Biometrika (2024), **111**, 1, *pp.* 355–363 https://doi.org/10.1093/biomet/asad027
Advance Access publication 3 May 2023

Power and sample size calculations for rerandomization

By ZACH BRANSON

Department of Statistics and Data Science, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, Pennsylvania 15213, U.S.A.
zach@stat.cmu.edu

XINRAN LI

Department of Statistics, University of Illinois at Urbana-Champaign, 605 E. Springfield Avenue, Champaign, Illinois 61820, U.S.A.
xinranli@illinois.edu

and PENG DING

Department of Statistics, University of California, 425 Evans Hall, Berkeley, California 94720, U.S.A.
pengdingpku@berkeley.edu

Summary

Power analyses are an important aspect of experimental design, because they help determine how experiments are implemented in practice. It is common to specify a desired level of power and compute the sample size necessary to obtain that power. Such calculations are well known for completely randomized experiments, but there can be many benefits to using other experimental designs. For example, it has recently been established that rerandomization, where subjects are randomized until covariate balance is obtained, increases the precision of causal effect estimators. This work establishes the power of rerandomized treatment-control experiments, thereby allowing for sample size calculators. We find the surprising result that, while power is often greater under rerandomization than complete randomization, the opposite can occur for very small treatment effects. The reason is that inference under rerandomization can be relatively more conservative, in the sense that it can have a lower Type-I error at the same nominal significance level, and this additional conservativeness adversely affects power. This surprising result is due to treatment effect heterogeneity, a quantity often ignored in power analyses. We find that heterogeneity increases power for large effect sizes, but decreases power for small effect sizes.

Some key words: Covariate balance; Design-based inference; Dispersive ordering; Experimental design; Treatment effect heterogeneity.

1. Introduction

We consider two-arm randomized experiments, with the aim of estimating causal effects. Randomized experiments are frequently considered the gold standard of causal inference because even simple estimators, such as the average difference in outcomes between groups, are unbiased if subjects are completely randomized between the two groups (Neyman, 1923; Imbens & Rubin, 2015). However, it is often beneficial to use covariate information when randomizing subjects. For example, estimators are usually more precise if subjects are grouped into similar blocks and randomized within blocks (Fisher, 1935; Imai, 2008; Miratrix et al., 2013; Pashley & Miratrix, 2021; Bai, 2022; Tabord-Meehan, 2022). More generally, one can use rerandomization (Morgan & Rubin, 2012), where subjects are

randomized until covariate balance is obtained. Block randomization is a special case, where subjects are randomized until the blocking variable is balanced. However, it is unclear how to form blocks when there are many, possibly continuous, covariates (Bruhn & McKenzie, 2009). Rerandomization can naturally incorporate categorical and continuous covariates in the covariate balance criterion.

Since Morgan & Rubin (2012), many works have established the benefits of rerandomization; this includes experiments with tiers of covariates (Morgan & Rubin, 2015), sequential experiments (Zhou et al., 2018), factorial experiments (Branson et al., 2016; Li et al., 2020), stratified experiments (Wang et al., 2021), experiments with clusters (Lu et al., 2023) and experiments with high-dimensional covariates (Branson & Shao, 2021; Wang & Li, 2022; Zhang et al., 2023). A common theme is that causal effect estimators are more precise under rerandomization than complete randomization as long as covariates are associated with outcomes. In particular, Li et al. (2018) established that, asymptotically, the mean-difference estimator has narrower symmetric quantile ranges under rerandomization than under complete randomization. Thus, confidence intervals for average causal effects are narrower under rerandomization.

Intuitively, because rerandomization increases estimation precision, we would expect it to also increase testing power. While this has been alluded to in the literature (Morgan & Rubin, 2012), the power of rerandomized experiments has not been established. Power analyses are an important aspect of experimental design, because they help determine how experiments are conducted in practice. For example, when designing an experiment, it is common to specify a desired level of power, and then compute the sample size necessary to obtain that level of power (Maxwell et al., 2008; Chow et al., 2017). There are many publicly available sample size calculators for completely randomized experiments (Dupont et al., 1990; Kang, 2021).

This work establishes testing power under rerandomization, thereby allowing for sample size calculators. We focus on the mean-difference estimator, such that we can leverage results from Li et al. (2018). We establish the surprising result that, while power is often greater under rerandomization than complete randomization, the opposite can occur when the average treatment effect is small. The main reason is that inference under rerandomization can be more conservative when there is treatment effect heterogeneity. More precisely, variance estimators for individual effects overestimate the true variance by the same amount under both designs, which results in a larger proportion of overestimation under rerandomization. Specifically, at the same nominal level, testing under rerandomization can have a smaller actual Type-I error than that under complete randomization. This additional conservativeness has an adverse effect on power that can outweigh the precision benefits of rerandomization, but only for small treatment effects.

To compare power and sample size between complete randomization and rerandomization, we establish a dispersive ordering between their respective normal and nonnormal distributions. Our results also quantify how power and sample size are affected by treatment effect heterogeneity, which is often ignored in power analyses. More generally, this work adds to the literature on power analyses for complex experiments, such as two-stage randomized experiments (Jiang et al., 2022), regression discontinuity designs (Schochet, 2009) and difference-in-differences designs (Schochet, 2022). In the Supplementary Material, we illustrate the sample size gains practitioners would see from rerandomization under various scenarios, and also show how to implement power and sample size calculations in our `R` package `rerandPower`.

2. Notation for treatment-control experiments

Consider a treatment-control experiment with N subjects indexed by $i = 1, \ldots, N$. Let $Z = (Z_1, \ldots, Z_N)^{\mathrm{T}}$ denote the binary group indicator, where $Z_i = 1$ denotes treatment and $Z_i = 0$ denotes control, and define $X = (X_1, \ldots, X_N)^{\mathrm{T}}$ as the $N \times K$ covariate matrix. Finally, let $Y_i(1)$ and $Y_i(0)$ denote the potential outcomes for subject i, where $Y_i(1)$ denotes the outcome subject i yields if assigned to $Z_i = 1$, and $Y_i(0)$ is analogously defined. Throughout, we assume that $Y_i(1)$ and $Y_i(0)$ are fixed; observed outcomes are random only to the extent that Z is random.

Because only $Y_i(1)$ or $Y_i(0)$ is observed for each subject, none of the individual treatment effects $\tau_i = Y_i(1) - Y_i(0)$ are fully observed. Nonetheless, average treatment effects, and other estimands,

can still be estimated. We assume that the goal is to well estimate the average treatment effect, defined as $\tau = N^{-1}\sum_{i=1}^{N}\tau_i = \bar{Y}(1) - \bar{Y}(0)$. There are many possible estimators for τ; for simplicity, we focus on the mean-difference estimator $\hat{\tau} = N_1^{-1}\sum_{i=1}^{N} Z_i Y_i(1) - N_0^{-1}\sum_{i=1}^{N}(1 - Z_i) Y_i(0)$, where N_1 and N_0 denote the numbers of treated and control subjects, respectively. We study testing power based on $\hat{\tau}$ under rerandomization. Power will depend on τ, the variance of potential outcomes, and the variance of individual treatment effects, which are respectively defined as

$$S_z^2 = (N-1)^{-1}\sum_{i=1}^{N}\{Y_i(z) - \bar{Y}(z)\}^2 \quad \text{for } z = 0, 1, \qquad S_\tau^2 = (N-1)^{-1}\sum_{i=1}^{N}(\tau_i - \tau)^2.$$

3. Power and sample size under rerandomization

3.1. *Inference and power*

In completely randomized experiments, a fixed number of N_1 subjects is assigned to treatment and a fixed number of N_0 subjects is assigned to control, completely at random (Imbens & Rubin 2015, Ch. 4). Alternatively, the covariates X can inform the design, which will affect inference and power. We consider the rerandomization scheme of Morgan & Rubin (2012), where subjects are completely randomized to treatment until $M \leqslant a$ for a prespecified threshold a, with M the Mahalanobis distance:

$$M = \frac{N_1 N_0}{N}(\bar{X}_1 - \bar{X}_0)^{\mathrm{T}}(S_X^2)^{-1}(\bar{X}_1 - \bar{X}_0).$$

Here $\bar{X}_1$ and $\bar{X}_0$ are K-length vectors of covariate means in the treatment and control groups, respectively, and $S_X^2 = (N-1)^{-1}\sum_{i=1}^{N}(X_i - \bar{X})(X_i - \bar{X})^{\mathrm{T}}$ denotes the covariance matrix of X.

Under complete randomization, the mean-difference estimator $\hat{\tau}$ is unbiased and asymptotically normally distributed (Neyman, 1923; Li & Ding, 2017), i.e.,

$$V^{-1/2}N^{1/2}(\hat{\tau} - \tau) \sim \mathcal{N}(0, 1) \quad \text{where} \quad V = p_1^{-1}S_1^2 + p_0^{-1}S_0^2 - S_\tau^2 \tag{1}$$

with $p_1 = N_1/N$ and $p_0 = N_0/N$. Meanwhile, under rerandomization, $\hat{\tau} \mid M \leqslant a$ follows a non-normal asymptotic distribution that depends on covariates' association with potential outcomes. Although the covariates and potential outcomes are fixed, a linear projection of potential outcomes on covariates can account for some of the potential outcomes' variance, and thus covariates can still have an association with the potential outcomes. To quantify this association, for $z \in \{0, 1\}$, we define the quantities $S_{z|X}^2 = S_{z,X}(S_X^2)^{-1}S_{X,z}$, where $S_{z,X} = S_{X,z}^{\mathrm{T}} = (N-1)^{-1}\sum_{i=1}^{N}\{Y_i(z) - \bar{Y}(z)\}(X_i - \bar{X})^{\mathrm{T}}$, and $S_{\tau|X}^2 = S_{\tau,X}(S_X^2)^{-1}S_{X,\tau}$ where $S_{\tau,X} = S_{X,\tau}^{\mathrm{T}} = (N-1)^{-1}\sum_{i=1}^{N}(\tau_i - \tau)(X_i - \bar{X})^{\mathrm{T}}$. Here, $S_{z|X}^2$ is the variance of the linear projection of potential outcomes on X and $S_{\tau|X}^2$ is analogously defined for treatment effects. The asymptotic distribution of $\hat{\tau}$ under rerandomization is (Li et al., 2018)

$$V^{-1/2}N^{1/2}(\hat{\tau} - \tau) \mid M \leqslant a \sim (1 - R^2)^{1/2}\epsilon_0 + RL_{K,a}, \tag{2}$$

where ϵ_0 and $L_{K,a}$ are independent, $\epsilon_0 \sim \mathcal{N}(0, 1)$, R^2 is the squared multiple correlation between X and potential outcomes, i.e.,

$$R^2 = \frac{p_1^{-1}S_{1|X}^2 + p_0^{-1}S_{0|X}^2 - S_{\tau|X}^2}{p_1^{-1}S_1^2 + p_0^{-1}S_0^2 - S_\tau^2}, \tag{3}$$

and $L_{K,a} \sim \chi_{K,a}S\beta_K^{1/2}$ with the following independent random variables:

$$\chi_{K,a}^2 \sim \chi_K^2 \mid \chi_K^2 \leqslant a, \qquad S \sim -1 + 2 \cdot \mathrm{Ber}(1/2), \qquad \beta_K \sim \mathrm{Be}\{1/2, (K-1)/2\}. \tag{4}$$

The rerandomization distribution (2) and complete randomization distribution (1) are identical only if $R^2 = 0$ or $a = \infty$. Otherwise, (2) will be a nonnormal distribution that has less variance than a standard normal distribution (Li et al., 2018). Both distributions involve V, which depends on the proportions p_1 and p_0 and variances S_1^2, S_0^2 and S_τ^2. The proportions are fixed; meanwhile, S_1^2 and S_0^2 can be consistently estimated, but S_τ^2 cannot without additional assumptions. In practice, we can estimate V conservatively. Neyman (1923) proposed $\hat{V}_{\mathrm{N}} = p_1^{-1}s_1^2 + p_0^{-1}s_0^2$, where s_1^2 and s_0^2 are sample versions of S_1^2 and S_0^2. The estimator $\hat{V}_{\mathrm{N}}$ implicitly estimates treatment effect heterogeneity as $\hat{S}_\tau^2 = 0$ and is thus conservative, in the sense that $E(\hat{V}_{\mathrm{N}} - V) = S_\tau^2 \geqslant 0$. Ding et al. (2019) noted that $S_\tau^2 \geqslant S_{\tau|X}^2$, which can be consistently estimated by $s_{\tau|X}^2 = (s_{1,X} - s_{0,X})(S_X^2)^{-1}(s_{X,1} - s_{X,0})$ (Li et al., 2018), and thus proposed an improved variance estimator $\hat{V}_{\mathrm{DFM}} = \hat{V}_{\mathrm{N}} - s_{\tau|X}^2$. Here, $s_{1,X} = s_{X,1}^{\mathrm{T}}$ and $s_{0,X} = s_{X,0}^{\mathrm{T}}$ are sample analogs of $S_{1,X}$ and $S_{0,X}$, respectively. We consider both estimators and use the generic notation $\hat{V}$. As demonstrated in Li & Ding (2017) and Ding et al. (2019), $\hat{V}$ has a probability limit $\tilde{V}$ no less than the true variance V, i.e., $\hat{V} = \tilde{V} + o_{\mathrm{pr}}(1)$ with $\tilde{V} \geqslant V$. The probability limits of $\hat{V}_{\mathrm{N}}$ and $\hat{V}_{\mathrm{DFM}}$ are respectively

$$\tilde{V}_{\mathrm{N}} = p_1^{-1}S_1^2 + p_0^{-1}S_0^2, \qquad \tilde{V}_{\mathrm{DFM}} = p_1^{-1}S_1^2 + p_0^{-1}S_0^2 - S_{\tau|X}^2.$$

Thus, $\hat{V}_{\mathrm{N}}$ is consistent when $S_\tau^2 = 0$, and $\hat{V}_{\mathrm{DFM}}$ is consistent more broadly when the individual effects can be linearly explained by the covariates.

