010). In the design stage, researchers use only background information of the subjects without accessing the outcome values. Ideally, they need to equate or balance the distribution of covariates in all treated and control groups. After determining the experiment design, researchers could conduct an outcome analysis through the outcome comparison of subjects in the treated and control groups.

Neyman (1923,1990) invented the first potential outcomes framework in an agricultural experiment. Rubin (1974) then extended the notation and idea of potential outcomes by applying to treatment effects, which became the mainstream approach for causal inference studies. Holland (1986) later labeled the framework as the Rubin Causal Model (RCM).

Consider a randomized experiment assigned with binary treatment. Let $T_i$ be the treatment received by subject *i*. The subjects are either assigned to treatment ($T_i = 1$) or control ($T_i = 0$) group. If subject *i* is assigned to the treatment group ($T_i = 1$) or the control group ($T_i = 0$), outcomes $Y_{i1}$ or $Y_{i0}$ will be observed, respectively. However, only one outcome ($Y_{i1}$ or $Y_{i0}$) can be observed, which is the fundamental problem of causal inference (Rubin, 1978). The missing potential outcomes are also known as counterfactuals (Greenland, Pearl & Robins, 1999). The individual causal effect for each subject is presented as: $ITE_i =$

$Y_{i1} - Y_{i0}$, also known as the Individual Treatment Effect (ITE). Due to the fundamental problem, the Average Treatment Effect (ATE) is of interest in a population of such subjects:

$$ATE = \tau = E(ITE_i) = E(Y_{i1}) - E(Y_{i0}) \quad (2.1)$$

ATE refers to the average value of all ITEs and is generally represented by $\tau$ (Imbens, 2004; Schafer & Kang, 2008). As defined, it is the average difference between the outcome values where the whole population receives and does not receive the treatment. However, it is imperative to note that ATE is not always a theoretically preferred quantity. Heckman, Smith and Clements (1997) pointed out that ATE for the treated is of substantive interest in several policy contexts, as it is unlikely to assign most policies to all individuals. Moreover, researchers would only be interested in whether the policy works to the advantage of individuals who are assigned. For instance, if the treatment $T_i$ could only be applied to a part of the population, ATE may not be meaningful. Winship and Morgan (1999) also indicated that ATE is not separately identified from that of the treated in many cases. Thus, it is important to consider an alternative to ATE, i.e., Average Treatment Effect among the Treated (ATT), represented by $\tau_T$.

$$ATT = \tau_T = E[Y_{i1} \mid T_i = 1] - E[Y_{i0} \mid T_i = 1] \quad (2.2)$$

Similarly, there is also the Average Treatment Effect of the Untreated (ATU), represented by $\tau_U$.

$$ATU = \tau_U = E[Y_{i1} \mid T_i = 0] - E[Y_{i0} \mid T_i = 0] \quad (2.3)$$

If the expected estimator value of a given parameter is equal to its true value, the estimator is considered an unbiased estimator. In a randomized experiment in which the subjects are independently assigned to the treatment or control groups, the assignment $T_i$ is

independent of the potential outcomes. In this case, $\tau$ and $\tau_T$ are equivalent, since the treated subjects would be similarly distributed as the whole population. If all potential outcomes can be observed, the unbiased estimator for $\tau$ and $\tau_T$ would be:

$$\widehat{ATE} = \hat{\tau} = \frac{1}{N}\sum_{i=1}^{N}(Y_{i1} - Y_{i0}) = \frac{1}{N}\sum_{i=1}^{n} Y_{i1} - \frac{1}{N}\sum_{i=1}^{n} Y_{i0} \tag{2.4}$$

And

$$\widehat{ATT} = \hat{\tau}_T = \frac{\sum_i T_i\,(Y_{i1} - Y_{i0})}{\sum_i T_i} \tag{2.5}$$

However, because of the fundamental problem of causal inference (Rubin, 1978), $Y_{i1}$ and $Y_{i0}$ cannot be observed simultaneously for the same subject. As a result, researchers may examine the mean difference between the treatment and control groups. According to Holland (1986), the mean difference between the outcomes of treated and controlled cases is the *prima facie* Treatment Effect (PFE) which is presented as follows:

$$PFE = E[Y_{i1} \mid T_i = 1] - E[Y_{i0} \mid T_i = 0] \tag{2.6}$$

The estimator of PFE is the difference between the means of the observed group, which is:

$$\widehat{PFE} = \frac{\sum_i T_i\, Y_{i1}}{\sum_i T_i} - \frac{\sum_i (1 - T_i)\, Y_{i0}}{\sum_i (1 - T_i)} \tag{2.7}$$

In a randomized experiment in which the subjects are randomly assigned to the treatment or control groups, the assignment $T_i$ is independent of the potential outcomes. Hence, the expected value of $Y_{i1}$ and $Y_{i0}$ would be:

$$E[Y_{i1}] = E[Y_{i1} \mid T_i = 1]\ and\ E[Y_{i0}] = E[Y_{i0} \mid T_i = 0] \tag{2.8}$$

Therefore, the *prima facie* estimator remains unbiased when the treatment is assigned randomly. Even when the treatment is not assigned randomly, it could be possible to develop

an unbiased estimator given that some assumptions are met (Rosenbaum & Rubin, 1983).

Two assumptions are required to identify the ATEs. First, the strong ignorability assumption (Rosenbaum & Rubin, 1983) requires the random assignment of subjects with the same pre-treatment covariates matrix $X = (X_1, X_2, \dots\dots, X_p)^T$ to the treatment or control groups. Specifically:

$$\{Y_{i1}, Y_{i0}\} \perp T_i \mid X \tag{2.9}$$

$$0 < P(T_i \mid X) < 1 \tag{2.10}$$

Practically, it only holds when all confounders associated with both treatment assignment and potential outcomes are observed and the reliability is measured (Steine, Cook & Shadish, 2011). The empirical distributions of the propensity scores of the treatment group overlap with that of the control group.

Second, the stable unit treatment value assumption (SUTVA) specifies that each subject will experience the same treatment and the values of their potential outcomes are independent of the pattern of assignment in *T* (Rubin, 1980). This assumption, in general, is usually a part of the matching and stratification framework and simplifies the estimation of treatment effect. Upon the violation of SUTVA, various subjects may get access to different treatments and it will vary the potential outcome values (Imbens & Rubin, 2015). SUTVA may sometimes be violated by the interactions between subjects (Rubin, 1990). Hong and Raudenbush (2006) presented a relaxed version of SUTVA that enables the potential outcomes to rely on both current and past schools along with class assignments.

**2.2 Estimation Bias of Causal Effect Estimators**

Randomized Controlled Trials (RCT) are regarded as the 'Gold Standard' for causal

inference studies (IES, 2003). They validate a balance between treatment and control groups on all observed and unobserved variables in expectation. When assessing observational data, several methods can be used to remove bias in estimating ATE, including regression estimators, matching estimators, propensity score methods, and the combinations of methods. As this literature review particularly focuses on the design stage of causal studies, it will only discuss the methods and techniques used to remove bias in estimating ATE in the design stage. The possible options include simple random treatment assignment; random treatment assignment within a block; matching after data collection. All these designs and treatments are targeted to reduce the bias and the variance of causal effect estimation. Several ways are also available to examine causal inference, such as approaches focusing on instrumental variables, regression discontinuity designs, randomized experiments, and non-randomized quasi-experiments. This paper will only focus on the assignment of random treatment.

Consider a sample with $n$ units randomly drawn from a population of $N$ units, where $N \gg n$. For convenience, set *n* as an even number, enabling the sizes of both treatment and control groups to be $n/2$. There are two potential outcomes, $Y_{i1}$ and $Y_{i0}$, for each unit based on all covariates *X*. They represent the fixed values of the outcome variable when the treatment is given or not. The unobserved treatment effect for subject $i$ is presented as:

$$ITE_i \equiv Y_{i1} - Y_{i0} \tag{2.11}$$

Here, $ITE_i$ is likely to be a function of both observed $X_i$ and unobserved $U_i$ in the sample (Imai, King & Stuart, 2008). However, only the covariates $X_i$, not $U_i$, are observable. Thus, it might not be a good option to directly estimate the $ITE_i$ from the observed covariates in most cases, as the unobserved $U_i$ may also contribute to the potential

outcomes. Practically, researchers would prefer estimating either based on the samples of observations or the population rather than $ITE_i$.

If the treatment effect estimation is based on the sample average, Sample Average Treatment Effect (SATE) is generated:

$$SATE \equiv \frac{1}{n} \sum_{i \in \{I_i = 1\}} ITE_i \,, \tag{2.12}$$

where $I_i$ indicates the selection of a subject from the population or not. If $I_i = 1$, the subject is in our observational study, while it is not included when $I_i = 0$.

If the treatment effect estimation of the population is based on population average, Population Average Treatment Effect (PATE) is generated: (Imbens, 2004)

$$PATE \equiv \frac{1}{N} \sum_{i=1}^{N} ITE_i \tag{2.13}$$

Researchers may get SATE for certain research questions of a specific study. However, for the majority of research, PATE would be the ultimate goal, as it unveils the truth across the whole population. To estimate SATE, we could simply use the estimator:

$$\widehat{SATE} = D \equiv \frac{1}{n/2} \sum_{i \in \{I_i = 1, T_i = 1\}} Y_i - \frac{1}{n/2} \sum_{i \in \{I_i = 1, T_i = 0\}} Y_i \,, \tag{2.14}$$

where $T_i$ indicates whether the subject is treated or not.

Based on our data, we can define the estimation error $\Delta$ as:

$$\Delta \equiv PATE - D \tag{2.15}$$

Imai, King, and Stuart (2008) decompose $\Delta$ into two parts. The estimation error particularly results from two reasons in the design stage, including sample selection and treatment imbalance.

$$\Delta = \Delta_S + \Delta_T = \Delta_{S_X} + \Delta_{S_U} + \Delta_{T_X} + \Delta_{T_U} \quad (2.16)$$

Here, $\Delta_S$ and $\Delta_T$ represent sample selection error and bias due to treatment imbalance that could be decomposed into observed ($X$) and unobserved ($U$) covariates respectively. Under various forms of estimation error, Imai, King, and Stuart (2008) further decompose the components. These forms include measurement error, missing data, post-treatment bias, and a lack of compliance with treatment assignments.

### *2.2.1 Bias from Sample Selection*

Sample selection bias, a type of systematic bias, results from the selection of a non-random sample. On the non-random assignment of subjects to treatments, PFE will not be an unbiased estimator of ATE or ATT.

We can present the sample selection error as:

$$\Delta_S \equiv PATE - SATE = \frac{N-n}{N}(NATE - SATE), \quad (2.17)$$

where NATE is the non-sample average treatment effect and presented as:

$$NATE \equiv \sum_{i \in \{I_i = 0\}} \frac{ITE_i}{N-n} \quad (2.18)$$

The formula indicated that we are adapting the same method to the subjects in the population but not in the sample. As a result, we established various methods to eliminate the estimation error due to selection bias.

1) We choose the entire population as the sample
2) Sample ATE and non-sample ATE are equal
3) To ensure that SATE is equivalent to PATE, we find a sample that is equivalent to our population of interest

Notably, when there is a constant treatment effect over the subjects, SATE would be equal to NATE. However, in the context of heterogeneous treatment effects, random sampling would only ensure the absence of sample selection bias rather than the absence of sample selection error, i.e., $E(\Delta_S) = 0$ (Imai, King & Stuart, 2008). If the empirical distributions of observed covariates *X* in the sample and population are assumed to be identical, the estimation error from the observed covariates *X* is likely to disappear. Similarly, the estimation error from the unobserved covariates *U* would disappear if the empirical distributions of all unobserved covariates *U* between the sample and population are identical.

Various options are available for reducing the estimation error due to sampling bias. One of the options is finding a representative sample from the population. Generalizability indices such as propensity score difference (Stuart, Cole, Bradshaw & Leaf, 2011), standardized mean difference (Stuart et al., 2011), and Tipton's Beta (Tipton, 2014) indicate how a sample is representative of the population. If the generalizability indices for the sample fulfil some threshold based on the rule of thumb, the selected sample represents a particular population of interest. Alternatively, randomly sample from the population repeatedly and obtain the empirical distribution of the estimator. It reduces the systematic bias due to sampling bias.

### *2.2.2 Bias from Imbalanced Treatment*

From formulas (2.15), (2.16) and (2.17), we could derive the following:

$$\Delta_T \equiv SATE - D \tag{2.19}$$

Similarly, if the treatment and control groups are assumed to be perfectly balanced, i.e., the empirical distributions for observed covariates *X* and unobserved covariates *U* of both groups are identical, then the bias due to covariate imbalance across both groups would

disappear (Imai, King & Stuart, 2008). By measuring covariate balance between these groups, researchers could determine whether the assumption for observed covariates *X* is met, whereas it is not possible to adjust unobserved covariates *U*. Researchers must consequently attempt to achieve optimal randomization for the treatment assignment to remove the estimation error due to imbalance treatment.

## 2.3 Rerandomization

As discussed, imbalance in treatment assignment may result in the estimation error of treatment effect. Rerandomization gives more accurate estimates of treatment effect through the improvement of covariate balance between treatment and control groups (Morgan & Rubin, 2012).

### *2.3.1 History of Rerandomization*

Considering an experiment with *k* independent covariates and $\alpha$ level of significance, the probability of at least one covariate showing a notable difference between treatment and control groups would be $1-(1-\alpha)^k$. For instance, for 10 covariates and a 5% significance level, the probability would be more than 40%. Fisher (1926) stated, "Most experiments on carrying out a random assignment of plots will be shocked to find how far from equally the plots distribute themselves". Consequently, pure randomization may not aid researchers in completely eliminating estimation bias. Rubin (2008) describes his conversation with his Ph.D. advisor Bill Cochran on rerandomization. Cochran stated that if the selected randomized allocation showed significant imbalance on a prognostically essential baseline covariate and inefficient blocking in a randomized experiment due to numerous baseline covariates, Fisher would suggest rerandomizing the subjects before initiating the experiment.

Rosenberger and Sverdlov (2008) suggested a covariate-adaptive randomization method, as they recognized that pure randomized trials are likely to cause substantial baseline covariate imbalances. Additionally, Rubin (2008) suggested that if propensity score diagnostics demonstrate the presence of crucial imbalances in some experiment assignments, researchers must rerandomize and repeat the procedure until satisfied while recording the reasons for discarding certain randomizations. Worrall (2010) also mentioned the need to rerandomize clinical trials to prevent the occurrence of baseline imbalances. However, it is only possible for observed covariates.

Certain research provides several reasons not to rerandomize. As the rerandomization lowers the true standard error to change the distribution of test statistics, it may lead to relatively conservative inferences. Some may also consider rerandomization as an unnecessary procedure in the context of large samples. With an increase in the sample size, the law of large numbers will result in a better balance between the groups. Discussions are still in progress on whether researchers should purposefully design balanced assignment or employ pure randomization.

Several researchers employed rerandomization in their papers (Urbach, 1985; Imai, King & Stuart, 2008; Bruhn & McKenzie, 2009). On the contrary, Morgan and Rubin (2012) became the first researchers to establish a theoretical framework of rerandomization. They presented a rerandomization procedure that ensure covariate balance, in addition to providing the benefits of randomization. The proposed rerandomization procedure has undergone several changes and transformations. Subsequently, they proposed another rerandomization approach that balances covariates within tiers according to the importance of covariates. Li,

Ding, and Rubin (2018) investigated the asymptotic properties of rerandomization and applied them to experiments with non-Gaussian distributed covariates. In contrast to complete randomization, rerandomization lowers the asymptotic quantile ranges of the difference-in-means estimator, which enables a more precise treatment effect estimation.

