Striking the Right Balance: Why Standard Balance Tests Over-Reject the Null, and How to Fix It

Jason T. Kerwin
Nada Rostom
Olivier Sterck*

December 2025

Abstract

We show that standard sampling-inference-based balance tests—particularly omnibus F-tests, where the treatment is regressed on every baseline covariate—systematically over-reject the null, especially when many covariates are included. Using simulations and empirical examples, we demonstrate that omnibus F-tests with randomization inference deliver correct size and strong power across individually- and cluster-randomized designs. This approach is both statistically valid and conceptually aligned with the goals of balance testing. Replicating balance tests in two prominent RCTs, we find that sampling-based inference indicates balance problems while randomization inference shows treatment arms are actually balanced. We provide code for implementing appropriate balance tests.

JEL Classification: C1, C9, O12
Keywords: Balance Tests, Power, Size, Randomization Inference

*Kerwin: Department of Economics, University of Washington, IZA, and J-PAL (jkerwin@uw.edu); Rostom: Department of Economics, University of Antwerp (nada.rostom@uantwerpen.be); Sterck: ODID, University of Oxford and IOB, University of Antwerp. (olivier.sterck@uantwerpen.be). We thank Ellen Anderson, Francis Annan, Susan Athey, Marc Bellemare, David Card, Dean Eckles, Alan Griffith, Rachel Heath, Guido Imbens, Ethan Ligon, David McKenzie, Mateo Montenegro, Austin Nichols, and seminar participants at CSSS and the Department of Economics at the University of Washington, the CEGA Research Retreat at UC Berkeley, the WHY Seattle Symposium, and Refine.ink for helpful comments. Nada Rostom and Olivier Sterck gratefully acknowledge financial support from BOF DOCPRO Fast Track 2023. All remaining errors are our own.

1 Introduction

Balance tests are a common diagnostic tool in randomized controlled trials (RCTs), used to assess whether treatment and control groups are comparable at baseline. In a review of all RCTs published in top-five economics journals between 2021 and 2023 (69 papers in total), we found that 90% reported results from balance tests (Table A.1).

Balance tests are especially important in RCTs where the randomization may not have actually been implemented correctly—for example, in situations where the research team lacks full control over the randomization process such as public lotteries (e.g. Hanna, Duflo, and Greenstone 2016, Kerwin and Thornton 2021, Gazeaud, Mvukiyehe, and Sterck 2023, Barker et al. 2024, Franklin et al. 2024). This is a particular concern when the implementers of a program may have an incentive to manipulate the allocation process. Even in RCTs in which the research team is in charge of randomization, balance tests may be useful to demonstrate that randomization did not result in an unlucky draw, which can skew the results of the study (Leamer 1983).[1] Some scholars have argued in favor of re-randomizing if balance tests identify serious imbalances (Bruhn and McKenzie 2009; Zhao and Ding 2024) although this can lead to complications for inference (Athey and Imbens 2017). Balance tests also play an essential role in the analysis of natural experiments, to demonstrate that compared groups are similar before the quasi-random intervention or shock (as in e.g., Diamond, McQuade, and Qian 2019, Depetris-Chauvin, Durante, and Campante 2020, Jones et al. 2022, McGuirk, Hilger, and Miller 2023, Bruhn et al. 2024).

Economists tend to use two methods to assess balance, often together. First, 82% of the papers we reviewed use pairwise t-tests (or groupwise F-tests if there are more than two study arms) with a series of baseline variables and argue that treatment and control groups are balanced if few tests reject the null hypotheses at conventional thresholds. Normalized differences are sometimes reported alongside t-tests to show that any differences are small in

[1] See Mutz, Pemantle, and Pham (2019) for arguments against this practice. See also Freedman (2008) and Lin (2013) on regression adjustments to address imbalances.

size. Second, 32% of papers use omnibus tests of joint orthogonality, in which the treatment dummy is regressed on the full list of baseline covariates, and conclude that experimental groups are balanced if the test statistic is below a conventional significance threshold. Most papers reporting the p-value of an omnibus F-test of joint orthogonality use an OLS regression with robust standard errors to address heteroskedasticity in the linear probability model (LPM), or cluster-robust standard errors to account for clustered randomization.

