Philosophy of Science (2026), **93**, 873–897
doi:10.1017/psa.2026.10203

PSA PHILOSOPHY OF SCIENCE ASSOCIATION

PHILOSOPHY OF SCIENCE

ARTICLE

What Is Wrong with Rerandomization in Randomized Field Experiments?

Mariusz Maziarz

The Institute of Philosophy, Faculty of Philosophy, Jagiellonian University, Krakow, Poland
Email: mariusz.mm@gmail.com

(Received 20 March 2025; revised 17 February 2026; accepted 19 February 2026; first published online 09 March 2026)

Abstract

Rerandomization is a new technique employed in randomized field experiments to enhance the balance of covariates across experimental groups. Rerandomization is recommended because it reduces imbalances in known covariates, thereby enhancing the precision of the average treatment effect estimates. However, employing rerandomization necessitates adjusting observed p-values, using regression-based inference, and selecting predictive covariates. All these amendments increase the number of researcher degrees of freedom, that is, methodological decisions involved in designing and analyzing experiments. I argue that this increased analytical flexibility may be misused to p-hack for statistically significant or preferred results, thereby reducing the credibility of the results.

1. Introductory remarks

The method of randomized experiments, which involves randomly assigning experimental units to treatment and control groups and then administering treatments to the trial arms to estimate the average treatment effect, seems homogeneous across the social and biomedical sciences. Social sciences, particularly economics, are among the main areas where the design of medical randomized controlled trials (RCTs) has been recently imported, mainly to deliver evidence for policy decisions by evaluating the efficacy of alternative programs. The movement of evidence-based policy reflects the earlier methodological debates concerning the quality of evidence in medicine, with the distinguishing characteristic of the Evidence-Based Medicine (EBM) and Evidence-Based Policy (EBP) being the prioritization of randomized trials and the combined results of such studies on the ground that the risk of bias is typically lowest for effect size estimates reported by such study types (La Caze 2009). Several philosophers have challenged the prioritization of RCTs (e.g., Worrall 2007; Cartwright and Hardie 2012), with most philosophical discussions concentrated on medical RCTs. Randomized field experiments (RFEs) used by social scientists to assess the efficacy of various development

© The Author(s), 2026. Published by Cambridge University Press on behalf of Philosophy of Science Association. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

https://doi.org/10.1017/psa.2026.10203 Published online by Cambridge University Press

programs and economic policies "have not attracted much attention [from philosophers] until recently" (Nagatsu and Favereau 2020).

Despite substantial similarities, RCTs and RFEs differ in several ways (Favereau 2016; Jiménez-Buedo and Russo 2021; Favereau and Nagatsu 2020). One such difference is how the baseline (pretreatment) balance between the treatment and control groups is evaluated and the subsequent solutions applied when such an imbalance occurs. In medicine, baseline characteristics of treatment and control groups are reported, but statistical tests of imbalance in covariates are currently frowned upon on methodological grounds (Altman and Dore 1990; Senn 1994). Furthermore, even if baseline imbalances are discovered in a medical trial, repeating randomized assignment (rerandomizing) is infeasible because patients are typically randomly assigned to treatments immediately after being recruited into a trial, which means that imbalances (such as a higher proportion of women receiving treatment) can only be identified after the trial concludes. The CONsolidated Standards of Reporting Trials (CONSORT) guidelines (Moher et al. 2014), followed by all major medical journals, explicitly prohibit such practices, even though testing for baseline imbalances was common in the 1990s and earlier.

Instead, social scientists typically experiment on units that can be recruited into a trial and randomized before the treatment and control interventions are given. The opportunity to test for baseline imbalances before administering interventions has led to a methodological amendment, as economists standardly check experimental groups for covariate imbalances and repeat randomized assignments after identifying such imbalances. For example, if twenty biomarkers predict the primary outcome in a clinical trial, simple randomization will, on average, lead to an assignment of patients to treatment and control groups such that the difference in the average value of one biomarker between the two groups will be statistically significant. The relatively high number of these covariates makes alternative assignment procedures, such as stratified randomization, infeasible. Rerandomization, which involves repeating randomized assignments until there are no statistically significant differences between experimental groups or generating many assignments and selecting the one that minimizes the differences in covariates, seems to be a useful way to balance groups and improve estimate precision.

Some prominent economists, including those who have won the Sveriges Riksbank Prize in Memory of Alfred Nobel, encourage such a practice on the grounds that rerandomizing enhances the precision of average treatment effect estimates (Peter and Soetevent 2019; Banerjee and Duflo 2017), though the extent of the practice and the degree to which it is currently endorsed are unknown. The existing methodological literature that appraises rerandomization because it improves balance between treatment and control groups and allows for reporting average treatment effect estimates with narrower confidence intervals. Rubin (2008) reported that Ronald Fisher and Bill Cochran both supported rerandomization, and he believed too that "[i]f a randomized allocation likely will lead to an imprecise result (i.e., the allocation is one with substantial potential for conditional bias given the observed values of covariates), then one should rerandomize" (p. 1351).

Later, Morgan and Rubin (2012, 2015) developed a rerandomization procedure aimed at minimizing a measure of imbalance and demonstrated that such

https://doi.org/10.1017/psa.2026.10203 Published online by Cambridge University Press

rerandomization increases estimate precision and enhances balance. Bruhn and McKenzie (2009) simulated the impact of rerandomization on covariate and outcome balance using economics data. Li and Ding (2020) analytically demonstrated precision gains from rerandomization and showed that rerandomized studies report more precise estimates of the average treatment effect. Banerjee et al. (2020) analyzed the trade-off between balancing covariates expected to be correlated with the outcome and robust inference independent from one's prior knowledge. So far, existing methodological literature has focused on the benefits of rerandomization, emphasizing better balance in experimental groups and improved precision. However, rerandomization affects observed p-values, which need to be adjusted; otherwise, their estimates become too conservative. This adjustment is a step often omitted by researchers using rerandomization (Mutz and Pemantle 2015).

In this article, I argue that the rerandomization used in RFEs enhances estimate precision at the expense of unbiasedness. This is because testing for baseline imbalances, rerandomizing, and adjusting statistical analysis complicate research design and data analysis. This heightened researcher degrees of freedom (Wicherts et al. 2016) creates more opportunities for researchers to misuse analytical flexibility to p-hack for statistically significant results, thereby diminishing the quality of evidence from RFEs utilizing rerandomization. P-hacking "is the intentional or unintentional misuse of statistical techniques, through the exploitation of researcher degrees of freedom, that may lead to artificially significant results for a predetermined hypothesis of interest" (Erasmus 2025, 2934).

P-hacking involves "trying multiple things" (Nuzzo 2014, 152), that is, running many statistical analyses based on different choices about design and analysis, then selectively reporting the results. To p-hack, researchers may try changing the dependent and independent variables, optionally stopping data collection, using different outlier exclusion procedures, including or changing covariates, redefining scales and transforming variables, discretizing variables in various ways, running alternative hypothesis tests, handling missing data favorably, dividing the sample into subgroups, and applying different rounding procedures (Stefan and Schönbrodt 2023). Because p-hacking involves misusing analytical flexibility, the effect of this questionable research practice on false-positive results depends on the researcher degrees of freedom: Simpler designs that require fewer analytical decisions allow for fewer statistical tests, while more complex designs enable more comparisons, increasing the chance of reporting a false-positive result. P-hacking commits a researcher to the problem of multiple comparisons and invalidates the assumptions of null hypothesis statistical testing (NHST) (Eisma and de Winter 2025). However, as Erasmus (2025, 2933) rightly notes in a footnote, p-hacking, despite the name suggesting otherwise, "can occur regardless of what statistical inference tools are being used, including other frequentist or Bayesian approaches." Less common terms include "hunting with a shotgun" (Mayo 1996, 300–1), data-dredging, or B-hacking/posterior hacking to distance it from NHST.

In the first study on analytical flexibility and p-hacking, Simmons et al. (2011) examined how the choice of the dependent variable, sample size, including covariates, and subgroup analyses could affect research outcomes. The simulation shows that the risk of false-positive results depends on the number of researcher degrees of freedom

https://doi.org/10.1017/psa.2026.10203 Published online by Cambridge University Press

and can be inflated to more than 60 percent if these four researcher degrees of freedom are used in a p-hacking attempt. However, their estimate is conservative because there are more than four simulated methods to push the p-value past the threshold of statistical significance (Stefan and Schönbrodt 2023). Despite some mixed findings, a growing body of evidence shows that p-hacking is common in published scientific literature. John et al. (2012) reported that 15–63 percent of surveyed US psychologists used various p-hacking strategies. The prevalence of p-hacking in published results can be examined by analyzing deviations in reported p-values from the expected distribution (Simonsohn et al. 2014). A p-curve analysis of articles indexed in Scopus shows a disproportionately high number of p-values between 0.4 and 0.5, suggesting possible p-hacking or other fraudulent practices such as data fabrication (De Winter and Dodou 2015). Using a similar method, Head et al. (2015) found that p-hacking is widespread in the life sciences.

The perspective or error statistics (Mayo 1996, 2018; Mayo and Spanos 2006) comes in handy for explaining the grave consequences of p-hacking, which diminishes the strictness of hypothesis tests: With enough motivation and unlimited analytical flexibility, nearly any hypothesis can be shown to have some statistically significant effect (Lakens 2019). This invalidates learning from failed predictions and correcting hypotheses and experimental design in response to discovered errors (see Allchin 2001 for a discussion of error types in science). In effect, empirical literature contains false-positive results that cannot be reliably replicated (Romero 2019). Preregistration, which involves planning research design and data analysis before conducting an experiment, was suggested to limit researcher degrees of freedom and thus p-hacking (Nosek et al. 2018). Rerandomization, by increasing analytical flexibility and p-hacking risks, poses a threat of introducing bias into randomized experiments.