The rerandomization distribution (2) suggests the $(1-\alpha)$-level confidence interval

$$\hat{\tau} \pm v_{1-\alpha/2}(\hat{R}^2)\hat{V}^{1/2}N^{-1/2}, \tag{5}$$

where $v_{1-\alpha/2}(\rho^2)$ denotes the $(1-\alpha/2)$-quantile of the distribution $(1-\rho^2)^{1/2}\epsilon_0 + \rho L_{K,a}$. Representation (4) makes it simple to approximate this quantile via Monte Carlo simulation after specifying $\hat{R}^2$. In (5), $\hat{R}^2 = (p_1^{-1}s_{1|X}^2 + p_0^{-1}s_{0|X}^2 - s_{\tau|X}^2)/\hat{V}$, where $s_{1|X}^2$, $s_{0|X}^2$, and $s_{\tau|X}^2$ are sample analogues of $S_{1|X}^2$, $S_{0|X}^2$ and $S_{\tau|X}^2$, as defined in Li et al. (2018). The estimator $\hat{R}^2$ has probability limit $\tilde{R}^2 = VR^2/\tilde{V} \leqslant R^2$; thus, $\hat{R}^2$ is conservative to the extent that $\hat{V}$ is conservative. Because $v_{1-\alpha/2}(R^2) \leqslant z_{1-\alpha/2}$ for all $\alpha \in (0,1)$, interval (5) is narrower than the analogous interval for a completely randomized experiment if the covariates have any linear association with the outcomes (Li et al. 2018, Theorem 2).

Interval (5) implies the following test for the null $H_0\colon \tau = 0$ against the alternative $H_A\colon \tau > 0$:

$$\begin{cases} \text{reject } H_0\colon \tau = 0 & \text{if } \hat{\tau} > v_{1-\alpha}(\hat{R}^2)\hat{V}^{1/2}N^{-1/2}, \\ \text{fail to reject } H_0\colon \tau = 0 & \text{otherwise.} \end{cases} \tag{6}$$

Similar to most power analyses (Lachin, 1981; Lerman, 1996; Wittes, 2002), we focus on the one-sided test here because it simplifies sample size calculations. For any α-level two-sided test, we can always bound its power and sample size requirement by that of an $\alpha/2$-level one-sided test. Furthermore, throughout the paper, we assume that the significance level α of the test in (6) is below or equal to 0.5, which is the case for most if not all applications. The following theorem establishes the power of the above test.

Theorem 1. *Under rerandomization, the power of test (6) is asymptotically*

$$\overline{\mathcal{V}}_{R^2}\left\{\frac{v_{1-\alpha}(\tilde{R}^2)\tilde{V}^{1/2} - \tau N^{1/2}}{V^{1/2}}\right\},$$

where $\overline{\mathcal{V}}_{R^2}(\cdot)$ denotes the survival function of $(1-R^2)^{1/2}\epsilon_0 + RL_{K,a}$, V denotes the variance in (1), $\tilde{V}$ denotes the probability limit of its corresponding estimator, R^2 denotes the squared multiple correlation in (3) and $\tilde{R}^2 = VR^2/\tilde{V}$ denotes the probability limit of $\hat{R}^2$ in (5).

Theorem 1 is quite similar to classical power calculations (e.g., Lachin 1981), with two important differences. First, power for completely randomized experiments relies on the standard normal quantile $z_{1-\alpha}$ and survival function $\overline{\Phi}(\cdot)$, whereas Theorem 1 relies on the nonnormal quantile $v_{1-\alpha}(\tilde{R}^2)$ and survival function $\overline{\mathcal{V}}_{R^2}(\cdot)$. Second, Theorem 1 involves the ratio $\tilde{V}/V$, which depends on the treatment effect heterogeneity S^2_τ. Such a term does not appear in previous power results (Lachin, 1981; Cohen, 1992; Lerman, 1996; Wittes, 2002), which typically assume a superpopulation framework that does not involve S^2_τ. However, under a finite-population framework, S^2_τ appears in the true variance of $\hat{\tau}$ and thereby affects power calculations. Thus, Theorem 1 allows for less conservative variance estimators, such as $\hat{V}_{\text{DFM}}$, and demonstrates how treatment effect heterogeneity affects testing power.

We can show that, for fixed S^2_1, S^2_0, τ, and $\tilde{R}^2$, power in Theorem 1 is increasing in S^2_τ as long as $\tau \geqslant v_{1-\alpha}(\tilde{R}^2)\tilde{V}^{1/2}N^{-1/2}$; otherwise, it is decreasing in S^2_τ. Thus, treatment effect heterogeneity has a beneficial effect on power for large effect sizes, but an adverse effect for small effect sizes. As a result, standard power calculations that assume that $S^2_\tau = 0$ may underestimate or overestimate power, depending on the effect size. We discuss this dynamic further in the Supplementary Material.

3.2. *Sample size calculations*

Theorem 1 establishes the testing power of a rerandomized experiment. The following theorem establishes the relationship between the sample size N and a prespecified degree of power γ when we use test (6) to conduct inference for a rerandomized experiment.

THEOREM 2. *Let $\gamma \geqslant \alpha$ denote a prespecified probability that we correctly reject $H_0\colon \tau = 0$ using test (6) under rerandomization. Then, the relationship between the sample size N and power γ is*

$$N = \left\{ \frac{v_{1-\alpha}(\tilde{R}^2)\tilde{V}^{1/2} - v_{1-\gamma}(R^2)V^{1/2}}{\tau} \right\}^2, \tag{7}$$

where V denotes the variance in (1), $\tilde{V}$ denotes the probability limit of its corresponding estimator, R^2 denotes the squared multiple correlation in (3) and $\tilde{R}^2$ denotes the probability limit of $\hat{R}^2$ in (5).

The sample size in Theorem 2 is (a) increasing in γ, (b) decreasing in τ, (c) increasing in S^2_1 and S^2_0 if $\gamma \geqslant 0.5$ and (d) decreasing in S^2_τ if $\gamma \geqslant 0.5$. These results also hold under complete randomization, which is a special case when $a = \infty$ or $R^2 = 0$. The first three observations are well known, but, to our knowledge, the fourth has remained largely unacknowledged. However, it is intuitive: if S^2_τ is larger then the variance of the mean-difference estimator is smaller. This additional precision is propagated into the sample size necessary to achieve power γ. This demonstrates that assuming that $S^2_\tau = 0$ is conservative, not only for confidence intervals, but also for sample size calculations. However, there is usually no knowledge about S^2_τ before the experiment; furthermore, if increased heterogeneity also increases outcome variation, this could adversely affect sample size. Nonetheless, after the experiment, one could test whether $S^2_\tau > 0$ (Ding et al., 2019). We discuss this further in the Supplementary Material, where we study via simulation how sample size is affected by S^2_1, S^2_0 and S^2_τ.

We also show how to compute the sample size in (7) with our `R` package `rerandPower` (R Development Core Team, 2024). One must specify the parameters γ, τ, α; the proportions p_1, p_0; the association R^2; and S^2_1, S^2_0 and S^2_τ, which define V, $\tilde{V}$ and $\tilde{R}^2 = VR^2/\tilde{V}$. Our package approximates the quantiles $v_{1-\alpha}(\tilde{R}^2)$ and $v_{1-\gamma}(R^2)$ via Monte Carlo simulation using representation (4).

Remark 1. Technically, the sample size N is also on the right-hand side of (7), because N is involved in the definitions of S^2_1, S^2_0 and S^2_τ. This is a by-product of adopting a finite-population framework. Nonetheless, we write N in terms of S^2_1, S^2_0 and S^2_τ, because it is common to specify these quantities when making sample size calculations (e.g., Lachin 1981; Wittes 2002). With a slight abuse of notation, we can view S^2_1, S^2_0 and S^2_τ in Theorem 2 as limits of potential outcome variances and treatment effect heterogeneity, thereby allowing for sample size calculators under a finite-population framework.

4. Comparing rerandomization to complete randomization

4.1. *Dispersive ordering of normal and nonnormal distributions*

A natural question to ask is how power and sample size compare for rerandomization and complete randomization. Li et al. (2018) established that two-sided confidence intervals under rerandomization are narrower than those under complete randomization. To do this, they compared the lengths of symmetric quantile ranges $[v_\alpha(R^2), v_{1-\alpha}(R^2)]$ and $[z_\alpha, z_{1-\alpha}]$ for $\alpha \in (0, 0.5]$. However, such a comparison is not sufficient for comparing power between rerandomization and complete randomization, which involves lengths of asymmetric quantiles ranges, and in particular, gaps between two quantiles on the same side of the origin.

To make this comparison, we establish a dispersive ordering of normal and nonnormal distributions involved in complete randomization and rerandomization. For two random variables X and Y with quantile functions F^{-1} and G^{-1}, X is said to be less dispersed than Y if $F^{-1}(\beta) - F^{-1}(\alpha) \leqslant G^{-1}(\beta) - G^{-1}(\alpha)$ for any $0 < \alpha < \beta < 1$ (Shaked, 1982). The following theorem establishes that the rerandomization distribution (2) is less dispersed than a standard normal distribution.

Theorem 3. *Let $\varepsilon_0 \sim \mathcal{N}(0,1)$ and $L_{K,a}$ be defined in (4), such that ε_0 and $L_{K,a}$ are independent. Then, for any $a \in [0,\infty]$, integer $K \geqslant 1$ and $\rho \in [0,1]$, $(1-\rho^2)^{1/2}\varepsilon_0 + \rho L_{K,a}$ is less dispersed than ε_0.*

Theorem 3 generalizes Li et al. (2018, Theorem 2), which showed that the symmetric quantile range of $(1-\rho^2)^{1/2}\varepsilon_0 + \rho L_{K,a}$ is less than or equal to that of the standard normal distribution. Theorem 3 further shows that the gap between the α and β quantiles of $(1-\rho^2)^{1/2}\varepsilon_0 + \rho L_{K,a}$, such as any nonsymmetric quantile range, is less than or equal to that of the standard normal distribution. Theorem 3 is crucial for establishing the later theorems on power and sample size comparisons.

4.2. *Power and sample size comparisons*

The following theorem quantifies when power is greater under rerandomization.

Theorem 4. *If $\tilde{V} = V$ then, for any $\tau \geqslant 0$,*

$$\overline{\mathcal{V}}_{R^2}\left(\frac{v_{1-\alpha}(\tilde{R}^2)\tilde{V}^{1/2} - \tau N^{1/2}}{V^{1/2}}\right) \geqslant \overline{\Phi}\left(\frac{z_{1-\alpha}\tilde{V}^{1/2} - \tau N^{1/2}}{V^{1/2}}\right). \tag{8}$$

Meanwhile, if $\tilde{V} > V$, (8) still holds when $\tau \geqslant v_{1-\alpha}(\tilde{R}^2)\tilde{V}^{1/2}N^{-1/2}$; otherwise, (8) may be violated.

From Theorem 4, if inference is not conservative, such that $\tilde{V} = V$, power is greater under rerandomization than under complete randomization. However, when inference is conservative, such that $\tilde{V} > V$, rerandomization may exhibit less power than complete randomization. This result is surprising: confidence intervals are always narrower under rerandomization, so we would suspect power to always be greater. However, for power, we must consider not only confidence intervals' precision, but also their conservativeness. Equation (2) shows that the true asymptotic distribution of $N^{1/2}(\hat{\tau} - \tau)$ under rerandomization is $V^{1/2}\{(1-R^2)^{1/2}\epsilon_0 + RL_{k,a}\}$; when conducting inference, this is asymptotically estimated as $\tilde{V}^{1/2}\{(1-\tilde{R}^2)^{1/2}\epsilon_0 + \tilde{R}L_{k,a}\}$. By recognizing that $\tilde{V}\tilde{R}^2 = VR^2$, one can show that this estimated distribution is equivalent to the convolution of the true distribution of $N^{1/2}(\hat{\tau} - \tau)$ and an independent normal distribution $\mathcal{N}(0, \tilde{V} - V)$. Because the true distribution of $N^{1/2}(\hat{\tau} - \tau)$ is more concentrated around zero under rerandomization than complete randomization, but distribution $\mathcal{N}(0, \tilde{V} - V)$ is the same for both designs, inference under rerandomization can be more conservative. As a result, the test under rerandomization can have a smaller Type-I error; see the Supplementary Material. This additional conservativeness has an adverse effect on power, but the additional precision from rerandomization has a beneficial effect. Theorem 4 establishes that the beneficial effect outweighs the adverse effect as long as τ is not too small. We illustrate this in the Supplementary Material.