In this paper, we use simulations to show that both of these approaches have poor statistical properties in term of size and power. Simple pairwise or groupwise t-tests and F-tests for individual variables pose different statistical challenges. With pairwise or groupwise t-tests, it is unclear how many rejections should lead to the conclusion that there is a balance problem or a randomization failure. Authors are left to subjectively assess whether an excessive number of tests have been rejected or whether one or more t-statistics are unreasonably large. One approach that is sometimes used is "vote counting", in which authors conclude that there is imbalance if e.g. more than 10 percent of tests reject the null at the 10% level. This approach is known to have low power in meta-analyses (Hedges and Olkin 1980). We show that it also has incorrect size: it rejects the null at very high rates. This happens because the fraction of p-values below 0.10 is itself a random variable. Even for independent tests with the correct size it is centered at 10% under the null, and so is greater than 10% nearly half the time.[2]

The omnibus tests of joint orthogonality typically deployed in the literature—which use sampling-based inference—have incorrect size, substantially over-rejecting the null hypothesis. This problem is worst when many baseline variables are included in the test and when heteroskedasticity-robust or cluster-robust standard errors are used. This over-rejection of the null means that omnibus tests of joint orthogonality wrongly indicate imbalance issues where none are present. The over-rejection problem is very large under realistic conditions. For example, with 500 observations and 50 covariates that are independent and normally dis-

[2]Asymptotically it exceeds 10% exactly half of the time. For datasets with finite samples, some fraction have a rejection rate of exactly 10%.

tributed, robust omnibus F-tests of joint orthogonality reject the null hypotheses at $\alpha = 0.10$ approximately 50% of the time, instead of the expected 10%.

In individually-randomized designs, the size problems are driven by the heteroskedasticity-robust variance estimators typically used in the literature. For balance tests, the error terms are actually homoskedastic under the null, and so the standard sandwich estimator performs better. In cluster-randomized designs, by contrast, over-rejection happens for all the usual variance estimators, including if we (falsely) assume homoskedasticity.

We propose and compare three alternative methods to assess balance: (1) an omnibus F-test of joint orthogonality with randomization inference p-values, (2) the minimum sharpened q-value to adjust p-values from pairwise t-tests and thereby control the false discovery rate (Benjamini, Krieger, and Yekutieli 2006; Anderson 2008), and (3) a Kolmogorov–Smirnov test to assess whether p-values from pairwise t-tests are uniformly distributed. We compare the performance of these methods in terms of statistical size and power using simulations. We conclude that omnibus F-tests of joint orthogonality with randomization inference p-values exhibit excellent performance in terms of both statistical power and size, for both individual and cluster RCTs. Randomization inference is also the conceptually correct method for calculating F-test p-values for balance tests in RCTs: under the null of balance, the primary source of uncertainty is from the random assignment of the treatment. Randomization inference computes the exact reference distribution of the F-statistic under that assignment mechanism, rather than relying on an asymptotic sampling approximation (Abadie et al. 2020). In individually-randomized designs, the minimum sharpened q-value from pairwise t-tests also performs well, providing the best statistical power to detect few large imbalances (even though it is conservative in terms of empirical size). The other approaches we test have problems with size or statistical power, especially for clustered RCTs.

We therefore recommend assessing balance using omnibus F-tests of joint orthogonality with randomization inference as this method is more reliable and flexible with different datasets, and it offers a more intuitive justification than alternative methods. This method

can be complemented with the minimum sharpened q-value from pairwise t-tests if treatment is randomized at the individual level. We discuss how to implement these tests with multiple treatments as well.

To illustrate the importance of these improved methods, we re-assess the balance of two RCTs whose results were recently published in top-five journals (Garbiras-Díaz and Montenegro 2022, Auriol et al. 2020). With pairwise t-tests and vote counting, 6% and 17% of tests are rejected at the 10% level in Garbiras-Díaz and Montenegro (2022) and Auriol et al. (2020) respectively, suggesting possible balance issues in the latter RCT. If we use typical omnibus F-tests of joint orthogonality with sampling-based inference, we reject the null of overall balance for some treatments in both papers. However, we find no significant balance issues in either paper when using an appropriate omnibus F-test of joint orthogonality with randomization inference. These findings strengthen the internal validity of the two papers in question. They also have broader implications for the RCT literature: while these two studies were not selected randomly, they are representative of the issues that affect balance tests in RCTs: 82% of recent randomized trials published in top five journals used pairwise t-tests or groupwise F-tests and vote counting and 32% used omnibus tests of joint orthogonality, and 100% of omnibus tests relied on sampling-based inference.

This paper contributes to four strands of the literature in empirical research in social science. First, it adds to existing work on the use of balance tests in randomized controlled trials. Senn (1994) prominently argued that balance tests should not be used at all; based on his work, using statistical tests to conclude that there are balance problems is commonly referred to as the "Table 1 Fallacy", particularly in health research (e.g. Sherry et al. 2023). In social science, however, balance tests are more widely supported. Leamer (1983) points out that unlucky draws in randomized trials can lead to exactly the same treatment assignments as would happen outside of an experiment, leading to the same concerns about a particular study's results being incorrect. Unbiasedness guarantees that those errors will cancel out on average, but offers no such promises about the results of any specific random assignment.