### *2.3.2 The Rerandomization Procedure*

Morgan and Rubin (2012) present a detailed procedure of rerandomization:

1) Collect covariate data of the population and sample
2) Specify a particular covariate balance criterion to determine the acceptability of a randomization
3) Randomly assign the subjects to treatment and control groups
4) Check the balance criterion and if met, proceed to step 5; otherwise, return to step 3
5) Perform the experiment using the randomization assignment obtained in step 4
6) Assess the results with a randomization test, but only retain the simulated randomizations that meet the balance criterion specified in step 2

Morgan and Rubin (2012) assume that the sample collection has been completed and focus more on the rerandomization procedure. They recommend using the Mahalanobis distance for the criterion specified in step 2. It is aimed at calculating the mean differences between covariates in treatment and control groups. Working with mean differences enables us to depend on the Central Limit Theorem and assume normality for reasonable sample sizes. It further allows the calculation for covariate balance to be tractable, interpretable, and comparable (Morgan, 2011).

Assume a matrix $X$ to be an $n \times k$ matrix, where $k$ is covariates and $n$ is subjects in the

experiment. Let $\bar{X}_T - \bar{X}_C$ be the *k*-dimensional vector of the mean differences between covariates in treatment and control groups. Then, we have:

$$\bar{X}_T - \bar{X}_C = \frac{W^T X}{np_w} - \frac{(1-W)^T X}{n(1-p_w)} = \frac{X^T(W-p_w)}{np_w(1-p_w)}, \tag{2.20}$$

where $p_w$ is the proportion of subjects in treatment group and *W* is the random assignment vector.

$$p_w \equiv \frac{\sum_{i=1}^{n} W_i}{n} \tag{2.21}$$

The Mahalanobis distance is presented as:

$$M \equiv (\bar{X}_T - \bar{X}_C)^T [cov(\bar{X}_T - \bar{X}_C)]^{-1} (\bar{X}_T - \bar{X}_C) \tag{2.22}$$

From formulas (2.25) and (2.27), we assume $p_w$ to be fixed in a randomized experiment,

$$M = np_w(1-p_w)(\bar{X}_T - \bar{X}_C)^T cov(X)^{-1} (\bar{X}_T - \bar{X}_C) \tag{2.23}$$

In an experimental design, as we know *n*, $p_w$, and the covariate matrix *X*, we only need to find out the assignment vector *W*. If $cov(X)$ is singular, i.e., if the number of covariates need to be balanced is larger than the sample size, $cov(X)^{-1}$ can be replaced by the pseudo-inverse of $cov(X)$ (Morgan & Rubin, 2012).

With the procedure and selected covariate balance criterion, we should set a specific proportion of acceptable randomizations. Taking the Mahalanobis distance *M* as an example, set $p_a$ to be the proportion and any *M* falling in the region which is smaller than *a* would be regarded as an acceptable balanced experimental design:

$$P(M \leq a) = p_a \tag{2.24}$$

For all randomizations, the most balanced $p_a$ will be kept among all rerandomizations. Intuitively, the smaller the $p_a$, the harder it would be to obtain a satisfied randomization, as

the most randomized assignment would be deemed unacceptable. Furthermore, when covariates and the potential outcomes are correlated, rerandomization would raise the precision of estimated treatment effect. It would result in more precise estimates and powerful tests along with narrower confidence intervals (Morgan & Rubin, 2012).

Theoretically, Morgan and Rubin (2012) have demonstrated the lowered amount of variance from the rerandomization procedure. Assume performing a rerandomization which is considered acceptable with $p_w = \frac{1}{2}$ and the covariate means are multivariate normal. Then we have:

$$cov(\bar{X}_T - \bar{X}_C | X, acceptable) = v_a cov(\bar{X}_T - \bar{X}_C | X), \tag{2.25}$$

where

$$v_a \equiv \frac{2}{k} \times \frac{\gamma(k/2+1, a/2)}{\gamma(k/2\,, a/2)} = \frac{P(\chi^2_{k+2} \leq a)}{P(\chi^2_k \leq a)} \tag{2.26}$$

where *k* represents the number of covariates, *a* represents the threshold for rerandomization of covariate balance criterion, and $\gamma$ denotes an incomplete gamma function: $\int_0^c y^{b-1} e^{-y} dy$. Subsequently, the percent reduction in variance is presented as:

$$\begin{aligned} &100\left(\frac{var\left(\bar{X}_{j,T} - \bar{X}_{j,C} \middle| X\right) - var\left(\bar{X}_{j,T} - \bar{X}_{j,C} \middle| X, acceptable\right)}{var\left(\bar{X}_{j,T} - \bar{X}_{j,C} \middle| X\right)}\right) \\ &= 100(1 - v_a) \end{aligned} \tag{2.27}$$

This formula is a function of *k* and $p_a$. With the lower proportion of acceptance and fewer balanced covariates, the percent reduction in variance will be larger (Morgan & Rubin, 2012).

Additionally, if the covariates are correlated with the potential outcomes, the treatment effect is considered additive, then the percent reduction in variance can be defined as:

$$100(1 - v_a)R^2 \tag{2.28}$$

where $R^2$ indicates the squared multiple correlation between *y* and *x* within the treatment group. If the sample size is very small and $\bar{X}_T - \bar{X}_C$ is not normally distributed, formula 2.26 will no longer be valid and we need to define *a* according to the empirical distribution of covariate balance criterion. It is not possible to calculate the reduction of percent in variance, and it can only be accepted as an approximation. Alternatively, we can estimate $v_a$ according to the comparison between pure randomization and rerandomization procedure with all possible simulations (Morgan 2011).

### *2.3.3 Sequential Rerandomization*

In the majority of experimental studies, researchers can always access all subjects before carrying out the experiments. Consequently, it is highly desirable to determine whether researchers could sequentially conduct the rerandomization procedure. Zhou et al. (2018) provide mathematical proof for sequential rerandomization. It shows that sequential rerandomization gains better covariate balance as compared to rerandomization at one time under some assumptions.

Consider a sequential trial with 2*N* subjects and *K* sequential groups containing $2n_1$, $2n_2$,…, $2n_K$ subjects, respectively. As the subjects enter the experiment sequentially, we cannot change the assignments for groups 1 to $t - 1$ at time *t*. Meanwhile, we should consider all covariate balance conditions of earlier assignments during the calculation of the covariate balance criterion.

For the first batch of $2n_1$ subjects, $n_1$ and the rest $n_1$ are randomly assigned to the treatment and control groups, respectively. In contrast to assignment vector *W* in (2.25), there

is a vector $W_1^* = (W_{1,1}^*, \dots\dots, W_{1,2n_1}^*)$, where $W_{1,i}^* = 1$ and $W_{1,i}^* = 0$ represent the *i*th subjects of the first batch assigned to the treatment and control groups, respectively. Therefore, based on (2.28), the Mahalanobis distance corresponding to $W_1^*$ would be defined as:

$$M_1^* = \frac{n_1}{2}\left(\bar{X}_{T,1}^* - \bar{X}_{C,1}^*\right)^T cov(X_1)^{-1}\left(\bar{X}_{T,1}^* - \bar{X}_{C,1}^*\right), \tag{2.29}$$

where $\bar{X}_{T,1}^*$ and $\bar{X}_{C,1}^*$ indicate the mean vectors of treatment and control groups for the first batch of subjects considering steps (3) and (4) of the rerandomization procedure. The subjects will be randomly assigned until we achieve an acceptable Mahalanobis distance, signifying well-balanced covariates in treatment and control groups. After deciding the assignment, we proceed to the second batch of $2n_2$ subjects.

With the ongoing process, for the *k*th group of units, we randomly assign $n_k$ subjects to both treatment and control groups. As mentioned earlier, we cannot change the assignment from previous batches of subjects and need to consider them when estimating covariate balance. Thus, the assignment matrix for the first batch to the (*k-1)*th group is fixed:

$$W_{1:(k-1)} = (W_1^T, \dots\dots, W_{k-1}^T)^T \tag{2.30}$$

Subsequently, we can derive the Malanobis distance for the first *k* groups which is presented as:

$$M_k^* = \frac{n_{1:k}}{2}\left(\bar{X}_{T,1:k}^* - \bar{X}_{C,1:k}^*\right)^T cov(X_{1:k})^{-1}\left(\bar{X}_{T,1:k}^* - \bar{X}_{C,1:k}^*\right), \tag{2.31}$$

where $cov(X_{1:k})$ represents the covariance matrix of $X_{1:k} = (X_1, \dots\dots, X_k)$. This process is repeated until the last group of subjects $2n_K$ is assigned. However, like other researchers, Zhou et al. (2018) do not recommend obtaining a deterministic assignment plan for all subjects by reducing the Mahalanobis distance. For larger size of the sample, the construction

is not practical and the procedure requires some degree of randomness to prevent the selection bias. Thus, researchers must be more careful when formulating the construction.

# Methods

## 3.1 Covariate Balance Criteria

As discussed in the previous chapter, imbalance treatment would result in estimation error. Therefore, we would be interested in how to measure the covariate balance. From Morgan & Rubin's perspectives (2012), the Mahalanobis distance is likely to be a good option for evaluating covariate balance in the rerandomization procedure. In this session, we would discuss the topic in-depth, provide other possible options for measuring covariate balance, and compare their performances in different sizes of the sample.

### *3.1.1 Covariate Balance in Causal Inference*

Based on the definition of balanced treatment, covariate balance criteria is aimed at determining whether the empirical distributions of covariates between treatment and control groups are similar. The selected criteria should have three essential properties. First, they should measure the balance for all covariates to simplify the comparison. Second, their performance should be stable when the sample size is small samples. Third, their performance should be measurable within each criterion, enabling us to access an empirical distribution of the criterion and determine the cut-off point. The criteria are generally divided into three types: logistic regression-based criteria, non-parametric tests, and the Mahalanobis distance, and the Mahalanobis distance is taken as the reference criterion. Traditional methods for measuring covariate balance criteria include standardized bias, t-test, average standardized absolute mean difference (ASAMD), and others. In this paper, we are particularly selecting balancing criteria that have been demonstrated to function effectively with respect to small samples and the criteria should be "comparable", for example, smaller

Mahalanobis distance indicates better balance. We are also measuring balance for multivariate distributions rather than univariate distribution. This paper will not consider the criteria that are intended to measure balance for univariate distribution and the classification permutation test (CPT; Gagnon-Bartsch & Shem-Tov, 2019). This is because the rerandomization will generate an empirical distribution of covariate balance criteria and the procedure is similar to a permutation test. Furthermore, we will present a hypothetical approach to measure covariate balance according to Super Learner (Van der Laan, Polley, & Hubbard, 2007).

#### 3.1.1.1 Logistic Regression-Based Tests

***Logistic Regression Test***. According to Rosenbaum and Rubin (1983), the balancing score $b(X)$ is a function of the observed covariates *X*, ensuring the conditional distribution of *X* given by $b(X)$ is the same between subjects in both treatment and control groups. It is presented as:

$$X \perp T \mid b(X), \tag{3.1}$$

where the propensity score $e(X)$ indicates the coarsest balancing score (Rosenbaum & Rubin, 1983). It is the probability of a subject receiving treatment rather than control given to observed covariates. The logistic regression of the assignment variable is often estimated to use on the observed covariates. It is presented as:

$$e(X) = P(T = 1|X) \tag{3.2}$$

Intuitively, if an observational study is balanced on the observed covariates, these covariates should not contribute to or predict treatment assignment (Imai, 2005). It indicates that when the propensity score $e(X)$ is conditioned, adjustments for covariate *X* should not

give additional assignment information, i.e.,

$$E\big(T\big|X, e(X)\big) = E(T|e(X)) \tag{3.3}$$

Imai (2005) employed the following procedure for measuring covariate balance. First, for predicting the assignment of each treatment with all covariates and their first-order interactions, he employed a logistic regression. He then carried out a residual deviance test to investigate if these covariates remarkably aid in predicting the assignment of treatment. The advantage of this method includes providing one p-value to balance all covariates rather than one p-value for each covariate. In conclusion, we would achieve balance if covariates do not present any extra information on the treatment assignment.

***Biased-Reduced Logistic Regression***. Logistic regression has an advantage in providing the researchers with a global p-value for an overall balance test on all covariates, in addition to giving access to p-values for each covariate of interest. On the contrary, Hansen and Bower (2008) revealed that the strict requirements of sample size for logistic regression would challenge the use of the method to test covariate balance when the sample size is small. Specifically, the logistic regression method to test covariate balance has a relatively high rate of type I error, indicating that it may reject the null hypothesis of covariate balance between treatment and control groups under random assignment. According to Harrell (2015), the logistic regression model requires about 10 observations per confounder. Kleyman's (2009) bias-reduced logistic regression is considered a preferred option to address this issue. It has a better performance in measuring covariate balance when the sample size is small.

Consider a set of $p$ independent variables $X^T = (x_1, x_2, \dots\dots, x_p)$, a vector of parameters

$\beta$ with length *m*, and a binary response variable *Y*. We employ logistic regression to model the presence of conditional probability of the outcome with the given covariates. The model is presented as:

$$Prob(y_i = 1|x_i, \beta) = \pi_i = \frac{1}{1 + e^{-x_i^T \beta}} \tag{3.4}$$

The likelihood function for logistic regression can be written as:

$$L(\beta|y_i, x_i) = \prod_{i=1}^{n} (\frac{e^{x_i^T \beta}}{1 + e^{x_i^T \beta}})^{y_i} (\frac{1}{1 + e^{x_i^T \beta}})^{1-y_i}$$
$$= \prod_{i=1}^{n} \frac{(e^{x_i^T \beta})^{y_i}}{1 + e^{x_i^T \beta}} \tag{3.5}$$

The maximum likelihood estimates of $\beta$ are the solutions to the score equation $U(\beta_j) = \frac{\partial logL}{\partial \beta_j} = 0, (j = 1,2, \dots \dots, p)$ that appears to be:

$$U(\beta_j) = \sum_{i=1}^{n} [y_i x_{ij} - x_{ij} \frac{e^{x_{ij}^T \beta_j}}{1 + e^{x_{ij}^T \beta_j}}] = \sum_{i=1}^{n} x_{ij}(y_i - \pi_i) = 0 \tag{3.6}$$

The Jeffreys prior is an uninformative prior distribution with a density function proportional to the square root of Fisher Information (Jeffreys, 1946). This prior can be used as a penalty to the maximum likelihood function for logistic regression. Firth (1993) showed that through the modification of the score function with Jeffery's prior, we could remove the first-order term from the asymptotic bias of the maximum likelihood estimates. To lower the occurrence of bias in these estimates with respect to small samples, Firth (1993) recommended employing the Jeffreys prior to penalize the likelihood.

$$L(\beta)^* = L(\beta)|I(\beta)|^{1/2}, \tag{3.7}$$

where $|I(\beta)|^{1/2}$ is Jeffreys' invariant prior.

It has been observed that utilizing the Jeffreys prior to penalize the maximum likelihood

function in small samples is not computationally costly and ensures better performance in small samples than other traditional methods (Kleyman, 2009). For this test, the covariate balance criterion will be the difference in AIC between two models (null model with the only intercept on treatment assignment and full model with all covariates on treatment assignment).

***Bayesian Modeling with Cauchy Prior***. Kleyman (2009) also proposed another logistic-based regression approach for measuring covariate balance. In logistic regression, separation occurs upon the accurate prediction of a response by a predictor or linear combination of predictors. The likelihood converges in this case; however, the estimate of at least one parameter diverges to plus or minus infinity. Moreover, separation mostly occurs when the sample size is small. Gelman, Jakulin, Pittau & Su (2008) presented an adaption of iteratively weighted least squares algorithm for estimating logistic regression coefficients with independent *t* prior distributions. When applied to sparse data, the algorithm produces relatively stable estimates (Gelman et al, 2008). It involves two steps: (1) standardize each input variable based on the scale of the selected prior distribution. (2) Gelman et al. (2008) suggested using student *t* family of distributions with mean zero, degrees of freedom which is equal to one, and a scale parameter of 2.5. It corresponds to a Cauchy distribution with center zero and a scale parameter of 2.5. We assigned these Cauchy priors to each coefficient in the logistic. The constant term, on the other hand, is assigned to a Cauchy distribution with center zero and a scale parameter of 10.