The article's structure is outlined as follows. In section 2, I discuss how randomized assignment ensures unbiasedness but does not guarantee estimate precision, that is, the actual effect of the intervention under test can differ from its expected value. In section 3, I summarize the arguments favoring rerandomization on the grounds that it improves estimate precision. However, as I argue in section 4, there is a trade-off between estimate precision and result credibility because rerandomization inflates researcher degrees of freedom and thus creates more opportunities for p-hacking. In section 5, I analyze the trade-off between improved estimate precision and reduced result credibility and point toward situations when either is preferable. Section 6 concludes.

2. Randomized assignment guarantees unbiasedness but not necessarily precision

Despite significant variability among various philosophical accounts of causality, causal effects are typically defined by the difference between a situation when a cause C occurs and an alternative when it does not ($\sim C$). For example, Cartwright's (2010, 1983) probabilistic theory states that C causes E when the probability of the effect E occurring is greater when C is present than when it is absent, considering other factors influencing E (K_i):

https://doi.org/10.1017/psa.2026.10203 Published online by Cambridge University Press

$$P(E \mid C\&K_i) > P(E \mid \sim C\&K_i)$$

The causal effect can be deductively inferred from the so-called ideal RCTs, that is, studies where all other covariates are evenly distributed across the treatment and control groups (Cartwright 2007). This requirement to balance the joint effect of all other covariates (K_i) (see Maziarz 2025; Philippi 2022; Fuller 2019) can be easily shown using the potential outcome framework (POA). POA defines individual treatment effect ($TE(n)$) as the difference in outcomes observed under treatment ($Y_T(n)$) and control ($Y_C(n)$). If we could observe individual outcomes in hypothetical scenarios in which treatment and control are administered, then the average treatment effect at the population level could be calculated by summarizing individual treatment effects:

$$ATE = \frac{1}{n}\sum_{i=1}^{n} TE(n)$$

Given that observing outcomes for the same individuals in counterfactual scenarios is impossible (Rubin 2005, 1974), and that the arithmetic mean of the differences equals the difference in the means, the solution to the fundamental problem of causal inference is to estimate the average treatment effect by comparing average outcomes in the treatment and control groups:

$$\widehat{ATE} = \frac{1}{n}\sum_{i=1}^{n} Y_T(n) - \frac{1}{m}\sum_{j=1}^{m} Y_C(m))$$

If the only difference between the causal factors determining outcomes in the treatment and control groups is the intervention, $\widehat{ATE}$ reflects the true average causal effect. Randomization ensures that potential outcomes are statistically independent of any covariates (Deaton and Cartwright 2018); that is, the only difference between the treatment and control groups is, in the limit, the treatment, and hence $\widehat{ATE}$ is an unbiased estimate of the true causal effect. Unfortunately, the perfect balance of the joint effect of confounders could only be observed in an ideal RCT with infinite sample size or, alternatively, in an infinite series of actual trials with finite sample sizes.

In actual experiments with finite samples, the average treatment effect estimate ($\widehat{ATE}$) remains an unbiased estimate of the true causal effect (ATE_{true}) (Greenland 1990; Neyman 1990[1923]), but the price for the stochastic distribution of confounders is the introduction of noise (ε) (Mutz et al. 2019): The average treatment effect may differ from the true causal effect by an arbitrarily large amount (see Worrall 2007), but the distribution of this difference can be estimated. It arises from an "unlucky" random assignment, which led to imbalanced treatment and control groups. For instance, more young people with better prognoses or classes with more gifted students were accidentally assigned to the treatment group, resulting in an overestimated effect size.

$$ATE_{true} = \widehat{ATE} + \varepsilon$$

$$\varepsilon = \sum_{j=1}^{z}(K_{ci}) - \sum_{i=1}^{y}(K_{ti})$$

https://doi.org/10.1017/psa.2026.10203 Published online by Cambridge University Press

The difference between ideal and actual RCTs is reflected in how inferences about causal effects are drawn in actual trials. While ideal RCTs report estimates that match the true causal effects, that is, causal conclusions follow deductively, the probability that an actual RCT might encounter an unlucky draw with imbalanced groups can be determined by examining outcome variance and sample size, which involves testing for a statistically significant difference in outcomes observed in the treatment and control groups (see Cook and De Mets 2007 for details). The variance of $\widehat{ATE}$ reported by an actual trial arises from the random differences in the distribution of the joint effect of covariates among treated and control instances. Although the probability of a false-positive result is always equal to the threshold of statistical significance (usually 0.05), estimate precision—described by the narrowness of the confidence interval—is positively related to sample size and negatively related to the heterogeneity of individual treatment responses. Researchers can enhance estimate precision by increasing sample size (which typically incurs ethical and/or financial costs) or managing the balance of covariates across trial arms.

3. Repeating randomized assignment improves balance and estimate precision

Researchers can employ several assignment procedures other than simple randomization to control for balance. These include blocked random assignment, matched pair random assignment, random subsampling, stratified randomization, and rerandomization. Some methods, such as stratified randomization, have long been used to balance known covariates at the design stage, and do not raise methodological concerns except for the limitation of the number of prognostic factors that can be balanced (Taves 1974). Other approaches, like rerandomization, have been developed more recently and await an in-depth methodological and philosophical assessment. In addition to designing randomized assignments to balance selected covariates, researchers can also reduce noise and improve estimate precision by statistically controlling for known prognostic factors. This can be achieved by using ANCOVA (analysis of covariance) instead of ANOVA (analysis of variance) or by incorporating covariates into a regression.

The rationale for improving the balance of covariates—whether in assigning patients to treatments or adding a covariate in a regression—is that balancing a covariate K_n will reduce the portion of outcome variance (${\sigma^2}_{outcome}$) caused by that covariate. Given that the remaining part of ${\sigma^2}_{outcome}$ is orthogonal to K_n, it will remain unaffected (see Bruhn and McKenzie 2009, 217). For instance, if age accounts for 10% of ${\sigma^2}_{outcome}$, balancing age will reduce the variance by 10 percent and will have no impact on the remaining 90 percent of ${\sigma^2}_{outcome}$. In other words, balancing an observable covariate does not affect the balance of unobserved covariates, neither positively nor negatively. Bruhn and McKenzie (2009) demonstrated this by simulating the rerandomization of RFE data and examining the balance in covariates that were assumed to be unobservable by experimenters. Li and Ding (2020) analytically proved this result, demonstrating that, in terms of estimate precision, "[r]erandomization trumps complete randomization in the design stage, and regression adjustment trumps the simple difference-in-means estimator in the analysis stage" (241).

https://doi.org/10.1017/psa.2026.10203 Published online by Cambridge University Press

Rerandomization improves $\widehat{ATE}$ estimate precision and has no negative consequences for the balance of unobserved covariates but claiming it has no negative effect on result credibility is, as I argue later, too quick of a conclusion. Credibility has been discussed as long as philosophy has existed, but I use this notion in the commonsensical sense. Accordingly, credibility denotes "the fact that . . . something can be believed or trusted" (Credibility 2026). When applied to experimental research, such a concept implies that studies are more credible when reported results reflect the actual effects of interventions rather than false-positive outcomes or results that emerge from the use of questionable or fraudulent practices.

Unfortunately, the RFE reports do not address the assignment procedure, suggesting that economists might anticipate reviewers or the audience for their RFEs to frown upon repeating randomized assignment. Based on their survey, Bruhn and McKenzie (2009) observed that only one out of eighteen papers published in prestigious venues described the rerandomization procedure, even though its use, as reported by economists participating in their survey, was much more common. The situation has not changed much, and my crude search for "rerandomization" or "rerandomized" in the articles published in the *American Economic Review* over the last four years returned only two results. One is a methodological article (Banerjee et al. 2020), and the other mentioned rerandomization in the context of a two-stage randomized controlled trial design (Leaver et al. 2021). Underreporting also affects more mature experimental sciences. For example, the type of randomization procedure (simple, stratified, block, etc.) was reported in fewer than a quarter of phase III clinical trials (Lai et al., 2021) and in approximately 70 percent of clustered randomized controlled trials testing public health interventions in developing countries (Shaw et al. 2025).

In December 2007, Bruhn and McKenzie (2009) surveyed thirty-five economists conducting field experiments. Although these results may not be representative of current practices among experimental economists, given the small sample and the field's evolution in response to criticism (see Ogden 2020), there is no better evidence on the extent to which rerandomization is used, as no more recent surveys have been conducted. According to the survey, economists rerandomize assignment to treatment and control groups mainly in two ways: either by deciding to repeat randomization when group assignment is not balanced (what is assessed subjectively or based on a prespecified criterion such as a failed balance test), or by conducting a substantial number of randomizations (a hundred or a thousand) and then selecting the assignment that minimizes a measure of balance among the treatment and control arms. This approach resembles the rerandomization procedure proposed by Morgan and Rubin (2012, 2015), who used Mahalanobis distance to measure the balance of chosen covariates.

Regardless of which of the two ways of rerandomization was used, the statistical inference should reflect the significant improvement in balance across trial arms compared to the simple randomization procedure (i.e., a single and unconstrained random assignment of experimental units). This is so because the distribution of the difference in means under the null hypothesis no longer accords with the Student's t distribution. As randomization succeeds in improving (expected) balance across trial arms, it "changes the distribution of the test statistic, most notably by decreasing the

https://doi.org/10.1017/psa.2026.10203 Published online by Cambridge University Press

true standard error" (Morgan and Rubin 2012, 1264). This undermines the usual approach to hypothesis testing by making it too conservative, that is, reported observed p-values are larger than actual probabilities of obtaining such or more extreme results.