In a similar vein, Eckles (2021) argues that balance tests are important for verifying that the randomization actually took place. Imperfect randomization and failure to comply with treatment assignment are common in social science RCTs.[3] Our results help to clarify how to test for aggregate balance problems of the sort emphasized by Eckles. An individual t-test or comparison of standardized differences is sufficient for seeing whether there is imbalance on a specific variable, as in Leamer. To know whether the randomization protocol may have been violated, however, overall balance tests are needed.

We extend this literature by comparing the performance of various tests for both individually-randomized and clustered designs, considering a wide range of sample sizes and number of covariates, and assessing both the statistical size and power of the tests.[4] We show that the omnibus F-tests of joint orthogonality that most economics papers currently use—based on sampling inference—substantially over-reject the null of balance. In contrast, the same F-tests have the correct size and high power with randomization inference. Our analysis therefore builds on the literature that uses randomization inference for testing overall balance, in particular Hansen and Bowers (2008), who propose a test of joint balance that performs well under randomization inference. Their test is uncommon in economics. We show that the standard omnibus F-test with randomization inference has similar statistical power to the Hansen and Bowers test in individually-randomized designs and in clustered designs when the number of covariates is small, and better power in clustered designs when the number is large.

Second, our paper contributes to the literature on methods for the design and analysis of randomized trials. Bruhn and McKenzie (2009) use simulations to show how to optimize balance in the design of RCTs and how to analyze the data conditional on specific designs, and Athey and Imbens (2017) provide an overall guide to the analysis of data from randomized experiments. Abadie et al. (2020) discuss how to correctly conduct inference for

[3]For example, in the Perry Preschool program, some children were reassigned to different treatment statuses (Heckman, Pinto, and Shaikh 2024).

[4]Our study focuses on standard tests of differences used to assess covariate balance; methods for testing equivalence rather than difference are examined in Hartman and Hidalgo (2018).

data-generating processes like RCTs where the uncertainty comes from the assignment process rather than random sampling. A related line of work discusses the value of pre-specifying one's plan for analyzing the data from RCTs ahead of time (Casey, Glennerster, and Miguel 2012). We build on this body of work by providing specific guidance on how to conduct tests for overall balance. Omnibus tests of joint orthogonality have been used in applied research for some time, and are recommended by McKenzie (2015). But there is limited evidence on how to do inference on the F-statistics from this test. We show, in line with Abadie et al., that the p-values for these omnibus F-tests of joint orthogonality should properly be constructed using randomization inference.

Third, we contribute to an extensive body of research on the validity of empirical work in economics. Brodeur et al. (2016) find that there is substantial "missing mass" in the distribution of test statistics, suggesting that researchers are engaging in p-hacking in the vein of Simmons, Nelson, and Simonsohn (2011). Eble, Boone, and Elbourne (2017) assess randomized trials in economics by the standards used in medical research, showing that there is substantial risk of bias in the reporting of economics RCTs. Young (2019) shows that statistical analyses of data from RCTs over-reject the null hypothesis due to the inappropriate use of sampling- (rather than design-) based inference. A related line of work shows that multiple testing problems mean that many RCT results are false positives (Anderson 2008). In contrast with this previous work, we show that randomized experiments typically perform *better* than the literature might suggest: sampling-based approaches to inference skew the results toward (incorrectly) rejecting the null of balance. However, our findings also suggest the possibility of another sort of selective reporting. Since the standard omnibus test of joint orthogonality rejects the null so often, authors may be running it but not reporting it, as observed by Snyder and Zhuo (2024). Our suggested alternative approach, which uses randomization inference instead, can help head off this issue.

Finally, this paper complements the literature in econometrics on inference in models where there are many covariates. There is a large theoretical literature showing that classical

asymptotic tests tend to over-reject in finite samples. Early demonstrations in MacKinnon and White (1985) and Davidson, MacKinnon, et al. (1993) showed that heteroskedasticity-robust t-test over-reject in small samples, and suggested what is now known as HC3 standard errors. Cattaneo, Jansson, and Newey (2018) show that in linear regression models with heteroskedasticity, standard heteroskedasticity-robust standard errors become inconsistent when the number of covariates grows at the same rate as the sample size. Anatolyev (2012) and Anatolyev and Sølvsten (2023) show the poor asymptotic performance of omnibus F-tests with numerous restrictions, both in homoskedastic and heteroskedastic linear regression models. This problem has been often traced back to 'leverage', a measure of how individual explanatory observations can take extreme values influencing outcomes. We build on this work by showing that the balance tests commonly used in economics over-reject the null—often severely—when the number of restrictions is large compared to the size of the sample. Our simulations show that this problem affects logits and probits in addition to OLS, and that using HC3 standard errors does not correct the issue. We also show that this problem arises in clustered RCTs, not just individually-randomized experiments. Moreover, we illustrate how this issue can be corrected using randomization inference. Our approach also extends to the analysis of balance in multi-treatment RCTs via multinomial logits.