Intuitively, when rerandomization increases power, it requires a smaller sample size to achieve a certain degree of power. Let N_{rr} denote the sample size necessary to achieve power γ under rerandomization, as provided by Theorem 2. Let N_{cr} denote the sample size necessary under complete randomization, which is the same as Theorem 2, but with the nonnormal quantiles replaced with normal quantiles. The following theorem establishes sufficient conditions for $N_{\text{rr}} \leqslant N_{\text{cr}}$.

THEOREM 5. *Let $\gamma \geqslant \alpha$ denote a desired level of power, and let N_{rr} and N_{cr} denote the sample sizes required to achieve power γ under rerandomization and complete randomization, respectively. We have two separate results. First, if $\tilde{V} = V$ then $N_{\text{rr}}/N_{\text{cr}} \leqslant 1$. Second, if $\gamma \geqslant 0.5$ then $N_{\text{rr}}/N_{\text{cr}}$ is*

(a) *less than or equal to* 1,
(b) *increasing in the number of covariates K and the rerandomization threshold a, and*
(c) *decreasing in* R^2.

Otherwise, if $\tilde{V} > V$ and $\gamma < 0.5$ then $N_{\text{rr}}/N_{\text{cr}}$ may be greater than 1.

Theorem 5 establishes that rerandomization requires a smaller sample size to achieve the same amount of power when either inference is not conservative or the desired power is greater than or equal to 0.5. The condition $\tau \geqslant \nu_{1-\alpha}(\tilde{R}^2)\tilde{V}^{1/2}N^{-1/2}$ in Theorem 4 implies the condition $\gamma \geqslant 0.5$ in Theorem 5, and thus these conditions are analogous. In most standard power calculations, the desired power is greater than 0.5.

The smaller the ratio $N_{\text{rr}}/N_{\text{cr}}$, the larger the sample size benefits of rerandomization over complete randomization. The sample size reduction depends on R^2, K, a and S_τ^2. In the Supplementary Material, we conduct a simulation study to assess how $N_{\text{rr}}/N_{\text{cr}}$ changes for different K, R^2, a and S_τ^2, thereby allowing practitioners to understand the sample size benefits of rerandomization. When there is no treatment effect heterogeneity, the median of $N_{\text{rr}}/N_{\text{cr}}$ is 0.75 for $K \in [1, 100]$ and $R^2 \in [0, 0.9]$ using a rerandomization threshold a such that $\text{pr}(M \leqslant a) = 0.001$, as recommended by Li et al. (2018). If, furthermore, $R^2 \geqslant 0.3$ and $K \leqslant 50$, the median is 0.58. Meanwhile, when there is treatment effect heterogeneity, these sample size benefits are dampened, because inference is conservative. When S_τ is half the size of S_1 and S_0, $N_{\text{rr}}/N_{\text{cr}}$ is on average 2.4% greater than when there is no heterogeneity; when S_τ is 50% greater than S_1 and S_0, $N_{\text{rr}}/N_{\text{cr}}$ is on average 23.7% greater. More details, such as how to implement power and sample size calculations with our `R` package `rerandPower` are discussed in the Supplementary Material.

5. Discussion and conclusion

Our results focus on rerandomized experiments with two groups. Rerandomization theory has been extended to more than two groups (Branson et al., 2016; Li et al., 2020), and thus we posit that similar results hold for multiarm experiments. However, multiple causal estimands arise in this setting, making power analyses more complex. Power depends on the potential outcome variance in each group, as well as the effect heterogeneity and rerandomization criterion for each estimand. Thus, power analyses will be notationally complex, but can rely on the same conceptual framework as here.

One could compare rerandomization to designs beyond complete randomization, such as block randomization, where blocks are constructed using, e.g., matching (Greevy et al., 2004; Bai, 2022). Power would then depend on the association between blocking variables and outcomes; however, blocking can be less efficient than complete randomization if the blocks are poorly chosen (Pashley & Miratrix, 2022). Rerandomization is always at least as efficient as complete randomization and can be combined with blocking to improve block randomization (Wang et al., 2021). Comparing block randomization and block rerandomization is analogous to our comparison.

Our results also focus on the Mahalanobis distance on covariate means, but other rerandomization criteria could be used. For example, Morgan & Rubin (2015) proposed different Mahalanobis distance criteria for tiers of covariates that vary in importance. Again, power analyses will be notationally complex in order to incorporate the criterion for each tier. Other examples include criteria

modified by ridge penalties (Branson & Shao, 2021) or principal component analysis (Zhang et al., 2023), which have been shown to increase precision in high-dimensional settings. We suspect that testing power may increase as well.

Acknowledgement

We thank the associate editor and two reviewers for insightful comments that improved this paper. Ding gratefully acknowledges support from the U.S. National Science Foundation.

Supplementary material

The Supplementary Material includes proofs for the theorems, a simulation study and example code for our R package `rerandPower`.

References

Bai, Y. (2022). Optimality of matched-pair designs in randomized controlled trials. *Am. Econ. Rev.* **112**, 3911–40.

Branson, Z., Dasgupta, T. & Rubin, D. B. (2016). Improving covariate balance in 2^K factorial designs via rerandomization with an application to a New York City Department of Education high school study. *Ann. Appl. Statist.* **10**, 1958–76.

Branson, Z. & Shao, S. (2021). Ridge rerandomization: an experimental design strategy in the presence of covariate collinearity. *J. Statist. Plan. Infer.* **211**, 287–314.

Bruhn, M. & McKenzie, D. (2009). In pursuit of balance: randomization in practice in development field experiments. *Am. Econ. J. Appl. Econ.* **1**, 200–32.

Chow, S.-C., Shao, J., Wang, H. & Lokhnygina, Y. (2017). *Sample Size Calculations in Clinical Research*. Boca Raton, FL: CRC Press.

Cohen, J. (1992). A power primer. *Psychol. Bull.* **112**, 155.

Ding, P., Feller, A. & Miratrix, L. (2019). Decomposing treatment effect variation. *J. Am. Statist. Assoc.* **114**, 304–17.

Dupont, W. D. & Walton D. P. Jr. (1990). Power and sample size calculations: a review and computer program. *Controlled clinical trials* **11**, 116–128.

Fisher, R. A. (1935). *Design of Experiments*, vol. 1. Edinburgh: Oliver and Boyd.

Greevy, R., Lu, B., Silber, J. H. & Rosenbaum, P. (2004). Optimal multivariate matching before randomization. *Biostatistics* **5**, 263–75.

Imai, K. (2008). Variance identification and efficiency analysis in randomized experiments under the matched-pair design. *Statist. Med.* **27**, 4857–73.

Imbens, G. W. & Rubin, D. B. (2015). *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge: Cambridge University Press.

Jiang, Z., Imai, K. & Malani, A. (2022). Statistical inference and power analysis for direct and spillover effects in two-stage randomized experiments. *Biometrics*, doi: 10.1111/biom.13782.

Kang, H. (2021). Sample size determination and power analysis using the G* Power software. *Journal of educational evaluation for health professions* 18.

Lachin, J. M. (1981). Introduction to sample size determination and power analysis for clinical trials. *Contr. Clin. Trials* **2**, 93–113.

Lerman, J. (1996). Study design in clinical research: sample size estimation and power analysis. *Can. J. Anaesth.* **43**, 184–91.

Li, X. & Ding, P. (2017). General forms of finite population central limit theorems with applications to causal inference. *J. Am. Statist. Assoc.* **112**, 1759–69.

Li, X., Ding, P. & Rubin, D. B. (2018). Asymptotic theory of rerandomization in treatment–control experiments. *Proc. Nat. Acad. Sci.* **115**, 9157–62.

Li, X., Ding, P. & Rubin, D. B. (2020). Rerandomization in 2^K factorial experiments. *Ann. Statist.* **48**, 43–63.

Lu, X., Liu, T., Liu, H. & Ding, P. (2023). Design-based theory for cluster rerandomization. *Biometrika* **110**, 467–83.

Maxwell, S. E., Kelley, K. & Rausch, J. R. (2008). Sample size planning for statistical power and accuracy in parameter estimation. *Ann. Rev. Psychol.* **59**, 537–63.

Miratrix, L. W., Sekhon, J. S. & Yu, B. (2013). Adjusting treatment effect estimates by post-stratification in randomized experiments. *J. R. Statist. Soc.* B **75**, 369–96.

Morgan, K. L. & Rubin, D. B. (2012). Rerandomization to improve covariate balance in experiments. *Ann. Statist.* **40**, 1263–82.

Morgan, K. L. & Rubin, D. B. (2015). Rerandomization to balance tiers of covariates. *J. Am. Statist. Assoc.* **110**, 1412–21.
Neyman, J. (1923). On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Statist. Sci.* **5**, 465–72.
Pashley, N. E. & Miratrix, L. W. (2021). Insights on variance estimation for blocked and matched pairs designs. *J. Educ. Behav. Statist.* **46**, 271–96.
Pashley, N. E. & Miratrix, L. W. (2022). Block what you can, except when you shouldn't. *J. Educ. Behav. Statist.* **47**, 69–100.
R Development Core Team (2024). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0, http://www.R-project.org.
Schochet, P. Z. (2009). Statistical power for regression discontinuity designs in education evaluations. *J. Educ. Behav. Statist.* **34**, 238–66.
Schochet, P. Z. (2022). Statistical power for estimating treatment effects using difference-in-differences and comparative interrupted time series estimators with variation in treatment timing. *J. Educ. Behav. Statist.* **47**, 367–405.
Shaked, M. (1982). Dispersive ordering of distributions. *J. Appl. Prob.* **19**, 310–20.
Tabord-Meehan, M. (2022). Stratification trees for adaptive randomization in randomized controlled trials. *arXiv:* 1806.05127v7.
Wang, X., Wang, T. & Liu, H. (2021). Rerandomization in stratified randomized experiments. *J. Am. Statist. Assoc.*, doi: 10.1080/01621459.2021.1990767.
Wang, Y. & Li, X. (2022). Rerandomization with diminishing covariate imbalance and diverging number of covariates. *Ann. Statist.* **50**, 3439–65.
Wittes, J. (2002). Sample size calculations for randomized controlled trials. *Epidemiol. Rev.* **24**, 39–53.
Zhang, H., Yin, G. & Rubin, D. B. (2023). PCA rerandomization. *Can. J. Statist.* doi: 10.1002/cjs.11765.
Zhou, Q., Ernst, P. A., Morgan, K. L., Rubin, D. B. & Zhang, A. (2018). Sequential rerandomization. *Biometrika* **105**, 745–52.

[*Received on* 6 *January* 2022. *Editorial decision on* 21 *April* 2023]

Biometrika (2023), **xx**, x, *pp*. 1–16
Printed in Great Britain

Supplementary material for "Power and sample size calculations for rerandomization"

BY ZACH BRANSON
Department of Statistics, Carnegie Mellon University, Pittsburgh, PA 15213, U.S.A.
zach@stat.cmu.edu

XINRAN LI
Department of Statistics, University of Illinois, Champaign, IL 61820, U.S.A.
xinranli@illinois.edu

PENG DING
Department of Statistics, University of California, Berkeley, CA 94720, U.S.A.
pengdingpku@berkeley.edu

Appendix A provides the proof of Theorem 1, which characterizes the asymptotic power under rerandomization.

Appendix B provides the proof of Theorem 2, which characterizes the sample size necessary to achieve a given power under rerandomization.

Appendix C provides the proof of Theorem 3, which establishes the dispersive ordering of the Normal and non-Normal distributions involved in the asymptotic approximations for complete randomization and rerandomization.

Appendix D provides the proof of Theorem 4, which determines when asymptotic power is greater under rerandomization than under complete randomization.

Appendix E provides the proof of Theorem 5, which characterizes the ratio of the rerandomization sample size and complete randomization sample size necessary to achieve a given power.

Appendix F compares the type-I error rates of rerandomization and complete randomization under the null hypothesis of $\tau = 0$.

Appendix G uses numerical examples to illustrate that rerandomization can be less powerful than complete randomization.

Appendix H presents a simulation study exploring how $N_{\text{rr}}/N_{\text{cr}}$ changes for various experimental settings, thereby allowing practitioners to understand the sample size gains of rerandomization compared to complete randomization.

Appendix I provides additional numerical examples that illustrate how treatment effect heterogeneity affects power and sample size under complete randomization and rerandomization.

Appendix J provides example code to implement power and sample size calculations for completely randomized and rerandomized experiments with our `R` package `rerandPower`.

A. PROOF OF THEOREM 1

Below we first give a rigorous statement of Theorem 1 in the main paper. We will conduct finite population asymptotic analyses; see, e.g., Li & Ding (2017) for a review. Specifically, we embed the N units into a sequence of finite populations and impose the following regularity conditions analogous to Li et al. (2018, Condition 1) along this sequence of finite populations.

Condition S1. As $N \to \infty$,

(i) the proportions of treated and control units, p_1 and p_0, have positive limits;
(ii) the finite population variances and covariances, $S_1^2, S_0^2, S_\tau^2, S_{1,X}, S_{0,X}$ and S_X^2 have finite limiting values, the limit of S_X^2 is nonsingular, and the limit of $V = p_1^{-1}S_1^2 + p_0^{-1}S_0^2 - S_\tau^2$ is positive.
(iii) $\max_{1\le i\le N}\{Y_i(z) - \bar{Y}(z)\}^2/N \to 0$ for $z = 0, 1$, and $\max_{1\le i\le N}\|X_i - \bar{X}\|^2/N \to 0$.