In effect, if standard statistical tests are used despite rerandomization, null hypotheses are rejected less often than the nominal significance level and the power of the test is lower. Bruhn and McKenzie (2009) observed that "the correct statistical methods for covariate-dependent randomization schemes, such as minimization, are still a conundrum in the statistical literature" (219), and the situation did not change much. For example, Zhao and Ding (2024) recently reported Monte Carlo results suggesting that standard regressions are not suitable for estimating $\widehat{ATE}$ when randomized assignment was repeated after discovering baseline imbalances in covariates and suggested some plausible improvements, but these amended regression-based inference methods are far from being widely endorsed and accepted as a standard of data analysis in rerandomized experiments.

4. Rerandomization, researcher degrees of freedom, and the problem of credibility

Wicherts et al. (2016) listed all methodological decisions required to design, execute, and analyze a psychological experiment, but the concept applies equally well to medical RCTs or RFEs in economics. These "researcher degrees of freedom" refer to the various choices and decisions that researchers make during the process of designing, conducting, and analyzing experiments. These degrees of freedom can include choices about study design, measuring outcomes, preprocessing data, statistical analyses, and how results are interpreted. Researchers might explore various paths and make subjective but methodologically sound decisions. Researcher degrees of freedom can introduce bias, as different teams analyzing the same data or attempting replication may report varying results. Many analyst studies vividly illustrate this issue (e.g., Silberzahn et al. 2018). A well-known example of such a situation concerning randomized field experiments is the so-called deworming controversy, where the efficacy of antiparasitic drugs administered to the entire population of children in the developing world (Miguel and Kremer 2004) was challenged by a reanalysis of data using an alternative statistical approach (Davey et al. 2015). Generally speaking, inflating the number of methodological decisions creates more analytical flexibility that may be misused by researchers to report statistically significant or favorable results, while limiting researcher degrees of freedom improves the trustworthiness and replicability of results. This is the purpose of preregistration, which is currently required by CONSORT guidelines in medicine and all American Economic Association journals.

Even though rerandomization succeeds in improving estimate precision, it inflates the number of researcher degrees of freedom involved in planning and executing RFE and analyzing data. It does so by creating the need to:

1. Choose predictive covariates, which is not straightforward (section 4.1.);
2. Adjust the observed p-values, a task that is poorly understood by current statistical theory (section 4.2); and

https://doi.org/10.1017/psa.2026.10203 Published online by Cambridge University Press

3. Analyze data in a way that accounts for the covariates used in a rerandomization procedure (section 4.3).

This increased analytical flexibility reduces the reliability of the results, even when the rerandomization procedure is fully preregistered because different analysts may commit to alternative forking paths that yield different results despite similar data. However, increasing evidence indicates that preregistration and preanalysis plans do not fully determine methodological decisions in RFEs (Brodeur et al. 2024) or medical RCTs (van Drimmelen et al. 2024). For this reason, employing rerandomization and data analysis implied by this assignment procedure improves estimate precision, although it diminishes the credibility of the results reported by rerandomized experiments because inflated analytical flexibility may yield more p-hacking opportunities.

As correctly observed by one of the anonymous referees for *Philosophy of Science*, the standard methods used to improve balance across experimental groups pose similar challenges concerning inflated researcher degrees of freedom. Blocking to improve balance across experimental groups requires identifying confounders that are predictive of the primary outcome, defining the block size based on the number of controlled confounders, and randomizing treatment assignment within blocks. Stratified randomization involves selecting predictive covariates and defining strata and is often combined with block random assignment. These assignment procedures also violate the assumption that experimental units are assigned to the treatment and control groups independently of their characteristics and therefore require a more complex statistical analysis that adjusts for balanced confounders (Broglio 2018). To do so, researchers need to choose the correct multivariate model, considering the data-generating process (Kernan et al. 1999).

When these methods were in their infancy, their use provided more freedom in research design and greater flexibility during analysis because statistical theory had not yet fully understood the effects of blocked randomization on the distribution of outcomes (Ogawa 1961). The advances in mathematical statistics and a wider use of these methods to enhance experimental group balance led to a reduction in researcher degrees of freedom. Nevertheless, medical researchers have favored complete randomization because it produces experimental results that are less likely to be biased compared to other assignment methods used to improve balance in the experimental groups (Lachin et al. 1988). Stratification and block randomization can be contrasted with matched pair randomization, which appears to introduce more researcher degrees of freedom, as several alternative matching algorithms and distinct statistical estimation procedures are currently in use. These conflicting approaches to pair matching and statistical analysis "have inspired a heated debate in the literature for over sixty years" (Balzer et al. 2015). Using covariate adjustment either as a separate tool to balance experimental groups or together with these assignment procedures, which is required due to their invalidation of the assumptions underlying randomization-based inference, additionally inflates researcher degrees of freedom and may lead to a higher rate of false-positive results when a data-driven method is used for selecting covariates (Kahan et al. 2014).

Using rerandomization instead of standard methods like blocking, stratification, and matching introduces the same types of researcher degrees of freedom involved in

https://doi.org/10.1017/psa.2026.10203 Published online by Cambridge University Press

designing the assignment process, making decisions about its details, and then analyzing data using model-based inference. However, due to the novelty of rerandomization, the decisions involved in designing the experiment and analyzing data appear less constrained. This may increase the risk of bias if researchers exploit this flexibility to achieve particular results, especially when their actions are influenced by a conflict of interest. Unfortunately, except for some medical researchers noting that simple randomization reduces bias risk (e.g., Lachin et al. 1988), the potential trade-off between improving experimental group balance through blocking, stratification, and matching and diminished result credibility resulting from inflated researcher degrees of freedom has not been studied systematically. This trade-off is also missing from the debate about rerandomization.

4.1. The choice of predictive covariates

The first type of researcher degrees of freedom introduced by using rerandomization in an experiment is the choice of covariates that will be balanced. A rule of thumb is to use covariates that are the most predictive of outcomes because balancing covariates orthogonal to the outcome measure has an effect on estimate precision that is contrary to what is expected; namely, it will lower estimate precision and increase the length of the confidence interval. This is so because the covariates considered in a rerandomization procedure should be included in the statistical analysis of data, which increases degrees of freedom and, therefore, positively affects the observed p-value.

Unfortunately, selecting the predictive covariates is not straightforward, particularly in fields where the theoretical understanding of phenomena is underdeveloped or where multiple competing theories exist. This is often the case in the social sciences, where rerandomization is primarily used. For example, Goldfarb (1997, 232) observed that one of the explanations for conflicting empirical results is the development of economic theory and economists' engagement in data dredging, which aims to confirm current theory. Moosa (2019, 353) argued that conflicting econometric models support alternative theories in growth economics. However, choosing predictive covariates can also pose a challenge in the natural sciences. Medical research often proceeds based on guesses and hypotheses that are later confirmed through empirical research. A theoretical understanding of why a drug works and which factors interact with treatment effect is frequently lacking (Varga 2024), as demonstrated by examples such as fruits curing scurvy with the mechanism of action involving vitamin C remaining unknown long after efficacy was proved (Baxby 1997) and selective serotonin reuptake inhibitors (SSRIs), which are effective in a subgroup of depression patients but predictive factors remain unknown (Healy 2015).

In essence, the selection of predictive covariates to balance in a rerandomization procedure allows for considerable flexibility, providing the opportunity to balance different covariates based on whether researchers utilize theoretical insights (sometimes stemming from competing theories) of the phenomenon in question or rely on statistical analyses. Assman et al. (2000), in their criticism of reporting regression-based results in clinical trials, observed that "[r]esearchers commonly cannot predeclare the strong predictors, so the choice of a covariate-adjusted analysis

https://doi.org/10.1017/psa.2026.10203 Published online by Cambridge University Press

is determined by a variable selection procedure." (1067). A practice common in medical RCTs is to report some standard characteristics of the experimental groups such as average age, sex ratio, comorbidities, and so forth. Those researchers who rerandomize after observing a failed baseline test (ReP rerandomization) may consider those standardly reported characteristics even if their correlation to outcomes is limited. However, such a procedure would involve balancing covariates with unknown predictive power and can be frowned upon on the ground that one tests for the hypothesis that experimental units were randomly assigned (which is known to be true) and hence only false-positive results can be discovered (Senn 1994; Mutz et al. 2019).

Furthermore, at least in principle, the additional degrees of freedom stemming from covariate choice may be used in a p-hacking attempt. In his defense of rerandomization, Rubin (2008) argued that rerandomization does not create an opportunity to fraudulently bias results because, at the stage when it is planned and executed, no outcome data are available. However, data from rerandomized experiments require more complex statistical analysis (see sections 4.2 and 4.3), which increases the opportunities for p-hacking. Additionally, experimental data are analyzed after observing treatment outcomes. Furthermore, it seems that this opinion underestimates researchers' creativity in their attempts at reporting publishable results, and the asymmetry of information available to researchers conducting a trial and the audience.

There seem to be two different ways a researcher could use rerandomization to bias the effect size estimate. First, they might use rerandomization to report balancing selected covariates but imbalance other, more predictive factors to produce spurious results. Such a misuse of rerandomization is particularly plausible given that rerandomization often goes unreported (Bruhn and McKenzie 2009). Schulz and Grimes (2002) provided evidence suggesting that either medical researchers can accurately predict which baseline characteristics are balanced by simple randomization, or that some studies only report baseline characteristics of treatment and control groups that do not differ significantly. Second, they might honestly rerandomize to improve experimental group balance and then use the more complicated and less constrained data analysis required by rerandomization to push the estimate's p-value past the threshold of statistical significance. The latter method of biasing results during analysis, which aligns with Rubin's view that the assignment procedure is used when treatment outcomes are unknown, appears especially convincing, considering the observation of Barrett and Carter (2010) that in the social sciences "[t]hese crucial details of how design deviates from implementation are almost never reported in papers that employ experimental methods, unlike in the natural sciences, where the exact details of experiments are systematically recorded and shared with reviewers and made publicly available to readers for the purpose of exact replication" (523). Furthermore, registration reports of randomized field experiments often lack a preanalysis plan and are insufficiently detailed to limit p-hacking (Brodeur et al. 2024).