Kleyman (2009) adapted this Bayesian-based logistic regression for estimating covariate balance and used likelihood ratio test statistic between two models as the covariate balance

criterion (null model with the only intercept on treatment assignment and full model with all covariates on treatment assignment).

**3.1.1.2 Non-parametric Tests**.

***Cross-Match Test***. Rosenbaum (2005) suggested a non-parametric test, also known as the Cross-Match test, to investigate covariate balance between two multivariate distributions. We will correspond the subjects to non-overlapping pairs without taking the treatment condition into account. With the covariate information of the subjects, we will compute a distance matrix for each pair and select the set of pairs with the smallest sum of statistical distances. The test statistic for Cross-Match test can be demonstrated as the number of matched pairs containing one subject each from the treatment and control groups. With an increase in the test statistic, the covariates between groups will be more balanced.

Consider the total sample size *N* to be an even number and the size of treatment group *n*. Correspondingly, there will be non-overlapping pairs presented by $I = N/2$. Let $A_k$ be the number of pairs with *k* treated subjects, where *k* is equal to 0,1,2. We have:

$$A_0 + A_1 + A_2 = I \tag{3.8}$$

And,

$$A_1 + 2A_2 = n \tag{3.9}$$

Subsequently, the number of treatment assignments with exactly $A_1$ pairs that contains one treated and one controlled subjects will be:

$$2^{A_1} I!/A_0!\, A_1!\, A_2! \tag{3.10}$$

The null distribution for $A_1$ is given by:

$$\Pr(A_1 = a_1) = \frac{2^{a_1} I!}{\binom{N}{n} a_0! \, a_1! \, a_2!} \tag{3.11}$$

The null hypothesis of the same distribution of treatments and controls will be rejected, if $A_1$ is too small (Rosenbaum, 2005).

***Hansen-Bower Test***. Hansen and Bower (2008) recommended using the probability distribution fitting the hypothetical study as a standard to examine balance in the actual study. Randomization inference is employed to examine how the matched samples from treatment and control groups are close to a block-randomized design in terms of covariate distribution (Hansen & Bower, 2008). The researchers used precision-weighted averages of differences on each covariate within the sets of matched samples to measure balance of the covariate distributions between the treatment and control groups. These comparisons will yield permutation tests and then combine with $\chi^2$ statistics comparing this study to a randomized experiment. The null hypothesis of Hansen-Bower test is that the data can be compared to a blocked randomized experiment.

### 3.2 Procedure of Rerandomization Experimental Design

To reduce the estimation error in the design stage, our randomized design procedure will attempt to reduce the estimation error resulting from selection bias as well as treatment imbalance. Consequently, the experimental design will mainly involve two stages – sampling from population and assigning to treatment and control groups. The paper particularly focuses on the assignment stage.

In addition to focusing on the rerandomization procedure with small samples, this paper will discuss the sampling procedure in the design stage. To remove the systematic estimation

error due to sample selection, we will randomly select 10,000 samples with varying sample sizes from the population that will lower the effect of finite sample bias.

After selecting the samples from the population, we will randomly assign the subjects for each sample to the treatment and control groups evenly. Here, we will refer to Morgan and Rubin's (2012) framework of rerandomization.

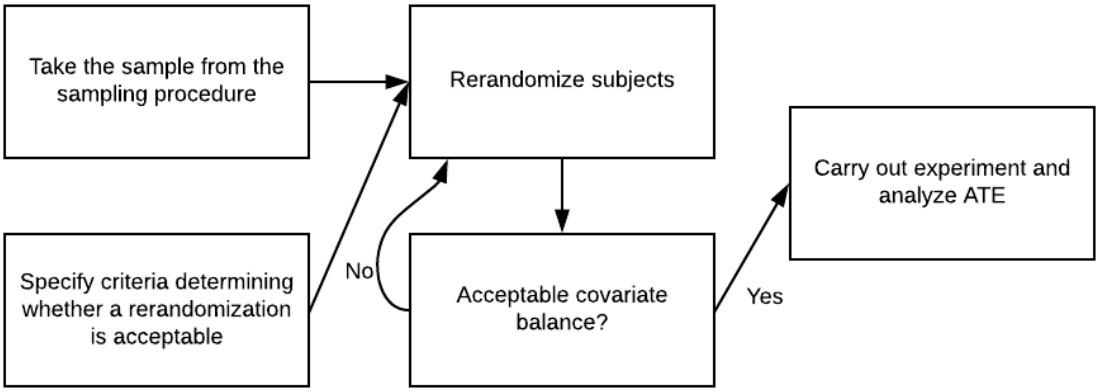



Figure 3. 1 Rerandomization Design

We will initially carry out a comparison between rerandomization and pure randomization procedure to determine the rerandomization performance. Secondly, we will examine various covariate balance criteria with respect to small samples to identify and determine those with relatively good performance. Lastly, considering the selected criterion, we will carry out sequential rerandomization to find out whether the estimation of ATE becomes more precise and stable.

### 3.3 Monte-Carlo Simulation Design

Using data from the Early Childhood Longitudinal Study, Kindergarten Class of 1998-1999 (ECLS-K) as a basis for simulation, we evaluate the performance of various covariate balance criteria. This observational study primarily aims to investigate the effect of exposure

to special education services on academic and social outcomes in fifth grade. Morgan, Frisco, Farkas & Hibel (2010) originally described the ECLS-K data consisting of 34 pretreatment covariates that include the information of demographic, academic, school composition, family context, health, and parent rating of Child. All 34 covariates will be balanced simultaneously during the rerandomization procedure. For the outcome variable, we are interested in the ECLS-K revised item response theory (IRT) scaled math achievement test score in 2004 (fifth grade of the subjects). The exposure depends on determining if the students obtained any special education services in grades K-4. We removed the data with missing covariate information. Subsequently, among the total of 7362 observations, only 429 subjects obtained special education services. To produce complete data for the simulation, we will impute the missing counterfactuals of all students with the use of predicted values from a flexible neural network model fit to all covariates and the exposure variable. In the presence of both potential outcomes, we can calculate the true ITEs and PATEs and use them to evaluate the balance criteria performance. Based on the aforementioned sampling procedure, the random selection of 10,000 samples from the ECLS-K data will be followed by the rerandomization. This process is called Monte-Carlo Simulation.

During simulation, a neural network model will be adapted in the process of data generation. In this study, we will employ neural network models for predicting potential outcomes. The missing counterfactuals for all subjects will be measured with all covariates and the intervention variable and their potential outcomes will be generated. Sample size is an important factor of the simulation design, as it focuses on the performance of covariate balance criteria with respect to small samples. We will randomly select 10,000 samples with

sizes of 20, 30, 40, 50, 100, and 200 from the ECLS-K dataset. After selecting the sample, it sample will be fixed during the rerandomization procedure. Subsequently, we will analyze various covariate balance criteria in the procedure according to different sample sizes. The rerandomization procedure for each covariate balance criterion will undergo simulation of 400 times. We will retain only the top 10 simulations with the best balancing scores which will help us in deciding the cut-off point for covariate balance criterion $a$, i.e., $p_a = 2.5\%$. Given a total sample size of 20, there are in total $\binom{20}{10} = 184756$ possible assignments for equal-sized design. Therefore, it is unrealistic to provide all possible assignments. However, for a relatively smaller sample such as 10, there are in total $\binom{10}{5} = 252$ possible assignments. In this case, we can reduce the number of rerandomization to some extent. For each rerandomization assignment, there will be the same sizes for both treatment and control groups, indicating that $p_w = 50\%$. Subsequently, for each sample, we will consider one acceptable rerandomized assignment and calculate the estimated SATE and average estimated SATE according to the available potential outcomes. As we know the response surfaces of all subjects, the true PATE can be measured. The estimation error due to treatment imbalance $\Delta_T \equiv SATE - \widehat{SATE}$ is thus accessible. This paper will consider the standard deviation of the estimated SATE and estimation bias (expected estimation error) along with the root mean square error (RMSE) of the estimation error as the measurement of the covariate balance criteria performance.

The following list presents a complete procedure of the simulation study:

1) Generate missing counterfactuals for all subjects' revised IRT scaled math score in ECLS-K data using neural network model.

2) Randomly select 10,000 samples from the data.

3) Specify a certain covariate balance criterion and rerandomize the sample given for 400 times to determine the empirical distribution of covariate balance criterion.

4) Set the proportion of acceptance to 2.5%, consider 10 assignments with relatively good balance score, and select the cut-off point of covariate balance criterion for acceptable assignments.

5) Consider one acceptable rerandomization for each sample and measure the average estimated SATE according to the potential outcomes.

6) To evaluate the covariate balance criterion performance, compare true SATE (PATE) with average estimated SATE. Take estimation bias, standard deviation of estimated SATE, and RMSE of the estimation error as the measurement of the covariate balance criteria performance.

$$SATE = \frac{1}{n}\sum_{i=1}^{n} ITE_i$$

$$\widehat{SATE} = \frac{1}{n/2}\sum_{i\in\{T_i=1\}} Y_i - \frac{1}{n/2}\sum_{i\in\{T_i=0\}} Y_i$$

$$Bias = E[SATE - \widehat{SATE}]$$

$$SD = \sqrt{Var(\widehat{SATE})} = \sqrt{E[(\widehat{SATE} - E[\widehat{SATE}])^2]}$$

$$RMSE = \sqrt{Bias^2 + Var(\widehat{SATE})} \qquad (3.12)$$

### 3.4 Sequential Rerandomization Based on Small Samples

After selecting the relatively good criterion from the procedure presented in 3.2, we will adapt it to the experiment design of sequential rerandomization. The Mahalanobis distance will be consequently adapted as the reference criterion to the procedure. In the sequential

design, this criterion will remain the same for each selected sample with sizes of 20, 30, 40, 50, and 100. However, we will only have access to a small batch of the sample initially, which is different from rerandomization design. After assigning this batch, we can access and assign another batch of subjects (cf., Section 2.3.3). These non-overlapping batches will ultimately become a partition of the whole sample correspondingly. This procedure imitates the real situation for several experiments in the field of psychology and education where the subjects take the experiments sequentially and not uniformly. It is worth noting that we can rerandomized the 'newcomers'; however, for previous subjects, we cannot modify the existing assignment plan.

To evaluate the sequential rerandomization performance, this paper will assume that subjects are coming in the batches of 4, 8, and 16 at a time. Each batch of subjects have an even number, as it will simply equal-sized assignment. Morgan and Rubin (2012) recommended equal-sized rerandomization, as it simplifies simulation and balance measurement. For some batches, an unequal-sized assignment plan may be more balanced than s equal-sized assignment plan, considering the information of all covariates and previous assignments. However, researchers may find difficulty in making an overall equal-sized assignment plan for the whole sample with unequal-sized assignment within each batch. To compare this assignment with the complete rerandomization design, we will only conduct equal-sized assignment plans and observe the results.

The equal-sized sequential rerandomization procedure is presented as:

1) Randomly assign a series of ordered indices from 1 to *n* to each subject in the sample, where *n* denotes the sample size. This indicates the subjects' order of

attending the experiment.

2) Take a batch of 4 in the sample with 100 subjects as an example, carry out the rerandomization procedure for the first batch of 4 subjects, and consider the assignment plan with the best covariate balance as an initial group. Here, the numbers of treated and controlled subjects must not be equal.
3) Repeat step 2 with the second batch of four subjects without changing the assignment for the previous batch.
4) Repeat step 3 until the 25th batch of four subjects.
5) Estimate the SATE according to the equal-sized sequential rerandomization and compare its performance with pure sequential randomization procedure.

# Results

Referring to earlier discussion, we outlined the methods for measuring covariate balance and the model for imputing missing counterfactuals. This chapter will first explain the ECLS-K data and the final cleaned-data population for this simulation study. Based on the proposed procedure, we will calculate covariate balance with various types of criteria to determine a relatively good one with respect to small samples. Lastly, the best criteria will be adapted to the sequential rerandomization design to compare its performance with the Mahalanobis distance.

## 4.1 Data Description

We obtained the data from the Early Childhood Longitudinal Study, Kindergarten Class of 1998-1999 (ECLS-K) for evaluating the effect of a special education service in 2002 on math and reading scores of students in 2004. The sample includes more than 21,000 children in Kindergarten from 1,200+ public and private schools across the United States in the base year of 1998-1999 school semester. It also includes the interviews of their parents and teachers (IES, NCES). Morgan et al. (2010) employed 35 pretreatment covariates and four interaction terms to model the propensity score of children receiving the special education services. Keller and Tipton (2016), on the other hand, utilized 34 selected covariates to estimate the propensity scores and give a review of various *R* packages to implement propensity score analysis. In this dissertation, we also used 34 pretreatment covariates of all types of information such as demographic, academic, school composition, family context, health, and parent rating of child. We removed the subjects with missing covariate information and obtained 7,362 anonymized cases in our dataset. Of these cases, only 429

students received special education services in 2002 in their third grade. The selected outcome variable is the ECLS-K revised IRT scaled math achievement test score in 2004, when the students are in their fifth grade. Table 4.1 presents the descriptions of each variable, which include information of variable names, contents of variable and means, and standard deviation of variables if continuous. The table also include the mean and standard deviation for each continuous variable.

Table 4. 1 Variable names, Descriptions, Range of values or Categories, Mean (SD) or Proportions

| Variable Name | Content of Variable | Values | Proportion/ Mean (SD) |
|---|---|---|---|
| **Demographic** | | | |
| GENDER | Male | 0, 1 | 49.67% |
| WKWHITE | White | 0, 1 | 75.05% |
| WKSESL | Socioeconomic Status | [-4.75, 2.75] | 0.17 (0.76) |
| **Academic** | | | |
| RIRT | Kindergarten Reading Score | [23.17, 139.36] | 42.18 (11.52) |
| MIRT | Kindergarten Math Score | [11.85, 98.99] | 33.35 (10.09) |
| S2KPUPRI | Public School | 0, 1 | 78.23% |
| P1EXPECT | Parental Expectations | Integers 1-6 | 4.10 (1.03) |
| P1FIRKDG | First-Time Kindergartener | 0, 1 | 96.66% |
| P1AGEENT | Child's Age at K Entry (Months) | [54, 79] | 65.75 (4.16) |
| apprchT1 | Approaches to Learning Rating | Integers 1-4 | 3.09 (0.64) |
| P1HSEVER | Attended Head Start | 0, 1 | 12.40% |

| | | | |
|---|---|---|---|
| chg14 | Ever Changed Schools | 0, 1 | 5.60% |
| **School Composition** | | | |
| avg_RIRT | Reading IRT | [27.91, 79.98] | 42.18 (5.96) |
| avg_MIRT | Math IRT | [16.06, 66.07] | 33.35 (5.47) |
| avg_SES | Socioeconomic Status | [-2.22, 2.50] | 0.17 (0.51) |
| avg_apprchT1 | Approaches to Learning | [1.5, 4.0] | 3.09 (0.31) |
| S2KMINOR | Percent Minority Students | Integers 1-5 | 2.35 (1.45) |
| **Family Context** | | | |
| P1FSTAMP | Received Food Stamps | 0, 1 | 12.66% |
| ONEPARENT | One-Parent Family | 0, 1 | 17.18% |
| STEPPARENT | Stepparent Family | 0, 1 | 5.70% |
| P1NUMSIB | Number of Siblings | [0, 10] | 1.45 (1.08) |
| P1HMAFB | Mother's Age at First Birth | [12, 45] | 24.77 (5.41) |
| WKCAREPK | Nonparental Pre-K Child Care | 0, 1 | 84.12% |
| **Health** | | | |
| P1EARLY | Number of Days Premature | [0, 112] | 4.28 (11.89) |
| wt_ounce | Birth Weight (Ounces) | [17,214] | 119.50 (20.51) |
| C1FMOTOR | Fine Motor Skills | Integers 0-9 | 6.10 (1.94) |
| C1GMOTOR | Gross Motor Skills | Integers 0-8 | 6.45 (1.77) |
| **Parent Rating of Child** | | | |
| P1HSCALE | Overall Health | Integers 1-5 | 1.60 (0.77) |
| P1SADLON | Sad/Lonely | Integers 1-4 | 1.53 (0.38) |

| | | | |
|---|---|---|---|
| P1IMPULS | Impulsive | Integers 1-4 | 1.90 (0.64) |
| P1ATTENI | Attentive | Integers 1-4 | 1.84 (0.62) |
| P1SOLVE | Problem Solving | Integers 1-4 | 1.70 (0.58) |
| P1PRONOU | Verbal Communication | Integers 1-4 | 1.73 (0.63) |
| P1DISABL | Child has Disability | 0, 1 | 12.78% |
| **Treatment Variable** | | | |
| F5SPECS | Special Education Services | 0, 1 | 5.83% |
| **Outcome Variable** | | | |
| C6R4MSCL | Fifth Grade Math Score | [50.86, 170.66] | 127.07 (23.37) |

### 4.2 Impute Missing Counterfactuals

The proposed procedure requires the missing counterfactuals to be generated according to the information of all covariates as the first step. In this paper, we will use a neural network model to produce the potential outcomes from the group that subjects are not assigned to. As the development of neural network model needs less formal statistical training, researchers only require to determine the variables that might potentially be related to the outcomes. The neural network model can also implicitly identify both linear and complex non-linear relationships between explanatory variables and outcome variable. It can further find every possible interaction between explanatory variables (Tu, 1996). As the data are relatively clean and the pretreatment covariates are assumed to be highly correlated with the academic performance of children, we will fit 34 covariates and the above-listed treatment variable to a neural network model with one or two hidden layers. The whole dataset will be regarded as the population. The steps of imputing steps are provided in the

following:

1) Rescale the whole data using min-max scaling to improve the accuracy and efficiency of the neural network model. For all covariates to range from 0 to 1, the formula is presented as:

$$X_{scaled} = \frac{X - X_{min}}{X_{max} - X_{min}} \tag{4.1}$$

2) Randomly split the population into training set and test set in the proportion of 80%/20%.