In Condition S1, (i) is a natural requirement, and (ii) assumes stable finite population variances and covariances along the sequence of finite populations. In addition, (ii) intuitively assumes that the covariates are not colinear, and that the variance of $\hat{\tau}$ under complete randomization is not zero. Meanwhile, (iii) assumes that the potential outcomes and covariates are not too heavy-tailed, and it will hold with probability 1 if the units are i.i.d. samples from some superpopulation with $2 + \eta$ moments, for any $\eta > 0$ (Li & Ding, 2017).

The theorem below is a rigorous version of Theorem 1 in the main paper. In this asymptotic power analysis, we consider the finite population asymptotics with Condition S1 and the local alternative where the true average treatment effect is on the scale of $N^{-1/2}$, analogous to usual power analysis.

THEOREM S1. *Consider a sequence of finite populations with increasing sizes. Assume that Condition* S1 *holds and the true average treatment effects satisfy $\tau = cN^{-1/2}$ for all N and some finite constant c. Under rerandomization, the power of the test* (8) *satisfies*

$$\text{pr}\left(\hat{\tau} > \nu_{1-\alpha}(\hat{R}^2)\hat{V}^{1/2}N^{-1/2}\right) = \overline{\mathcal{V}}_{R^2}\left(\frac{\nu_{1-\alpha}(\tilde{R}^2)\tilde{V}^{1/2} - \tau N^{1/2}}{V^{1/2}}\right) + o(1),$$

where $\overline{\mathcal{V}}_{R^2}(\cdot)$ denotes the survival function of $(1-R^2)^{1/2}\epsilon_0 + RL_{K,a}$, V denotes the variance in (2), $\tilde{V}$ denotes the probability limit of its corresponding estimator, R^2 denotes the squared multiple correlation in (4), and $\tilde{R}^2 = VR^2/\tilde{V}$ denotes the probability limit of $\hat{R}^2$ in (7).

Now we prove Theorem 1. Under the test (8), we reject the null hypothesis if $\hat{\tau} > \nu_{1-\alpha}(\hat{R}^2)\hat{V}^{1/2}N^{-1/2}$. Thus, the power of the test is

$$\begin{aligned}\text{pr}\left(\hat{\tau} > \nu_{1-\alpha}(\hat{R}^2)\hat{V}^{1/2}N^{-1/2}\right) &= \text{pr}\left(N^{1/2}(\hat{\tau}-\tau) - \nu_{1-\alpha}(\hat{R}^2)\hat{V}^{1/2} > -N^{1/2}\tau\right)\\ &= \text{pr}\left(N^{1/2}(\hat{\tau}-\tau) - \nu_{1-\alpha}(\hat{R}^2)\hat{V}^{1/2} > -c\right),\end{aligned}$$

where the last equality holds due to the fact that $\tau = N^{-1/2}c$. From Li et al. (2018), under rerandomization, $\sqrt{N}(\hat{\tau}-\tau) \dot{\sim} V^{1/2}\{(1-R^2)^{1/2}\varepsilon_0 + RL_{K,a}\}$, where $\dot{\sim}$ denotes that the two sequences of distributions or random variables converging weakly to the same distribution. Besides, $\tilde{V} - V = o_{\text{pr}}(1)$ and $R^2 - \tilde{R}^2 = o_{\text{pr}}(1)$. By Slutsky's theorem, we have

$$N^{1/2}(\hat{\tau}-\tau) - \nu_{1-\alpha}(\hat{R}^2)\hat{V}^{1/2} \dot{\sim} V^{1/2}\{(1-R^2)^{1/2}\varepsilon_0 + RL_{K,a}\} - \nu_{1-\alpha}(\tilde{R}^2)\tilde{V}^{1/2}.$$

Consequently, we have, as $N \to \infty$,

$$\begin{aligned}\text{pr}\left(\hat{\tau} > \nu_{1-\alpha}(\hat{R}^2)\hat{V}^{1/2}N^{-1/2}\right) &= \text{pr}\left(V^{1/2}\{(1-R^2)^{1/2}\varepsilon_0 + RL_{K,a}\} - \nu_{1-\alpha}(\tilde{R}^2)\tilde{V}^{1/2} > -c\right) + o(1)\\ &= \overline{\mathcal{V}}_{R^2}\left(\frac{\nu_{1-\alpha}(\tilde{R}^2)\tilde{V}^{1/2} - c}{V^{1/2}}\right) + o(1)\\ &= \overline{\mathcal{V}}_{R^2}\left(\frac{\nu_{1-\alpha}(\tilde{R}^2)\tilde{V}^{1/2} - \tau N^{1/2}}{V^{1/2}}\right) + o(1).\end{aligned}$$

Therefore, Theorem 1 holds.

B. Proof of Theorem 2

Let γ denote a prespecified degree of power desired for a rerandomized experiment, where γ is the probability we reject the null hypothesis using the test (8). Then, by Theorem 1, we have:

$$\gamma = \overline{\mathcal{V}}_{R^2}\left(\frac{\nu_{1-\alpha}(\tilde{R}^2)\tilde{V}^{1/2} - \tau N^{1/2}}{V^{1/2}}\right) = 1 - \mathcal{V}_{R^2}\left(\frac{\nu_{1-\alpha}(\tilde{R}^2)\tilde{V}^{1/2} - \tau N^{1/2}}{V^{1/2}}\right),$$

where $\mathcal{V}_{R^2}(\cdot)$ denotes the distribution function of the distribution $(1-R^2)^{1/2}\epsilon_0 + RL_{K,a}$ defined in (3). Then, solving for N, we have:

$$\begin{aligned} & \frac{\nu_{1-\alpha}(\tilde{R}^2)\tilde{V}^{1/2} - \tau N^{1/2}}{V^{1/2}} = \nu_{1-\gamma}(R^2) \\ \Longrightarrow \quad & \nu_{1-\alpha}(\tilde{R}^2)\tilde{V}^{1/2} - \tau N^{1/2} = \nu_{1-\gamma}(R^2)V^{1/2} \\ \Longrightarrow \quad & N = \left(\frac{\nu_{1-\alpha}(\tilde{R}^2)\tilde{V}^{1/2} - \nu_{1-\gamma}(R^2)V^{1/2}}{\tau}\right)^2. \end{aligned}$$

C. Proof of Theorem 3

To prove Theorem 3, we need the following five lemmas.

Lemma S1. *For any integer $K \geq 1$ and threshold $a \in (0, \infty)$, the probability density function of $L_{K,a}$ is*

$$g_{K,a}(x) = \phi(x)\frac{F_{K-1}(a - x^2)}{F_K(a)},$$

where $\phi(\cdot)$ is the probability density of $\mathcal{N}(0,1)$ and $F_K(\cdot)$ is the distribution function of χ^2_K, with $F_0(x) = \mathbb{1}(x \geq 0)$ being the distribution function of a point mass at 0.

Proof of Lemma S1. Lemma S1 follows from Li et al. (2018, Proof of Proposition 2). For completeness, we give a proof below. Let $D = (D_1, \ldots, D_K)^{\mathrm{T}} \sim \mathcal{N}(0, I_K)$. For any $x \in \mathbb{R}$, we have

$$\begin{aligned} \text{pr}(L_{K,a} \leq x) &= \text{pr}(D_1 \leq x \mid D^{\mathrm{T}}D \leq a) \\ &= \frac{\text{pr}(D_1 \leq x, D^{\mathrm{T}}D \leq a)}{\text{pr}(D^{\mathrm{T}}D \leq a)} \\ &= \frac{1}{F_K(a)}\int_{-\infty}^{\infty} \text{pr}\left(t \leq x, t^2 + \sum_{j=2}^{K} D_j^2 \leq a\right)\phi(t)\mathrm{d}t \\ &= \frac{1}{F_K(a)}\int_{-\infty}^{x} F_{K-1}(a - t^2)\phi(t)\mathrm{d}t \\ &\equiv \int_{-\infty}^{x} g_{K,a}(t)\mathrm{d}t, \end{aligned}$$

where $g_{K,a}(t) \equiv F_{K-1}(a-t^2)\phi(t)/F_K(a)$. Therefore, $g_{K,a}(\cdot)$ must be the probability density function of $L_{K,a}$, i.e., Lemma S1 holds. □

Lemma S2. *For any $a \in (0, \infty)$, $K \geq 1$ and $c \in \mathbb{R}$, define*

$$h_{K,a,c}(x) = \log g_{K,a}(x) - \log \phi(x + c), \qquad (-\sqrt{a} < x < \sqrt{a}).$$

Then $\mathrm{d}^2 h_{K,a,c}(x)/\mathrm{d}x^2 \leq 0$ for $x \in (-\sqrt{a}, \sqrt{a})$.

Proof of Lemma S2. From Lemma S1,

$$h_{K,a,c}(x) = \log g_{K,a}(x) - \log \phi(x+c) = \log \phi(x) + \log F_{K-1}(a-x^2) - \log F_K(a) - \log \phi(x+c)$$
$$= \log F_{K-1}(a-x^2) + cx + c^2/2 - \log F_K(a).$$

When $K = 1$, $h_{1,a,c}(x)$ reduces to $cx + c^2/2 - \log F_1(a)$, which is a linear function of x. Consequently, $\mathrm{d}^2 h_{1,a,c}(x)/\mathrm{d}x^2 = 0$ for all $x \in (-\sqrt{a}, \sqrt{a})$, i.e., Lemma S2 holds for $K = 1$. Below we consider only the case with $K > 1$.

Let $f_K(x)$ be the density of χ^2_K, and $\dot{f}_K(x) = \mathrm{d}f_K(x)/\mathrm{d}x$ be its derivative over x. We have

$$\dot{f}_K(x) = f_K(x) \cdot \left(\frac{K/2-1}{x} - \frac{1}{2} \right) = f_K(x) \cdot \frac{K-2-x}{2x}. \tag{S.1}$$

Consequently, for $K > 1$, the second derivative of $h_{K,a,c}$ reduces to

$$\begin{aligned}
\frac{\mathrm{d}^2}{\mathrm{d}x^2} h_{K,a,c}(x) &= \frac{\mathrm{d}^2}{\mathrm{d}x^2} \log F_{K-1}(a-x^2) \\
&= \frac{\mathrm{d}}{\mathrm{d}x} \left\{ \frac{f_{K-1}(a-x^2) \cdot (-2x)}{F_{K-1}(a-x^2)} \right\} \\
&= \frac{\dot{f}_{K-1}(a-x^2) \cdot (-2x)^2 \cdot F_{K-1}(a-x^2) - \{f_{K-1}(a-x^2) \cdot (-2x)\}^2}{\{F_{K-1}(a-x^2)\}^2} \\
&= \frac{4x^2 \cdot f_{K-1}(a-x^2)}{\{F_{K-1}(a-x^2)\}^2} \left\{ \frac{K-1-2-(a-x^2)}{2(a-x^2)} \cdot F_{K-1}(a-x^2) - f_{K-1}(a-x^2) \right\} \\
&\equiv \frac{4x^2 \cdot f_{K-1}(a-x^2)}{\{F_{K-1}(a-x^2)\}^2} \cdot \Delta_{K-1}(a-x^2),
\end{aligned}$$

where

$$\Delta_K(x) = \frac{K-2-x}{2x} \cdot F_K(x) - f_K(x).$$

Thus, to prove Lemma S2, it suffices to prove $\Delta_K(x) \le 0$ for all $K > 0$ and $x \in (0, \infty)$. Note that $\Delta_K(x) \le -f_K(x) \le 0$ when $x \ge K-2$. It suffices to show $\Delta_K(x) \le 0$ for all $K > 2$ and $x \in (0, K-2)$.

For $K > 2$ and $x \in (0, K-2)$, define

$$\tilde{\Delta}_K(x) = \frac{2x}{K-2-x} \Delta_K(x) = F_K(x) - \frac{2x}{K-2-x} f_K(x).$$

It then suffices to show $\tilde{\Delta}_K(x) \le 0$ for all $K > 2$ and $x \in (0, K-2)$. By some algebra and (S.1), for $K > 2$ and $x \in (0, K-2)$,

$$\frac{\mathrm{d}}{\mathrm{d}x} \tilde{\Delta}_K(x) = f_K(x) - \frac{2(K-2)}{(K-2-x)^2} f_K(x) - \frac{2x}{K-2-x} \dot{f}_K(x) = -\frac{2(K-2)}{(K-2-x)^2} f_K(x) \le 0.$$

We can verify that $\lim_{x \to 0+} \tilde{\Delta}_K(x) = 0$ for any $K > 2$. Thus, we must have $\tilde{\Delta}_K(x) \le 0$ for all $K > 2$ and $x \in (0, K-2)$.

From the above, Lemma S2 holds. □

For a real function ψ defined on $\mathcal{I} \subset \mathbb{R}$, define the number of sign changes of ψ in $\mathcal{I}$ as

$$S^-_{\mathcal{I}}(\psi) = S^-_{\mathcal{I}}(\psi(x)) = \sup S^-[\psi(x_1), \psi(x_2), \ldots, \psi(x_m)] \tag{S.2}$$

where $S^-(y_1, y_2, \ldots, y_m)$ is the number of sign changes of the sequence $(y_1, y_2, \ldots, y_m)$ with the zero terms being discarded, and the supremum in (S.2) is over all sets $x_1 < x_2 < \cdots < x_m$ with $x_i \in \mathcal{I}$ and $m < \infty$. For any $c \in \mathbb{R}$ and function ψ, define $\psi_c(x) = \psi(x-c)$.