In the limit, studies that balance different subsets of predictive covariates in a rerandomization procedure will report the same results (that will also agree with $\widehat{ATE}$ reported by a study relying on simple randomization). However, focusing on balancing various subsets of predictive covariates may negatively impact the

https://doi.org/10.1017/psa.2026.10203 Published online by Cambridge University Press

replicability and, thus, the credibility of results in individual experiments. Furthermore, simulation studies show that even rerandomizing highly predictive covariates balances future outcomes to a limited degree compared to the effect of balancing baseline outcomes (Bruhn and McKenzie 2009). This suggests that a workaround for the problem of choosing predictive covariates is to focus balancing efforts on the baseline outcome, which is a univariate measure of the joint effect of all (known and unknown) confounders (Fuller 2019). The feasibility of this solution hinges on the specific characteristics of the domain, with greater applicability in social sciences compared to medical trials, mainly due to the nature of certain outcome measures employed in health sciences. In this context, a simpler randomization procedure, such as stratified randomization, may be preferable to rerandomization. Such a solution should be preferred as rerandomization involves other methodological decisions not determined by existing statistical theory.

4.2. Adjusting observed p-values

Because rerandomization is successful at improving experimental group balance and treatment effect estimate precision, "the appropriate analysis given the rerandomization is different from the one based on ignoring the rerandomization. Specifically, p-values need to be adjusted" (Banerjee et al. 2017, 109). However, if rerandomization is not reported in papers describing RFEs, economists do not account for it in the data analysis stage. In effect, researchers do not adjust observed p-values because, without reporting rerandomization, doing so would not be justified and could be perceived by the audience as an attempt at p-hacking. As Bruhn and McKenzie (2009) reported, "it appears that many researchers ignore the method of randomization in inference.... [T]his leads to hypothesis tests with incorrect size [...and] the standard errors are overly conservative when the method of randomization is not controlled for in the analysis, implying that researchers may not detect treatment effects that they would detect if the inference did take into account the randomization method" (202). Even though reporting too conservative p-values may not seem a particularly harmful problem if one is mainly concerned with the risk of false-positive results, simulation shows that treating some rerandomized experiments as relying on simple randomized assignment may lead to "less than 10 percent of replications having p-values under 0.10" (ibid., 219).

However, even if researchers using rerandomization to enhance covariate balance are willing to adjust p-values accordingly, the process remains ambiguous due to the underdevelopment of current statistical theory. As I mentioned previously, rerandomization comes in two flavors. It can be done by randomly assigning experimental units a specified number of times and then selecting the assignment that minimizes imbalance measured as a function of chosen covariate values (usually Mahalanobis distance concerning chosen covariates), which is the main rerandomization approach discussed in the statistical literature (referred to as ReM). Alternatively, researchers rerandomize when a baseline imbalance in a covariate(s) of interest is detected either through a statistical test (ReP) or observed subjectively.

The latter approach to rerandomization (ReP) is currently insufficiently studied, and statistical theory lacks an understanding of how such a procedure impacts p-values beyond that it inflates the type-II error rate and reduces the type-I error

https://doi.org/10.1017/psa.2026.10203 Published online by Cambridge University Press

rate. Apart from the issue of applying statistical testing to a problem, where only a false-positive result can be discovered (provided that one trusts their assistants' abilities to randomize properly) (Altman and Dore 1990; Senn 1994), such a procedure "only results in narrow improvements in balance over a single random draw" (Bruhn and McKenzie 2009, 212). This suggests that ReM may be a preferable rerandomization procedure. Zhao and Ding (2024) is the first attempt to estimate the impact of the "test baseline imbalance in covariates and rerandomize" (ReP) procedure on estimate precision. They did not establish a statistical theory for testing the null hypothesis based on data produced by such rerandomization; but merely noted that such an inference method can be imported and adjusted from their earlier paper concerning the adjustment of the Fisher randomization test for the ReM rerandomization (Zhao and Ding 2021). This gives significant flexibility to researchers, who may potentially use it to shift the reported p-values beyond the threshold of statistical significance, given the evidence that more analytical flexibility in economic research leads to more p-hacking (Brodeur et al. 2020). Zhao and Ding (2024) simulated nine different rerandomization and inference schemes, which add to the several decisions concerning the choice of balanced covariates and decision rules about when to rerandomize. Among the experimental economists surveyed by Bruhn and McKenzie (2009), almost one-third decided subjectively to rerandomize at least once, while only 12 percent used a statistical rule based on testing for baseline imbalances to redraw. If researchers decide to rerandomize based on a subjective, informal assessment of imbalance, they will be unable to adjust observed p-values due to their subjective and ad hoc nature, with unknown implications of repeating randomized assignment for the distribution of potential trial results.

The impact of the minimization-based rerandomization procedure (ReM) on observed p-values is now better understood; thus, adjusting for rerandomization is more straightforward. However, it still introduces significant analytical flexibility, which creates a potential for p-hacking. Zhao and Ding (2021) advanced the theoretical framework of the Fisher randomization test to test both sharp and weak null hypotheses on data from ReM rerandomized experiments. However, the use of other inferential procedures, such as linear regression, which is popular in economics, necessitates employing Monte Carlo methods to simulate the distribution of potential trial results to adjust the values of observed p-values. Simulation-based inference involves numerous decisions about probability distribution, preprocessing data used in simulation, and choosing algorithms relying on approximations that will effectively run the simulation. Each of these decisions may have an effect on the p-value adjustment, potentially leading to severe consequences when the adjusted p-value approaches the threshold of statistical significance.

The lack of theoretical foundations for statistical hypothesis testing of rerandomized experimental data led Glennerster and Takavarasha (2014), the authors of the first textbook on randomized field trials, to recommend against repeating randomized assignments, at least until statistical theory provides a clearer understanding of its benefits and drawbacks. Even though statisticians are currently developing theory and statistical tests for rerandomized experiments, procedures beyond ReP and ReM have also been proposed, such as using a rank-based covariate balance measure (Johansson and Schultzberg 2020). Other rerandomization procedures will require developing new approaches to p-value adjustment. This

https://doi.org/10.1017/psa.2026.10203 Published online by Cambridge University Press

further adds to the researcher degrees of freedom of experiments using rerandomization.

One of the referees for *Philosophy of Science* suggested that (1) although homemade p-value adjustments could enable p-hacking, all statistical methods pose this risk, and (2) proposed a workaround involving reporting unadjusted, conservative p-value estimates, which is common in randomized field experiments. These two potential counterarguments against the claim that p-value adjustment in rerandomized experiments inflates researcher degrees of freedom could be a serious objection to the main argument of the paper and should be addressed.

I believe that the first objection is correct in claiming that all statistical inferences involve some degree of analytical flexibility. Often, statistical methodology and best practices do not fully constrain methodological decisions, leading to the situation in which different teams of analysts may commit to different choices and potentially obtain different or even conflicting results based on the same data (Sterner and Lidgard 2024; Gelman and Loken 2013). However, the extent of analytical flexibility or the number of researcher degrees of freedom may differ across statistical methods, depending on the complication of the method (and the number of decisions it involves) and the degree to which methodological decisions are constrained. For example, calculating an average from a full set of observations involves fewer methodological decisions, which are more constrained, meaning that fewer choices are permissible at each step, compared to the endeavor of estimating a multivariate nonlinear model based on a multidimensional dataset. In a similar vein, analyzing a dataset produced by an experiment using simple randomized assignment with randomization-based inference involves fewer decisions compared to analyzing a dataset from a rerandomized experiment. While it is true that every statistical analysis involves some flexibility, the extent of that flexibility differs and is higher when data (and the experimental design) are more complex. Using any assignment procedure that is not fully random violates the assumptions of randomization-based inference, thereby complicating statistical analysis. Because rerandomization is more effective than simple randomization at balancing experimental groups, it requires adjusting p-values to accurately reflect the true probabilities of observing a given or more extreme result if the null hypothesis is true. Considering that p-value adjustment is not straightforward (at least given the current state of statistical theory), rerandomizing inflates the researcher degrees of freedom involved in, among others, p-value calculation.

4.3. Issues with regression-based inference from experimental data

While experiments using simple randomization can be analyzed with randomization-based inference without controlling for covariates at the analysis stage, data from rerandomized experiments require regression-based inference and controlling for covariates used in the rerandomization procedure. Although statistical control for covariates during the analysis stage may appear to be a distinct method for balancing experimental groups, it is closely related to rerandomization. Bruhn and McKenzie (2009), following Scott et al. (2002), advised that "researchers include all the variables used to check balance as linear covariates in the regression" (Bruhn and McKenzie 2009, 220). This advice arises from Taves's (1974) observation that the minimization

https://doi.org/10.1017/psa.2026.10203 Published online by Cambridge University Press

method (a nonrandom assignment procedure that balances selected covariates) does not fulfill the assumptions of randomization-based inference. Given that inferences like testing for a difference in means require each unit to be assigned equiprobably to treatment and control groups, which is only achievable through simple randomization, "concerns over the validity of the analysis surround not just minimization but all other allocation methods" (Scott et al. 2002, 668). In particular, rerandomization also alters the probabilities of experimental units being assigned to treatment and control groups based on their covariate values to some extent. Rubin (2008) also advised using model-based adjustments for the covariates balanced in a rerandomization procedure; he pointed to ANCOVA as an example of such an adjustment. He argued that the adjustment "will not change the point estimates much because within the balanced groups, the distributions of the [balanced] covariates are so similar-there are only small . . . differences for which to adjust" (1352). Rubin's suggestion indicates that the choice between model-based and randomization-based inference is an additional researcher degree of freedom in rerandomized experiments.