3) In the training set, fit 34 covariates and the treatment variable to neural network model with different combinations of layers and neurons.

4) Fit neural network model to the test set and obtain predicted result of outcome variable. Calculate RMSE according to the prediction and original outcomes from the test sets. Formula is presented as:

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}} \tag{4.2}$$

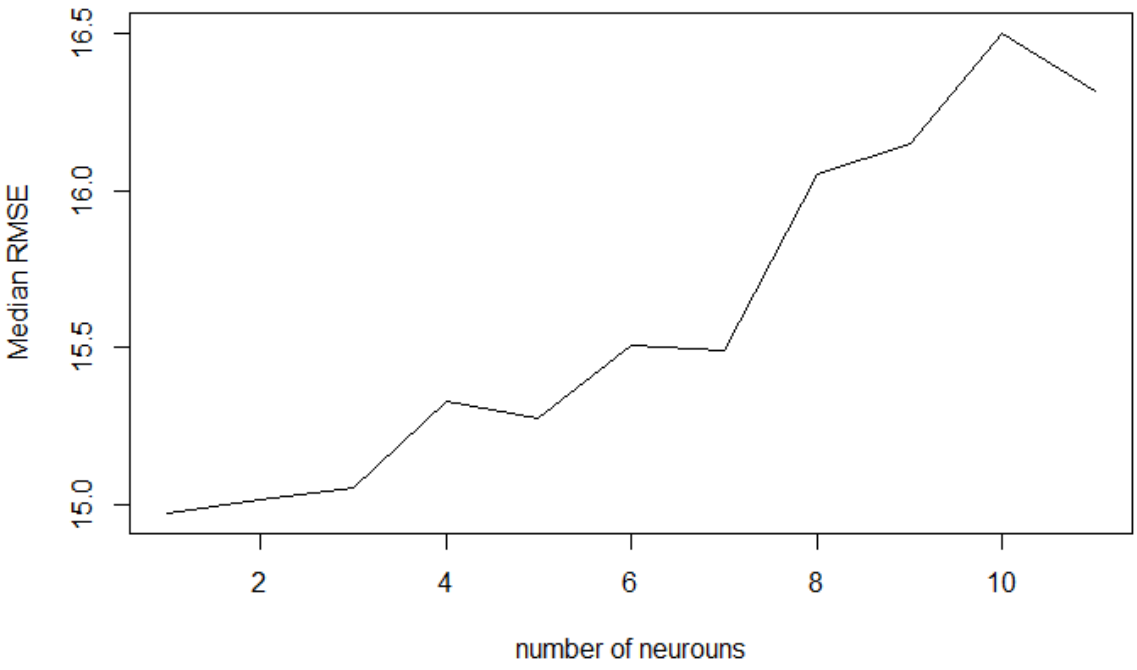
Note that before calculating RMSE, it is important to unscale the predicted outcome.

5) Repeat steps 2 to 4 several times and consider the median of RMSE as the performance criterion for the corresponding neural network model.

For one hidden layer models, we selected 1 to 11 neurons and fitted each model for 20 times repeatedly. Based on the median RMSE line plot (Figure 4.1) for a neural network model with one hidden layer, we could observed that the performance of the model gets worse with an increase in the number of neurons. One hidden layer with a neuron performs better than all 11 models. We can hypothesize that the ECLS-K data does not need numerous

neurons in hidden layers. Consequently, we could experiment smaller number combinations when tuning the number of neurons for models with two hidden layers.

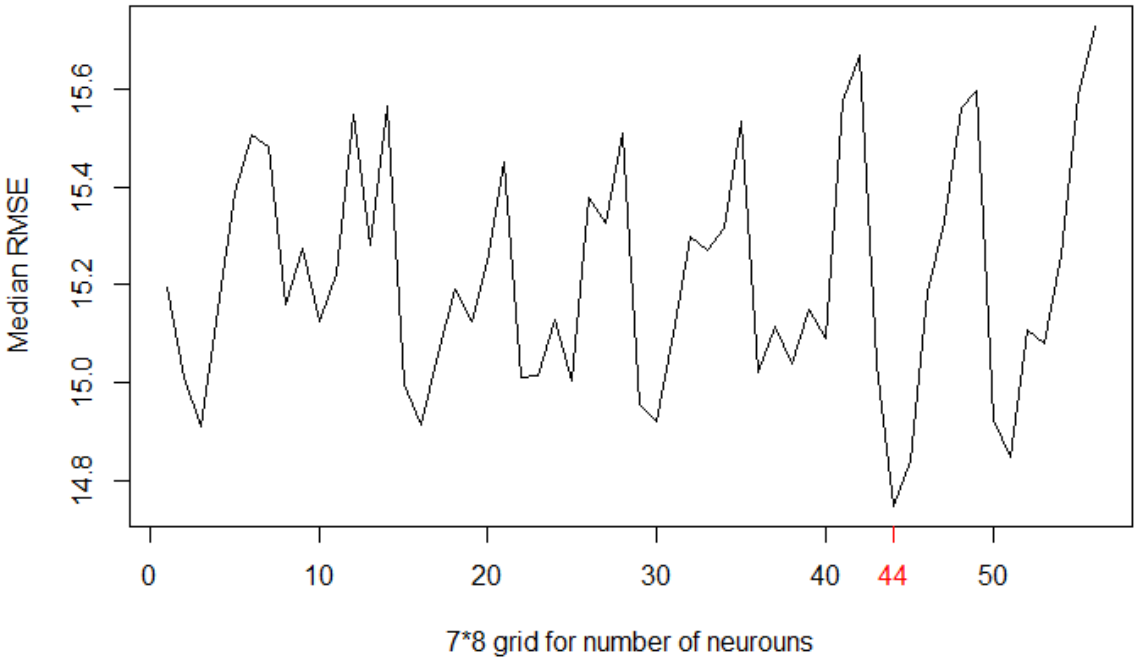
Figure 4. 1 Median RMSE for neural network model with one hidden layer




For two hidden layer models, a 7 by 8 grid for the first and second layers of the neural network model is created with each starting from 1 neuron, respectively and fitted 10 times. Based on the median RMSE line plot (Figure 4.2) for neural network model with two hidden layers, we could observe that the model with 2 and 7 neurons in the first and second hidden layers, respectively, has the best performance. Additionally, it outperforms the single hidden layer neural network model. However, there is no much difference in median RMSE between these two models. Thus, we will not attempt more combinations of neurons and model with three hidden layers. We will treat the neural network model with 2 and 7 neurons in the first and second hidden layers, respectively, as the predicting model, with the learning rate as 0.1

and threshold for the partial derivatives of the error function as 0.1. With sigmoid function as the activation function, the trained neural network model visualization is presented in Appendix.

Figure 4. 2 Median RMSE for neural network model with two hidden layers




We fit the opposite treatment variable to the model to obtain the response surfaces for all subjects in our dataset. The True PATE is -2.477 (14.947). To increase the ATE between treatment and control groups without any changes in the estimate variability, we add 15 to all potential outcomes for the treated, which adjust the True PATE to 12.523 (14.947).

### 4.3 Random Sampling and Measuring Covariate Balance

Our simulation study involves two main sampling strategies. First, randomly select the sample from the population repeatedly and conduct the rerandomization procedure for each sample. Second, randomly select the sample once and perform the rerandomization procedure

but with multiple acceptable rerandomizations. We select the first approach, as it corresponds to real case studies. Also, multiple samples eliminate the estimation error due to selection bias.

To identify and determine the effect of covariate balance criteria on different sizes of sample, 10,000 samples are selected from the population with sizes of 20, 30, 40, 50, 100, and 200. Choosing 10,000 samples for each size of the sample would significantly eliminate the systematic bias due to sampling bias (cf., Section 2.2.1). For each sample, we repeatedly randomize subjects into treatment and control groups for 400 times. Based on the proposed procedure, set the proportion of acceptance to 2.5%, consider the assignments with relatively good balancing score, and select the cut-off point of covariate balance criterion for acceptable assignments. Record all calculated criteria and cut-off values which will determine the distribution for each covariable balance criterion and the cut-off point of all sizes of sample such that for each sample, all acceptable rerandomizations is guaranteed to achieve a better balance among covariates than the cut-off points.

Different criteria might also exist for a single measurement. For logistic regression-based tests, there are three criteria to measure the way to balance the assignment. First, the difference in AIC between full and null models, where full model and null model represent for regressing assignment vector on all 34 covariates and an intercept, respectively. Ideally, if no extra information can be obtained from the covariates, the null model should perform better compared to the full model. Hence, the difference in AIC is defined as the AIC for null model minus the AIC for full model. The assignment will be more balanced with a decrease in the value. The second criterion is the test statistics for likelihood ratio test between full and

null models. With smaller test statistics, we are more likely to accept the null hypothesis that no difference is observed between the full model and the null model, signifying covariate balance between the groups. The last criterion is the area under the curve (AUC) of receiver operating characteristic curve (ROC curve) for predicting the assignment with all covariates. With perfectly balanced assignment, there is no extra information to be obtained from covariates. The predicting should also be close to random guessing, i.e., the AUC should be nearly 0.5. Therefore, if the model predicting the assignment gets worse, the assignment will be more balanced. We have two criteria for Hansen-Bower test and cross-match test: (1) the corresponding test statistic and (2) p-value for the test. In this study, the difference in AIC between full model and null model is considered a criterion for all logistic regression-based tests. We also take Chi-square test statistic and p-value as the criteria for Hansen-Bower test and cross-match test, respectively.

Table 4. 2 RMSE for estimated SATE compared to True SATE and PATE

| RMSE | 20 | | 30 | | 40 | |
|---|---|---|---|---|---|---|
| | SATE | PATE | SATE | PATE | SATE | PATE |
| md | 8.354 | 9.046 | 6.902 | 7.381 | 5.618 | 6.072 |
| log | 8.905 | 9.511 | 7.301 | 7.750 | 5.684 | 6.132 |
| brlog | 8.847 | 9.431 | 7.329 | 7.857 | 5.722 | 6.193 |
| bayeslog | 7.130 | 7.898 | 5.949 | 6.549 | 5.195 | 5.671 |
| hb | 9.063 | 9.625 | 7.454 | 7.930 | 5.612 | 6.107 |
| cm | 8.290 | 8.932 | 6.924 | 7.446 | 5.926 | 6.406 |
| pr | 8.838 | 9.463 | 7.343 | 7.795 | 6.333 | 6.770 |

| RMSE | 50 | | 100 | | 200 | |
|---|---|---|---|---|---|---|
| | SATE | PATE | SATE | PATE | SATE | PATE |
| md | 4.821 | 5.249 | 3.290 | 3.602 | 2.250 | 2.506 |
| log | 5.025 | 5.457 | 3.245 | 3.555 | 2.265 | 2.510 |
| brlog | 4.892 | 5.344 | 3.200 | 3.544 | 2.260 | 2.492 |
| bayeslog | 4.656 | 5.126 | 3.252 | 3.600 | 2.273 | 2.503 |
| hb | 4.861 | 5.305 | 3.291 | 3.607 | 2.270 | 2.504 |
| cm | 5.292 | 5.692 | 3.736 | 4.056 | 2.655 | 2.867 |
| pr | 5.617 | 5.979 | 4.004 | 4.275 | 2.803 | 3.002 |

On the basis of the RMSE results for simulation study, the performances of all rerandomization procedure in estimating both SATE and PATE are better than that of pure randomization. We could thus conclude that rerandomization will increase the precision in estimating sample ATE. In the scale of sample sizes, the differences in performance decrease with an increase in the sample size. Morgan & Rubin (2012) stated that several views suggest that large samples do not require rerandomizations. This is due to a decrease in the difference in covariate means between groups with increasing sample size because of the law of large numbers.

Figure 4. 3 RMSE of SATE for (re)randomization procedures based on different covariate