LEMMA S3. *Let F and G be two absolutely continuous distributions having intervals as their support, in the sense that each of F and G has a probability density function that takes positive values on an interval and zero values otherwise, and let f and g be the corresponding densities. If $\mathcal{S}^-_{\mathbb{R}}(f_c - g) \leq 2$ for all $c \in \mathbb{R}$, with the sign sequence being $-,+,-$ in case of equality, then F is less dispersed than G.*

Proof of Lemma S3. Lemma S3 follows from Shaked (1982, Theorem 2.5). □

LEMMA S4. *For any $a \in [0,\infty]$ and integer $K \geq 1$, $L_{K,a}$ is less dispersed than ε_0.*

Proof of Lemma S4. Lemma S4 holds obviously when a equals zero or infinity. Below we consider only the case with $a \in (0,\infty)$.

Let $g_{K,a}$ and ϕ be the densities of $L_{K,a}$ and ε_0. We have derived the form of $g_{K,a}$ in Lemma S1. Furthermore, we define $g_{1,a}(x)$ to be zero when $x^2 = a$; obviously, $g_{1,a}(x)$ is still the density of $L_{K,a}$. Let $\mathcal{I} = (-\sqrt{a}, \sqrt{a})$ be the support of $L_{K,a}$. For any $c \in \mathbb{R}$, define $g_{K,a,c}(x) = g_{K,a}(x-c)$, $\mathcal{I}_c = (-\sqrt{a}+c, \sqrt{a}+c)$, and $h_{K,a,c}$ the same as in Lemma S2. We then have, for any $x \in \mathcal{I}_c$,

$$\operatorname{sign}\{g_{K,a,c}(x) - \phi(x)\} = \operatorname{sign}\{g_{K,a}(x-c) - \phi(x-c+c)\} = \operatorname{sign}\{h_{K,a,c}(x-c)\}.$$

Therefore, $S^-_{\mathcal{I}_c}(g_{K,a,c} - \phi) = S^-_{\mathcal{I}}(h_{K,a,c})$. By Lemma S2, $h_{K,a,c}$ is a concave function on $\mathcal{I}$. This then implies that

$$S^-_{\mathcal{I}_c}(g_{K,a,c} - \phi) = S^-_{\mathcal{I}}(h_{K,a,c}) = \begin{cases} 0, & \text{with sign being } + \text{ or } -, \\ 1, & \text{with sign sequence being } (-,+) \text{ or } (+,-), \\ 2, & \text{with sign sequence being } (-,+,-). \end{cases}$$

Note that $g_{K,a,c}(x) = 0 < \phi(x)$ for $x \notin \mathcal{I}_c$. We can then verify that $S^-_{\mathbb{R}}(g_{K,a,c} - \phi)$ must have the following forms:

$$S^-_{\mathbb{R}}(g_{K,a,c} - \phi) = \begin{cases} 0, & \text{with sign being } -, \\ 2, & \text{with sign sequence being } (-,+,-). \end{cases}$$

By Lemma S3, $L_{K,a}$ is less dispersed than ε_0. Therefore, Lemma S4 holds. □

LEMMA S5. *Assume X is less dispersed than Y. Let W be a random variable independent of X and Y. Let $f(w)$ be the density of W. If $f(w) > 0$ and $\mathrm{d}^2 \log f(w)/\mathrm{d}w^2 \leq 0$ for all w, then $X + W$ is less dispersed than $Y + W$.*

Proof of Lemma S5. Lemma S5 follows from Lewis & Thompson (1981, Theorem 7). □

Equipped with the above lemmas, we now prove Theorem 3.

Proof of Theorem 3. Let $\varepsilon_1 \sim \mathcal{N}(0,1)$ be independent of $(\varepsilon_0, L_{K,a})$. From Lemma S4, $L_{K,a}$ is less dispersed than ε_1, which immediately implies that $\rho L_{K,a}$ is less dispersed than $\rho\varepsilon_1$. Thus, Theorem 3 holds when $\rho = 1$. Below we consider only the case where $0 \leq \rho < 1$.

By some algebra, the second derivative of the log-density of $\sqrt{1-\rho^2}\varepsilon_0 \sim \mathcal{N}(0, 1-\rho^2)$ is a constant $-(1-\rho^2)^{-1} < 0$. Thus, by Lemma S5, $\sqrt{1-\rho^2}\varepsilon_0 + \rho L_{K,a}$ is less dispersed than $\sqrt{1-\rho^2}\varepsilon_0 + \rho\varepsilon_1 \sim \mathcal{N}(0,1) \sim \varepsilon_0$.

From the above, Theorem 3 holds. □

D. PROOF OF THEOREM 4

Let β_{rr} and β_{cr} denote the left and right hand sides of (10), respectively. We first consider the case with $\tilde{V} = V$, which implies $\tilde{R}^2 = R^2$. We can then verify that

$$\nu_{1-\alpha}(R^2) - \nu_{1-\beta_{\text{rr}}}(R^2) = V^{-1/2}N^{1/2}\tau = z_{1-\alpha} - z_{1-\beta_{\text{cr}}}.$$

Assuming $\tau \geq 0$, we have $1-\alpha \geq 1-\beta_{\text{rr}}$. From Theorem 3, $\nu_{1-\alpha}(R^2) - \nu_{1-\beta_{\text{rr}}}(R^2) \leq z_{1-\alpha} - z_{1-\beta_{\text{rr}}}$. This then implies that $z_{1-\alpha} - z_{1-\beta_{\text{cr}}} \leq z_{1-\alpha} - z_{1-\beta_{\text{rr}}}$. Consequently, we must have $\beta_{\text{rr}} \geq \beta_{\text{cr}}$, i.e., the inequality in (10) holds.

We then consider the case where $\tau \geq \nu_{1-\alpha}(\tilde{R}^2)\tilde{V}^{1/2}N^{-1/2}$. We can verify that

$$\nu_{1-\alpha}(\tilde{R}^2)\tilde{V}^{1/2} - \nu_{1-\beta_{\text{rr}}}(R^2)V^{1/2} = N^{1/2}\tau = z_{1-\alpha}\tilde{V}^{1/2} - z_{1-\beta_{\text{cr}}}V^{1/2}.$$

Assuming $\tau \geq \nu_{1-\alpha}(\tilde{R}^2)\tilde{V}^{1/2}N^{-1/2}$, we have $\beta_{\text{rr}} \geq 1/2$. Note that the distribution $(1-R^2)^{1/2}\epsilon_0 + RL_{K,a}$ is symmetric around zero. From Li et al. (2018, Theorem 2), we can verify that $-\nu_{1-\beta_{\text{rr}}}(R^2) = \nu_{\beta_{\text{rr}}}(R^2) \leq z_{\beta_{\text{rr}}} = -z_{1-\beta_{\text{rr}}}$ and $\nu_{1-\alpha}(\tilde{R}^2) \leq z_{1-\alpha}$. These imply that

$$z_{1-\alpha}\tilde{V}^{1/2} - z_{1-\beta_{\text{cr}}}V^{1/2} = \nu_{1-\alpha}(\tilde{R}^2)\tilde{V}^{1/2} - \nu_{1-\beta_{\text{rr}}}(R^2)V^{1/2} \leq z_{1-\alpha}\tilde{V}^{1/2} - z_{1-\beta_{\text{rr}}}V^{1/2}.$$

Consequently, we must have $\beta_{\text{rr}} \geq \beta_{\text{cr}}$, i.e., the inequality in (10) holds.

From the above, Theorem 4 holds.

E. Proof of Theorem 5

Let N_{cr} and N_{rr} denote the sample sizes necessary to achieve power γ under complete randomization and rerandomization, respectively, as provided by Theorem 2. We have:

$$\frac{N_{\text{rr}}}{N_{\text{cr}}} = \left(\frac{\nu_{1-\alpha}(\tilde{R}^2)\tilde{V}^{1/2} - \nu_{1-\gamma}(R^2)V^{1/2}}{z_{1-\alpha}\tilde{V}^{1/2} - z_{1-\gamma}V^{1/2}}\right)^2. \tag{S.1}$$

We first consider the case with $\tilde{V} = V$. In this case, the ratio simplifies to

$$\frac{N_{\text{rr}}}{N_{\text{cr}}} = \left(\frac{\nu_{1-\alpha}(R^2)V^{1/2} - \nu_{1-\gamma}(R^2)V^{1/2}}{z_{1-\alpha}V^{1/2} - z_{1-\gamma}V^{1/2}}\right)^2 = \left(\frac{\nu_{1-\alpha}(R^2) - \nu_{1-\gamma}(R^2)}{z_{1-\alpha} - z_{1-\gamma}}\right)^2. \tag{S.2}$$

From Theorem 3, when $\gamma \geq \alpha$, we have $\nu_{1-\alpha}(R^2) - \nu_{1-\gamma}(R^2) \leq z_{1-\alpha} - z_{1-\gamma}$. This implies that $N_{\text{rr}}/N_{\text{cr}} \leq 1$ by (S.2).

We then consider the case with $\gamma \geq 0.5$. From Li et al. (2018, Theorem 2), both $\nu_{1-\alpha}(R^2)$ and $\nu_{\gamma}(R^2)$ are decreasing in R^2 and increasing in K and a, and they are less than or equal to $z_{1-\alpha}$ and z_{γ}, respectively. Consequently,

$$\begin{aligned}
\nu_{1-\alpha}(\tilde{R}^2)\tilde{V}^{1/2} - \nu_{1-\gamma}(R^2)V^{1/2} &= \nu_{1-\alpha}(\tilde{R}^2)\tilde{V}^{1/2} + \nu_{\gamma}(R^2)V^{1/2} \\
&\leq z_{1-\alpha}\tilde{V}^{1/2} + z_{\gamma}V^{1/2} \\
&= z_{1-\alpha}\tilde{V}^{1/2} - z_{1-\gamma}V^{1/2},
\end{aligned}$$

which implies that $N_{\text{rr}}/N_{\text{cr}} \leq 1$ by (S.1). Moreover, $\nu_{1-\alpha}(\tilde{R}^2)\tilde{V}^{1/2} - \nu_{1-\gamma}(R^2)V^{1/2} = \nu_{1-\alpha}(\tilde{R}^2)\tilde{V}^{1/2} + \nu_{\gamma}(R^2)V^{1/2}$ is decreasing in R^2 and increasing in K and a. This then implies that $N_{\text{rr}}/N_{\text{cr}}$ is decreasing in R^2 and increasing in K and a.

From the above, Theorem 5 holds.

F. Type-I Error Rates under Rerandomization and Complete Randomization

From Theorem 1, the type-I error rates of the α-level tests under rerandomization and complete randomization are, respectively,

$$\alpha_{\text{rr}} = \overline{\mathcal{V}}_{R^2}\left(\frac{\nu_{1-\alpha}(\tilde{R}^2)\tilde{V}^{1/2}}{V^{1/2}}\right) \quad \text{and} \quad \alpha_{\text{cr}} = \overline{\Phi}\left\{\frac{z_{1-\alpha}\tilde{V}^{1/2}}{V^{1/2}}\right\}.$$

It is challenging to compare α_{rr} and α_{cr} theoretically, due to the possible difference between R^2 and $\tilde{R}^2$. Nevertheless, we conjecture that $\alpha_{\text{rr}} \leq \alpha_{\text{cr}}$, due to the same reason discussed in §4.2 of the main paper:

We conservatively estimate the true distribution of $N^{1/2}(\hat{\tau}-\tau)$ by the same amount under both designs, and the true distribution under rerandomization is more concentrated around zero. See Appendix G for a numerical study.

Below we consider the limiting case with $a=0$, which can be a good approximation for rerandomization with a small threshold as suggested in Morgan & Rubin (2012). When $a=0$, $\overline{\mathcal{V}}_{R^2}$ simplifies to the survival function of $\mathcal{N}(0,1-R^2)$, and $\nu_{1-\alpha}(\tilde{R}^2)$ simplifies to the $(1-\alpha)$-quantile of $\mathcal{N}(0,1-\tilde{R}^2)$. Consequently, the type-I error rate under rerandomization simplifies to

$$\alpha_{\text{rr}} = \overline{\Phi}\left\{\frac{z_{1-\alpha}\tilde{V}^{1/2}(1-\tilde{R}^2)^{1/2}}{V^{1/2}(1-R^2)^{1/2}}\right\} = \overline{\Phi}\left\{\frac{z_{1-\alpha}(\tilde{V}-VR^2)^{1/2}}{(V-VR^2)^{1/2}}\right\}, \tag{S.1}$$

where the last equality holds because $VR^2=\tilde{V}\tilde{R}^2$. Because

$$\frac{\tilde{V}-VR^2}{V-VR^2}-\frac{\tilde{V}}{V}=\frac{VR^2(\tilde{V}-V)}{V(V-VR^2)}\geq 0,$$

we have $\alpha_{\text{rr}}\leq\alpha_{\text{cr}}$. Moreover, if $R^2>0$ and $\tilde{V}>V$, α_{rr} is strictly less than α_{cr}.

When rerandomization has a small threshold and the treatment effect has a small size, the power of rerandomization can be close to that in (S.1). This implies that the power under rerandomization can be smaller than that under complete randomization.

