Optimization-Assisted Rerandomization*

Connor T. Jerzak† Rebecca Goldstein‡

October 15, 2018

Abstract

How much should experimenters rerandomize? This paper quantifies the idea that too much rerandomization can lead to wide exact confidence intervals, but too little rerandomization can result in less accurate point estimates. We then develop an algorithm, which we call Optimization-Assisted Rerandomization (OAR), for generating an approximately uniform set of acceptable rerandomization samples. OAR allows researchers to push the acceptance rate to new lows, while the resulting samples can be used for non-parametric inference or for assigning subjects to treatment conditions. We apply this procedure to simulated datasets to illustrate advantages of the framework.

*Our thanks to Donald Rubin and Marie-Abèle Bind for helpful comments.

†PhD Candidate and Carl J. Friedrich Fellow, 1737 Cambridge Street, Harvard University, Cambridge MA 02138, ConnorJerzak.com, cjerzak@g.harvard.edu.

‡PhD Candidate, 1737 Cambridge Street, Harvard University, Cambridge MA 02138.

1 Introduction

Randomized controlled experiments are a cornerstone of causal inference. However, in finite samples, random assignment will often result in imbalance between the treated and control groups on covariates predictive of the outcome. This imbalance will generate conditional bias while adding noise to treatment effect estimates. This issue is not uncommon: with 6 covariates and a significance level of 0.05, the probability that treated and control group means will differ significantly is greater than 0.25 (because $1 - [1 - 0.05]^6 = 0.26$; Morgan and D. B. Rubin (2012)). In other words, 1 out of every 4 experiments would be expected to have significant imbalance even if only 6 background variables were collected. With more variables, the probability of imbalance would be higher still.

Morgan and D. B. Rubin (2012) and Morgan (2011) propose rerandomization as a solution to this problem of covariate imbalance. They argue that, if experimenters provide a quantitative definition of imbalance in advance, they can safely discard randomizations that are too imbalanced in order to obtain more precise estimates of treatment effects. In this context, it is necessary to adjust exact significance tests and fiducial intervals after rerandomization: experimenters can perform exact significance tests by adjusting permutation-based inference, excluding randomizations that would not have been accepted according to the initial acceptance threshold. Researchers can also calculate fiducial intervals by exploiting the duality between intervals and tests.

Here, we observe that it is possible to take rerandomization too far. In particular, problems occurs if researchers use an acceptance threshold so high that the set of acceptable randomizations becomes too small to construct exact confidence intervals.[1] Some, such

[1] Neyman (parametric) confidence intervals are also invalid, because parametric confidence intervals also depend on variation in the treatment assignment vector (Imbens and D. Rubin, 2015).

as Kasy (2016), rightly argue that there exits a strictly optimal treatment assignment in terms of covariate balance, and that this optimum can in principle be found. However, viewed through the lense of Morgan and Rubin's rerandomization approach, there is an equivalence between identifying the optimally balanced treatment assignment and rerandomizing so much as to exclude all but one randomization scheme. This situation conflicts with the assumptions underlying statistical inference as laid out in the work of Neyman and especially Fisher. Thus, Kasy's line of argumentation raises a crucial question: how *much* should experimenters rerandomize?

In this paper, we consider this question of where experimenters should set the acceptance threshold for rerandomization, especially in small experiments. Our argument proceeds in three parts. First, we show that the expected minimum p-value achievable under a rerandomization scheme is a function of the acceptance probability, p_a. Then, we show that the minimum p-value determines the width of fiducial intervals when these intervals are generated using permutation-based tests that account for the rerandomization procedure. Finally, we propose a framework for generating a set of acceptable randomizations which preserves enough randomness while involving significant computational gains over previous approaches. We call this approach optimization-assisted rerandomization (OAR).

2 Preliminaries

2.1 Notation & Assumptions

Before we present our main arguments, it is useful to introduce the notation we will use going forward. Following the notation in Morgan and D. B. Rubin (2012), we assume that the experiment of interest is 2^1, meaning there is one experimental factor with two levels (i.e. a treatment and a control level). Let $W_i = 1$ if unit i is treated and $W_i = 0$ if unit i is control. Say $i \in \{1, ..., n\}$. The vector $\mathbf{W}_{n\times 1}$ denotes treatment assignment. Let $\mathbf{X}_{n\times k}$ denote the data matrix containing k pre-treatment covariates relevant to the outcome of interest. We are interested in the outcome vector $\mathbf{Y}_{n\times 2}$, where $\mathbf{Y}(0)_{n\times 1}$ denotes the complete potential outcomes under control and $\mathbf{Y}(1)_{n\times 1}$ denotes the complete potential outcomes under treatment. If $(\mathbf{Y}_{\text{Obs}}(\mathbf{W}))_{n\times 1}$ denotes the vector of observed outcome values, $Y_{\text{Obs},i} = Y_i(1)W_i + Y_i(0)(1 - W_i)$. If we assume the sharp null hypothesis (zero treatment effect for every unit) and if we leave "$\mathbf{Y}_{\text{Obs}}$ fixed and simulating many acceptable randomization assignments, $\mathbf{W}$, we can empirically create the distribution of any estimator," $g(\mathbf{X}, \mathbf{W}, \mathbf{Y}_{Obs}(\mathbf{W}))$ conditional on the null hypothesis (Morgan and D. B. Rubin, 2012). We make the additional assumptions that the experimenter is interested in using Fisherian inference and that the treatment effects are constant and additive.

2.2 Defining Randomization-based Fiducial Intervals

We can also exploit the duality between intervals and tests to form a fiducial interval from a randomization test. In this case, we can produce an interval by finding the set of all null hypotheses that the observed data would fail to reject. If we recall the constant and

additive treatment effect assumption ($\mathbf{Y}(1) = \mathbf{Y}(0) + \tau$), then the hypotheses can be written as

$$\begin{aligned} H_0 &: \tau = \tau_0 \\ H_1 &: \tau \neq \tau_0. \end{aligned} \tag{1}$$

An α-level fiducial interval for τ consists of the set of τ_0 such that the observed test statistic would not lead to a rejection of the null hypothesis at significance-level α. Here, when $\tau_0 \neq 0$, the randomization test is conducted by constructing

$$Y_i(0)^* = (Y_{\text{Obs},i} - \tau_0)\, W_i + Y_{\text{Obs},i}(1 - W_i).$$

Then, keeping $Y_i(0)^*$ fixed, we permute the treatment assignment vector, and calculate $\mathbf{Y}^*_{\text{Obs}}$ by adding τ_0 to the treatment group outcomes under the permutation. This procedure generates a distribution of $\hat{\tau}$ under the null $H_0 : \tau = \tau_0$. We can use this distribution to calculate a p-value for $\hat{\tau}_{\text{Obs}}$ under the null. We can form an $100(1-\alpha)\%$ interval for τ by finding the values of τ_0 which generate a p-value greater than or equal to α. Garthwaite (1996) discuss an efficient algorithm for obtained randomization-based fiducial intervals, which searches for the interval endpoints using a procedure based on the Robbins-Monro search process.