However, analyzing experimental data with linear regression, a common practice in randomized field experiments, poses challenges. The issue with using regression methods to draw inferences from (re-)randomized experiments is that randomization alone does not validate these methods (Freedman 2008). While testing for a difference in means is straightforward and does not require overly stringent assumptions about the data-generating process, regression-based adjustments involve greater commitments (Blackburn and Neumark 1993). Simple randomization asserts that the unadjusted average treatment effect (the difference in average outcomes between the treatment and control groups) is an unbiased estimate under the Neyman model, requiring fewer assumptions compared to linear regression. In particular, "It relies on the fact that the mean is a linear operator so that the difference in means is the mean of the differences. No model or assumptions about other variables are required. The treatment effects can be heterogeneous, requiring nothing to be stated about the shapes of statistical distributions" (Sekhar and Thapa 2024, 65).

The Ordinary Least Squares (OLS) estimation of linear regression, commonly used in randomized field experiments, can be demonstrated to yield an unbiased estimate of the average treatment effect only when no covariates are included in the regression. Adding covariates (explanatory variables) to a regression requires checking the accuracy of the assumptions OLS relies on, such as exogeneity, linear independence, homoscedasticity, no autocorrelation, and normality (see Wooldridge 2016, 149 et seq.). Using model-based inference in RFEs validates Heckman's (2020) criticism that "advocates of randomization have overstated their case for having avoided arbitrary assumptions" (304). Testing whether OLS assumptions are met is not straightforward, and, particularly in some areas of social sciences, they may remain unfulfilled (e.g., Bun and Harrison 2019; Blackburn and Neumark 1993). Other estimation methods rely on different assumptions, and testing those assumptions, along with selecting the appropriate estimator, is a forking path with considerable extent of analytical flexibility. Moreover, different estimation approaches may produce varying results. For instance, the Least Absolute Deviation (LAD) estimation places greater importance on outliers than the OLS method (Maziarz 2024). The unfulfilled assumptions of model-based inference led Freedman (2008) to comment that "[r]egression adjustments are often made to experimental data. Since

https://doi.org/10.1017/psa.2026.10203 Published online by Cambridge University Press

randomization does not justify the models, almost anything can happen" because estimators do not remain unbiased when any of their assumptions are not true about the data.

However, the primary issue with using linear regression to adjust for covariates in the context of how rerandomization inflates researcher degrees of freedom is that it introduces substantial analytical flexibility, which may be exploited for p-hacking to achieve a statistically significant result or confirm/disconfirm a theory under study. In principle, if covariate adjustment in a linear model is preregistered and the model's assumptions are fulfilled, the risk of false positive results should remain constant and equal to the threshold of statistical significance. However, despite the American Economic Association's (AEA's) requirement for preregistration, only 35 percent of articles published in the AEA journals are preregistered (Brodeur et al. 2024). Furthermore, preregistration and especially statistical analysis plans are often too vague to fully limit researcher degrees of freedom (ibid.). This, along with simulation results suggesting that attempting model specifications with various sets of covariates can yield any results if a sufficient number of covariates are available (Pham 2016), indicates that using regression with covariates balanced through a rerandomization procedure—especially if the details of the procedure and statistical analysis are not preregistered—may lead to flipping the observed p-value past the threshold of statistical significance, resulting in false-positive results. In this regard, the ReP rerandomization procedure may resemble the practice of testing for baseline imbalances in covariates and adjusting for any imbalanced covariates. Mutz et al. (2019) criticized this on the grounds that it jeopardizes scientific integrity by generating misleading confidence statements and permitting arbitrary results through the selection of covariates.

Although there are few studies comparing the results of randomization-based and model-based inference from experimental data, current evidence confirms that the increased researcher degrees of freedom resulting from using linear regression to adjust for covariates leads to a higher rate of false-positive results. Heckman (2020) discussed two studies that used nonexperimental methods to analyze the efficacy of some unemployment interventions tested in the mid-1970s in The National Supported Work Demonstration (NSW) study. He observed that LaLonde (1986) and Fraker and Maynard (1987) "produce a wide array of estimates of impacts of the same program using different non-experimental methods. They claim that there is no way to choose among competing non-experimental estimators. All non-experimental methods are based on some maintained, untestable assumption." In a recent study examining the impact of applying regressions to experimental data, Young (2019) reanalyzed data from fifty-three RFEs published in the American Economic Association journals using randomization-based inference and reported that this approach yields up to 49 percent fewer positive results.

Recently, in an arXiv preprint, Lu and Ding (2025) argued, using mathematical statistics and simulation, that rerandomization reduces p-hacking opportunities, which seems to directly conflict with the argumentation of this article. The disagreement seems to stem from two differences: (1) using a too narrow definition of p-hacking and (2) comparing p-hacking opportunities between model-based inference without recognizing that simple randomization allows for using randomization-based

https://doi.org/10.1017/psa.2026.10203 Published online by Cambridge University Press

inference without the need to control for covariate imbalances by including them in a statistical model like regression.

According to Lu and Ding (2025), p-hacking is "the practice of strategically selecting covariates to get more significant p-values" (henceforth, LD-p-hacking). Such a definition is too narrow and does not encompass the full range of methods researchers use to push the observed p-values beyond the threshold of statistical significance. Instead, researchers may use other degrees of freedom involved in "flexibility in data collection, analysis, and reporting" (Simmons et al. 2011, 1359), which is a much broader concept than simply selecting covariates for a regression. Therefore, Lu and Ding (2025) only demonstrated that "rerandomization can mitigate [LD-]p-hacking caused by strategically selecting covariates in regression adjustment" (2), but they did not address other types of p-hacking involving flexibility in different aspects of research design and data analysis.

Furthermore, Lu and Ding (2025) compared the LD-p-hacking potential of a fully randomized experiment to that of ReM and ReP rerandomized experiments, but their comparison focuses on the regression-based inference from the two types of experiments. Because rerandomization effectively balances covariates, these covariates have less influence on the ATE estimate when included in a regression, which explains their conclusion. However, rerandomization requires regression-based adjustments, whereas simple randomization allows for the use of randomization-based inference, such as testing for a difference in mean outcomes observed in the treatment and control groups. Such an inference is more constrained than the model-based inference with covariates. In fact, testing for a difference in means does not permit the inclusion of any covariates, so the asymptotic type I error rate of the LD-p-hacked p-value under ReM should be compared to no LD-p-hacking. Although Lu and Ding (2025) does not present such a comparison, their conclusion that "under rerandomization, [LD-]p-hacking will not change the p-values too much" (8) indicates a nonzero (even if negligible) effect and suggests that even such restrictively interpreted LD-p-hacking risk is greater in rerandomized experiments than in those using simple randomization to assign units to groups.

5. The trade-off between credibility and precision of the effect size estimate

Previously, I argued that rerandomization effectively improves the balance of experimental groups, thereby enhancing the precision of estimates. This provides a substantial epistemic benefit, allowing for smaller sample sizes while keeping the same levels of type I and type II error rates. However, the benefit comes with an epistemic cost, that is, reduced credibility of the results. The reason is that more complex data analysis increases researcher degrees of freedom and creates more opportunities for p-hacking. Choosing between the improved precision of estimates in rerandomized experiments and the credibility of studies using simple randomization involves balancing epistemic reasons and nonepistemic values, with decisions likely depending on the context and purpose of the experiment.

Furthermore, the choice between assigning experimental units with either simple randomization or rerandomization depends on weighing the gain in estimate precision against the risk of bias from increased analytical flexibility of rerandomized

https://doi.org/10.1017/psa.2026.10203 Published online by Cambridge University Press

data. If rerandomization did not inflate researcher degrees of freedom, then the change in estimate precision can be described as follows:

$$\triangle_{\varepsilon} = \varepsilon_{rerandomization} - \varepsilon_{simple\ randomization}$$

Where:

$\triangle_{\varepsilon}$—the change in sampling error due to rerandomization;

$\varepsilon_{rerandomization}$—sampling error emerging from the rerandomization procedure;

$\varepsilon_{simple\ randomization}$—sampling error emerging from the simple randomization.

Further considerations require assuming the distribution of bias introduced by the additional researcher degrees of freedom. When the same conflict of interest affects research teams in a field, or they share similar assumptions about expected results, it is likely that all results will be biased in the same direction, with an average bias size of (x). In such a situation, if a decision or inference about effect size is made based on only one experiment, comparing the gain in estimate precision and the size of bias can inform the decision regarding the assignment procedure. However, if an experiment is likely to be repeated, unbiasedness might be preferred because the estimate's precision will eventually improve as more published results enable meta-analyses of the effect size estimate.

Otherwise, if both conflict of interest and guesses about expected results differ among researchers, the inflated analytical flexibility may cause researchers to use different analysis pipelines and report varying results. In such a case, the researcher degrees of freedom in experiments can be seen as an additional source of random error.

$$\triangle_{\varepsilon+\zeta} = \left(\varepsilon_{rerandomization} - \varepsilon_{simple\ randomization}\right) + \left(\zeta_{rerandomization} - \zeta_{simple\ randomization}\right)$$

Where:

$\zeta_{rerandomization}$—sampling error resulting from different analysis pipelines in a rerandomized experiment;

$\zeta_{simple\ randomization}$—sampling error resulting from different analysis pipelines in an experiment using simple randomization.