2 Methods to assess test size and power

We use simulations with four different data generating processes (DGPs) to assess the size and power of different omnibus tests of joint orthogonality.[5] We conduct all our simulations using Stata. Stata code for our simulations is available in the replication package for our paper.

[5]As robustness checks, we also considered two additional DGPs: (1) a DGP in which we first sample observations and then randomize them, and (2) a DGP with stratified randomization, where treatment probabilities vary across strata. In both cases, our main results hold. For stratified designs, it is important to control for strata fixed effects in the balance regressions, and to exclude the coefficients on these fixed effects from the omnibus test of joint orthogonality—particularly when treatment probabilities differ by strata.

2.1 Test Size

The size of a balance test refers to its probability of rejecting the null hypothesis of equal means when this null is true. A test has correct size if it rejects at the nominal rate (typically 1%, 5%, or 10%) under repeated realizations of the assignment or sampling process.[6] In balance tests, correct size corresponds to p-values being uniformly distributed on $[0, 1]$ under the null. Over-rejection—p-values skewed toward zero—indicates that empirical size exceeds nominal size, while conservative tests have empirical size below nominal size.[7]

We consider four DGPs to assess the size of balance tests. The first two DGPs use simulated data. We consider a simple DGP with N observations, k independent variables that are normally distributed $\sim N(0, 1)$,[8] and an independent treatment that is randomly assigned to half of the N observations (DGP 1). We also consider a more complex clustered design, in which the N observations are split into $C = 100$ clusters of equal size, the k variables are correlated within clusters (average intra-cluster correlation=0.2) and jointly correlated (average correlation = 0.2), and treatment is randomly assigned to half of the clusters (DGP 2).[9]

In our benchmark estimates, we assess test size by generating 500 simulated datasets with $N = 500$ observations and $n = 50$ covariates. To examine how test size varies with the number of regressors and sample size, we also vary the number of observations N from 200 to 5000, and the number of variables k from 10 to 100 for DGP 1 and from 10 to 80 for DGP 2.[10]

The two other DGPs use original datasets from existing randomized controlled trials

[6]The nominal size is the pre-specified rejection threshold, while the empirical size is the actual rejection frequency under repeated realizations of data generated under the null.

[7]Overly conservative tests are typically viewed as less problematic, but may indicate that a more powerful test is possible (Fisher and Robbins 2019).

[8]As a robustness check, we also considered a DGP with variables that are distributed following uniform, chi-squared, or binary distributions. Results suggest the distribution of variables has only minor impact on test size.

[9]We also explore allowing C to vary with the sample size as a robustness check.

[10]For the clustered design, the number of variables needs to remain below the number of clusters in order to retain enough degrees of freedom for the regressions to run.

but randomly (re-)assign the treatment variable using the same assignment rule as in the original paper. We selected two papers from our review of recent RCTs published in top five journals. These papers were selected intentionally, rather than at random, because we wanted to explore key features of the DGP that were shown to be problematic in our simulations. We selected one individually-randomized trial and one cluster-randomized trial, as the results of our simulations show that balance tests perform differently in these two types of RCTs.

The individually-randomized trial we study is Garbiras-Díaz and Montenegro (2022), which has a relatively small sample size and large number of covariates; our simulations show that this combination exacerbates the spurious imbalance issues with traditional omnibus F-tests of joint orthogonality. Garbiras-Díaz and Montenegro (2022) evaluate whether crowdsourced election monitoring via a Facebook ad campaign reduced electoral fraud during Colombia's 2019 mayoral elections. In 698 municipalities, citizens were randomized into four groups: a placebo group received a basic election reminder; the *"information"* group received a link to report irregularities; the *"call-to-action"* group received a motivational message to report irregularities; and the *"information + call-to-action"* group received both. In our simulations, we reassign the 698 municipalities into a "placebo" treatment group and a control group and use balance tests with 33 covariates to compare observations in both groups (DGP 3).

The cluster-randomized trial we analyze is Auriol et al. (2020) because one of the omnibus F-test of joint orthogonality reported in the paper has a p-value below 0.1, and we suspected that this was not a genuine imbalance problem but rather an issue with the test. In a lab-in-the-field experiment in Ghana, Auriol et al. (2020) test whether religious donations serve as a form of insurance, with believers giving in hopes of protection against future shocks. Clusters of participants from a Pentecostal church were randomized into one of three treatment arms: (i) the "Insurance" arm received a funeral insurance policy; (ii) the "Insurance Information" arm received only information about the policy; and (iii) the "No Insurance" arm received neither. All participants then played a dictator game to measure donations to the church

and other charities. In our simulations, we randomly reassign clusters into three groups and use balance tests with 10 covariates to compare two of them (DGP 4).