Because fiducial intervals are less familiar to many readers than their more familiar cousin confidence intervals, we briefly review their similarities and differences. Jerzy Neyman is credited with originating the idea of classical confidence intervals (Neyman, 1941). Intuitively, confidence intervals should capture the true θ_1 value in $100 \times (1-\alpha)\%$ of repeated experiments. Conversely, randomization-based intervals, which originated in the work of Neyman's great rival, Ronald A. Fisher, are not justified with reference to

the properties of an estimator over multiple experiments, but instead are used to facilitate inferences *after* when condition on the sample. Put differently, randomization intervals incorporate randomness only through variation in the treatment vector, and do not explicitly make a reference to repeated experiments. Nevertheless, randomization intervals can have good frequentist properties. Even so, it is important distinguish randomization-based from classical frequentist intervals by using the term "fiducial intervals" to denote the former. The adjective, "fiducial," is meant to convey the idea that we are interested here in making inferences about the actually observed sample, without reference to hypothetical repeats of the experiment.

3 Rerandomization and the Limits of Inference

Is it possible to take rerandomization "too far"? We argue that it is indeed possible, in the sense of accepting so few randomizations that it is no longer possible to perform a meaningful randomization test. This important method of assessing uncertainty becomes degenerate.

Consider the following example: there are 8 experimental units, and we would like to test the effect of one factor at two levels (one treatment group and one control group). There is one background covariate, h, which denotes the hour a test is conducted. For simplicity, assume that there are 8 observed values for h, ranging from 1 to 8. We can quantify the randomization balance with the square root of the quadratic loss:

$$m = \sqrt{\left[(\bar{h}|W_i = 1) - (\bar{h}|W_i = -1)\right]^2} \tag{2}$$

where smaller values of m are more desirable because they indicate better balance on h. If the experiment is completely randomized and no rerandomization done, the smallest possible p-value is $1 \div \binom{8}{4} = 0.015$ and the average m value is 1.4. If the experiment is pair matched and complete randomization occurs within pairs for whom $h \in \{1, 2\}, ..., h \in \{7, 8\}$, the average m improves to 0.375. However, the minimum possible p-value in randomization inference is now $1 \div \binom{2}{1}^4 = 0.063$. The minimum possible p-value does not vary linearly with the maximum acceptable m in completely randomized experiments (see Table 1).

Table 1: A simple example: How does one's balance threshold influence uncertainty calculations?

Maximum acceptable balance score	Minimum p-value
0	0.125
1	0.028
2	0.018
3	0.015
4	0.014

This example illustrates that low acceptance thresholds can invalidate the estimation of uncertainty, in the sense that the minimum possible p-value will eventually increase above 0.05. In addition, the minimum possible p-value varies nonlinearly with the level of strictness the researcher adopts accepting randomizations. These considerations suggest that it may be possible to systematically establish an acceptance criterion which will minimize the minimum possible p-value while maximizing the balance improvement from rerandomization.

An important theoretical principle here is that

$$\text{Minimum } p\text{-value} = \frac{1}{\text{\# acceptable randomizations}}. \tag{3}$$

This expression is true because the exact p-value is defined as the fraction of hypothetical randomizations which would show results as or more extreme than the observed value. If there are r acceptable randomizations, then at least 1 of the r randomizations is as or more extreme than the observed value, so the minimum fraction is $1/r$. When the set of acceptable randomizations gets smaller, the minimum p-value invariably gets larger.

This discussion leads naturally to **Theorem 1**. **Theorem 1** finds the expected minimum p-value from rerandomization. The expectation averages over randomness caused by the fact that the number of acceptable randomizations is not a fixed quantity. Indeed, unconditionally, the decision to accept or reject a randomization is itself random and governed by the acceptance probability, p_a. More informally, with a specified imbalance criterion, we do not before the sample is drawn know how many randomizations are acceptable given that criterion. But we can specify a threshold such that, *a priori*, the probability of acceptance is known and fixed (see Morgan and D. B. Rubin (2012)).

Theorem 1. The expected minimum p-value from rerandomization is

$$\begin{aligned}\mathbb{E}\left[\text{Minimum } p\text{-value}\right] &= \mathbb{E}\left[\frac{1}{\text{\# acceptable randomizations}}\right] \\ &= n_{\text{Cand}} p_a (1-p_a)^{n_{\text{Cand}}-1}\left(1-(1-p_a)^{n_{\text{Cand}}}\right) \\ &\quad \times {}_3F_2\Big(\{1,1,1-n_{\text{Cand}}\},\ \{2,2\},\ p_a/(p_a-1)\Big)\end{aligned} \tag{4}$$

where p_a denotes the acceptance probability,

where n_{Cand} denotes the total number of candidate randomizations ($n_{\text{Cand}} = 2^k$),

and where ${}_3F_2$ denotes the generalized hypergeometric function with $p = 3$ and $q = 2$.

Proof of Theorem 1. See Appendix.

Setting $n_{\text{Cand}} = \binom{10}{5}$ in Figure 1, we see the clear non-linear relationship between p_a and the expected minimum p-value. This figure is consistent with the later Monte Carlo results of Figure 2.i.

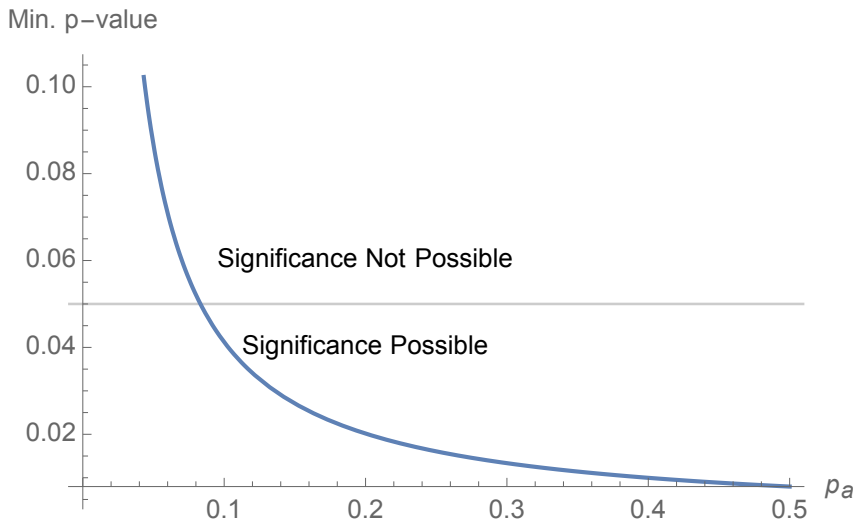


Figure 1: With n_{Cand} held fixed, we can plot the relationship between p_a and the expected minimum p-value.