This suggests that, regardless of the decision maker's values and the experimental context, rerandomization should be favored if the improvement in estimate precision is larger than the variability produced by the additional analytical flexibility of rerandomized experiments $\left(\Delta_{\varepsilon+\zeta} < 0\right)$. While the significant gains in estimate precision resulting from rerandomization are well understood, simulation studies or multiverse analyses of actual rerandomized experiments (Steegen et al., 2016) are necessary to quantify how much the permissible forking paths resulting from more complex data analysis of rerandomized experiments may bias the results.

Scientific inferences and clinical and policy decisions are primarily based on amalgamated evidence rather than single studies (Fletcher 2022), and hence, such context prioritizes the unbiasedness of individual effect size estimates because precision increases eventually when more studies addressing the same research question are published. Furthermore, Banerjee et al. (2020) argued that fully random assignment should be preferred in cases of "adversary experimentation" when the audience does not trust the experimenter. Given that most scientific contexts, such as

https://doi.org/10.1017/psa.2026.10203 Published online by Cambridge University Press

a regulatory agency (e.g., the Food and Drug Administration) using clinical results for drug approval decisions or competing teams testing theoretical predictions, resemble adversarial experimentation, rerandomization may currently not be the best assignment procedure because the increased analytical flexibility can lead to p-hacking opportunities. In contrast, rerandomization is more appropriate when the experimenter uses the results internally, such as during quality control tests within a company or an in vivo study of a new drug candidate.

The problem of additional analytical flexibility introduced by rerandomization can be mitigated to some extent. If statisticians and research communities develop standardized methods for analyzing data from rerandomized experiments, researcher degrees of freedom will decrease, thereby increasing the credibility of these experiments while preserving the benefits of this assignment procedure. For example, there are currently no standardized methods for adjusting p-values (see section 4.2), and homemade adjustments developed by experimenters can risk p-hacking. This problem could be minimized if a standardized approach were developed, other than simply not reporting rerandomization and calculating unadjusted p-values.

The credibility of rerandomized experiments, especially in the social sciences, can be increased more quickly by improving the reporting practices. Publishing a preanalysis plan with a sufficiently detailed preregistration report will limit the researcher degrees of freedom introduced by rerandomization and thus preserve the results credibility. Greater transparency about: (1) how rerandomized assignment was designed and executed, and (2) which statistical analyses were conducted to analyze the data, and the results they produced, may limit researchers' opportunities to use the additional researcher degrees of freedom emerging from rerandomization to report desired results.

6. Concluding remarks

Rerandomization is an effective method for enhancing the precision of average treatment effect estimates ($\widehat{ATE}$). Reducing the standard deviation of the estimate while keeping the sample size fixed reduces both type-I and type-II error rates, decreases exaggeration of effect size estimates (type-M errors), and results in fewer incorrect sign results (type-S errors), which are epistemically significant benefits. Ethical considerations regarding the burden of involving additional patients or laboratory animals in a trial, the cost-effectiveness of experimental samples, and policy makers' need for more accurate evidence all deem rerandomization a valuable methodological improvement. Such considerations led several statisticians and economists to favor rerandomization and advocate for its wider use (Banerjee et al. 2017; Morgan and Rubin 2012).

However, the necessity to employ alternative and more complex methods of data analysis inflates researcher degrees of freedom and thus creates more chances for researchers to misuse this additional flexibility to produce results that are statistically significant or align with their preferred theory or political agenda. While refraining from adjusting p-values seems promising, it has a detrimental effect on replicability: In such cases, replicator teams will remain unaware of rerandomization, which reduces standard deviation, and thus may conduct an underpowered replication attempt. When analyzing the difference in false positive

https://doi.org/10.1017/psa.2026.10203 Published online by Cambridge University Press

result probability across various experimental and quasiexperimental research methods in economics, Brodeur et al. (2020) hypothesized that "some methods offer researchers different degrees of freedom than others" (3). Applying this to the problem at hand indicates that experiments using rerandomization, while providing more precise estimates of the average treatment effect ($\widehat{ATE}$), carry a greater risk of bias. Therefore, the quality of evidence produced by such studies should be rated lower than otherwise. This conclusion is accurate as long as further methodological work does not develop some widely accepted standards for analyzing data from rerandomized experiments, thereby limiting analytical flexibility. However, the study of rerandomized experiments may indicate that considering researcher degrees of freedom involved in designing and executing a study may help in assessing the quality of evidence stemming from that study. Future philosophical work is needed to develop a more nuanced concept of credibility that applies across the full range of empirical research.

The analysis of rerandomization points toward a trade-off between estimate precision and the risk of bias. Depending on the context and purpose of experimental research, the preference for unbiasedness may outweigh the epistemic benefits of more precise estimates, or vice versa; sometimes, the need for a precise estimate despite small samples may take priority over concerns about credibility. More efforts from statisticians and philosophers are needed to fully understand the impact of inflated degrees of freedom on the credibility of rerandomized experiments, and to guide decisions on designing these studies and interpreting their outcomes.

Data availability statement. Not applicable.

Acknowledgments. The author acknowledges the helpful comments from the editor and numerous reviewers for Philosophy of Science, as well as the suggestions from participants in the workshop "Causality in Complex Systems," organized by Tobias Henschen at the University of Cologne.

Author contribution. The sole author responsible for the research and writing of the manuscript is Mariusz Maziarz.

Funding. The research received support from the National Science Centre, Poland under grant no. 2022/45/B/HS1/00183 ("Evidential Pluralism in Philosophy of Economics").

Competing interests. The author does not have any conflict of interests.

Informed consent. n/a.

Ethical approval. n/a.

References

Allchin, Douglas. 2001. "Error Types." *Perspectives on Science* 9 (1):38–58. https://doi.org/10.1162/10636140152947786.

Altman, Douglas G., and Caroline J. Dore. 1990. "Randomisation and Baseline Comparisons in Clinical Trials." *The Lancet* 335 (8682):149–53.

Assmann, Susan F., Stuart J. Pocock, Laura E. Enos, and Linda E. Kasten. 2000. "Subgroup Analysis and Other (Mis) Uses of Baseline Data in Clinical Trials." *The Lancet* 355 (9209):1064–69.

Balzer, Laura B., Maya L. Petersen, Mark J. van der Laan, and Search Consortium. 2015. "Adaptive Pair-Matching in Randomized Trials with Unbiased and Efficient Effect Estimation." *Statistics in Medicine* 34 (6):999–1011. https://doi.org/10.1002/sim.6380.

https://doi.org/10.1017/psa.2026.10203 Published online by Cambridge University Press

Banerjee, Abhijit V., Sylvain Chassang, Sergio Montero, and Erik Snowberg. 2020. "A Theory of Experimenters: Robustness, Randomization, and Balance." *American Economic Review* 110 (4):1206–30. https://doi/org/10.1257/aer.20171634.

Banerjee, Abhijit V., Sylvain Chassang, and Erik Snowberg. 2017. "Decision Theoretic Approaches to Experiment Design and External Validity." In *Handbook of Economic Field Experiments*, vol. 1, 141–74. North-Holland.

Banerjee, Abhijit V., and Esther Duflo. 2017. "An Introduction to the 'Handbook of Field Experiments'." In *Handbook of Economic Field Experiments*, vol. 1, edited by Duflo Esther, and Abhijit Banerjee, 1–24. Elsevier.

Barrett, Christopher B., and Michael R. Carter. 2010. "The Power and Pitfalls of Experiments in Development Economics: Some Non-random Reflections." *Applied Economic Perspectives and Policy* 32 (4):515–48. https://doi.org/10.1093/aepp/ppq023.

Baxby, Derrick. 1997. "Lind's Clinical Trial and the Control of Scurvy." *Journal of the Royal Society of Medicine* 90 (9):526–27. https://doi.org/10.1177/014107689709000.

Blackburn, McKinley L., and David Neumark. 1993. "Are OLS Estimates of the Return to Schooling Biased Downward? Another Look." NBER Working Paper. https://doi.org/10.3386/w4259.

Brodeur, Abel, Nikolai Cook, and Anthony Heyes. 2020. "Methods Matter: P-hacking and Publication Bias in Causal Analysis in Economics." *American Economic Review* 110 (11):3634–60. https://doi.org/10.1257/aer.20190687.

Brodeur, Abel, Nikolai M. Cook, Jonathan S. Hartley, and Anthony Heyes. 2024. "Do Preregistration and Preanalysis Plans Reduce p-Hacking and Publication Bias? Evidence from 15,992 Test Statistics and Suggestions for Improvement." *Journal of Political Economy Microeconomics* 2 (3):527–61. https://doi.org/10.1086/730455.

Broglio, Kristine. 2018. "Randomization in Clinical Trials: Permuted Blocks and Stratification." *JAMA* 319 (21):2223–24. https://doi:10.1001/jama.2018.6360.

Bruhn, Miriam, and David McKenzie. 2009. "In Pursuit of Balance: Randomization in Practice in Development Field Experiments." *American Economic Journal: Applied Economics* 1 (4):200–32. https://doi.org/10.1257/app.1.4.200.

Bun, Maurice J. G., and Teresa D. Harrison. 2019. "OLS and IV Estimation of Regression Models Including Endogenous Interaction Terms." *Econometric Reviews* 38 (7):814–27. https://doi.org/10.1080/07474938.2018.1427486.

Cartwright, Nancy. 1983. *How the Laws of Physics Lie*. Oxford: Oxford University Press.