We use the four DGPs to assess the size of balance tests. For each DGP and each balance test, we generate 500 datasets and run the test separately in each dataset. We then estimate the cumulative distribution of p-values, i.e. the proportion of p-values (out of 500) that are below a threshold t, for t ranging from 0 to 1 in steps of 0.01. By construction, the treatment indicators are independent from the baseline covariates. Therefore, the balance tests have the correct size if p-values are uniformly distributed on [0,1], implying that the cumulative distribution of p-values is aligned with the 45 degree line. Deviations above or below the 45 degree line indicate over- or under-rejection, respectively. When interpreting the results of simulations, we will often consider DGPs 1 and 3 together, as they are both individually-randomized designs, and DGPs 2 and 4 together, as they are both clustered designs.

2.2 Statistical Power

To assess the power of balance tests, we add imbalances to some of the covariates of DGPs 1 and 2 and estimate how frequently the tests reject the null hypothesis that there is no imbalance between the treatment and control groups. We use the significance level of 10%, which is the highest threshold typically reported in economics.

We consider four approaches to generate imbalances of different magnitudes in different subsets of covariates. First, for one variable only, a very large imbalance is created by adding 0.25 to treated observations. Second, large imbalances in 10% of variables are created by adding 0.2 to treated observations. Third, for 20% of variables, medium imbalances are created by adding 0.15 to treated observations. Finally, small imbalances are created in 50% of variables by adding 0.1 to treated observations.

We consider simulated datasets with different numbers of observations, letting N ranging between 200 and 5,000. We estimate the proportion of p-values below 0.10. In our benchmark

estimates, we consider simulated datasets with 50 covariates. As a robustness check, we also assess how statistical power changes when the number of covariates varies between 10 and 100 for DGP 1 and between 10 and 80 for DGP 2.

When assessing statistical power, we focus on tests whose empirical sizes are equal or below their nominal sizes. Indeed, assessing the power of a test is misleading if its empirical size is above its nominal size, as a higher rate of rejecting the null hypothesis would likely be caused by more type I errors instead of fewer type II errors. If two balance tests have empirical size equal or below the nominal size, one should prefer the test with the highest statistical power to detect imbalance. On the contrary, if two tests have similar statistical power, one should prefer the test with the lowest empirical size, i.e. the lowest probability of type I error.

3 Balance tests in economics

To gain insights into how researchers approach balance tests in the economics literature, we systematically reviewed all papers that appeared when searching for the words "experiment", "field experiment", "field-experiment" , "randomized controlled trial", "randomized controlled trials", "randomised controlled trial", and "random" in the search engine of each of the top five journals in economics: the *American Economic Review*, *Econometrica*, the *Journal of Political Economy*, the *Review of Economic Studies*, and the *Quarterly Journal of Economics*. We also reviewed papers in the *American Economic Journal: Applied Economics* given the journal's focus on applied research and particularly RCTs. We identified 69 papers that were published between 2021 and 2023 and report original results from a randomized controlled trial. We analyzed these papers to identify whether and how the authors conducted balance tests. Our findings are summarized in Table A.1.

We find that 90% of the reviewed papers use balance tests, which we define as tests assessing the statistical significance or magnitude of the correlation between treatment status

and a vector of baseline covariates. Our review of the literature shows that economists employ a variety of balance tests. Pairwise t-tests are reported in 65% of the papers, followed by omnibus F-tests or chi-squared tests of joint orthogonality (32%), groupwise F-tests (26%), and pairwise normalized differences (6.5%).[11]

We investigate the statistical properties of these tests by categorizing them into two distinct groups based on whether they consider baseline covariates individually or jointly. Pairwise and groupwise tests examine the association between treatment status and the different baseline covariates, each considered independently. By contrast, omnibus tests of joint orthogonality consider the different baseline covariates jointly in a multivariate regression framework. These two categories of tests yield distinct statistical insights and challenges.

3.1 Pairwise and groupwise tests

Pairwise t-tests are by far the most-frequently used method to test for balance. Pairwise t-tests are testing the null hypothesis of equality of means in the treatment and control groups for each baseline covariate considered separately. Pairwise t-tests can also be estimated by regressing covariates on the treatment status. Regressions give authors the advantage of controlling for fixed effects and clustering at the randomization level. When there are more than two treatment arms, pairwise t-tests are sometimes replaced by groupwise F-tests obtained by regressing each baseline covariate on the full vector of treatment indicators and then testing the joint equality of the coefficient estimates with zero. Groupwise F-tests should not be confused with omnibus tests of joint orthogonality, where instead the treatment indicator is regressed on the full vector of baseline covariates; we discuss these in Section 3.2.