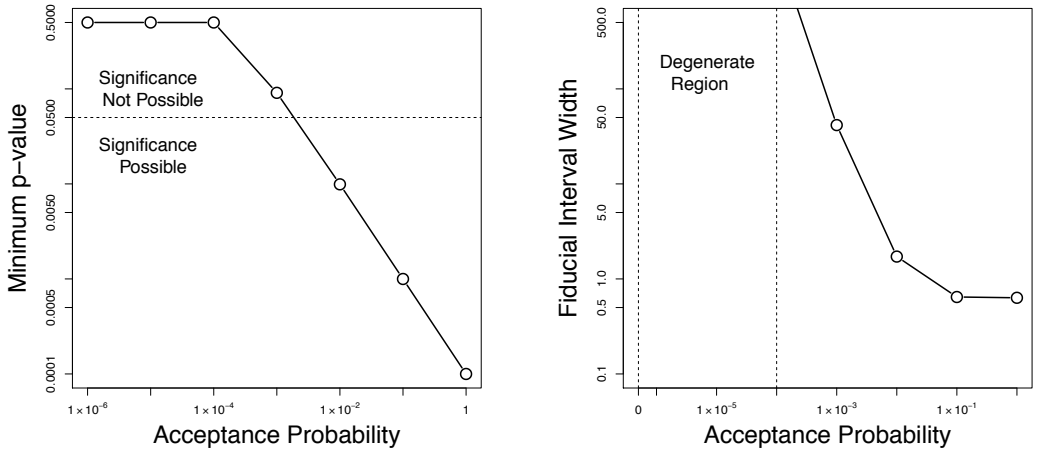
The expression in **Theorem 1** shows how, as experimenters reduce p_a, fiducial intervals get wider as well, since it becomes more and more difficult to reject $H_0 : \tau = \tau_0$ for any τ_0. In the extreme, if the minimum p-value is above α, randomization-based fiducial interval will degenerate to $(-\infty, \infty)$.

For the sake of illustration, we simulate the case where there are 6 covariates and $n = 10$ units in a completely randomized experiment. Let $Y_i(0) = f(x_i)$ and $Y_i(1) = f(x_i)+\tau$, where $f(x_i)$ is a linear function. For each candidate randomization, we calculate M, where

$$\begin{aligned} M &\equiv \left(\bar{\mathbf{X}}_T - \bar{\mathbf{X}}_C\right)' \left[\text{Cov}\left(\bar{\mathbf{X}}_T - \bar{\mathbf{X}}_C\right)^{-1}\right] \left(\bar{\mathbf{X}}_T - \bar{\mathbf{X}}_C\right); \\ &= np_w\left(1-p_w\right)\left(\bar{\mathbf{X}}_T - \bar{\mathbf{X}}_C\right)' \text{Cov}(\mathbf{x})^{-1} \left(\bar{\mathbf{X}}_T - \bar{\mathbf{X}}_C\right), \end{aligned} \tag{5}$$

where p_w denotes the fixed proportion of treated units and $\text{Cov}(\mathbf{x})$ denotes the sample covariance matrix of $\mathbf{X}$. Then, we accept M with probability p_a based on the inverse CDF of M. Due to the Multivariate Normality of $\mathbf{X}$, we know that $M \sim \chi^2_k$.

In Figure 2.i, the minimum p-value decreases non-linearly as a function of the acceptance probability (p_a), which mirrors the theoretical results obtained in **Theorem 1**. This phenomenon has implications for inference: Figure 2.ii shows how the width of the resulting fiducial intervals gets wider as p_a gets closer to 0. In fact, the interval width goes to ∞ because the size of the rerandomization set is now too small. The exact threshold when degeneracy is reached will vary with the experimental design, but the general pattern is theoretically inevitable: if we shrink the size of the set in which a random choice gets made, we also shrink the reference set for assessing uncertainty.



(a) (b)

Figure 2: (a) As the acceptance probability decreases, the minimum p-value increases until no statistical significance at $\alpha = 0.05$ is possible. (b) For experiments in the "degenerate region", the acceptance probabilities are such that the fiducial intervals have infinite width. Note the use of the log scale in both figures.

We can state these ideas more formally as follows.

Notation. Let $\mathcal{R}$ denote the set of all possible ways of assigning N observational units to treatment or control. Let $\mathcal{A}$ denote the set of all algorithms which take $\mathcal{R}$ as an input and give as an output a reduced set of acceptable assignments, $\mathcal{R}_A$. A subset of $\mathcal{A}$, denoted $\mathcal{A}^1$ produces an $\mathcal{R}_A$ such that $|\mathcal{R}_A| = 1$. This subset contains Kasy (2016)'s algorithm as a member, as well as any algorithm for returning a strictly optimal treatment vector, where the definition of "optimal" is arbitrary.

Theorem 2. Any member of the set, $\mathcal{A}^1$, yields infinitely wide fiducial intervals and has no ability to reject any null hypothesis using randomization tests.

Proof of Theorem 2. By construction, the minimum possible p-value of non-parametric randomization tests for any member of $\mathcal{A}^1$ is

$$\text{Minimum } p\text{-value} = \frac{1}{\text{\# acceptable randomizations}}. \tag{6}$$

When $|\mathcal{R}_A| = 1$ by definition, we have

$$\text{Minimum } p\text{-value when } |\mathcal{R}_A| \text{ is } 1 = \frac{1}{1} = 1 > 0.05. \tag{7}$$

Thus, all p-values will be greater than 0.05 for $\mathcal{A}^1$, directly implying no ability to reject any null hypothesis. Moreover, by the construction of intervals from tests, it follows that every candidate endpoint of the fiducial interval will fail to be rejected. As a consequence, the fiducial interval will span the entire range of the outcome variable and will be infinitely

wide for continuous outcomes. In sum, optimal treatment assignment leads to degenerate exact tests and confidence intervals.