balance criteria

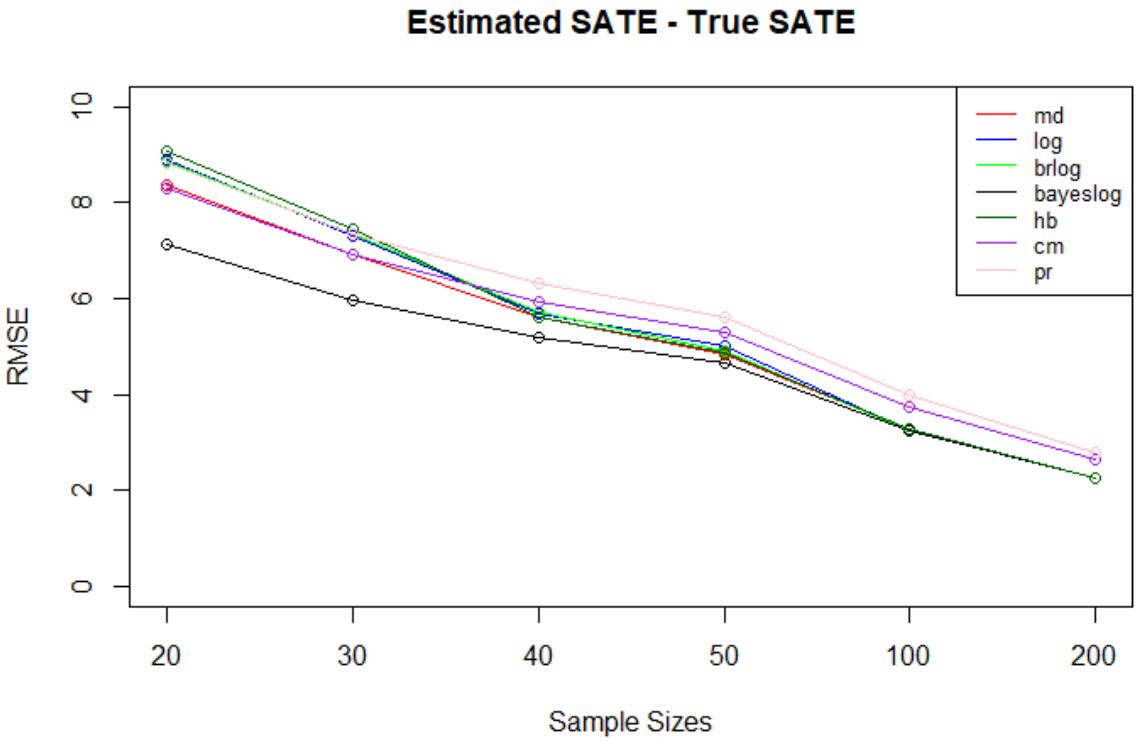
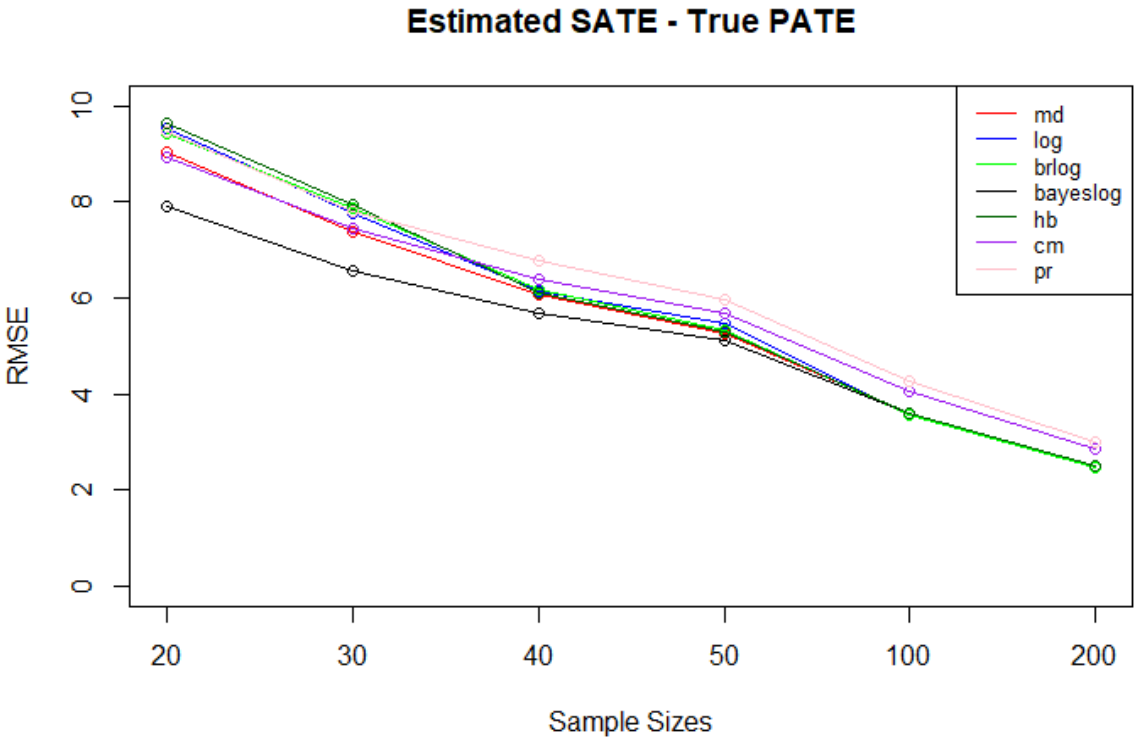



Figure 4. 4 RMSE of PATE for (re)randomization procedures based on different covariate balance criteria



The line plots demonstrate that the Bayesian modeling with Cauchy prior perform the best as the covariate balance criterion for relatively small size of sample. With an increase in the sample size, there is a decrease in the performance difference between criteria and their advantages that are suitable for small samples become insignificant. Upon transforming the reduced amount of RMSE into percentage, a much clearer result can be achieved. We present the formula as:

$$Reduced\ Percentage = \frac{RMSE_{Pure\ Randomization} - RMSE_{Criterion}}{RMSE_{Pure\ Randomization}} \quad (4.3)$$

Table 4. 3 Percentage Reduced of RMSE for estimated SATE compared to True SATE and PATE

| RMSE | 20 | | 30 | | 40 | |
|---|---|---|---|---|---|---|
| | SATE | PATE | SATE | PATE | SATE | PATE |
| md | 5.48% | 4.41% | 6.01% | 5.31% | 11.29% | 10.31% |
| log | -0.76% | -0.51% | 0.57% | 0.58% | 10.25% | 9.42% |
| brlog | -0.10% | 0.34% | 0.19% | -0.80% | 9.65% | 8.52% |
| bayeslog | 19.33% | 16.54% | 18.98% | 15.98% | 17.97% | 16.23% |
| hb | -2.55% | -1.71% | -1.51% | -1.73% | 11.38% | 9.79% |
| cm | 6.20% | 5.61% | 5.71% | 4.48% | 6.43% | 5.38% |
| pr | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% |

| RMSE | 50 | | 100 | | 200 | |
|---|---|---|---|---|---|---|
| | SATE | PATE | SATE | PATE | SATE | PATE |
| md | 14.17% | 12.21% | 17.83% | 15.74% | 19.73% | 16.52% |

| | | | | | | |
|---|---|---|---|---|---|---|
| log | 10.54% | 8.73% | 18.96% | 16.84% | 19.19% | 16.39% |
| brlog | 12.91% | 10.62% | 20.08% | 17.10% | 19.37% | 16.99% |
| bayeslog | 17.11% | 14.27% | 18.78% | 15.79% | 18.91% | 16.62% |
| hb | 13.46% | 11.27% | 17.81% | 15.63% | 19.02% | 16.59% |
| cm | 5.79% | 4.80% | 6.69% | 5.12% | 5.28% | 4.50% |
| pr | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% |

As observed from the table, when the sample size is small, the Bayesian modeling with Cauchy prior outperforms other criteria in significantly lowering RMSE. Negative percentages demonstrate that RMSE increases under the rerandomization procedure with some covariate balance criteria. With a decrease in the proportion of acceptance $p_a$, the performance of lowering RMSE will become more significant. Morgan and Rubin (2012) found that with a decrease in acceptance probability $p_a$, balanced covariates decrease and the percent reduction in variance increases. Our finding is consistent with Morgan's (2011) rerandomization study on the vocabulary training data from Shadish, Clark and Steiner (2008). There are in total 445 subjects in the vocabulary data, 10 covariates are balanced. With Mahalanobis distance as the covariate balance criterion, rerandomization decreases variance by 88% compared to pure randomization design given $p_a = 0.1\%$.

Table 4. 4 Average Bias for estimated SATE compared to True SATE and PATE

| Bias | 20 | | | 30 | | | 40 | | |
|---|---|---|---|---|---|---|---|---|---|
| | SATE | PATE | SYSBIAS | SATE | PATE | SYSBIAS | SATE | PATE | SYSBIAS |
| md | -0.091 | -0.041 | 0.050 | -0.003 | 0.005 | 0.008 | -0.037 | -0.055 | -0.018 |
| log | -0.061 | -0.012 | 0.050 | -0.080 | -0.071 | 0.008 | -0.002 | -0.02 | -0.018 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| brlog | 0.085 | 0.135 | 0.050 | -0.024 | -0.015 | 0.008 | -0.033 | -0.052 | -0.018 |
| bayeslog | 0.012 | 0.062 | 0.050 | -0.013 | -0.005 | 0.008 | 0.020 | 0.002 | -0.018 |
| hb | 0.200 | 0.250 | 0.050 | -0.179 | -0.170 | 0.008 | 0.094 | 0.076 | -0.018 |
| cm | -0.076 | -0.027 | 0.050 | -0.042 | -0.033 | 0.008 | -0.113 | -0.131 | -0.018 |
| pr | 0.020 | 0.070 | 0.050 | -0.120 | -0.111 | 0.008 | -0.001 | -0.019 | -0.018 |

| Bias | 50 | | | 100 | | | 200 | | |
|---|---|---|---|---|---|---|---|---|---|
| | SATE | PATE | SYSBIAS | SATE | PATE | SYSBIAS | SATE | PATE | SYSBIAS |
| md | -0.028 | -0.068 | -0.040 | -0.040 | -0.070 | -0.030 | 0.030 | 0.013 | -0.017 |
| log | 0.068 | 0.029 | -0.040 | 0.024 | -0.006 | -0.030 | 0.010 | -0.006 | -0.017 |
| brlog | 0.033 | -0.007 | -0.040 | -0.047 | -0.077 | -0.030 | 0.018 | 0.001 | -0.017 |
| bayeslog | -0.016 | -0.055 | -0.040 | -0.015 | -0.045 | -0.030 | -0.008 | -0.024 | -0.017 |
| hb | -0.036 | -0.076 | -0.040 | 0.001 | -0.029 | -0.030 | 0.007 | -0.009 | -0.017 |
| cm | 0.044 | 0.004 | -0.040 | -0.031 | -0.062 | -0.030 | -0.002 | -0.019 | -0.017 |
| pr | 0.074 | 0.035 | -0.040 | -0.111 | -0.141 | -0.030 | -0.034 | -0.051 | -0.017 |

However, from the bias table, we could only conclude that the systematic bias due to random sampling is nearly zero. All estimated SATE in rerandomization and pure randomization procedures are the unbiased estimates of either SATE or PATE. Pure randomization is more likely to result in a relatively high bias. However, there are no significant differences as compared to the rerandomization procedures with various covariate balance criteria. For the differences between estimated SATE and true SATE, boxplots may

provide an insight.

Figure 4. 5 Boxplots for SATE-(SATE)ˆ based on 10,000 samples

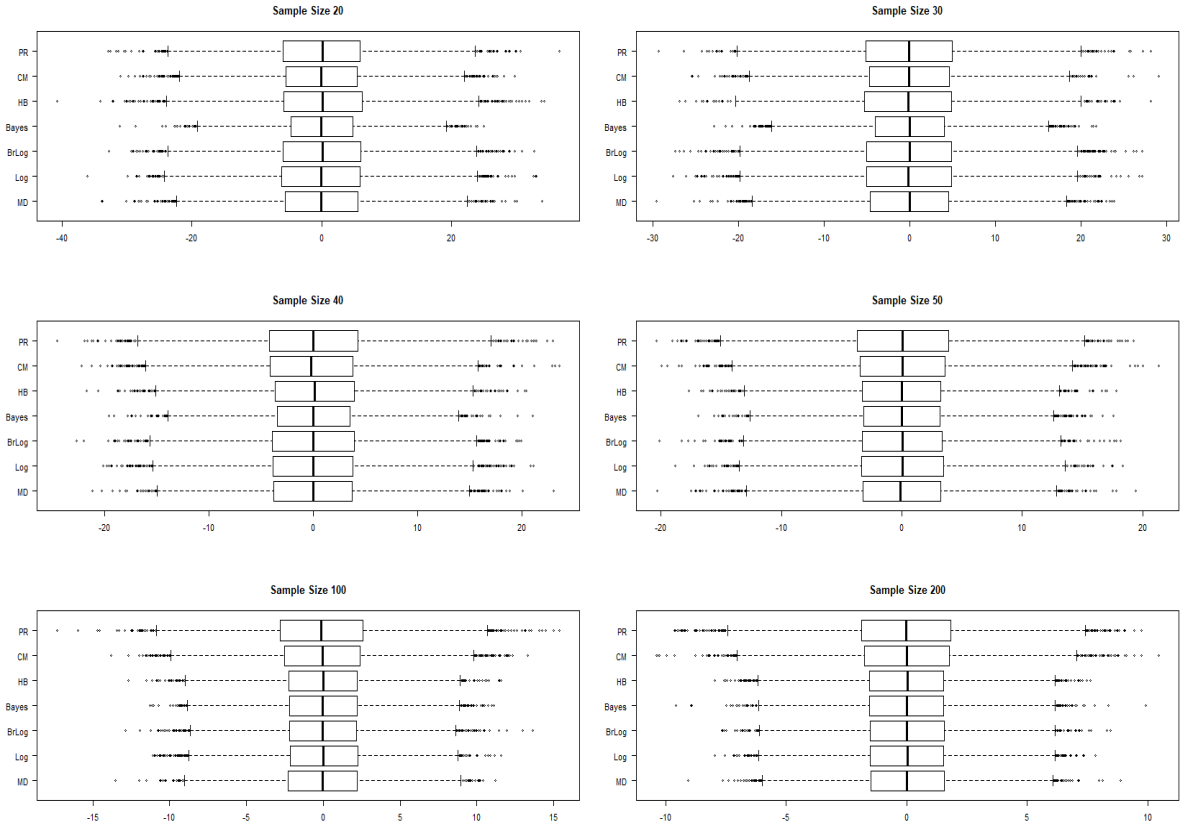
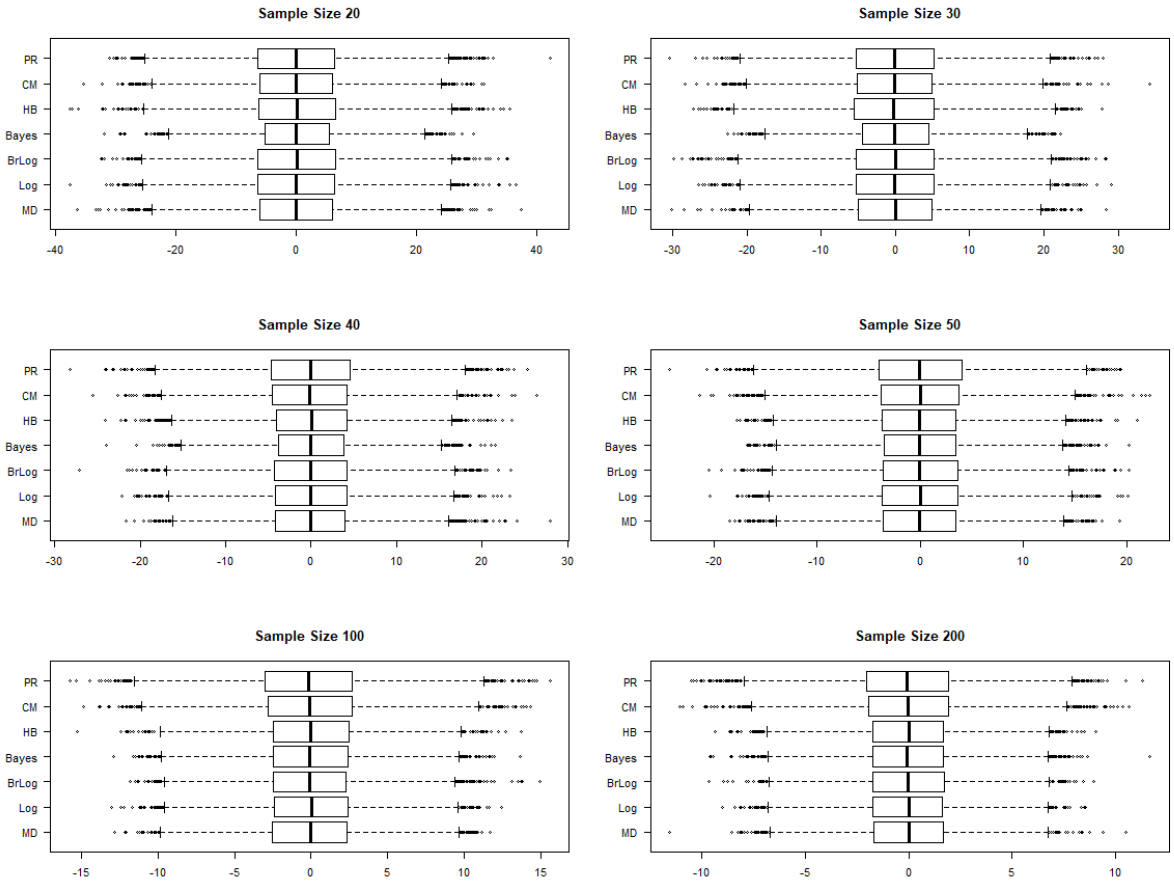


Figure 4. 6 Boxplots for PATE-(SATE)ˆ based on 10,000 samples




As observed in the boxplots, the medians for all rerandomization procedures with various covariate balance criteria are all centered around zero and there are no significant differences even when the sample size is small. However, in the same context, the interquartile range (IQR) for Bayesian modeling with Cauchy prior is the smallest out of all criteria. The difference disappears with an increase in the sample size.