Cartwright, Nancy. 2007. "Are RCTs the Gold Standard?" *BioSocieties* 2 (1):11–20. https://doi.org/10.1017/S1745855207005029.

Cartwright, Nancy. 2010. "What Are Randomised Controlled Trials Good For?" *Philosophical Studies* 147 (1):59–70. https://doi.org/10.1007/s11098-009-9450-2.

Cartwright, Nancy, and Jeremy Hardie. 2012. *Evidence-Based Policy: A Practical Guide to Doing it Better*. Oxford University Press.

Cook, Thomas D., and David L. DeMets. 2007. *Introduction to Statistical Methods for Clinical Trials*. Delft, Netherlands: Chapman and Hall/CRC. https://doi.org/10.1201/9781420009965.

Credibility. 2026. *Cambridge Dictionary*. Cambridge: Cambridge University Press.

Davey, Calum, Alexander M. Aiken, Richard J. Hayes, and James R. Hargreaves. 2015. "Re-analysis of Health and Educational Impacts of a School-Based Deworming Programme in Western Kenya: A Statistical Replication of a Cluster Quasi-randomized Stepped-Wedge Trial." *International Journal of Epidemiology* 44 (5):1581–92. https://doi.org/10.1093/ije/dyv128.

De Winter, Joost C. F., and Dimitra Dodou. 2015. "A Surge of p-Values between 0.041 and 0.049 in Recent Decades (But Negative Results Are Increasing Rapidly Too)." *PeerJ* 3: e733. https://doi.org/10.7717/peerj.733.

Deaton, Angus, and Nancy Cartwright. 2018. "Reflections on Randomized Control Trials." *Social Science & Medicine* 210 (C).

Eisma, Yke Bauke, and Joost de Winter. 2025. "Preregistration in Scientific Research: A Synthesis of Perspectives."

Erasmus, Adrian. 2025. "P-Hacking: Its Costs and When It Is Warranted." *Erkenntnis* 90 (7):2929–50. https://doi.org/10.1007/s10670-024-00834-3.

https://doi.org/10.1017/psa.2026.10203 Published online by Cambridge University Press

Favereau, Judith. 2016. "On the Analogy Between Field Experiments in Economics and Clinical Trials in Medicine." *Journal of Economic Methodology* 23 (2):203–22.
Favereau, Judith, and Michiru Nagatsu. 2020. "Holding Back from Theory: Limits and Methodological Alternatives of Randomized Field Experiments in Development Economics." *Journal of Economic Methodology* 27 (3):191–211.
Fletcher, Samuel C. 2022. "Replication Is for Meta-analysis." *Philosophy of Science* 89 (5):960–69. https://doi.org/10.1017/psa.2022.38.
Fraker, Thomas, and Rebecca Maynard. 1987. "The adequacy of comparison group designs for evaluations of employment-related programs." *Journal of Human Resources*: 194–227.
Freedman, David A. 2008. "On Regression Adjustments to Experimental Data." *Advances in Applied Mathematics* 40 (2):180–93. https://doi.org/10.1016/j.aam.2006.12.003.
Fuller, Jonathan. 2019. "The Confounding Question of Confounding Causes in Randomized Trials." *The British Journal for the Philosophy of Science* 70 (3):901–26. https://doi.org/10.10932Fbjps2015.
Gelman, Andrew, and Eric Loken. 2013. "The Garden of Forking Paths: Why Multiple Comparisons Can Be a Problem, Even When There Is No 'Fishing Expedition' or 'p-Hacking' and the Research Hypothesis Was Posited Ahead of Time." *Department of Statistics, Columbia University* 348 (1–17):3.
Glennerster, Rachel, and Kudzai Takavarasha. 2014. *Running Randomized Evaluations: A Practical Guide*. Princeton: Princeton University Press.
Goldfarb, Robert S. 1997. "Now You See it, Now You Don't: Emerging Contrary Results in Economics." *Journal of Economic Methodology* 4 (2):221–44.
Greenland, Sander. 1990. "Randomization, Statistics, and Causal Inference." *Epidemiology* 64 (1):21–9. https://www.jstor.org/stable/25759845.
Head, Megan L., Luke Holman, Rob Lanfear, Andrew T. Kahn, and Michael D. Jennions. 2015. "The Extent and Consequences of p-Hacking in Science." *PLOS Biology* 13 (3):e1002106. https://doi.org/10.1371/journal.pbio.1002106.
Healy, David. 2015. "Serotonin and depression." *Bmj*: 350.
Heckman, James J. 2020. "Epilogue: Randomization and Social Policy." *Randomized Control Trials in the Field of Development: A Critical Perspective*, 304.
Jiménez-Buedo, María, and Federica Russo. 2021. "Experimental Practices and Objectivity in the Social Sciences: Re-Embedding Construct Validity in the Internal–External Validity Distinction." *Synthese* 199 (3):9549–79.
Johansson, Per, and Mårten Schultzberg. 2020. "Rerandomization Strategies for Balancing Covariates Using Pre-Experimental Longitudinal Data." *Journal of Computational and Graphical Statistics* 29 (4):798–813. https://doi.org/10.1080/10618600.2020.1753531.
John, Leslie K., George Loewenstein, and Drazen Prelec. 2012. "Measuring the Prevalence of Questionable Research Practices with Incentives for Truth Telling." *Psychological Science* 23 (5):524–32. https://doi.org/10.1177/0956797611430953.
Kahan, Brennan C., Vipul Jairath, Caroline J. Doré, and Tim P. Morris. 2014. "The Risks and Rewards of Covariate Adjustment in Randomized Trials: An Assessment of 12 Outcomes from 8 Studies." *Trials* 15 (1):139. https://doi.org/10.1186/1745-6215-15-139.
Kernan, Walter N., Catherine M. Viscoli, Robert W. Makuch, Lawrence M. Brass, and Ralph I. Horwitz. 1999. "Stratified Randomization for Clinical Trials." *Journal of Clinical Epidemiology* 52 (1):19–26. https://doi.org/10.1016/S0895-4356(98)00138-3.
La Caze, Adam. 2009. "Evidence-Based Medicine Must Be . . . ." *Journal of Medicine and Philosophy* 34 (5):509–27. https://doi.org/10.1093/jmp/jhp034.
Lachin, John M., John P. Matts, and L. J. Wei. 1988. "Randomization in Clinical Trials: Conclusions and Recommendations." *Controlled Clinical Trials* 9 (4):365–74. https://doi.org/10.1016/0197-2456(88)90049-9.
Lai, Deborah, Daniel Wang, Matthew McGillivray, Shadi Baajour, Ali S. Raja, and Shuhan He. 2021. "Assessing the Quality of Randomization Methods in Randomized Control Trials." *Healthcare* 9 (4):100570. https://doi.org/10.1016/j.hjdsi.2021.100570.
Lakens, Daniel. 2019. "The Value of Preregistration for Psychological Science: A Conceptual Analysis." *Japanese Psychological Review* 62 (3):221–30. https://doi.org/10.24602/sjpr.62.3_221.

https://doi.org/10.1017/psa.2026.10203 Published online by Cambridge University Press

LaLonde, Robert J. 1986. "Evaluating the Econometric Evaluations of Training Programs with Experimental Data." *The American Economic Review* 76 (4):604–20. https://www.jstor.org/stable/1806062.

Leaver, Clare, Owen Ozier, Pieter Serneels, and Andrew Zeitlin. 2021. "Recruitment, Effort, and Retention Effects of Performance Contracts for Civil Servants: Experimental Evidence from Rwandan Primary Schools." *American Economic Review* 111 (7):2213–46. https://doi.org/10.1257/aer.20191972.

Li, Xinran, and Peng Ding. 2020. "Rerandomization and Regression Adjustment." *Journal of the Royal Statistical Society Series B: Statistical Methodology* 82 (1):241–68. https://doi.org/10.1111/rssb.12353.

Lu, Xin, and Peng Ding. 2025. "Rerandomization for Covariate Balance Mitigates p-Hacking in Regression Adjustment." *arXiv preprint arXiv:2505.01137*. https://doi.org/10.48550/arXiv.2505.01137.

Mayo, Deborah G. 1996. *Error and the Growth of Experimental Knowledge*. Chicago: University of Chicago Press.

Mayo, Deborah G. 2018. *Statistical Inference as Severe Testing: How to Get beyond the Statistics Wars*. Cambridge: Cambridge University Press.

Mayo, Deborah G., and Aris Spanos. 2006. "Severe Testing as a Basic Concept in a Neyman–Pearson Philosophy of Induction." *The British Journal for the Philosophy of Science* 57 (2):323–57. https://doi.org/10.1093/bjps/axl003.

Maziarz, Mariusz. 2024. "Conflicting Results and Statistical Malleability: Embracing Pluralism of Empirical Results." *Perspectives on Science* 32 (6):701–28. https://doi.org/10.1162/posc_a_00627.

Maziarz, Mariusz. 2025. "Does Randomization Assert the Balance across Trial Arms? Revisiting Worrall's Criticism." *History and Philosophy of the Life Sciences* 47 (1):6. https://doi.org/10.1007/s40656-024-00655-1.

Miguel, Edward, and Michael Kremer. 2004. "Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities." *Econometrica* 72 (1):159–217. https://doi.org/10.1111/j.1468-0262.2004.00481.x.

Moher, David, Douglas Altman, Kenneth Schulz, Iveta Simera, and Elizabeth Wager, eds. 2014. *Guidelines for Reporting Health Research: A User's Manual*. Oxford: John Wiley and Sons.

Moosa, Imad A. 2019. "The Fragility of Results and Bias in Empirical Research: An Exploratory Exposition." *Journal of Economic Methodology* 26 (4):347–60. https://doi.org/10.1080/1350178X.2018.1556798.