Pairwise t-tests and groupwise F-tests have the same underlying logic: for each variable, they test whether the differences across arms are consistent with what one would expect from random chance, under the null hypothesis. Among the 69 papers we reviewed, 82% report the results of pairwise t-tests or similar groupwise F-tests.

[11] Omnibus F-tests can also be used for re-randomization, but since re-randomization procedures are not commonly documented in detail we do not know how common this is.

To complement significance tests, authors sometimes also report pairwise normalized differences (Imbens and Rubin 2015). This happens in 6.5% of papers. The normalized difference between study arms is the difference in means divided by the pooled standard deviation ($\sqrt{(\sigma_C^2 + \sigma_T^2)/2}$). This is normally compared against some cutoff value such as 0.15 or 0.25. Large normalized differences suggest imbalances between study arms that are substantively large compared with the sample variance, which implies that there could also be discrepancies between the estimated treatment effects and the true parameter value. The value of this approach is that it avoids failing to reject the null simply because of a small sample.

Pairwise and groupwise balance tests are problematic for three related reasons. First, there is no clear rule for determining how many rejections of the null in a balance table constitute a balance problem. Second, this ambiguity creates additional "researcher degrees of freedom" (Simmons, Nelson, and Simonsohn 2011), allowing authors to frame balance tables as problematic or not, depending on the authors' preferences and audience pressures. Researchers may downplay imbalance problems in order to ease publication. Conversely, stakeholders with a vested interest in a program continuing may wish to sweep inconvenient null results under the rug by claiming that the randomization had problems. Third, because of the lack of guidance, researchers sometimes use *ad hoc* rules of thumb like "vote counting", in which the balance table indicates an overall balance problem if more than a specific fraction of pairwise tests rejects the null (Hedges and Olkin 1980).

We examine the performance of pairwise t-tests and vote counting in Appendix Figure A.1. For 500 repetitions of DGPs 1 and 2 respectively, we calculate the proportion of t-tests' p-values that are below versus above 0.1, as economists doing vote counting typically use this threshold to conclude that the study arms are balanced.[12] We find that vote counting dramatically over-rejects the null hypothesis that treatment and control groups are balanced. In 37% of the datasets generated using DGP 1, strictly more than 10% of t-tests are significant at the 10% level when considering a heteroskedasticity-robust variance estimator (the

[12]To implement each t-test, we use the Stata command `regress`, with the options `robust` or `cluster` when relevant, and record the p-value of the regression coefficient.

percentage is 34% with a variance estimator assuming homoskedasticity). Note that the first calculation is almost exactly what we would expect based on the binomial distribution, which predicts a probability of 0.38 for more than 5 successes in 50 trials. This suggests that these results are not due to any problems with the variance estimators for the underlying t-tests, but rather are driven by problems with the vote-counting method itself.

For DGP 2, 32% of datasets yield strictly more than 10% of t-tests that are significant at the 10% level when considering a cluster-robust variance estimator (the percentage is as high as 64% with a variance estimator assuming homoskedasticity). These percentages are much higher than the nominal size of 10%, implying that vote counting—if applied strictly—would misleadingly lead researchers to over-estimate imbalance problems.

3.2 Omnibus tests of joint orthogonality

Omnibus tests of joint orthogonality aim to address the limitations of pairwise and group-wise tests by considering baseline covariates jointly in a unique test. Among the papers we reviewed, 32% report the results of one or more omnibus tests of joint orthogonality. These tests involve regressing the treatment dummy on the vector of baseline covariates and test the null hypothesis that all regression coefficients are jointly equal to zero. Specifically, consider the following linear regression:

$$T_i = \alpha + \sum_{j=1}^{k} \beta_j x_{i,j} + \varepsilon_i \tag{1}$$

where $x_{i,j}$ is a set of k baseline variables for observation i. An omnibus test of joint orthogonality is a test of the null hypothesis that $\beta_1 = \beta_2 = \cdots = \beta_k = 0$, i.e. that the true data-generating process has no treatment-control differences for any baseline variable. The F-statistic for this test can be written out, in general, as

$$F = \frac{1}{k} \cdot \hat{\boldsymbol{\beta}}' \left(\widehat{V}^{-1} \right) \hat{\boldsymbol{\beta}} \tag{2}$$

where $\boldsymbol{\beta}$ is the vector of coefficient estimates $\hat{\beta}_j$ and $\widehat{V}$ is the estimated variance-covariance matrix of for the coefficient estimates. Equation 2 simplifies to $F = \frac{SSR/k}{SSE/(n-k-1)}$ under homoskedastic standard errors, but also allows for alternative variance estimators; as we discuss below, it is typical in the literature to use heteroskedasticity-robust or cluster-robust standard errors.[13]