As the current result shows that Bayesian logistic regression with Cauchy prior performs the best when the sample size is small, we are interested in determining if different prior improves the performance of the rerandomization procedure. For each coefficient, the scale parameters for Cauchy distribution are changed to 5 and 10, respectively. We rerun the whole

procedure and found no significant difference in RMSE as compared to the original settings.

### 4.4 Adapting Different Covariate Criteria to Sequential Rerandomization

Following the results in section 4.3, we adapt the Bayesian modeling with Cauchy prior criterion to the sequential rerandomization design (cf., Section 3.4), followed by the comparison of its performance to the Mahalanobis distance and pure randomization in a sequential design. Under a sequential experimental design, researchers may not be able to access all information of covariates for all subjects. Hence, the comparison of sequential rerandomization and complete rerandomization is meaningless. However, in sequential rerandomization, the performances between different covariate balance criteria can be compared using a consistent framework.

To manage the factors that may affect the results, we fix the total sizes of sample while only modifying the number for each batch of subjects which are 4, 8, and 16 in a batch. We should also consider each batch of subjects attending the experiment as a small sample. Additionally, we will perform the sequential rerandomization procedure with the 10,000 samples chosen in the previous section. For every iteration, we will repeatedly randomize the subjects for each batch into treatment and control groups for 400 times with no changes in the existing assignment vector. Taking the most balanced 2.5% assignments, we detect cut-off point for covariate balance criterion where we consider the covariate balance for all existing subjects.

Table 4. 5 RMSE for estimated SATE compared to True SATE and PATE for Sequential

Rerandomization

| RMSE | Sample Size | 20 | | 30 | | 40 | |
|---|---|---|---|---|---|---|---|
| Batch Size | | SATE | PATE | SATE | PATE | SATE | PATE |
| 4 | md | 8.899 | 9.483 | 7.221 | 7.703 | 6.049 | 6.463 |
| | bayeslog | 6.808 | 7.594 | 5.686 | 6.262 | 4.811 | 5.390 |
| | pr | 8.911 | 9.506 | 7.247 | 7.701 | 6.235 | 6.639 |
| 8 | md | 8.916 | 9.484 | 7.227 | 7.702 | 5.949 | 6.431 |
| | bayeslog | 6.726 | 7.480 | 5.477 | 6.077 | 4.663 | 5.242 |
| | pr | 8.980 | 9.596 | 7.236 | 7.707 | 6.233 | 6.698 |
| 16 | md | 8.825 | 9.480 | 7.013 | 7.546 | 5.907 | 6.377 |
| | bayeslog | 6.884 | 7.644 | 5.644 | 6.291 | 4.766 | 5.317 |
| | pr | 8.930 | 9.533 | 7.290 | 7.791 | 6.311 | 6.733 |

| RMSE | Sample Size | 50 | | 100 | |
|---|---|---|---|---|---|
| Batch Size | | SATE | PATE | SATE | PATE |
| 4 | md | 5.158 | 5.543 | 2.856 | 3.213 |

| | | | | | |
|---|---|---|---|---|---|
| | bayeslog | 4.273 | 4.757 | 2.632 | 3.040 |
| | pr | 5.638 | 6.013 | 3.952 | 4.222 |
| | md | 5.068 | 5.520 | 2.722 | 3.098 |
| 8 | bayeslog | 4.144 | 4.659 | 2.545 | 2.953 |
| | pr | 5.696 | 6.081 | 3.959 | 4.217 |
| | md | 4.978 | 5.437 | 2.866 | 3.236 |
| 16 | bayeslog | 4.267 | 4.728 | 2.710 | 3.081 |
| | pr | 5.614 | 6.037 | 3.968 | 4.250 |

| RMSE | Sample Size | 20 | | 30 | | 40 | |
|---|---|---|---|---|---|---|---|
| Batch Size | Percent Reduced | SATE | PATE | SATE | PATE | SATE | PATE |
| | md | 0.13% | 0.24% | 0.36% | -0.03% | 2.98% | 2.65% |
| 4 | bayeslog | 23.60% | 20.11% | 21.54% | 18.69% | 22.84% | 18.81% |
| | pr | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% |
| | md | 0.71% | 1.17% | 0.12% | 0.06% | 4.56% | 3.99% |
| 8 | bayeslog | 25.10% | 22.05% | 24.31% | 21.15% | 25.19% | 21.74% |
| | pr | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% |
| 16 | md | 1.18% | 0.56% | 3.80% | 3.14% | 6.40% | 5.29% |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | bayeslog | 22.91% | 19.82% | 22.58% | 19.25% | 24.48% | 21.03% |
| | pr | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% |

| RMSE | Sample Size | 50 | | 100 | |
|---|---|---|---|---|---|
| Batch Size | Percent Reduced | SATE | PATE | SATE | PATE |
| | md | 8.51% | 7.82% | 27.73% | 23.90% |
| 4 | bayeslog | 24.21% | 20.89% | 33.40% | 28.00% |
| | pr | 0.00% | 0.00% | 0.00% | 0.00% |
| | md | 11.03% | 9.23% | 31.25% | 26.54% |
| 8 | bayeslog | 27.25% | 23.38% | 35.72% | 29.97% |
| | pr | 0.00% | 0.00% | 0.00% | 0.00% |
| | md | 11.33% | 9.94% | 27.77% | 23.86% |
| 16 | bayeslog | 23.99% | 21.68% | 31.70% | 27.51% |
| | pr | 0.00% | 0.00% | 0.00% | 0.00% |

According to the results, in the context of small samples, Bayesian modeling with Cauchy prior performs better than the Mahalanobis distance in reducing RMSE. In the same context, the Mahalanobis distance can sometimes fail to enhance estimation precision when both batch and sample sizes are small. Given a fixed batch size, both criteria gradually

perform better with an increase in the sample size. On the other hand, given a fixed sample size, with an increase in the batch size, the Mahalanobis distance performs better for relatively small sizes of sample, while the performance of Bayesian modeling with Cauchy prior remains stable across all batch sizes.

Table 4. 6 Average Bias for estimated SATE compared to True SATE and PATE for Sequential Rerandomization

| Bias | Sample Size | 20 | | | 30 | | | 40 | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Batch Size | | SATE | PATE | SYSBIAS | SATE | PATE | SYSBIAS | SATE | PATE | SYSBIAS |
| 4 | md | -0.138 | -0.088 | 0.05 | 0.064 | 0.073 | 0.008 | -0.138 | -0.157 | -0.018 |
| | bayeslog | 0.004 | 0.054 | 0.05 | -0.018 | -0.01 | 0.008 | 0 | -0.018 | -0.018 |
| | pr | -0.013 | 0.037 | 0.05 | -0.012 | -0.004 | 0.008 | 0.161 | 0.143 | -0.018 |
| 8 | md | 0.004 | 0.054 | 0.05 | 0.001 | 0.009 | 0.008 | -0.1 | -0.118 | -0.018 |
| | bayeslog | -0.025 | 0.025 | 0.05 | 0.014 | 0.022 | 0.008 | -0.036 | -0.054 | -0.018 |
| | pr | 0.094 | 0.144 | 0.05 | 0.124 | 0.132 | 0.008 | 0.04 | 0.022 | -0.018 |
| 16 | md | 0.051 | 0.101 | 0.05 | 0.024 | 0.032 | 0.008 | -0.074 | -0.092 | -0.018 |
| | bayeslog | -0.034 | 0.016 | 0.05 | 0.018 | 0.026 | 0.008 | -0.04 | -0.058 | -0.018 |
| | pr | -0.151 | -0.101 | 0.05 | 0.102 | 0.11 | 0.008 | -0.121 | -0.14 | -0.018 |

| Bias | Sample Size | 50 | 100 |
|---|---|---|---|

| Batch Size | | SATE | PATE | SYSBIAS | SATE | PATE | SYSBIAS |
|---|---|---|---|---|---|---|---|
| | md | -0.019 | -0.059 | -0.040 | 0.023 | -0.008 | -0.030 |
| 4 | bayeslog | 0.063 | 0.023 | -0.040 | -0.016 | -0.046 | -0.030 |
| | pr | -0.069 | -0.109 | -0.040 | 0.001 | -0.030 | -0.030 |
| | md | 0.083 | 0.043 | -0.040 | -0.028 | -0.058 | -0.030 |
| 8 | bayeslog | 0.032 | -0.007 | -0.040 | -0.001 | -0.031 | -0.030 |
| | pr | -0.068 | -0.107 | -0.040 | 0.083 | 0.053 | -0.030 |
| | md | 0.026 | -0.013 | -0.040 | -0.002 | -0.032 | -0.030 |
| 16 | bayeslog | 0.033 | -0.007 | -0.040 | -0.002 | -0.032 | -0.030 |
| | pr | 0.036 | -0.004 | -0.040 | 0.023 | -0.008 | -0.030 |

With respect to the same set of samples, the average biases in estimating SATE and PATE are nearly zero. As compared to pure randomized sequential design, the average bias for sequential rerandomization design is slightly smaller. However, no significant differences are observed between them.

### 4.5 Power Analysis for Rerandomization and Sequential Rerandomization

Bayesian logistic regression with Cauchy prior performs better compared to the Mahalanobis distance as covariate balance criterion for rerandomization as well as sequential rerandomization procedures. Therefore, it would be interesting to find out if the rerandomization design enhances the power of hypothesis testing and the resulting

spontaneous increase in power for both designs. Considering the null hypothesis of no treatment effect between groups, we can get the p-value and check if it should be rejected using t-test, permutation test, or ANCOVA power analysis. However, in the rerandomization procedure, there is a difference between the empirical distribution for test statistics, i.e., the estimated SATE, and T-distribution, as this procedure will avoid assignments that may result in a 'bad' balance between groups. With a decrease in the proportion of acceptance $p_\alpha$, the distribution will be steeper. For small size of sample with high dimensional covariates, ANCOVA will also lose the ability to detect power because of singularity problem. Not enough degrees of freedom will be available for small samples. Thus, there is a need to propose a restricted permutation test for evaluating the power.

Instead of random permutation of the assignment, the rerandomization procedure needs to generate assignments for some samples. Given below is the restricted permutation test for power analysis:

1) Take one acceptable (sequential) rerandomization for a sample.
2) Get potential outcomes vector $Y_0$ based on both response surfaces and the assignment vector and measure original estimated SATE as $\widehat{SATE}_0$.
3) Carry out (sequential) rerandomization procedure 400 times and achieve 400 acceptable assignment vectors as permutated assignments. Follow the same settings of simulation study. For sequential rerandomization, the batch size should be 8.
4) Measure 400 estimated SATE based on $Y_0$ as the restricted empirical distribution for test statistics.
5) Achieve p-value through the comparison of the absolute values for all estimated

SATE and the absolute value of $\widehat{SATE}_0$.

$$p\_value = \frac{I_{\{|\widehat{SATE}_i|>|\widehat{SATE}_0|\}}}{400}, \tag{4.4}$$

where $I_{\{|\widehat{SATE}_i|>|\widehat{SATE}_0|\}}$ indicates the total number of cases when $\left|\widehat{SATE}_i\right| > \left|\widehat{SATE}_0\right|$.

6) Based on the rerandomization procedure with the Mahalanobis distance and Bayesian logistic regression with Cauchy prior as covariate balance criteria, repeat step 1 to 5 for 400 samples sized 20 and 40 to achieve 400 p-values. Considering that the true PATE is not equal to 0, count the number of p-values that are smaller than the significance level of 0.05 to calculate power.

$$power = \frac{I_{\{|p\_value_i|<0.05\}}}{400} \tag{4.5}$$

To compare the values, t-test and permutation test are considered as reference for representing the pure randomization design. Conduce t-test for 400 sample to obtain 400 p-values for hypothesis testing. Permutate the assignments for each sample to obtain p-values for hypothesis testing using *aovp* function in *lmperm* package in *R*. After repeating the test for 100 times for each sample to account for the variability of p-values from permutation, take the average p-value. When the p-values are smaller than the significance level of 0.05, power can be achieved.

The results of complete rerandomization procedures are given below:

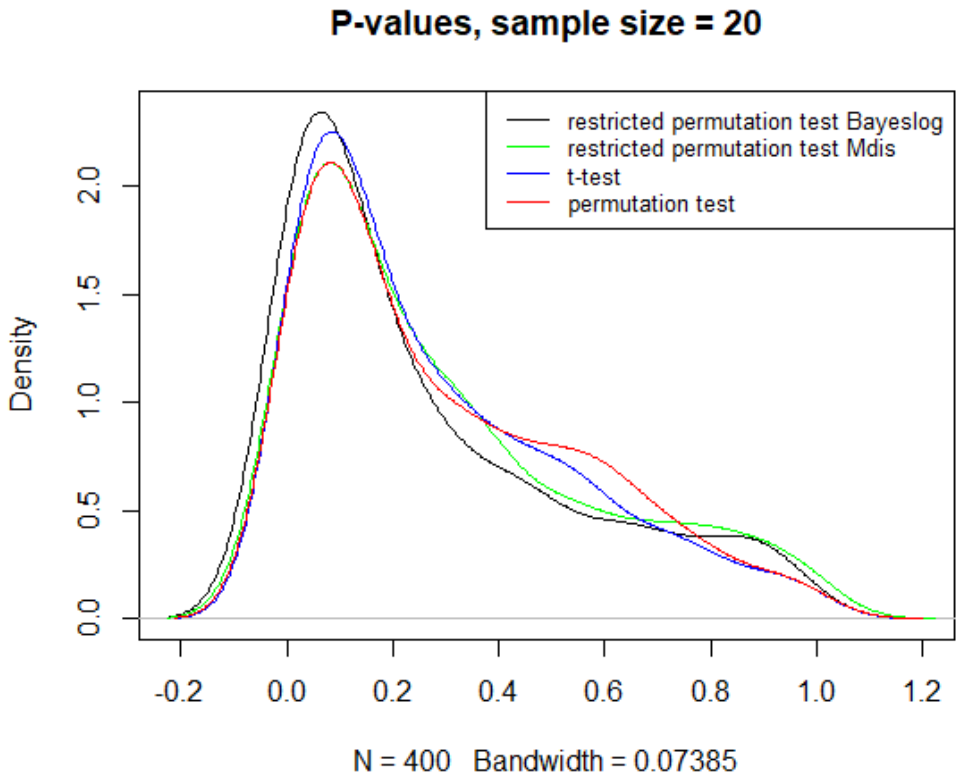
Table 4. 7 Power for rerandomization and complete randomization