To further illustrate this point, it is instructive to consider the literature on matching, which contains a degeneracy of a similar nature. In that context, researchers have to decide how much "dissimilarity" to allow between matches. This dissimilarity is often set with a caliper. Cochran and D. B. Rubin (1973), Austin (2011), and others have made recommendations about the optimal caliper size in the context of propensity score matching. These recommendations seek to reduce the imbalance as much as possible without ending up with no actual matches: if the caliper were exactly 0 and if the sample space were continuous, the analysis would be degenerate as no units could be matched. In both matching and in rerandomization, then, we would like to create a situation where treated and control units are as similar as possible to one another. In both cases, when we increase the similarity of treated and control units, we also incur a certain cost. The nature of this cost differs in the two cases: in matching, a tighter caliper leads to a smaller sample size and thus higher estimator variability, whereas in rerandomization, a lower acceptance probability leads to widening fiducial intervals. Nevertheless, it could be instructive for future work to consider such similarities in greater detail and to consider shared principles that may be at play.

4 Selecting p_a

With these results in mind, how should one go about selecting p_a? The choice of p_a involves a triple tradeoff between better balance, degeneracy risk, and computational time to randomization acceptance. The first two issues have been discussed already. The third

issue is relevant because the waiting time until an acceptable randomization is distributed according to a Geometric(p_a) distribution. In general, we would expect to generate $1/p_a$ randomizations before accepting a single one—an issue to which we shall later return.

4.1 An *a priori* threshold

As Morgan and Rubin write, "for small samples, care should be taken to ensure the number of acceptable randomizations does not become too small, for example, less than 1000" (p. 7). This view suggests that p_a should be set as low as possible so long as the number of acceptable randomizations is greater than a fixed threshold that is determined *a priori*. In this sense, one decision rule might be to set p_a such that

$$\mathbb{E}\left[\text{minimum } p\text{-value} \middle| p_a \right] = \beta. \tag{8}$$

The value of p_a which yields β can easily be found by inverting the results of **Theorem 1**. On the one hand, it is difficult to know whether a given threshold is reasonable given the number of units available, the number of covariates observed pre-treatment, and the computational resources at researchers' disposal. On the other hand, a threshold is easy to interpret and does not involve additional optimization steps.

4.2 Explicit Tradeoff

We could also propose an alternative procedure to determine a p_a which explicitly trades off the costs and benefits of rerandomization. Let $e_{p_a}(p_a) = e_a(a) \equiv \mathbb{E}[\text{minimum } p\text{-value}\big|p_a]$. Let $v_{p_a}(p_a) = v_a(a) \equiv v_a$, where v_a loosely denotes the remaining variance between treat-

ment and control covariate means on a percentage basis. This double notation allows us to emphasize the duality between setting a and p_a, where p_a is defined as the acceptance probability and a is defined as the acceptance threshold for M (see Morgan and D. B. Rubin (2012) for further details). We want both $\mathbb{E}[\text{minimum } p\text{-value} \big| p_a]$ and v_a to be small: when the minimum p-value is small, we can form tight fiducial intervals; when v_a is small, we achieve a large reduction in variance for each covariate. Furthermore, both $e_{p_a}(p_a)$ and $v_{p_a}(p_a)$ are bounded between 0 and 1. Thus, we want to set p_a (or, equivalently a) such that

$$p_a^* := \arg\min_{p_a} \quad w \cdot e_{p_a}(p_a) + (1 - w) \cdot v_{p_a}(p_a), \tag{9}$$

where w determines the tradeoff between the variance reduction and the minimum possible p-value. It is possible to derive the gradients for this optimization problem (using either automatic differentiation or analytical methods), which can be fed into gradient descent solvers. Moreover, given a target minimum p-value, we can also back out the w value that this implies. Thus, this framework may not only be useful for selecting the acceptance threshold, but also for evaluating threshold choices and examining their implications.

The optimization framework described here has several attractive features. First, when $w = 0.5$ and there is only one covariate, the expected minimum p-value will always be less than 0.05. Second, conditional on k, p_a^* describes the same relative place on the objetive function no matter the sample size, so in this sense is comparable across experiments. Finally, p_a^* has a nice interpretation in that it captures the point at which we have gained most of the variance-reducing benefits of rerandomization but have not incurred significant inferential costs.

5 A New Sampling Algorithm For Generating Rerandomization Sets: Introducing Optimization-Assisted Rerandomization (OAR)

The preceding discussion describes a theoretical problem—that of a vanishing ability to reject the null hypothesis using randomization tests. Computational speedups alone will not make this problem severe in large experiments. For example, in a sample of 1000 units, an *a priori* expected minimum p-value of $1/1000$ would imply that only 1000 out of all possible 2.70×10^{299} randomizations would be considered acceptable. If every atom in the observable universe could generate a hundred randomizations per second, it would still take 8.56×10^{190} years for all the randomizations to be enumerated.

We present in this section an approximate rerandomization algorithm which makes the vanishing power problem more relevant even for larger experiments. In this approach, we place guarantees on the uniformity of the resulting candidate randomizations while speeding up the waiting time to acceptance. The speedup comes through an explicit optimization step. The simple approach described below is an intuitive and attractive way to perform approximate rerandomization. Indeed, to perform an exact test following rerandomization, we must generate a large pool of acceptable randomizations. The protocol described below can greatly expedite that process.

To put the problem more formally, we are trying to uniformly sample from the set of all binary vectors which would meet the experimenter's acceptance criterion. Experimenters can then use this set to perform randomization tests or for assigning subjects to treatment conditions. Let $\mathcal{R}_A$ denote the set of all acceptable randomizations. What does

it mean to uniformly sample from $\mathcal{R}_A$, conditional on $\mathbf{X}$? It means that that, conditional on $\mathbf{X}$, we have a protocol to select any given acceptable randomization with probability $1/|\mathcal{R}_A|$, where acceptance is no longer probabilistic but deterministic based on some quality function, which will often be M as defined above.[2]

Standard rerandomization as described in the original rerandomization paper is a protocol which allows us to sample acceptable randomizations with uniform probability, $1/|\mathcal{R}_A|$. In that algorithm, we first select a candidate assignment vector, V. We then calculate the imbalance statistic, which we here denote as $M(V)$ to emphasize the dependance of the imbalance statistic, M, on the assignment vector V, conditional on $\mathbf{X}$. If $M(V)$ is less than some threshold value a, we place V into our sample rerandomization set, $\hat{\mathcal{R}}_A$. It is important to emphasize again that a is, through its CDF, associated with an unconditional acceptance probability, p_a. We can repeat this process until a rerandomization set of a target size has been reached. Denote this target size by T. The notation $\mathbf{X}_T(V)$ and $\mathbf{X}_C(V)$ is simply meant to emphasize that the background covariate matrices of the treated and control units depend on V.[3]