Researchers face different options when implementing such tests in practice, including the choice of methods for estimating regression coefficients and standard errors. Among papers employing an omnibus balance test, a large majority (86%) report the F-statistic and associated p-value resulting from an Ordinary Least Squares (OLS) regression. A minority of studies (14%) opted for a chi-squared test resulting from a logit or probit regression, or their multinomial equivalent when there are more than two treatment arms.

Researchers also need to determine how to do inference on the test statistics to compute p-values. Among the papers we reviewed, 100% rely on sampling-based inference (we study randomization inference in Section 4.1). As the dependent variable in omnibus tests of joint orthogonality is either binary or categorical, heteroskedasticity may be an issue when OLS is used to estimate a linear probability model (LPM). Statistical software packages usually provide several heteroskedasticity-consistent (HC) variance estimators, including the HC1, HC2, and HC3 estimators (MacKinnon and White 1985). Researchers also use cluster-robust variance estimators when treatment status is assigned at the cluster level. Among the reviewed papers that use omnibus tests of joint orthogonality, 4.5% used heteroskedasticity-robust estimators, 59% used cluster-robust estimators, and 27% used a variance estimator that assumes homoskedasticity.

In Figure 1, we assess the size of omnibus F-tests and chi-squared tests of joint orthogonality using sampling-based inference.[14] Our key finding is that all versions of the omnibus

[13] Hotelling (1931)'s T^2 statistic has also been proposed as an omnibus test of balance. In the two-group setting relevant for RCTs, Hotelling's T^2 is algebraically equivalent (up to a monotonic scaling) to the homoskedastic omnibus F-test of joint orthogonality: $F = \frac{n-k-1}{k(n-2)}T^2 \sim F_{k,n-k-1}$. As a result, the two tests yield identical p-values. We therefore focus on the omnibus F-test in the paper, which is more standard in economics.

[14] The results using a probit model are very similar to those from the logit and hence are only reported in

tests of joint orthogonality have empirical size above nominal size, rejecting the null hypothesis too frequently in some or all DGPs. When the DGP does not involve clustering (DGPs 1 and 3), correct test size is only obtained when the omnibus test is issued from an OLS regression with a variance estimator that assumes homoskedasticity. All heteroskedasticity-robust variance estimators dramatically over-reject the null hypothesis. For instance, under DGP 1, an omnibus F-test of joint orthogonality resulting from an OLS regression with the robust HC1 variance estimator rejects the null hypothesis in 50% of samples, instead of the expected 10%. When a clustered design is considered (DGPs 2 and 4), all omnibus tests of joint orthogonality over-reject the null, even the one assuming homoskedasticity. The problem remains if cluster-robust standard errors are used.

Why do heteroskedasticity-robust standard errors perform so poorly in the individually-randomized designs? One possible explanation comes from Cattaneo, Jansson, and Newey (2018), who study inference for linear regression models where there are large numbers of covariates and heteroskedastic errors. They show that when k is large relative to N, the usual asymptotics for HC1 and HC2 SEs do not hold, and that using them can lead to over-rejection of the null. However, they show that HC3 SEs should perform well and, if anything, under-reject the null, whereas we do not find that for omnibus F-tests. We also find almost-identical problems for probits and logits, which are not linear models and do not suffer from the heteroskedasticity problem that typifies LPMs.

In our setting, the error terms are not actually heteroskedastic under the null hypothesis. Heteroskedastic errors are a general feature of LPMs, because extreme values of covariates push the outcome variable up against the boundary of its support, causing it to approach 0 or 1, which creates a correlation between the covariates and the error term. However, in our simulations, the null hypothesis of zero treatment-control difference is true and the probability of treatment is independent of the covariates, so there is no heteroskedasticity problem to address.[15]

the Appendix (Figures A.5 and A.6).