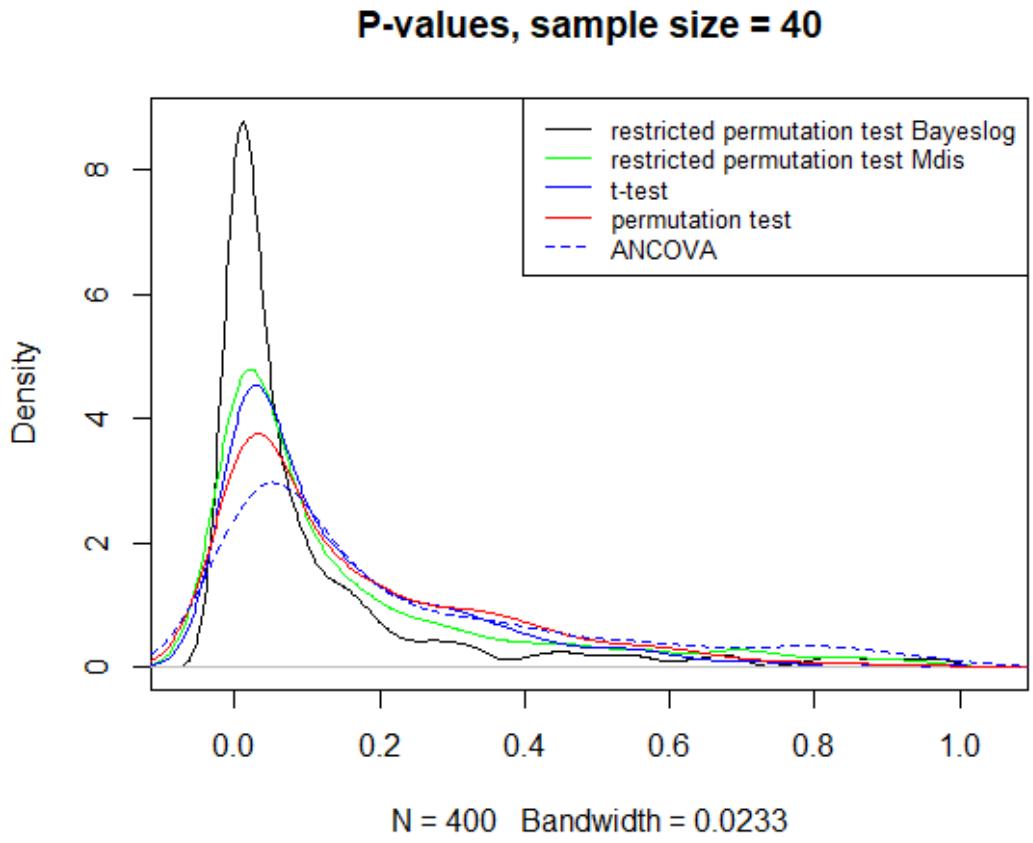
| Power | Sample Size | |
|---|---|---|
| | 20 | 40 |

| | | |
|---|---|---|
| Restricted Permutation with Bayesian Logistic | 0.2650 | 0.5775 |
| Restricted Permutation with Mahalanobis Distance | 0.1800 | 0.4450 |
| T-test | 0.1825 | 0.4075 |
| Permutation Test | 0.1800 | 0.3725 |
| ANCOVA Power Analysis | NA | 0.3000 |

Density curves for p-values are:

Figure 4. 7 Density curves of P-values for rerandomization and complete randomization






Unsurprisingly, the rerandomization procedure substantially increases the power of hypothesis testing. For sample size of 20, there is about 45.2% increase in the power of rerandomization procedure based on Bayesian logistic regression with Cauchy prior, the power of rerandomization procedure based on the Mahalanobis distance does not increase. For sample size of 40, the power of both rerandomization procedures increases significantly. The rerandomization procedure based on the Mahalanobis distance starts differentiating with a 9.2% increase in the power, while there is a 41.7% increase in the power of rerandomization procedure based on Bayesian logistic regression with Cauchy prior. ANCOVA power, however, remains the worst because of the absence of degrees of freedom. Consistent with our finding, the Mahalanobis distance might not be able to detect covariate

balance between groups when the sample size is small.

The results of sequential rerandomization procedures are given below:

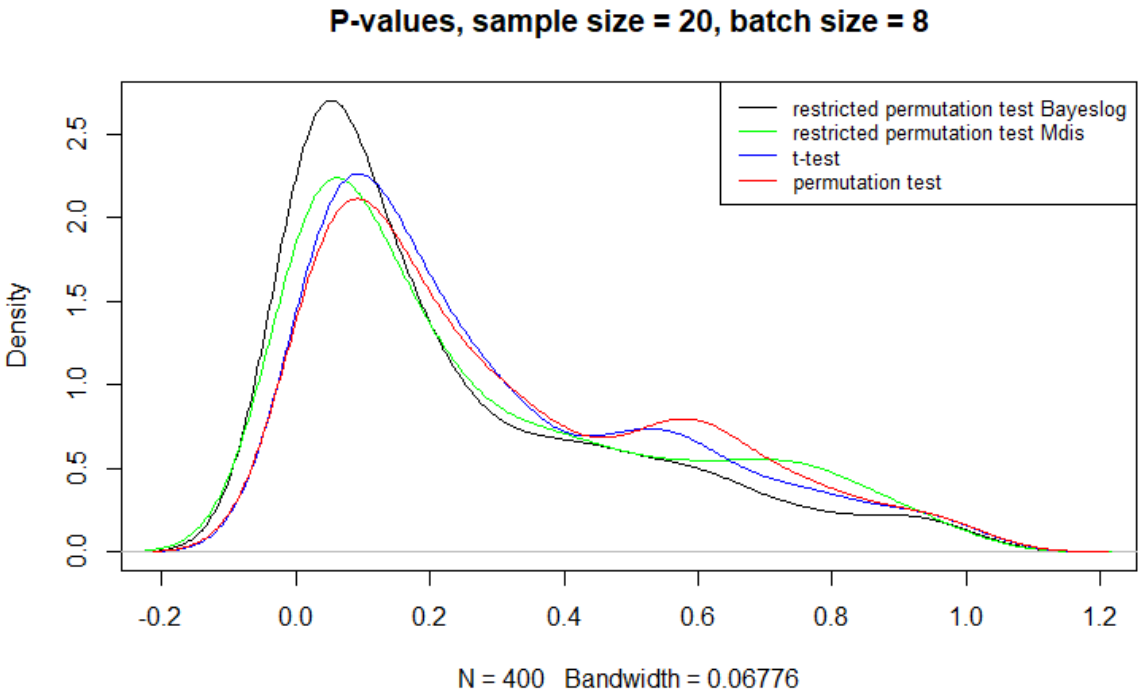
Table 4. 8 Power for sequential rerandomization and complete randomization in sequential design

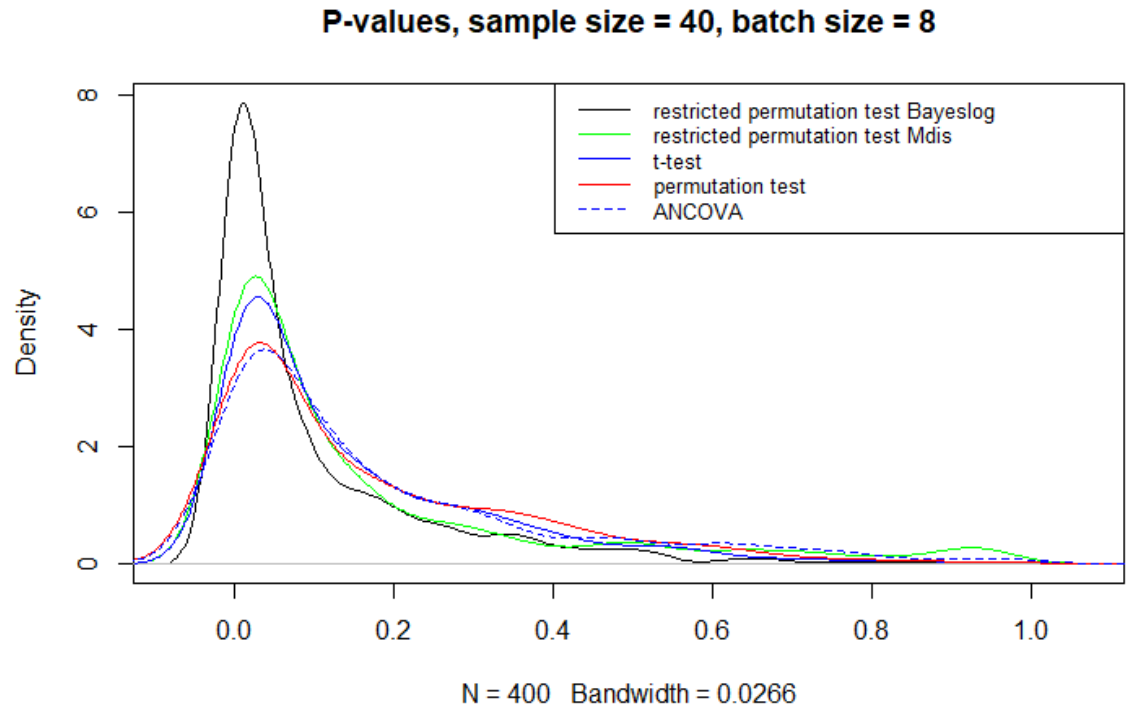
| Power | Sample Size | |
|---|---|---|
| | 20 | 40 |
| Restricted Permutation with Bayesian Logistic | 0.3050 | 0.5725 |
| Restricted Permutation with Mahalanobis Distance | 0.2575 | 0.4450 |
| T-test | 0.1600 | 0.4075 |
| Permutation Test | 0.1575 | 0.3950 |
| ANCOVA Power Analysis | NA | 0.3525 |

The batch sizes are set to 8 for the sample sizes of 20 and 40. Density curves for p-values are:

Figure 4. 8 Density curves of P-values for sequential rerandomization and complete

randomization in sequential design




The results are again consistent with the RMSE findings. Sequential rerandomization procedure performs better than complete rerandomization procedure in increasing the power

of hypothesis testing. For sample size of 20, there is around 90.6% increase in the power of sequential rerandomization procedure based on Bayesian logistic regression with Cauchy prior, while the power of sequential rerandomization procedure based on the Mahalanobis distance increases by around 60.9%. For sample size of 40, the power of both procedures increases significantly. Similar to the increase in complete rerandomization procedure, there is a 9.2% increase in the power of sequential rerandomization procedure based on the Mahalanobis distance increases. On the other hand, the power of sequential rerandomization procedure based on Bayesian logistic regression with Cauchy prior rises by 40.5%. Attributing to the limited computing power, we can test more combinations of various sample sizes, numbers of selected samples, and effect sizes with respect to more permutated assignments. Overall, we hypothesize that both sequential and complete rerandomization procedures will increase the power of hypothesis testing. With an increase in sample size, the power increases more slowly than before. When the sample size is small, the performance of rerandomization procedure based on Bayesian logistic regression remains better than that of rerandomization procedure based on the Mahalanobis distance.

## Conclusion

Rerandomization is intended to enhance the precision in estimating ATE for randomized experiments when the covariates information of subjects and potential outcomes are correlated. Rerandomization is advantageous in the improvement of covariate balance between treatment and control groups by simultaneously balancing all covariates or the covariates of research interest. Rerandomization also does not need asymptotic assumption and any assumptions for covariate distribution and the relationship between covariate information and potential outcomes (Morgan & Rubin, 2012). In the rerandomization procedure, the Mahalanobis distance is introduced to calculate covariate balance between treatment and control groups. Nonetheless, the covariate balance can be determined using other possible criteria in the rerandomization procedure. Researchers must also evaluate the rerandomization performance when the sample size is small.

For answering the research questions, we perform a Monte Carlo simulation and compare the performance of rerandomization procedure with various covariate balance criteria with respect to both small and moderate samples. We then randomly select 10,000 samples from the population with 34 covariate information, one treatment variable, and potential outcomes. To determine the cut-off score for each covariate balance criterion, we carry out 400 rerandomized experiments for each sample. We consider one acceptable rerandomization for each sample and subsequently calculate an estimate of sample ATE, while the true sample ATE exists at the mean time. The overall performance of each rerandomization procedure can be checked using two criteria – the average bias and root mean square error between estimated SATE and true SATE.

As per the results, there are no significant differences in bias between rerandomization procedures with various covariate balance criteria and pure randomization procedure and between all biases with 0. It signifies that all estimated SATE are unbiased which is reasonable due to a large number of selected samples, thereby eliminating the estimation error resulting from sample selection (cf., formula 2.17). Rerandomization and pure randomization procedures also give an overall balanced assignment between treatment and control groups, eliminating the estimation error due to imbalanced treatment (cf., formula 2.23).

With respect to RMSE, it decreases with an increase in the sample size, which is also intuitive. When the sample sizes are large, achieving better balance across all covariates becomes easier. In the context of relatively small sizes of sample, the performance of rerandomization procedure based on Bayesian logistic regression with Cauchy prior is significantly better than other criteria. This criterion comparatively lowers RMSE by 17% to 19% than RMSE for pure randomization procedure. Certain criteria, on the other hand, lose the ability to find covariate balance between groups in small samples but gradually achieve estimation precision with an increase in the sample size.

It has been proved that rerandomization procedures improve the power of hypothesis testing. We propose a restricted permutation test for evaluating the power of rerandomization procedure based on Bayesian logistic regression with Cauchy prior and the Mahalanobis distance. It appears that the rerandomization procedure based on Mahalanobis distance for small samples does not ensure an increase in the power of hypothesis testing. However, in the rerandomization procedure based on Bayesian logistic regression with Cauchy prior, there is

a 45.2% increase in the power of hypothesis testing. This result is consistent with our finding that the Mahalanobis distance does not outperform Bayesian logistic regression with Cauchy prior to differentiate covariate balance when the sample size is small.

When the Bayesian logistic regression with Cauchy prior is introduced to Zhou et al.'s (2018) sequential rerandomization design, it performs better than the sequential rerandomization based on the Mahalanobis distance, which is consistent with complete rerandomization procedures. Additionally, sequential rerandomization procedures give an unbiased estimate of SATE. When the samples are small or moderate in size, Bayesian logistic regression with Cauchy prior performs significantly better than the Mahalanobis distance as the covariate balance criterion for sequential rerandomization procedure. The reduced percentage of RMSE generally increases with an increase in the sample size.

With a fixed batch size, both criteria gradually perform better, as the sample size increases. However, when the sample size is small, the Mahalanobis distance lose the ability to find covariate balance between groups in sequential rerandomization procedure. With an increase in the sample size, it gradually achieves power, while the power for Bayesian logistic regression with Cauchy prior is substantial and relatively stable in improving the precision in estimating SATE for small samples.

With a fixed sample size, the increase in batch size may improve the performance of Mahalanobis distance for relatively smaller sample sizes, while Bayesian modeling with Cauchy prior has relatively stable performance across all batch sizes. As the sample increases, the performance of Mahalanobis distance gradually becomes stable and significantly improves precision. For extremely small or moderate samples, a small size of

sample is not helpful for the Mahalanobis distance to differentiate covariate balance under a sequential design.

This paper has some limitations which are listed below and can be further discussed or examined.

1. Morgan and Rubin (2012) suggested more rerandomization trials for small samples, about 1,000 times per sample. However, because of the limited computing power, we only carried out 400 rerandomization trials to obtain the empirical distributions for all covariate balance criteria. More trial can be performed for further tests in the context of small samples.
2. Morgan (2011) indicated that for small samples, there may be violation in the results that rely on the assumption that $\bar{X}_T - \bar{X}_C$ is multivariate normal. Therefore, there is a need to understand a way to theoretically prove the amount of percentage reduced in variance.
3. In this dissertation, we uniformly set the proportion of acceptance to 2.5% for all rerandomization procedures to manage the factors that may affect our conclusion. The possibility of variation in lowering RMSE could be evaluated using more proportion of acceptance.
4. In the simulation study, out of all 34 covariates, the types of variables include nominal variable with binary classes as well as ordinal, interval, and ratio variables. No nominal variable does not exist with multiple classes. The covariates with multiple classes including race and blood type can be converted into multiple dummy variables with binary classes. Subsequently, they can be adapted to the

rerandomization procedure.

5. We also use *prima facie* estimator for estimating SATE for all samples and sizes of sample. Focusing on the design stage of causal inference study and employing a direct estimation method will help remove other factors with the potential to influence the estimation error. However, more efficient or appropriate estimators could be available. After conducting the rerandomization procedure, various estimation methods that are most suitable for some studies could be used.
6. Given the fact that we achieve the results and conclusions from simulation studies based on ECLS-K data, it is worth discussing that whether the results can be generalized to other scenarios. When generating response surfaces for the subjects, the covariates are related to the potential outcomes. Intuitively, if the covariate information for data is related to potential outcomes, our results and conclusions will still hold. However, the distributions and variable types for ECLS-K data is rather clear. If there are nominal pretreatment variable with multiple categories, researchers should first make dummies and then adapt the rerandomization procedures. For specific conditions, my hypotheses are that:
    a) If the covariance matrix for covariates is singular, Bayesian logistic regression with Cauchy prior may perform better than Mahalanobis distance in (sequential) rerandomization procedure in reducing the estimation error.
    b) If there are many dummy variables or most covariates are nominal, Mahalanobis distance may perform better than Bayesian logistic regression with Cauchy prior. Because the prior distribution of covariates may not interpret the real

distributions. More research needs to be done.