Algorithm 1 Standard Rerandomization (SR)

function SR($\mathbf{X}, T$)
  while $|\hat{R}|_A$ **do** $< T$
    for $j \leftarrow 1$ to n **do**
      Draw $V_i \sim$ Bernoulli(0.5)
      $M(V) \leftarrow \left(\bar{\mathbf{X}}_T(V) - \bar{\mathbf{X}}_C(V)\right)' \left[\text{Cov}\left(\bar{\mathbf{X}}_T(V) - \bar{\mathbf{X}}_C(V)\right)^{-1}\right] \left(\bar{\mathbf{X}}_T(V) - \bar{\mathbf{X}}_C(V)\right)$
      if $M(V) < a$ **then**
        Add V to $|\hat{R}|^{\mathcal{A}}$
  return $|\hat{R}|_A$

[2]For completeness, note that $\mathbb{E}[|\mathcal{R}_A|] = p_a \cdot n_{\text{Cand}}$.

[3]There is a subtle interplay between statements made conditional and unconditional on $\mathbf{X}$, which we hope is clear throughout despite suppressing some of these statements to maintain notational simplicity.

Theorem 3a. Standard rerandomization yields samples of $\mathcal{R}_A$ that are uniform. The expected waiting time to reach k acceptable samples is k/p_a.

Proof of Theorem 3a. See Appendix.

The uniform sampling of acceptable randomizations is an important assumption in rerandomization-based tests. If it is violated, researchers in general have no guarantees on the resulting p-values from Fisherian inference.

Optimization-Assisted Rerandomization (OAR) is similar to standard rerandomization but contains a crucial extra step. In OAR, we randomly assign treatment to some number of units (κ) and assign the remaining treatments in a way that minimizes $M(V)$. In what follows, $BB(\cdot)$ denotes a Black Box optimization routine which takes as an input a partially completed assignment vector and then selects the remaining assignments which optimize $M(V)$. We show that OAR does no worse than rerandomization in gathering rerandomization samples (and experimentally show that it can be up to 90% faster). The precise magnitude of the waiting time speedup is difficult to quantify without imposing more structure on the nature of the optimization scheme. At the same time, we show that OAR has theoretical guarantees which prevent the OAR samples from diverging far from a uniform distribution over the set of acceptable randomizations.

Theorem 3b. OAR yields samples of $\mathcal{R}_A$ with probabilities no further in L_1 distance from uniform than the following:

$$\max\left(\frac{1}{\mathcal{R}_A},\ \left|\frac{1}{(\frac{|\mathcal{R}_A|}{n_{\text{Cand}}}-s)\cdot 2^{|\mathcal{U}|}}-1/|\mathcal{R}_A|\right|\right).$$

Algorithm 2 Optimization-Assisted Randomization (OAR).

function OAR($X, \kappa, t, B(\cdot)$)
  while $|\hat{R}|^{\mathcal{A}} < T$ **do**
    Uniformly draw $\mathcal{J}$ from $\{i\}_1^n$ such that $|\mathcal{J}| = \kappa$
    for j in $\mathcal{J}$ **do**
      Draw $V_j \sim$ Bernoulli(0.5)
    $V \leftarrow \text{BB}(\{V_j\}_{j \in \mathcal{J}})$
    $M(V) \leftarrow \text{Mahal}(\mathbf{X}({}_T V), \mathbf{X}_C(V))$
    if $M(V) \leq a$ **then**
      Add V to $|\hat{R}|_A$
  return $|R|^{\mathcal{R}}$

Moreover, the maximum sampling probability in OAR compared to the maximum sampling probability in uniform sampling of acceptable randomizations is less in L_1 distance than

$$\left| \frac{1}{(\frac{|\mathcal{R}_A|}{n_{\text{Cand}}} - s) \cdot 2^{|\mathcal{U}|}} - 1/|\mathcal{R}_A| \right|$$

Waiting time speedups are proportional to the quality of the optimization algorithm.

Proof of Theorem 3b. See Appendix.

In sum, OAR allows experimenters to push p_a very low while maintaining theoretical guarantees on the sampling properties of the algorithm. These properties are very important to show analytically since non-parametric tests hinge on researcher's ability to uniformly sample from the set of acceptable randomizations.

5.1 Experimental Validation

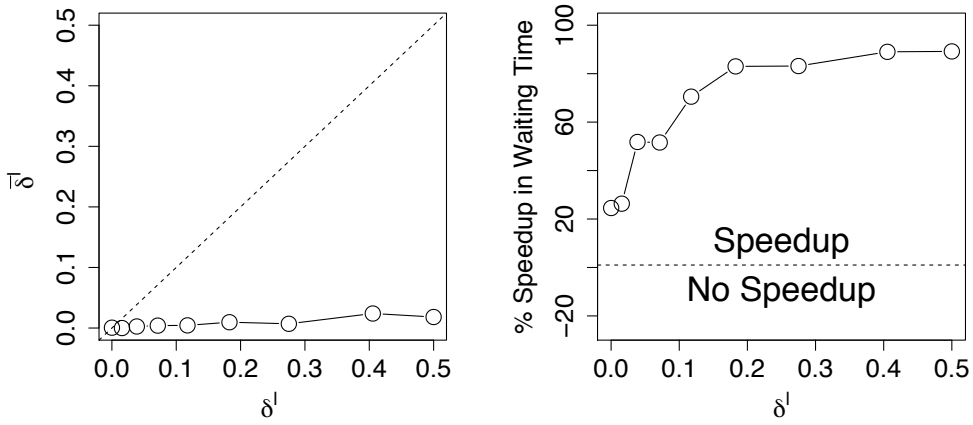
To illustrate this simple algorithm in action, we created a synthetic experiment which allows us to compare the difference between the optimization-assisted sampling and purely random acceptance sampling of assignment vectors. For 9 values of δ' ranging between

5×10^{-1} and 10^{-5}, we repeatedly generated pre-treatment covariate datasets, $\mathbf{X}$, from a Multivariate Gaussian with a mean parameter of $0_{3\times1}$ and an identity covariate matrix $I_{3\times3}$. The number of units in each synthetic experiment was 10, with 5 units receiving treatment and 5 receiving control. The treatment effect was a constant value of 0.25 and the background covariates were linearly related to the outcome. Then, we selected a sample of acceptable treatment assignment vectors ($p_a = 0.10$) using both standard rerandomization and OAR. In OAR, we used a k of n genetic algorithm for optimizing $M(V)$. We compared $\max_v \Pr(V = v|M(V) \leq a)$ in the two cases. The value for OAR should be within a known factor of that for standard rerandomization, which itself should equal to $1/|\mathcal{R}_A|$.