Morgan, Kari Lock, and Donald B. Rubin. 2012. "Rerandomization to Improve Covariate Balance in Experiments." *Annals of Statistics* 40 (2):1263–82. https://doi.org/10.1214/12-AOS1008.

Morgan, Kari Lock, and Donald B. Rubin. 2015. "Rerandomization to Balance Tiers of Covariates." *Journal of the American Statistical Association* 110 (512):1412–21. https://doi.org/10.1080/01621459.2015.1079528.

Mutz, Diana C., and Robin Pemantle. 2015. "Standards for Experimental Research: Encouraging a Better Understanding of Experimental Methods." *Journal of Experimental Political Science* 2 (2):192–215.

Mutz, Diana C., Robin Pemantle, and Philip Pham. 2019. "The Perils of Balance Testing in Experimental Design: Messy Analyses of Clean Data." *The American Statistician* 73 (1):32–42. https://doi.org/10.1080/00031305.2017.1322143.

Nagatsu, Michiru, and Judith Favereau. 2020. "Two Strands of Field Experiments in Economics: A Historical-Methodological Analysis." *Philosophy of the Social Sciences* 50 (1):45–77.

Neyman, Jerzy. 1990 [1923]. "On the Application of Probability Theory to Agricultural Experiments: Essay on Principles. Section 9." *Statistical Science* 5:465–72.

Nosek, Brian A., Charles R. Ebersole, Alexander C. DeHaven, and David T. Mellor. 2018. "The Preregistration Revolution." *Proceedings of the National Academy of Sciences* 115 (11):2600–6. https://doi.org/10.1073/pnas.1708274114.

Nuzzo, Rigina. 2014. "Statistical Errors: P Values, the 'Gold Standard' of Statistical Validity, Are Not as Reliable as Many Scientists Assume." *Nature* 506 (7487):150–53.

Ogawa, Junjiro. 1961. "The Effect of Randomization on the Analysis of Randomized Block Design." *Annals of the Institute of Statistical Mathematics* 13 (1):105–17. https://doi.org/10.1007/BF02868863.

Ogden, Timothy. 2020. "RCTs in Development Economics, Their Critics and Their Evolution." In *Randomized Control Trials in the Field of Development: A Critical Perspective*, edited by Florent Bédécarrats, Isabelle Guérin, François Roubaud, 126–51. Oxford: Oxford University Press.

Peter, Noemi, and Adriaan R. Soetevent. 2019. "Randomization in Field Experiments." In *Handbook of Research Methods and Applications in Experimental Economics*, 121–40. Northampton, Massachusetts: Edward Elgar Publishing.

https://doi.org/10.1017/psa.2026.10203 Published online by Cambridge University Press

Pham, Philip. 2016. "Just How Easy Is It to Cheat a Linear Regression?" PhD diss., University of Pennsylvania.
Philippi, Cristian Larroulet. 2022. "There Is Cause to Randomize." *Philosophy of Science* 89 (1):152–70. https://doi.org/10.1017/psa.2021.19.
Romero, Felipe. 2019. "Philosophy of Science and the Replicability Crisis." *Philosophy Compass* 14 (11): e12633. https://doi.org/10.1111/phc3.12633.
Rubin, Donald B. 1974. "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies." *Journal of Educational Psychology* 66 (5):688. https://psycnet.apa.org/doi/10.1037/h0037350.
Rubin, Donald B. 2005. "Causal Inference Using Potential Outcomes: Design, Modeling, Decisions." *Journal of the American Statistical Association* 100 (469):322–31. https://doi.org/10.1198/016214504000001880.
Rubin, Donald B. 2008. "Comment: The Design and Analysis of Gold Standard Randomized Experiments." *Journal of the American Statistical Association* 103 (484):1350–53. https://doi.org/10.1198/016214508000001011.
Schulz, Kenneth F., and David A. Grimes. 2002. "Allocation Concealment in Randomised Trials: Defending against Deciphering." *The Lancet* 359 (9306):614–18.
Scott, Neil W., Gladys C. McPherson, Craig R. Ramsay, and Marion K. Campbell. 2002. "The Method of Minimization for Allocation to Clinical Trials: A Review." *Controlled Clinical Trials* 23 (6):662–74. https://doi.org/10.1016/S0197-2456(02)00242-8.
Sekhar, C. S. C., and Namrata Thapa. 2024. "Experimental Approach to Development Economics: A Review of Issues and Options." *Journal of Economic Methodology* 31 (2):63–77. https://doi.org/10.1080/1350178X.2024.2341292.
Senn, Stephen. 1994. "Testing for Baseline Balance in Clinical Trials." *Statistics in Medicine* 13 (17):1715–26. https://doi.org/10.1002/sim.4780131703.
Shaw, Julia F., Cory E. Goldstein, Thais Mazzetti, Anna Catharina Vieira Armond, Yacine Marouf, Kyle Lamprecht, Eric Tran et al. 2025. "Randomization Procedures in Parallel-Arm Cluster Randomized Trials in Low- and Middle-Income Countries: A Review of 300 Trials Published between 2017–2022." *Journal of Clinical Epidemiology* 184:111825. https://doi.org/10.1016/j.jclinepi.2025.111825.
Silberzahn, Raphael, Eric L. Uhlmann, Daniel P. Martin, Pasquale Anselmi, Frederik Aust, Eli Awtrey, Štěpán Bahník et al. 2018. "Many Analysts, One Data Set: Making Transparent How Variations in Analytic Choices Affect Results." *Advances in Methods and Practices in Psychological Science* 1 (3):337–56. https://doi.org/10.1177/2515245917747646.
Simmons, Joseph P., Leif D. Nelson, and Uri Simonsohn. 2011. "False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant." *Psychological Science* 22 (11):1359–66. https://doi.org/10.1177/0956797611417632.
Simonsohn, Uri, Leif D. Nelson, and Joseph P. Simmons. 2014. "P-Curve: A Key to the File-Drawer." *Journal of Experimental Psychology: General* 143(2):534. https://psycnet.apa.org/doi/10.1037/a0033242.
Simonsohn, Uri, Leif D. Nelson, and Joseph P. Simmons. 2014. "P-Curve and Effect Size: Correcting for Publication Bias Using Only Significant Results." *Perspectives on Psychological Science* 9 (6):666–81. https://doi.org/10.1177/1745691614553988.
Steegen, Sara, Francis Tuerlinckx, Andrew Gelman, and Wolf Vanpaemel. 2016. "Increasing Transparency through a Multiverse Analysis." *Perspectives on Psychological Science* 11 (5):702–12. https://doi.org/10.1177/1745691616658637.
Stefan, Angelika M., and Felix D. Schönbrodt. 2023. "Big Little Lies: A Compendium and Simulation of p-Hacking Strategies." *Royal Society Open Science* 10 (2):220346. https://doi.org/10.1098/rsos.220346.
Sterner, Beckett, and Scott Lidgard. 2024. "Objectivity and Underdetermination in Statistical Model Selection." *The British Journal for the Philosophy of Science* 75 (3):717–39. https://doi.org/10.1086/716243.
Taves, Donald R. 1974. "Minimization: A New Method of Assigning Patients to Treatment and Control Groups." *Clinical Pharmacology & Therapeutics* 15 (5):443–53. https://doi.org/10.1002/cpt1974155443.
van Drimmelen, Tom, M. Nienke Slagboom, Ria Reis, Lex M. Bouter, and Jenny T. van der Steen. 2024. "Decisions, Decisions, Decisions: An Ethnographic Study of Researcher Discretion in Practice." *Science and Engineering Ethics* 30 (6): 59.
Varga, Somogy. 2024. "Understanding in Medicine." *Erkenntnis* 89 (8):3025–49. https://doi.org/10.1007/s10670-023-00665-8.

https://doi.org/10.1017/psa.2026.10203 Published online by Cambridge University Press

Wicherts, Jelte M., Coosje L.S. Veldkamp, Hilde E.M. Augusteijn, Marjan Bakker, Robbie C.M. van Aert, and Marcel A.L.M. van Assen. 2016. Degrees of Freedom in Planning, Running, Analyzing, and Reporting Psychological Studies: A Checklist to Avoid P-Hacking. *Frontiers in Psychology* 7:222767.

Wooldridge, Jeffrey M. 2016. *Introductory Econometrics: A Modern Approach*. Boston, Massachusetts: South-Western Cengage Learning.

Worrall, John. 2007. "Why There's No Cause to Randomize." *The British Journal for the Philosophy of Science* 58 (3):451–88. https://doi.org/10.10932Fbjps2Faxm024.

Young, Alwyn. 2019. "Channeling Fisher: Randomization Tests and the Statistical Insignificance of Seemingly Significant Experimental Results." *The Quarterly Journal of Economics* 134 (2):557–98. https://doi.org/10.1093/qje/qjy029.

Zhao, Anqi, and Peng Ding. 2021. "Covariate-Adjusted Fisher Randomization Tests for the Average Treatment Effect." *Journal of Econometrics* 225 (2):278–94. https://doi.org/10.1016/j.jeconom.2021.04.007.

Zhao, Anqi, and Peng Ding. 2024. "No Star Is Good News: A Unified Look at Rerandomization Based on p-Values from Covariate Balance Tests." *Journal of Econometrics* 241 (1):105724. https://doi.org/10.1016/j.jeconom.2024.105724

Cite this article: Maziarz, Mariusz. 2026. "What Is Wrong with Rerandomization in Randomized Field Experiments?" *Philosophy of Science* 93 (4):873–897. https://doi.org/10.1017/psa.2026.10203

https://doi.org/10.1017/psa.2026.10203 Published online by Cambridge University Press