## Discussion

This chapter focuses on possible future research beyond this topic and includes discussions about the framework with similar purposes as sequential rerandomization, several interesting results of this simulation study, and more general cases for experimental design including multi-treatment or unequal-sized experimental design.

### 6.1 Interesting Findings from the Simulation Study

In this dissertation, given the same sample size, both criteria for a batch size of 8 outperform the batch sizes of 4 and 16. The conclusion indicates the presence of a suitable batch size for sequential rerandomization design. However, future research may ensure a better understanding of the topic. Based on the ratio of the number of covariates and batch sizes, following a certain number of batches, the sequential rerandomization procedure may gradually achieve power. For a relatively small batch size, there may only be trial improvement in covariate balance between groups in every iteration. Alternatively, for a relatively large batch size, as the total sample size may not yield enough number of iterations, the whole process is likely to be similar to complete rerandomization.

Table 6. 1 Percentage Reduced of RMSE for estimated SATE compared to True SATE for both Sequential Rerandomization (SEQ) and Complete Rerandomization (MR)

| Percent Reduced | Sample Size | 20 | | 30 | | 40 | | 50 | | 100 | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Batch Size | | SEQ | MR | SEQ | MR | SEQ | MR | SEQ | MR | SEQ | MR |
| 4 | md | 0.13% | 5.48% | 0.36% | 6.01% | 2.98% | 11.29% | 8.51% | 14.17% | 27.73% | 17.83% |
| | bayeslog | 23.60% | 19.33% | 21.54% | 18.98% | 22.84% | 17.97% | 24.21% | 17.11% | 33.40% | 18.78% |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | pr | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% |
| 8 | md | 0.71% | 5.48% | 0.12% | 6.01% | 4.56% | 11.29% | 11.03% | 14.17% | 31.25% | 17.83% |
| | bayeslog | 25.10% | 19.33% | 24.31% | 18.98% | 25.19% | 17.97% | 27.25% | 17.11% | 35.72% | 18.78% |
| | pr | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% |
| 16 | md | 1.18% | 5.48% | 3.80% | 6.01% | 6.40% | 11.29% | 11.33% | 14.17% | 27.77% | 17.83% |
| | bayeslog | 22.91% | 19.33% | 22.58% | 18.98% | 24.48% | 17.97% | 23.99% | 17.11% | 31.70% | 18.78% |
| | pr | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% |

Another interesting finding is that if we insist comparing the performance between sequential rerandomization and complete rerandomization with the same sample size, the performance of the sequential rerandomization procedure is better than the complete rerandomization procedure for all batch and sample sizes when using Bayesian logistic regression with Cauchy prior as covariate balance criterion. For the sample size of 100, the conclusion is consistent with Bayesian logistic regression with Cauchy prior when using the Mahalanobis distance as covariate balance criteria. However, the study result is no longer true when the sample size is small. Zhou et al. (2018) stated under some assumptions, sequential rerandomization gains better covariate balance compared to rerandomization at one time. Based on our finding, this is true for small sizes of sample, but not with the Mahalanobis distance as covariate balance criterion. Again, it validates that Bayesian logistic

regression with Cauchy prior remains good at calculating covariate balance for (sequential) the rerandomization procedure when the sample size is small.

### 6.2 Designs with Multiple Treatments

Experimental designs with multiple treatments have two intuitive alternatives. First, we verify covariate balance for all pairwise comparisons between treatment groups. However, it is computationally costly, in addition to being strict. To safely get a threshold with the proportion of acceptance $p_\alpha$ for the whole experiment, with *j* treatment groups, $p_\alpha$ for each pairwise comparison should be set to $p_\alpha^{\binom{j}{2}}$. Second, we discover a univariate covariate balance criterion for measuring a multivariate balance between all treatment groups. According to a study conducted by Morgan (2011), the commonly used test statistics in multivariate analysis of variance (MANOVA) can also be employed as a univariate measure for balance across groups. Wilk's $\Lambda$, Lawley-Hotelling test, Pillai's test, and Roy's test are some of the possible tests or statistics. For Wilk's $\Lambda$, larger values represent better balance, while smaller values represent better balance for the three other tests. These test statistics essentially compare the variability within groups with the variability between groups. Larger variance between groups than the variance within groups significantly affects the balance between multiple treatment groups. Irrespective of the distribution of the test statistics, after determining the balance criterion to evaluate the covariate balance between groups, we will perform the rerandomization procedure as proposed to achieve an empirical distribution for these test statistics. Rencher (2003) provides the details for these statistics. It will also be worth conducting research to determine which covariate balance criteria perform the best when the sample size is small.

### 6.3 Designs with Unequal-sized Rerandomization Procedure

In this dissertation, for the convenience of measurement for complete and sequential rerandomization, we set $p_w = 0.5$. In most real studies, $p_w < 0.5$ particularly for high-budget experiments. Morgan (2011) stated that this goal can be achieved by dividing the population into multiple equal-sized groups, performing rerandomization procedure between groups, and merging the groups to obtain expected $p_w$. For instance, if $p_w = 0.33$, first, the population is randomly separated into three exchangeable groups and the rerandomization procedure is performed to balance these groups. Lastly, two of the three groups are merged to form the control group and the third group is regarded as the treatment group. If $p_w = 0.4$, i.e., $\frac{1-p_w}{p_w}$ is not integer, we can randomly divide the population into 5 equal-sized groups and repeat the same procedure.

Morgan (2011) also proposed another option that the sample size can be slightly reduced to create an equal-sized design. However, it has numerous disadvantages. First, when the sample size reduces, the precision may also decrease. Additionally, the rerandomization procedure based on the Mahalanobis distance may not efficiently work with small samples. The use of Bayesian logistic regression with Cauchy prior as the covariate balance criterion may somehow fix the problem. Generally speaking, the reduction in sample size will decrease estimation precision, while it increases with the rerandomization procedure. Therefore, it is important to evaluate the conditions under which the increased precision due to rerandomization should exceed the loss in precision due to reduced sample size. Second, it may not be plausible to reduce the sample size to force $p_w = 0.5$, as $p_w$ may intentionally be set to be larger than 0.5 for some experiments. The goal may be achieved by reducing

sample sizes. However, this method can be used if the treatment group size is limited due to the cost of experiment or the availability of subjects.

### 6.4 Extended Topics

Rerandomization is one of the promising experimental design procedures that could improve the power of randomized experiments and the precision in estimating treatment effect with almost no assumptions or limitation. Interestingly, according to Harshaw, Sävje, Spielman and Zhang (2019), there should be a trade-off between covariate balance and the robustness of experimental designs. Maximizing covariate balance does not always ensure an optimal estimation for ATE. It shows a high correlation with the setting of proportion of acceptance in the rerandomization procedure. As a topic, the way to define the proportion of acceptance accounting for sample size and robustness has high potential.

Both Morgan (2011) and Zhou et al. (2018) discussed the covariate-adaptive minimization method for sequential experimental design. Sequential rerandomization is not only adaptive but also allows for rerandomization. Thus, it suffers less from the selection bias than the minimization method (Berger, 2010), which remains one of its key advantages. Therefore, the comparison of the performance between sequential rerandomization and covariate-adaptive minimization method based on real data would be interesting. This topic also appears to be related to the trade-off between covariate balance and robustness. Certain randomizations in experimental designs between subjects remain a key area of interest.

An index with the ability to rank the covariate balance is essential when subjects are randomly assigned to treatment and control groups in the rerandomization procedure. However, the models or classifiers used for providing the index are not fixed. According to

Gagnon-Bartsch and Shem-Tov (2019), we can adapt various classifiers including logistic regression, random forest, and K-nearest neighbors to classification permutation test (CPT) for measuring covariate balance. Intuitively, as long as the approach can give a comparable index, it can be specified as the covariate balance criterion in the rerandomization procedure. Super learner (Van der Laan, Polley, & Hubbard, 2007), in particular, is a machine learning algorithm that uses cross-validation for the performance estimation of different models or same model with different settings. It will form an optimal set of weights for combining an initial set of candidate learners (models). It has also been proven to serve as an asymptotically optimal system for learning (Polley & Van der Laan, 2010). Polley and Van der Laan (2010) conclude that the practical performance of Super Learner is affirmed to be adaptive and robust in smaller samples. Consequently, researchers are expected to evaluate Super Learner's performance as covariate balance criterion. It will be interesting to see the performance of Super Learner algorithm to measure covariate balance in future research while experimenting different model combinations to determine the best framework for measuring covariate balance.

Lastly, we are interested in determining the possibility of using 'rerandomization' procedure in the estimation stage of causal inference study. Finding similar subjects in treatment and control group after propensity score matching is one of the traditional methods to estimate ATE. The questions arises whether we can reversely use the rerandomization procedure to identify the subjects with a relatively good match to estimate ATE. For an unequal-sized experiment where subjects in treatment group are lesser than those in control group, we randomly select a batch of subjects with the size of treated subjects from the

control group and calculate covariate balance based on a particular criterion. The procedure is then repeated to obtain an empirical distribution of the covariate balance criterion. A proportion of acceptance is consequently set and subjects whose balance between the treated subjects is relatively good as ‘matched’ subjects are considered. Finally, we estimate ATE according to a set of ‘matched’ subjected for the treated.

## References


Berger, V. W. (2010). Minimization, by its nature, precludes allocation concealment, and invites selection bias. *Contemporary clinical trials*, *31*(5), 406.

Branson, Z., Dasgupta, T., & Rubin, D. B. (2016). Improving covariate balance in 2k factorial designs via rerandomization with an application to a new york city department of education high school study. *The Annals of Applied Statistics*, *10*(4), 1958-1976.

Branson, Z., & Shao, S. (2018). Ridge rerandomization: An experimental design strategy in the presence of collinearity. *arXiv preprint arXiv:1808.04513*.

Bruhn, M., & McKenzie, D. (2009). In pursuit of balance: Randomization in practice in development field experiments. *American Economic Journal: Applied Economics*, *1*(4), 200-232.

Coalition for Evidence-Based Policy. (2003). *Identifying and implementing educational practices supported by rigorous evidence: A user-friendly guide*. US Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance. https://ies.ed.gov/ncee/pubs/evidence_based/randomized.asp

*Early Childhood Longitudinal Studies (ECLS) Program*. IES National Center for Education Statistics, from https://nces.ed.gov/ecls/kindergarten.asp

Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika*, *80*(1), 27-38.

Fisher, R. A. (1992). The arrangement of field experiments. In *Breakthroughs in statistics* (pp. 82-91). Springer, New York, NY.

Gagnon-Bartsch, J., & Shem-Tov, Y. (2019). The classification permutation test: A flexible approach to testing for covariate imbalance in observational studies. *The Annals of Applied Statistics*, *13*(3), 1464-1483.

Gelman, A., Jakulin, A., Pittau, M. G., & Su, Y. S. (2008). A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics*, *2*(4), 1360-1383.

Greenland, S., Pearl, J., & Robins, J. M. (1999). Causal diagrams for epidemiologic research. *Epidemiology*, 37-48.

Hansen, B. B., & Bowers, J. (2008). Covariate balance in simple, stratified and clustered comparative studies. *Statistical Science*, 219-236.

Harrell Jr, F. E. (2015). *Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis*. Springer.

Harshaw, C., Sävje, F., Spielman, D., & Zhang, P. (2019). Balancing covariates in randomized experiments using the Gram-Schmidt Walk. *arXiv preprint arXiv:1911.03071*.

Heckman, J. J., Smith, J., & Clements, N. (1997). Making the most out of programme evaluations and social experiments: Accounting for heterogeneity in programme impacts. *The Review of Economic Studies*, *64*(4), 487-535.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, *81*(396), 945-960.

Hong, G., & Raudenbush, S. W. (2006). Evaluating kindergarten retention policy: A case study of causal inference for multilevel observational data. *Journal of the American Statistical Association*, *101*(475), 901-910.

Imai, K. (2005). Do get-out-the-vote calls reduce turnout? The importance of statistical methods for field experiments. *American Political Science Review*, *99*(2), 283-300.

Imai, K., King, G., & Stuart, E. A. (2008). Misunderstandings between experimentalists and observationalists about causal inference. *Journal of the Royal Statistical Society: Series A (statistics in society)*, *171*(2), 481-502.

Imbens, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and Statistics*, *86*(1), 4-29.

Imbens, G. W., & Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, *186*(1007), 453-461.

Keller, B., & Tipton, E. (2016). Propensity score analysis in R: A software review. *Journal of Educational and Behavioral Statistics, 41*(3), 326-348.

Kleyman, Y. N. (2009). *Testing for Covariate Balance in Comparative Studies* (Doctoral dissertation).

Krause, M. S., & Howard, K. I. (2003). What random assignment does and does not do. *Journal of Clinical Psychology*, *59*(7), 751-766.

Li, X., Ding, P., & Rubin, D. B. (2018). Asymptotic theory of rerandomization in treatment–control experiments. *Proceedings of the National Academy of Sciences*, *115*(37), 9157-9162.

Li, X., & Ding, P. (2020). Rerandomization and regression adjustment. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*.

Morgan, P. L., Frisco, M. L., Farkas, G., & Hibel, J. (2010). A propensity score matching analysis of the effects of special education services. *The Journal of Special Education*, *43*(4), 236-254.

Morgan, K. L. (2011). Rerandomization to improve covariate balance in randomized experiments.

Morgan, K. L., & Rubin, D. B. (2012). Rerandomization to improve covariate balance in experiments. *The Annals of Statistics*, *40*(2), 1263-1282.

Morgan, K. L., & Rubin, D. B. (2015). Rerandomization to balance tiers of covariates. *Journal of the American Statistical Association*, *110*(512), 1412-1421.

Polley, E. C., & Van der Laan, M. J. (2010). Super learner in prediction.

Rencher, A. C. (2003). *Methods of multivariate analysis* (Vol. 492). John Wiley & Sons.

Rosenbaum, P. R., & 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. (2005). An exact distribution-free test comparing two multivariate distributions based on adjacency. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *67*(4), 515-530.

Rosenberger, W. F., & Sverdlov, O. (2008). Handling covariates in the design of clinical trials. *Statistical Science*, *23*(3), 404-419.

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

Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *The Annals of Statistics*, 34-58.

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. (1990). Comment: Neyman (1923) and causal inference in experiments and

observational studies. *Statistical Science*, *5*(4), 472-480.

Rubin, D. B. (1990). Formal mode of statistical inference for causal effects. *Journal of Statistical Planning and Inference*, *25*(3), 279-292.

Rubin, D. B. (2008). Comment: The design and analysis of gold standard randomized experiments. *Journal of the American Statistical Association*, *103*(484), 1350-1353.

Shadish, W. R., Clark, M. H., & Steiner, P. M. (2008). Can nonrandomized experiments yield accurate answers? A randomized experiment comparing random and nonrandom assignments. *Journal of the American statistical association*, *103*(484), 1334-1344.

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

Steiner, P. M., Cook, T. D., & Shadish, W. R. (2011). On the importance of reliable covariate measurement in selection bias adjustments using propensity scores. *Journal of Educational and Behavioral Statistics*, *36*(2), 213-236.

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical science: a review journal of the Institute of Mathematical Statistics*, *25*(1), 1.

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

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.

Tu, J. V. (1996). Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. *Journal of Clinical Epidemiology*, *49*(11), 1225-1231.

Urbach, P. (1985). Randomization and the design of experiments. *Philosophy of Science*, *52*(2), 256-273.

Van der Laan, M. J., Polley, E. C., & Hubbard, A. E. (2007). Super learner. *Statistical Applications in Genetics and Molecular Biology*, *6*(1).

Winship, C., & Morgan, S. L. (1999). The estimation of causal effects from observational data. *Annual Review of Sociology*, *25*(1), 659-706.

Worrall, J. (2010). Evidence: philosophy of science meets medicine. *Journal of Evaluation in Clinical Practice, 16*(2), 356-362.

Zhou, Q., Ernst, P. A., Morgan, K. L., Rubin, D. B., & Zhang, A. (2018). Sequential rerandomization. *Biometrika*, *105*(3), 745-752.

# Appendix

## Visualization of Trained Neural Network Model for Predicting Missing Potential Outcomes