The simulation results confirm the theoretical analysis and imply that even stronger guarantees on performance may be found in the future. In Figure 3.i, we see that there is very good match between the theory discussed in the previous section with the experimental results: the allowable deviance in OAR from uniform sampling is a close and conservative approximation for the true deviation from uniformity. In Figure 3.ii, we see that the waiting time to reach the 50^{th} randomization is greatly improved in OAR compared to standard rerandomization. OAR seems to perform better than the theoretical worst-case, so it is possible that this worst-case is not the highest upper bound but could be improved. We also observe a tradeoff between obtaining wait-time speedups and δ', meaning that faster generation of acceptable randomizations leads to weaker guarantees on the uniformity of the resulting sample. The weaker guarantees mean that exact tests using the randomization set could be less accurate.

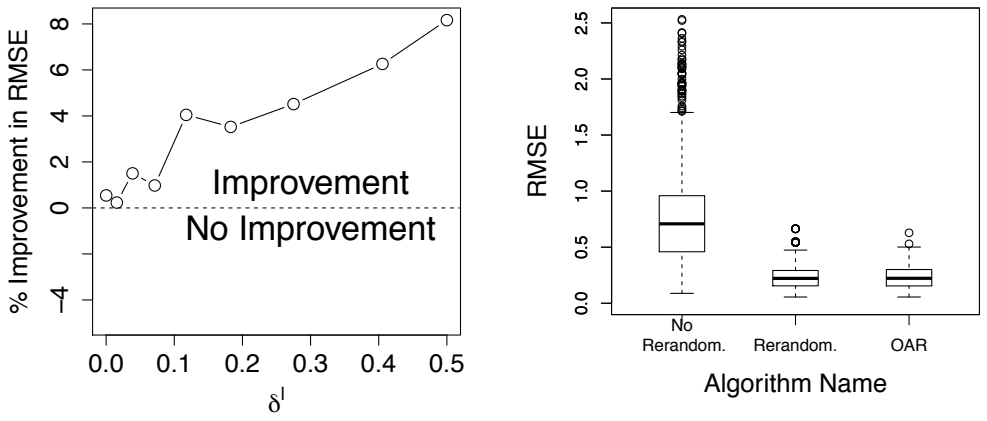
Finally, we can examine the Root Mean Squared Error (RMSE) in the treatment ef-

fect estimation. Figure 4.i shows that OAR yields lower RMSE compared to standard rerandomization, likely because the effective p_a was effectively lowered by the optimization step. Figure 4.ii reaffirms the general importance of rerandomization: both standard rerandomization and OAR lead to substantially lower RMSE compared to the case when no rerandomization is used. This illustrates how rerandomization can help experimenters arrive at better inferences, since it reduces noise around the estimated treatment effect. Taken together, these simulation results suggest that OAR can be profitably applied in future experiments because it allows researchers to quickly obtain large sets of acceptable randomizations.



(a) (b)

Figure 3: (a) The allowable deviance in OAR from uniform sampling is a close and conservative approximation for the true deviation from uniformity. (b) We find a significant speedup to obtain 50 acceptable samples using OAR compared to standard rerandomization.



(a) (b)

Figure 4: (a) We see gains in Root Mean Squared Error (RMSE) using OAR compared to standard rerandomization, especially when δ' grows (since in this case the imbalance between treatment classes will be optimized with more degrees of freedom). (b) Overall, we find a large decrease in RMSE when using standard rerandomization and OAR (with $\delta = 10^{-5}$) compared to the case when no rerandomization is done.

6 Conclusion

In this paper, we formalize the idea that rerandomization can be taken "too far," in the sense that the variance-reduction benefits of rerandomization can be outweighed by the costs of having undesirably wide permutation-based fiducial intervals when acceptance probabilities are low. We accomplish this formalization by deriving the expected minimim possible p-value, which turns out to be a function of the acceptance threshold and the number of units.

We then show that one attractive option for choosing an acceptance threshold is to set p_a equal to the value that minimizes the sum of v_a (which controls variance reduction) and the expected minimum p-value (which controls the width of fiducial intervals). We also devised an algorithm—optimization-assisted rerandomization (OAR)—for generating acceptable randomizations far faster than previous approaches while guaranteeing that this set is close to identical to the full set. We believe that this algorithm may make rerandomization an even more attractive approach for researchers, as they can push p_a to new lows (while understanding the risks of going too low).

We have contributed to the understanding of rerandomization as a kind of slippery slope: at the top of this slope, all randomizations are accepted and, at the bottom, only one randomization is accepted but it becomes impossible to estimate uncertainty using important non-parametric tests. Our paper is thus an attempt at helping experimenters to understand where they are on this slope, where they should be, and how they can get to that target.

7 Appendix

7.1 Proof of Theorem 1

Recall from the above that we can define the minimum p-value as

$$\text{Minimum } p\text{-value} = \frac{1}{\text{\# acceptable randomizations}}. \tag{10}$$

It is tempting to conclude that $(\text{\# acceptable randomizations}) \sim \text{Binomial}(p_a, n_{\text{Cand}})$ because n_{Cand}, which represents the number of candidate randomizations, is always a fixed quantity, as is p_a, which represents the unconditional acceptance probability of randomization acceptance.[4] However, this is not strictly speaking true, at least if we assume that at least one randomization is accepted so that the experiment is actually able to be conducted. We therefore condition on at least 1 randomization being accepted. To find the distribution of this quantity, we momentarily simplify the notation and consider

$$\begin{gathered} X \sim \text{Binomial}(n, p). \\ \& \\ \text{\# acceptable randomizations} = X|X > 1. \end{gathered} \tag{11}$$

[4] Recall that, following Morgan and D. B. Rubin (2012), we know $M \sim \chi^2_k$ under pure randomization and if n units is large enough. Thus, we can move between p_a and the threshold a using the χ^2_k inverse CDF, since $\Pr(M \leq a) = p_a$.