G. Testing Power Under Rerandomization Can Be Less Than Under Complete Randomization

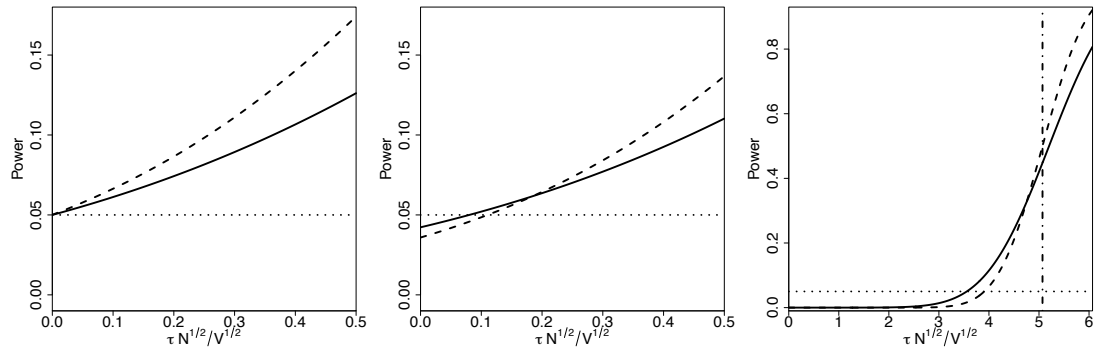
Theorem 4 establishes that testing power is greater under rerandomization than complete randomization when $\tilde{V}=V$ or when the treatment effect $\tau\geq\nu_{1-\alpha}(\tilde{R}^2)\tilde{V}^{1/2}N^{-1/2}$. Otherwise, rerandomization may exhibit less testing power than complete randomization, because inference under rerandomization can be more conservative than that under complete randomization, as discussed in Appendix F. This additional conservativeness decreases power, but the additional precision from rerandomization increases power. To illustrate this trade-off, we consider a simple numerical example below.

Suppose that $V=1$ and $R^2=0.5$. We consider two cases, which correspond to $\tilde{V}=V$ and $\tilde{V}>V$, as in Theorem 4. In Case (i), the probability limits of our estimators are the same as the corresponding truth, i.e., $\tilde{V}=V$ and $\tilde{R}^2=R^2$. Meanwhile, in Case (ii), inference is asymptotically conservative, in the sense that $\tilde{V}/V=1.1>1$ and $\tilde{R}^2=VR^2/\tilde{V}\approx 0.455$. Figure S1 shows the power (as in Theorem 1) of the 0.05-level one-sided test based on the mean-difference estimator under complete randomization and rerandomization, with the scaled average treatment effect $\tau N^{1/2}/V^{1/2}$ ranging from 0 to 0.5. From Fig. S1(a), when inference is not conservative, the power at $\tau=0$ equals the nominal level 0.05 under both designs, and rerandomization provides better power than complete randomization. From Fig. S1(b), when we can only conduct conservative inference, the power at $\tau=0$ is less than the nominal 0.05 under both designs, and moreover, the test is more conservative under rerandomization. However, the power of rerandomization quickly passes that of complete randomization when $N^{1/2}\tau/V^{1/2}$ is not too small, and the cutoff for $N^{1/2}\tau/V^{1/2}$ is much smaller than the theoretical cutoff $\nu_{1-\alpha}(\tilde{R}^2)\tilde{V}^{1/2}/V^{1/2}\approx 1.27$ in Theorem 4. In addition, in Fig. S1(c) we also consider Case (iii), which is the same as Case (ii), except that our inference is much more conservative with $\tilde{V}/V=10$. In this case, the power of rerandomization also passes that of complete randomization when $N^{1/2}\tau/V^{1/2}$ is not too small, and the cutoff becomes closer to the theoretical cutoff $\nu_{1-\alpha}(\tilde{R}^2)\tilde{V}^{1/2}/V^{1/2}\approx 5.07$.

H. Comparing Sample Size for Rerandomization and Complete Randomization

H.1. *Setup and Parameters*

Theorem 5 establishes, for any significance level α and power γ, the ratio between the sample size needed under rerandomization to achieve power γ and the sample size needed under complete random-



(a) Case (i): $\tilde{V} = V$. (b) Case (ii): $\tilde{V} = 1.1V$ (c) Case (iii): $\tilde{V} = 10V$

Fig. S1: The power for the 0.05-level one-sided test using the mean-difference estimator under complete randomization (solid line) and rerandomization (dashed line). The dotted horizontal line denotes 0.05. The dotdash vertical line in (c) refers to the threshold $\nu_{1-\alpha}(\tilde{R}^2)\tilde{V}^{1/2}$ as in Theorem 4.

ization:

$$\frac{N_{\text{rr}}}{N_{\text{cr}}} = \left\{ \frac{\nu_{1-\alpha}(\tilde{R}^2)\tilde{V}^{1/2} - \nu_{1-\gamma}(R^2)V^{1/2}}{z_{1-\alpha}\tilde{V}^{1/2} - z_{1-\gamma}V^{1/2}} \right\}^2 \tag{S.1}$$

where V denotes the variance defined in (2), $\tilde{V}$ denotes the probability limit of its corresponding estimator defined in (6), $\nu_\alpha(\rho^2)$ denotes the α-quantile of the distribution $(1-\rho^2)^{1/2}\epsilon_0 + \rho L_{K,a}$ in (3), and z_α denotes the α-quantile of the standard Normal distribution. This ratio depends on the number of covariates K, the correlation R^2, and the rerandomization threshold a. In this section, we present a simulation study to better understand how $N_{\text{rr}}/N_{\text{cr}}$ behaves for different K, R^2, and a, as well as varying levels of treatment effect heterogeneity. The smaller the ratio, the larger the benefits of rerandomization over complete randomization in terms of sample size.

We consider the dimension of covariates $K \in \{1, 10, 20, \ldots, 100\}$, correlation $R^2 \in \{0, 0.1, \ldots, 0.9\}$, and acceptance probabilities $p_a \in \{0.001, 0.01, 0.1\}$, where $p_a = \text{pr}(M \leq a)$, i.e., the probability that a given randomization fulfills the rerandomization criterion. For simplicity, we focus on significance level $\alpha = 0.05$ and power $\gamma = 0.8$, both of which are common values in the power analysis literature. We found that results were consistent across other values of α and γ. The sample size ratio $N_{\text{rr}}/N_{\text{cr}}$ in Theorem 5 also depends on the non-Normal quantiles $\nu_{1-\alpha}(\tilde{R}^2)$ and $\nu_{1-\gamma}(R^2)$, which in turn depend on K, R^2, and p_a. For each K, R^2, and p_a, we simulate 10^6 draws from the non-Normal distributions $(1-\tilde{R}^2)^{1/2}\epsilon_0 + \tilde{R}L_{K,a}$ and $(1-R^2)^{1/2}\epsilon_0 + RL_{K,a}$, in order to approximate the quantiles $\nu_{1-\alpha}(\tilde{R}^2)$ and $\nu_{1-\gamma}(R^2)$, respectively. Note that, when there is no treatment effect heterogeneity, $\tilde{R}^2 = R^2$; and when there is treatment effect heterogeneity, $\tilde{R}^2 = VR^2/\tilde{V}$.

We will first consider the case where there is no treatment effect heterogeneity, such that $S_\tau^2 = 0$ and thus $\tilde{V} = V$. As a result, the sample size ratio does not depend on the potential outcome variances S_1^2 and S_0^2. Then we will consider the case where there is treatment effect heterogeneity, and thus S_1^2, S_0^2, and S_τ^2 will affect the sample size ratio.

H.2. *Without Treatment Effect Heterogeneity*

Figure S2 displays $N_{\text{rr}}/N_{\text{cr}}$ for different combinations of K, R^2, and p_a. There are several observations from Fig. S2, all of which validate the statements made in Theorem 5. First, the ratio is always below 1. This confirms that there are always sample size benefits when running a rerandomized experiment, com-

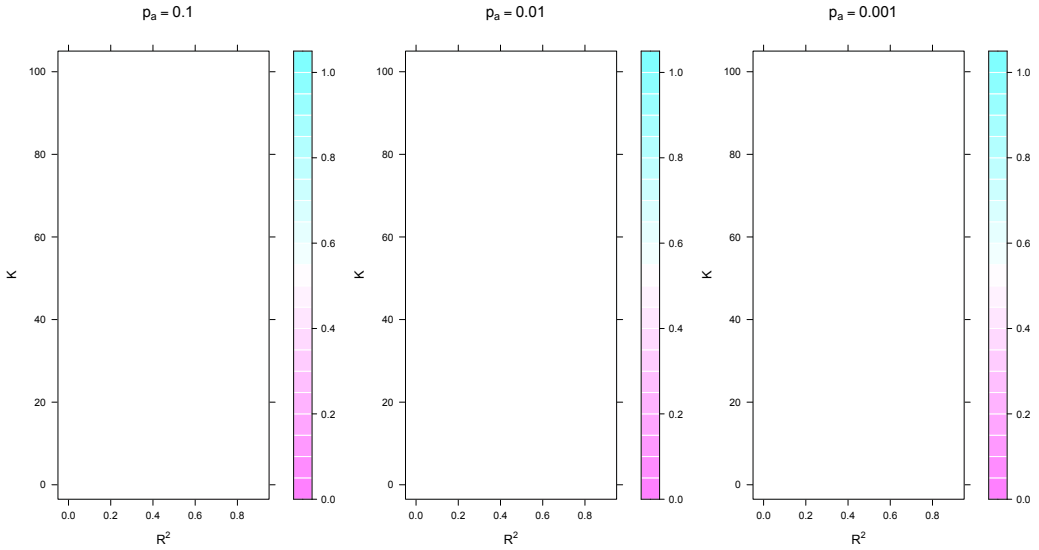


Fig. S2: The ratio $N_{\text{rr}}/N_{\text{cr}}$ for different K, R^2, and p_a, when there is no treatment effect heterogeneity.

Table S1: ANOVA Table Based on Results in Figure S2

Factor	Degrees of Freedom	Sum of Squares	Mean Sum of Squares	F value
p_a	2	0.450	0.225	1775.695
K	10	2.658	0.266	2096.851
R^2	9	7.059	0.784	6188.679
$p_a : K$	20	0.053	0.003	20.786
$p_a : R^2$	18	0.184	0.010	80.501
$K : R^2$	90	1.084	0.012	94.987
Residuals	180	0.023	0.000	

pared to a completely randomized experiment, at least when $\gamma \geq 0.5$. Furthermore, the ratio is decreasing in R^2 and increasing in K and p_a. This demonstrates that the sample size benefits of rerandomization are large when a stringent criterion is used to balance a few covariates that are strongly related with experimental outcomes. More generally, Fig. S2 shows that rerandomization can lead to substantial sample size gains: For example, if $p_a = 0.001$, the median of the ratios in Fig. S2 is 0.75, and if further $R^2 \geq 0.3$ and $K \leq 50$, the median of the ratios is 0.58. This suggests that rerandomization can reduce sample size by 25% to 40%, compared to complete randomization.

To better understand how K, R^2, and p_a impact the sample size ratio $N_{\text{rr}}/N_{\text{cr}}$, we used the results in Figure S2 to run an analysis of variance (ANOVA), where $N_{\text{rr}}/N_{\text{cr}}$ was used as the outcome and K, R^2, p_a, and their two-way interactions were included as factors. The resulting ANOVA table is in Table S1; all of the p-values for these factors were less than $2 \cdot 10^{-16}$. From the mean sum of squares and corresponding F values, it's clear that R^2 has the largest impact on the sample size ratio, followed by K, followed by p_a. This echoes simulation results in Li et al. (2018), who found that once the acceptance probability p_a is made adequately small, reducing it further does not have a large impact on variance reduction under rerandomization. In short, previous observations about how K, R^2, and p_a impact variance reduction under randomization seem to carry over for how they impact sample size reduction under rerandomization.

Because there is no treatment effect heterogeneity and thus $\tilde{V} = V$, the results in Fig. S2 hold for any degree of potential outcome variation S_1^2 and S_0^2 and any average treatment effect τ, as shown by (S.1). According to Theorem 2, the sample size needed to achieve power γ is increasing in S_1^2, S_0^2 and decreasing in τ for both rerandomization and complete randomization, which is a special case of rerandomization when $a = \infty$ or $R^2 = 0$. Thus, as S_1^2 and S_0^2 increase and as τ decreases, the nominal sample size gains from rerandomization can be arbitrarily large, at least as long as the desired power is greater than 50%. For example, consider conducting an experiment where we desire 80% power and $S_1 = S_0 = 4$. When $\tau = 2$, i.e. half a standard deviation, which is a medium effect according to a commonly used effect size rule-of-thumb by Cohen (2013), the necessary sample size under complete randomization is $N_{\text{cr}} \approx 99$. In this case, one may view the results in Fig. S2 as modest: If the covariates are modestly related to the outcomes ($R^2 = 0.3$), there are a moderate amount of covariates ($K = 50$), and we use a somewhat stringent rerandomization criterion ($p_a = 0.01$), we would expect only an approximately 13.3% reduction in sample size under rerandomization, or approximately 14 fewer subjects. However, when we consider a small effect $\tau = 0.8$, or one-fifth of a standard deviation, $N_{\text{cr}} \approx 619$. In this scenario, a 13.3% sample size reduction, or approximately 83 fewer subjects, may be considered quite large.