Moreover,

$$\begin{aligned}
\Pr(X = x | I\{X \geq 1\}) &= \frac{\Pr(I\{X \geq 1\}|X = x) \cdot \Pr(X = x)}{\Pr(I\{X \geq 1\})}; \\
&= I\{x \geq 1\} \frac{\Pr(X = x)}{\Pr(X \geq 1)}; \\
&= I\{x \geq 1\} \frac{\Pr(X = x)}{1 - \Pr(X < 1)} = I\{x \geq 1\} \frac{\Pr(X = x)}{1 - \Pr(X = 0)}; \\
&= I\{x \geq 1\} \frac{\Pr(X = x)}{1 - \binom{n}{0} p^0 (1-p)^{n-0}}; \\
&= I\{x \geq 1\} \frac{\binom{n}{x} p^x (1-p)^{n-x}}{1 - (1-p)^n};
\end{aligned} \tag{12}$$

Now, the expected minimum p-value will be equal to

$$\begin{aligned}
\mathbb{E}[1/X | X \geq 1] &= \sum_{x=0}^{\infty} \frac{1}{x} \cdot \Pr(X = x | X \geq 1); \\
&= \sum_{x=0}^{\infty} \frac{1}{x} \cdot I\{x \geq 1\} \frac{\binom{n}{x} p^x (1-p)^{n-x}}{1 - (1-p)^n}; \\
&= \sum_{x=1}^{\infty} \frac{1}{x} \cdot \frac{\binom{n}{x} p^x (1-p)^{n-x}}{1 - (1-p)^n}; \\
&= (1 - (1-p)^n) \cdot \sum_{x=1}^{\infty} \frac{1}{x} \binom{n}{x} p^x (1-p)^{n-x}; \\
&= (1 - (1-p)^n) \cdot \sum_{x=1}^{\infty} \frac{1}{x} \binom{n}{x} p^x (1-p)^{n-x}; \\
&= np(1-p)^{n-1} (1 - (1-p)^n) \, {}_3F_2\big(\{1, 1, 1-n\},\ \{2, 2\},\ p/(p-1)\big).
\end{aligned} \tag{13}$$

There is a closed-form expression for the generalized hypergeometric function used, but it is messy and we do not report it here (interested readers can consult Gottschalk and

Maslen (1988), especially p. 1996). From this, we conclude

$$\mathbb{E}\Big[\text{Minimum } p\text{-value}\Big] = n_{\text{Cand}} p_a (1-p_a)^{n_{\text{Cand}}-1}\big(1-(1-p_a)^{n_{\text{Cand}}}\big) \times {}_3F_2\big(\{1,1,1-n_{\text{Cand}}\},\ \{2,2\},\ p_a/(p_a-1)\big) \tag{14}$$

7.2 Proofs of Theorems 3a and 3b

7.2.1 Proof of Theorem 3a

For standard rerandomization, each accepted V is equally likely to have been drawn (since each V, accepted or not, was drawn with probability $1/n_{\text{Cand}}$). Recall that, if all outcomes are equally likely, the probability of the event $\{V = v\}$ given $M(V) \leq a$ is equal to the number of outcomes in which that event occurs (which is 1) divided by the total number of outcomes (which is $|\mathcal{R}_A|$). This discussion means

$$\Pr(V = v | I\{M(V) \leq a\}) = \frac{1}{|\mathcal{R}_A|}.$$

Another way to see this is to consider how

$$\begin{aligned}\Pr(V = v | I\{M(V) \leq a\}) &= \frac{\Pr(I\{M(V) \leq a\} | V = v) \times \Pr(V = v)}{\Pr(I\{M(V) \leq a\})}; \\ &= \frac{I\{M(v) \leq a\} \times 2^{-k}}{\frac{|\mathcal{R}_A|}{n_{\text{Cand}}}} = \frac{1 \times \frac{1}{n_{\text{Cand}}}}{\frac{|\mathcal{R}_A|}{n_{\text{Cand}}}} = 1/|\mathcal{R}_A|,\end{aligned}$$

where the last line uses the fact that $n_{\text{Cand}} = 2^k$. The average waiting time for the first acceptance is governed by a Geometric distribution with probability parameter p_a. By the independence of the sampling process, the waiting time until the k^{th} acceptance is given by k/p_a.

7.2.2 Proof of Theorem 3b

The probability of yielding any accepted sample, V, is more complex in OAR than for standard rerandomization. The notation, V_i is meant to indicate the i^{th} slot of V. Moreover, $\{j\}_{j\in\mathcal{D}}$ denotes the set of all indices which are found deterministically, conditional on $\{i\}_{i\in\mathcal{U}}$, which denotes the set of all indices which receive treatment or control status with equal (and therefore uniform) probability.

$$
\begin{aligned}
\Pr(V = v | I\{M(V) \le a\}) &= \Pr(V_1 = v_1, ..., V_n = v_n | I\{M(V) \le a\}) \\
&= \Pr(\{V_k = v_k\}_{k\in\mathcal{D}}\ ,\ \{V_j = v_j\}_{j\in\mathcal{U}} \ \big|\ I\{M(V) \le a\}); \\
&= \Pr(\{V_k = v_k\}_{k\in\mathcal{D}} \ \big|\ \{V_j = v_j\}_{j\in\mathcal{U}}, I\{M(V) \le a\}) \\
&\qquad \times \Pr(\{V_j = v_j\}_{j\in\mathcal{U}} \ \big|\ I\{M(V) \le a\}); \\
&= I\{M(v) \le a\} \times \frac{\Pr(M(V) \le a | \{V_j = v_j\}_{j\in\mathcal{U}}) \Pr(\{V_j = v_j\}_{j\in\mathcal{U}})}{\Pr(M(V) \le a)}; \\
&= I\{M(v) \le a\} \times \frac{\Pr(M(V) \le a \ \big|\ \{V_j = v_j\}_{j\in\mathcal{U}})\ 2^{-|\mathcal{U}|}}{\Pr(M(V) \le a)}; \\
&= I\{M(v) \le a\} \times \frac{\Pr(M(V) \le a \ \big|\ \{V_j = v_j\}_{j\in\mathcal{U}})\ 2^{-|\mathcal{U}|}}{\frac{|\mathcal{R}_A|}{n_{\text{Cand}}} - s}; \\
&\le \frac{1}{(\frac{|\mathcal{R}_A|}{n_{\text{Cand}}} - s) \times 2^{|\mathcal{U}|}},
\end{aligned}
$$

where s denotes the "speedup" factor in obtaining samples with $M(V) \le a$ using the selected Black Box optimizer. An implication of these results is that the L_1 distance between the uniform sampling probability and the OAR sampling probability can be written as

$$
\delta := d_{L_1}\big(\Pr(V = v | M(V) \le a),\ \ 1/|\mathcal{R}_A|\big) \le \overline{\delta} := \max\left(\frac{1}{\mathcal{R}_A},\ \ \left|\frac{1}{(\frac{|\mathcal{R}_A|}{n_{\text{Cand}}} - s) \times 2^{|\mathcal{U}|}} - 1/|\mathcal{R}_A|\right|\right)
$$