H.3. *With Treatment Effect Heterogeneity*

Now we consider the case where $S_\tau^2 > 0$, and thus power and sample size will depend on the potential outcome variances S_1^2 and S_0^2 in addition to S_τ^2. To our knowledge the literature has not discussed how treatment effect heterogeneity affects the power of completely randomized experiments, let alone rerandomized experiments. First we will discuss how treatment effect heterogeneity affects power and sample size for complete randomization and rerandomization, and then we will discuss how heterogeneity affects the sample size ratio $N_{\text{cr}}/N_{\text{rr}}$.

The asymptotic power for rerandomized experiments is characterized by Theorem 1; for fixed values of S_1^2, S_0^2, and τ, the power is increasing in S_τ^2 as long as $\tau \geq \nu_{1-\alpha}(\tilde{R}^2)\tilde{V}^{1/2}N^{-1/2}$; otherwise, it is decreasing in S_τ^2. Because complete randomization is a special case of rerandomization, a similar result holds for completely randomized experiments, where power is increasing in S_τ^2 as long as $\tau \geq z_{1-\alpha}\tilde{V}^{1/2}N^{-1/2}$. Thus, treatment effect heterogeneity has a beneficial effect on power for relatively large effect sizes but an adverse effect for relatively small effect sizes. Furthermore, because $\nu_{1-\alpha}(\tilde{R}^2) \leq z_{1-\alpha}$ for all $\alpha \in (0, 0.5]$, power is increasing in S_τ^2 for a wider range of effect sizes under rerandomization than under complete randomization. In other words, treatment effect heterogeneity is less likely to adversely affect power under rerandomization than under complete randomization. We illustrate this point further with numerical examples in Section I.

However, the results in the previous paragraph only hold when the variances S_1^2 and S_0^2 are fixed, and it's difficult to imagine a scenario where an increase in S_τ^2 does not also increase S_1^2, which adversely affects power. For example, previous works studying treatment effect heterogeneity have considered data-generating models like $Y_i(1) = Y_i(0) + \tau + \sigma_\tau Y_i(0)$ for some heterogeneity parameter σ_τ (Ding et al., 2016; Branson & Dasgupta, 2020). In this case, $S_1^2 = (1 + \sigma_\tau)^2 S_0^2$ and $S_\tau^2 = \sigma_\tau^2 S_0^2$, and thus more heterogeneity increases both S_1^2 and S_τ^2. Because power tends to be decreasing in S_1^2 and S_τ^2 for large S_1^2, this suggests that treatment effect heterogeneity generally has an adverse effect on power.

Meanwhile, from Theorem 2, the sample size necessary to achieve power γ is decreasing in S_τ^2 as long as $\gamma \geq 0.5$. Thus, for a fixed τ and power $\gamma \geq 0.5$, treatment effect heterogeneity has a beneficial effect on sample size for both completely randomized and rerandomized experiments. Indeed, this is analogous to the aforementioned results on power, because $\gamma \geq 0.5$ when $\tau \geq z_{1-\alpha}\tilde{V}^{1/2}N^{-1/2}$ for completely randomized experiments and when $\tau \geq \nu_{1-\alpha}(\tilde{R}^2)\tilde{V}^{1/2}N^{-1/2}$ for rerandomized experiments. However, if increased heterogeneity results in increased potential outcome variation, this may have an adverse affect on sample size, in the sense that increased S_1^2 will in turn increase sample size, as communicated in Theorem 2. These results are also illustrated further in Section I.

Finally, we consider how treatment effect heterogeneity affects the sample size ratio $N_{\text{rr}}/N_{\text{cr}}$. As communicated in Theorem 5, when $S_\tau > 0$, the sample size ratio depends on two conservative estimators: $\tilde{V}$, which impacts inference for complete randomization and rerandomization, and $\tilde{R}^2 = VR^2/\tilde{V}$, which only impacts inference for rerandomization. As a result, the sample size under rerandomization N_{rr} is

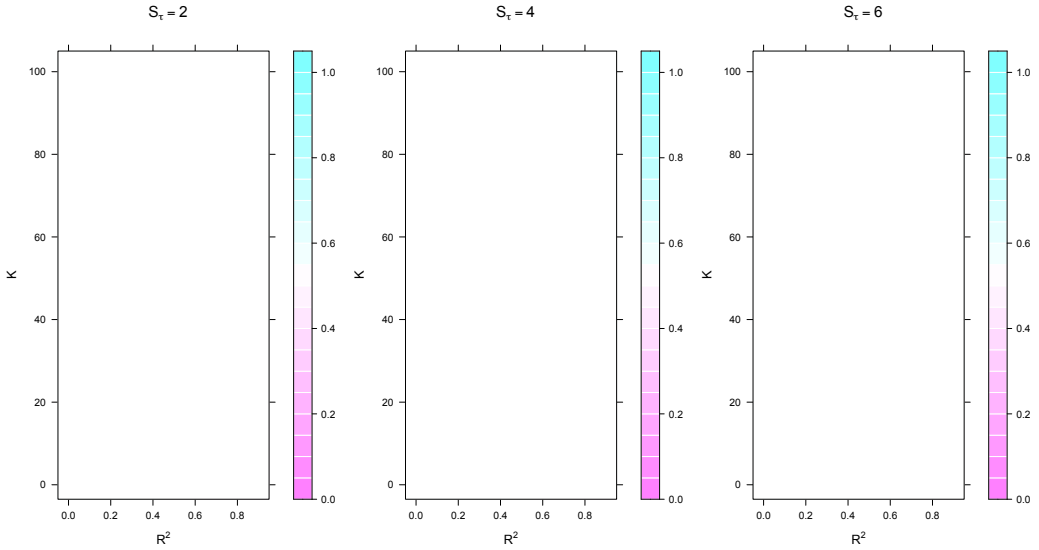


Fig. S3: The sample size ratio $N_{\text{rr}}/N_{\text{cr}}$ when running an experiment with $p_1 = p_0 = 0.5$ and $S_1 = S_0 = 4$, where $\alpha = 0.05$, $\gamma = 0.8$, and $p_a = 0.001$. The three panels correspond to heterogeneity $S_\tau \in \{2, 4, 6\}$.

doubly-impacted by the conservative estimator $\tilde{V}$, thereby diminishing the sample size benefits of rerandomization when there is treatment effect heterogeneity. To demonstrate, let's consider an experiment where $p_1 = p_0 = 0.5$ and $S_1 = S_0 = 4$, the significance level is $\alpha = 0.05$, the desired power is $\gamma = 0.8$, and acceptance probability is $p_a = 0.001$. Figure S3 shows the resulting $N_{\text{rr}}/N_{\text{cr}}$ for treatment effect heterogeneity $S_\tau \in \{2, 4, 6\}$ for different K and R^2. Many of the results from Fig. S2 still hold: $N_{\text{rr}}/N_{\text{cr}}$ is decreasing in R^2, increasing in K, and always below 1, as established by Theorem 5. However, we see that this ratio is increasing in the treatment effect heterogeneity S_τ^2; thus, rerandomization has less ability to reduce sample sizes when there is large treatment effect heterogeneity. When $S_\tau^2 = 2$, $N_{\text{rr}}/N_{\text{cr}}$ is on average 2.4% greater than when $S_\tau^2 = 0$; when $S_\tau^2 = 4$, $N_{\text{rr}}/N_{\text{cr}}$ is on average 9.9% greater; and when $S_\tau^2 = 6$, $N_{\text{rr}}/N_{\text{cr}}$ is on average 23.7% greater. However, $S_\tau = 6$ denotes unusually large effect heterogeneity, because it is larger than S_1 and S_0. Furthermore, it's important to remember that the ratio result in Theorem 5 holds for any τ; thus, as discussed in Section H.2, when τ is small, N_{cr} will be large, making even small multiplicative sample size reductions possibly worthwhile.

Furthermore, because the potential outcome variances S_1^2 and S_0^2 also impact power and sample size, they may also impact the ratio $N_{\text{rr}}/N_{\text{cr}}$. Let us consider the same example in Fig. S3, but where we fix $K = 10$ and vary S_1, S_0, S_τ, and R^2. Figure S4 shows the ratio for different values of S_1, S_0, S_τ, and R^2; in Fig. S4 we restricted the color scale to $[0.25, 1.0]$ to more easily see trends for this plot. We see that as S_1 and S_0 increase, $N_{\text{rr}}/N_{\text{cr}}$ somewhat decreases, signaling that rerandomization can lead to larger sample size reductions when potential outcome variances are high. However, it appears that treatment effect heterogeneity has a relatively larger adverse impact on these sample size reductions; in other words, there is more variation with respect to the vertical axis in Fig. S4 than the horizontal axis. Thus, if higher treatment effect heterogeneity in turn induces higher potential outcome variation, the adverse effects of heterogeneity will likely outweigh the beneficial effects of higher variation, thereby limiting the amount of sample size reductions we can expect from rerandomization.

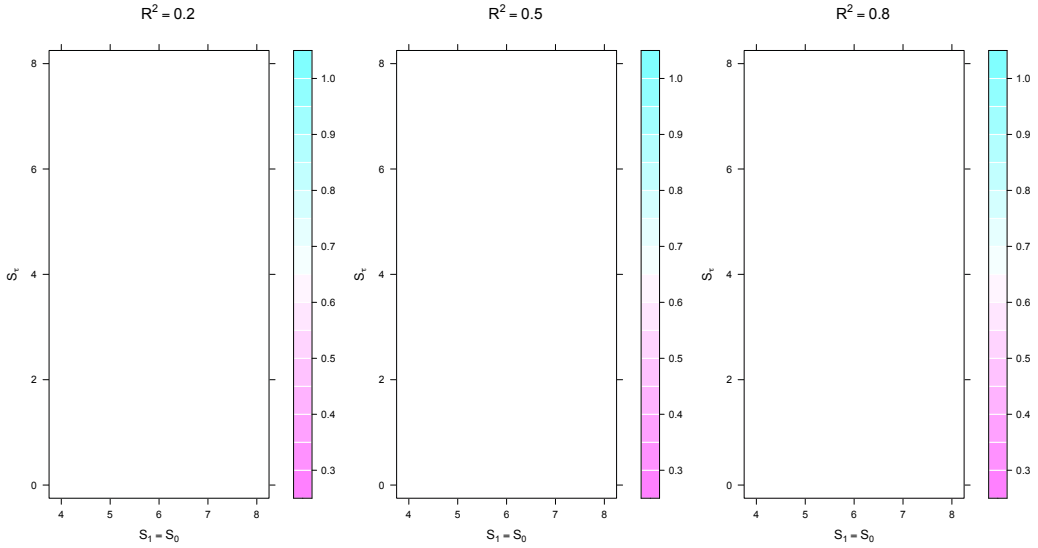


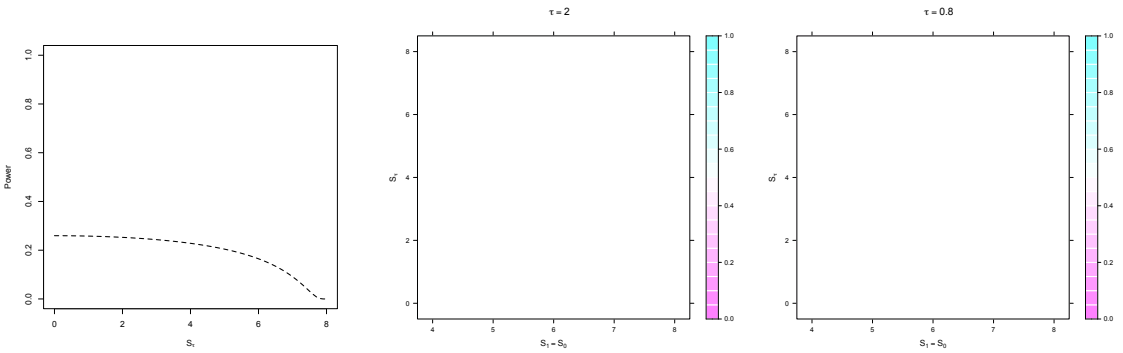
Fig. S4: The sample size ratio $N_{\text{rr}}/N_{\text{cr}}$ when running an experiment with $p_1 = p_0 = 0.5$ and $S_1 = S_0 = 4$, where $\alpha = 0.05$, $\gamma = 0.8$, $K = 10$, and $p_a = 0.001$. The three panels correspond to $R^2 \in \{0.2, 0.5, 0.8\}$.

I. Additional Numerical Examples: How Treatment Effect Heterogeneity Affects Power and Sample Size in Randomized and Rerandomized Experiments

In Section H.3, we discussed how treatment effect heterogeneity affects power and sample size for completely randomized and rerandomized experiments according to Theorems 1 and 2. In this section, we present numerical examples to supplement that discussion. First we will present examples for completely randomized experiments, because, to our knowledge, the literature has not discussed how treatment effect heterogeneity affects power for completely randomized experiments, let alone rerandomized experiments. Then, we will discuss how these examples apply to rerandomized experiments.