By controlling $|U|$, we can bring the sampling probability to within a small fraction of that which can be achieved under the uniform sampling of acceptable randomizations. By the same logic,

$$\delta' := d_{L_1}\big(\max_v \Pr(V = v|M(V) \leq a),\ \ 1/|\mathcal{R}_A|\big) \leq \overline{\delta}' := \left|\frac{1}{(\frac{|\mathcal{R}_A|}{n_{\text{Cand}}} - s) \times 2^{|\mathcal{U}|}} - 1/|\mathcal{R}_A|\right|$$

Again, by controlling $|U|$, we can bring the maximum sampling probability of OAR as close as desired to the maximum sampling probability implied by the uniform sampling of acceptable randomizations (which is just $1/|\mathcal{R}_A|$.

Moreover, consider the waiting time for the first acceptance under OAR, which is defined to be $\Pr(M(V) \leq a)$. For standard rerandomization, this waiting time is p_a. Here, we specifically optimize $M(V)$, implying by construction that $\Pr(M(V) \leq a) = p_a - s \leq p_a$.

7.3 Kasy (2016) in the Rerandomization Framework

We argued in the above that Kasy's optimal assignment procedure is a special case of rerandomization in which the only acceptable randomization is the one that provides optimal covariate balance. Let the $\mathbf{x}$ matrix include all observed covariates that the experimenter seeks to balance. Let $\mathbf{W}$ denote the n-dimensional treatment assignment vector indicating the treatment group for each unit. In the notation of Morgan and Rubin, the

rerandomization criterion can be written as

$$\varphi(\mathbf{x}, \mathbf{W}) = \begin{cases} 1, & \text{if } \mathbf{W} \text{ is an acceptable randomization,} \\ 0, & \text{if } \mathbf{W} \text{ is not an acceptable randomization.} \end{cases}$$

With the appropriate $\varphi(\cdot, \cdot)$, Kasy's non-deterministic assignment procedure can be considered a special case of rerandomization. That is, we will obtain Kasy's optimal assignment if we define $\varphi(\cdot, \cdot)$ as follows, letting $R(\cdot)$ denote either Bayesian or minimax risk, letting $\hat{\beta}$ denote the choice of estimator, and letting U denote a generic randomization procedure independent of $\mathbf{x}$ and $\mathbf{Y}$,

$$\varphi(\mathbf{x}, \mathbf{W}) = \begin{cases} 1, & \text{if } \mathbf{W} \text{ minimizes } R(\mathbf{W}, \hat{\beta} | \mathbf{x}, U), \\ 0, & \text{if } \mathbf{W} \text{ does not minimize } R(\mathbf{W}, \hat{\beta} | \mathbf{x}, U). \end{cases}$$

With n units, we will obtain a $\mathbf{W}$ such that $\varphi(\mathbf{x}, \mathbf{W}) = 1$ with probability $(1/2)^n$. The expected wait-time until obtaining this $\mathbf{W}$ is given by a Geometric distribution, with mean 2^n. This calculation implies that, as soon as $n \geq 20$, the expected wait-time will exceed 1 million rerandomization draws if standard rerandomization is used (Soares and Wu, 1985).

This section has put our discussion in closer dialogue Kasy's work in order to further illustrate how Kasy's deterministic procedure for optimal treatment assignment is a special case of rerandomization developed in Morgan and D. B. Rubin (2012).

References

Austin, Peter C (2011): "Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies". In: *Pharmaceutical Statistics*, no. 2, vol. 10, pp. 150–161. ISSN: 1539-1604. DOI: 10.1002/pst.433. URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3120982/ (visited on 12/03/2015).

Cochran, William G. and Donald B. Rubin (1973): "Controlling Bias in Observational Studies: A Review". In: *Sankhya: The Indian Journal of Statistics, Series A (1961-2002)*, no. 4, vol. 35, pp. 417–446. ISSN: 0581-572X. URL: http://www.jstor.org/stable/25049893 (visited on 12/03/2015).

Garthwaite, Paul H. (1996): "Confidence Intervals from Randomization Tests". In: *Biometrics*, no. 4, vol. 52, pp. 1387–1393. ISSN: 0006-341X. DOI: 10.2307/2532852. URL: http://www.jstor.org/stable/2532852 (visited on 12/03/2015).

Gottschalk, J.E. and E.N. Maslen (1988): "Reduction formulae for generalized hypergeometric functions of one variable". In: *Journal of Physics A: Mathematical and General*, no. 1, vol. 21, pp. 1983–1998.

Imbens, Guido W. and Donald Rubin (2015): *Causal Inference for Statistics, Social, and Biomedical Sciences*. Cambridge: Cambridge University Press. URL: http://www.cambridge.org/US/academic/subjects/statistics-probability/statistical-theory-and-methods/causal-inference-statistics-social-and-biomedical-sciences-introduction (visited on 12/03/2015).

Kasy, Maximilian (2016): "Why Experimenters Might Not Always Want to Randomize, and What They Could Do Instead". In: *Political Analysis*, no. 3, vol. 24, pp. 324–338.

Morgan, Kari Lock (2011): "Rerandomization to Improve Covariate Balance in Randomized Experiments". PhD thesis. Harvard University.

Morgan, Kari Lock and Donald B. Rubin (2012): "Rerandomization to improve covariate balance in experiments". EN. In: *The Annals of Statistics*, no. 2, vol. 40, pp. 1263–1282. ISSN: 0090-5364, 2168-8966. URL: http://projecteuclid.org.ezp-prod1.hul.harvard.edu/euclid.aos/1342625468 (visited on 12/03/2015).

Neyman, J. (1941): "Fiducial Argument and the Theory of Confidence Intervals". In: *Biometrika*, no. 2, vol. 32, pp. 128–150. ISSN: 0006-3444. DOI: 10.2307/2332207. URL: http://www.jstor.org/stable/2332207 (visited on 12/03/2015).

Soares, Jose F. and C.F.J. Wu (1985): "Optimality of random allocation design for the control of accidental bias in sequential experiments". en. In: *Journal of Statistical Planning and Inference*, no. 1, vol. 11, pp. 81–87. ISSN: 03783758. DOI: 10.1016/0378-3758(85)90027-8. URL: http://linkinghub.elsevier.com/retrieve/pii/0378375885900278 (visited on 12/03/2015).