As mentioned in Section H.3, for fixed values of S_1^2, S_0^2, and τ, testing power for completely randomized experiments is increasing in S_τ^2 as long as $\tau \geq z_{1-\alpha}\tilde{V}^{1/2}N^{-1/2}$, where $\tilde{V}$ is the probability limit of the estimator for V, defined in (2). As a toy example, consider a completely randomized experiment where the proportions of treatment and control subjects are $p_1 = p_0 = 0.5$ and there are $N = 100$ subjects. Furthermore, say $S_1 = S_0 = 4$ and we use the ubiquitous variance estimator $\hat{V}_{\text{N}} = p_1^{-1}s_1^2 + p_0^{-1}$ and thus $\tilde{V}_{\text{N}} = p_1^{-1}S_1^2 + p_0^{-1}S_0^2$. Thus, power is increasing in S_τ^2 as long as $\tau \geq z_{1-\alpha} \cdot 8 \cdot 0.1 \approx 1.3$ for $\alpha = 0.05$. Figure S5a shows power for this toy example when we vary S_τ for $\tau = 2$ and $\tau = 0.8$; we see that power is monotonically increasing in S_τ for the former but monotonically decreasing for the latter. This suggests that treatment effect heterogeneity has a beneficial effect on power for large effect sizes but an adverse effect for small effect sizes. Thus, if we incorrectly assume $S_\tau^2 = 0$, which is common in power analyses, then we may underestimate power for large effect sizes but overestimate power for small effect sizes. Note that, in Fig. S5a, $S_\tau = 8$ is very extreme; in this case, $V = 0$, and thus power is either 0 or 1, depending on whether $\tau \geq z_{1-\alpha}\tilde{V}^{1/2}N^{-1/2}$.

Alternatively, we can consider a fixed average treatment effect τ and study power when we vary S_1^2 and S_0^2 in addition to S_τ^2. Figures S5b and S5c show the power for the aforementioned toy example for $\tau = 2$ and $\tau = 0.8$, respectively, for different values of S_1, S_0 and S_τ. When $\tau = 2$, power is monotonically increasing in S_τ, as we saw in Fig. S5a, but only for small values of S_1 and S_0; otherwise, it is monotonically decreasing. Meanwhile, we see that power is always monotonically decreasing in S_τ when $\tau = 0.8$. Furthermore, we see in Fig. S5b and S5c that power is monotonically decreasing in S_1 and S_0,



(a) Power when varying S_τ for $S_1 = S_0 = 4$, when $\tau = 2$ (solid line) and $\tau = 0.8$ (dotted line).

(b) Power when varying S_τ as well as S_1 and S_0 for $\tau = 2$. Power ranges from 32.4% to 100.0%.

(c) Power when varying S_τ as well as S_1 and S_0 for $\tau = 0.8$. Power ranges from 0.0% to 26.0%.

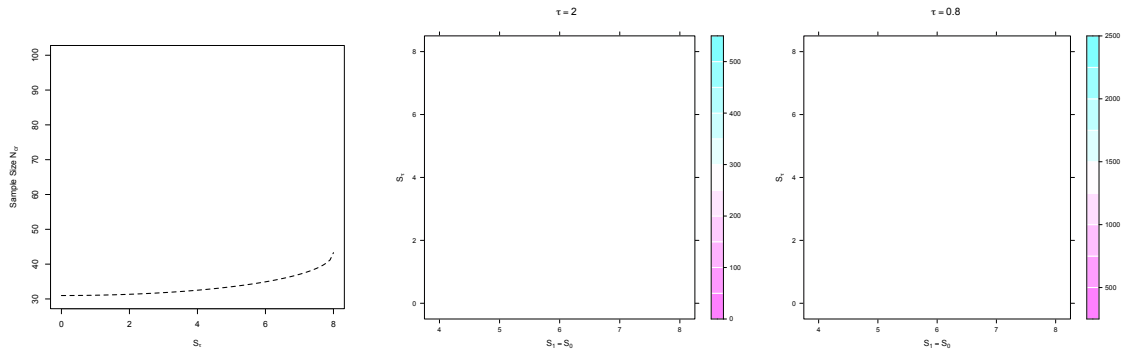
Fig. S5: Power of a completely randomized experiment under several scenarios when $p_1 = p_0 = 0.5$ and $N = 100$.

which is already a well-known phenomenon in power analyses. Taking all of Fig. S5 together, treatment effect heterogeneity can have an adverse effect on power if τ is small or the potential outcome variances are large. It also appears that the potential outcome variances tend to have a more consequential effect on power than treatment effect variation.

Now we consider sample size. When $\tau \geq z_{1-\alpha}\tilde{V}^{1/2}N^{-1/2}$, power will always be greater than or equal to 50%, as a consequence of Theorem 1. Thus, for fixed τ and power $\gamma \geq 0.5$, treatment effect heterogeneity actually has a beneficial effect on the required sample size to achieve power γ, in the sense that larger S_τ^2 leads to a smaller required sample size, as a consequence of Theorem 2. To demonstrate, let us again consider our toy example where $p_1 = p_0 = 0.5$, $S_1 = S_0 = 4$, and $\tau = 2$. Figure S6a displays the sample size N_{cr} to achieve power $\gamma = 0.8$ and $\gamma = 0.4$ under complete randomization for increasing values of S_τ. When $\gamma = 0.8$, sample size is increasing in S_τ, but it is decreasing in S_τ when $\gamma = 0.4$; this is analogous to our previous finding that power is increasing in S_τ only for treatment effects above a certain magnitude. Furthermore, note that in the extreme case when $S_\tau = 8$, $V = 0$, and thus N_{cr} is no longer a function of γ.

Again we can also consider varying S_1^2 and S_0^2, in addition to S_τ^2; the resulting sample size N_{cr} required to achieve power $\gamma = 0.8$ is shown in Figures S6b and S6c for $\tau = 2$ and $\tau = 0.8$, respectively. For both of these scenarios, the sample size is decreasing in S_τ^2, again suggesting that larger treatment effect heterogeneity can have a beneficial effect on the sample size N_{cr} if $\gamma \geq 0.5$. However, in Fig. S6b and S6c, we see that there is more variation in N_{cr} across the horizontal axis than the vertical axis. This suggests that potential outcome variances have a larger effect on sample size than treatment effect heterogeneity, validating common power analyses that focus on these quantities rather than treatment effect heterogeneity. In particular, if increased treatment effect heterogeneity in turn increases potential outcome variances, then in general heterogeneity may have an adverse effect on sample size.

Finally, we can also consider how treatment effect heterogeneity affects power and sample size for rerandomized experiments. According to Theorem 1, for fixed S_1^2, S_0^2, and τ, power under rerandomization is increasing in S_τ^2 as long as $\tau \geq \nu_{1-\alpha}(\tilde{R}^2)\tilde{V}^{1/2}N^{-1/2}$, where $\nu_{1-\alpha}(\tilde{R}^2)$ denotes the $(1-\alpha)$-quantile of the distribution $\sqrt{1-\tilde{R}^2}\epsilon_0 + \sqrt{\tilde{R}^2}L_{K,a}$ and $\tilde{R}^2 = VR^2/\tilde{V}$. Note that $\nu_{1-\alpha}(\tilde{R}^2) \leq z_{1-\alpha}$ for all $\alpha \in (0, 0.5)$, with equality only if $\tilde{R}^2 = 0$. Thus, the same conclusions made in this section for completely randomized experiments also hold for rerandomized experiments, but for smaller effect sizes. In other words, under rerandomization, a smaller τ is required in order for power to be increasing in S_τ^2; or



(a) N_{cr} when varying S_τ for $S_1 = S_0 = 4$ and $\tau = 2$, when $\gamma = 0.8$ (solid line) and $\gamma = 0.4$ (dotted line).

(b) N_{cr} when varying S_τ, S_1, S_0 for $\tau = 2$. N_{cr} ranges from approximately 44 to approximately 396.

(c) N_{cr} when varying S_τ, S_1, S_0 for $\tau = 0.8$. N_{cr} ranges from approximately 271 to approximately 2473.

Fig. S6: Sample size N_{cr} required to achieve power γ when running a completely randomized experiment with $p_1 = p_0 = 0.5$ under different scenarios. In (b) and (c), $\gamma = 0.8$.

conversely, a smaller sample size N_{rr} is required to achieve a certain level of power $\gamma \geq 0.5$, as established by Theorem 5.

J. Power and Sample Size Calculations using the R package rerandPower

Practitioners may be interested in implementing power and sample size calculations for completely randomized and rerandomized experiments based on the results presented in this paper. Our `R` package `rerandPower`, available on `CRAN`, has four functions: `power.rand()`, `power.rerand()`, `sampleSize.rand()`, and `sampleSize.rerand()`.

The functions `power.rand()` and `power.rerand()` compute power for given sample sizes N_1 and N_0, potential outcome standard deviations S_1 and S_0, treatment effect heterogeneity standard deviation S_τ, and average treatment effect τ. The significance level α can also be specified. Let's consider the toy example in Section I, where $N_1 = N_0 = 50$, $S_1 = S_0 = 4$, $\tau = 2$, and $S_\tau = 0$ or $S_\tau = 4$ for a completely randomized experiment. The following lines of code implement the power calculations for these two cases, presented in Figure S5:

```
> power.rand(N1 = 50, N0 = 50, s1 = 4, s0 = 4, tau = 2)
[1] 0.8037649
> power.rand(N1 = 50, N0 = 50, s1 = 4, s0 = 4, s.tau = 4, tau = 2)
[1] 0.838286
```

We see that power increases when $S_\tau > 0$ because $\tau \geq z_{1-\alpha}\tilde{V}^{1/2}N^{-1/2}$, as discussed in Section I. The calculation made in the first line of code, which by default sets $S_\tau = 0$, is widely available in other power analysis software; however, to our knowledge, other available software does not allow one to specify $S_\tau > 0$, as done in the second line of code.

Similar calculations can be made for rerandomized experiments using `power.rerand()`, except one also has to specify the number of covariates K, the correlation between covariates and potential outcomes R^2, and the acceptance probability $p_a = \text{pr}(M \leq a)$, where M is the Mahalanobis distance defined in (1). When designing an experiment in practice, one can control K and p_a, but of course one will not have knowledge about R^2 until the experiment has been conducted. Thus, R^2 should be specified based on subject-matter knowledge, or based on best- and worst-case scenarios. For example, consider conducting a rerandomized experiment where there are $K = 10$ covariates, $p_a = 0.01$, and there is a

moderate correlation of $R^2 = 0.3$. Then the power under rerandomization for the same toy example above, with $S_\tau = 4$, is:

```
> power.rerand(N1 = 50, N0 = 50, s1 = 4, s0 = 4, s.tau = 4, tau = 2,
                K = 10, pa = 0.01, R2 = 0.3)
[1] 0.901424
```

We see that, compared to the complete randomization example, power is higher in this case.

Meanwhile, the functions `sampleSize.rand()` and `sampleSize.rerand()` compute the sample size N necessary to achieve a prespecified level of power γ for given S_1, S_0, S_τ, τ, and sample size proportions $p_1 = N_1/N$ and $p_0 = N_0/N$. Let's again consider the toy example from Section I, where $p_1 = p_0 = 0.5$, $S_1 = S_0 = 4$, $\tau = 2$, and $S_\tau = 0$ or $S_\tau = 4$. The following lines of code implement sample size calculations for these two cases when power $\gamma = 0.8$, presented in Figure S6:

```
> sampleSize.rand(power = 0.8, s1 = 4, s0 = 4, tau = 2)
[1] 98.92092
> sampleSize.rand(power = 0.8, s1 = 4, s0 = 4, s.tau = 4, tau = 2)
[1] 90.15267
```

We see that the necessary sample size decreases when there is treatment effect heterogeneity, because $\gamma \geq 0.5$. Again, the calculation made in the first line of code is also widely available in other power analysis software, but to our knowledge, that in the second line of code is not.

We can again make similar calculations for a rerandomized experiment. Let's again consider the case where $K = 10$, $p_a = 0.01$, and $R^2 = 0.3$. Then, the sample size calculation for the case where $S_\tau = 4$ is:

```
> sampleSize.rerand(power = 0.8, s1 = 4, s0 = 4, s.tau = 4, tau = 2,
                    K = 10, pa = 0.01, R2 = 0.3)
[1] 72.6096
```

We see that, for this example, rerandomization requires a smaller sample size than complete randomization, which will always be the case when $\gamma \geq 0.5$, as established by Theorem 5. For this example, rerandomization decreases the sample size requirement by approximately 19.5%.

More details and examples are available in the documentation for `rerandPower` on `CRAN`.

REFERENCES

BRANSON, Z. & DASGUPTA, T. (2020). Sampling-based randomised designs for causal inference under the potential outcomes framework. *International Statistical Review* **88**, 101–121.

COHEN, J. (2013). *Statistical power analysis for the behavioral sciences*. Academic Press.

DING, P., FELLER, A. & MIRATRIX, L. (2016). Randomization inference for treatment effect variation. *Journal of the Royal Statistical Society: Series B: Statistical Methodology* , 655–671.

LEWIS, T. & THOMPSON, J. W. (1981). Dispersive distributions, and the connection between dispersivity and strong unimodality. *Journal of Applied Probability* **18**, 76–90.

LI, X. & DING, P. (2017). General forms of finite population central limit theorems with applications to causal inference. *Journal of the American Statistical Association* **112**, 1759–1769.

LI, X., DING, P. & RUBIN, D. B. (2018). Asymptotic theory of rerandomization in treatment–control experiments. *Proceedings of the National Academy of Sciences* **115**, 9157–9162.

MORGAN, K. L. & RUBIN, D. B. (2012). Rerandomization to improve covariate balance in experiments. *The Annals of Statistics* **40**, 1263–1282.

SHAKED, M. (1982). Dispersive ordering of distributions. *Journal of Applied Probability* **19**, 310–320.