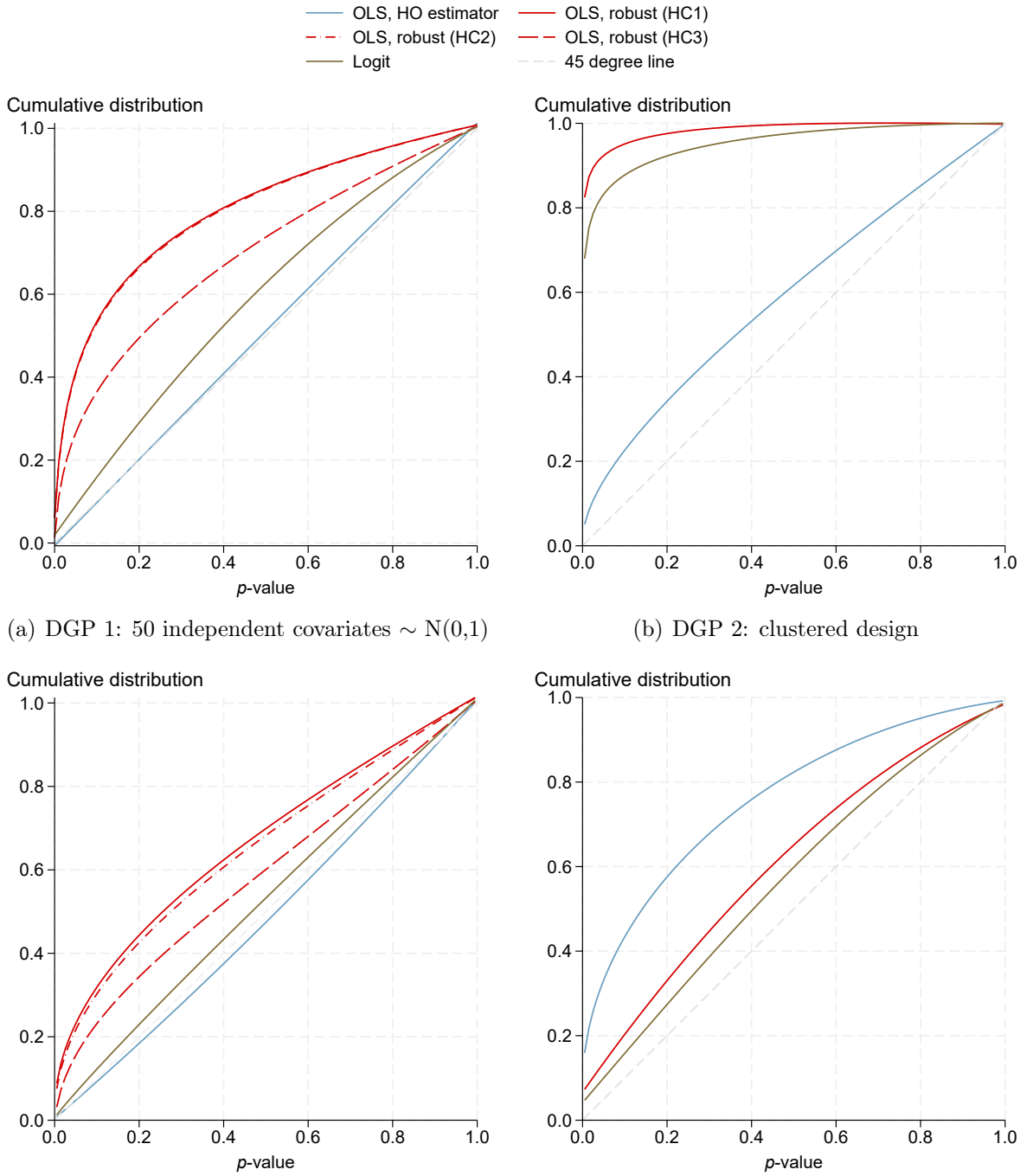
[15]We thank Dean Eckles for pointing this out, and Alan Griffith for helpful clarifying comments.

We conduct formal tests of the null of homoskedasticity in DGP 1 (Appendix Figure A.2), and find no evidence of heteroskedasticity in any of our simulations. Specifically, Breusch-Pagan tests never reject the null of homoskedastic errors at the 0.05 level, and only 0.4% of the time at the 0.10 level. This helps explain why the omnibus F-test assuming homoskedasticity performs well with DGPs 1 and 3. We also find that our HC1 SEs are actually slightly smaller on average than the HO SEs—the difference is just 0.2%, but it is statistically significant at the 0.01 level. This is in line with Angrist and Pischke (2008, Ch. 8), who show that, if there is no heteroskedasticity, then HC1 SEs are slightly downward-biased. Consistent with this, the heteroskedasticity-robust F-statistics are meaningfully larger than those that use homoskedastic SEs. Appendix Figure A.17 compares the empirical and theoretical distributions of the various F-statistics across three scenarios, which differ in the number of regressors ($k = 10$ or 50) and observations ($N = 500$ or 5000). In all three cases, the empirical distribution of the F-statistic under homoskedasticity closely matches its theoretical counterpart. In contrast, the empirical distributions of the F-statistics based on HC1, HC2, and HC3 heteroskedasticity-consistent estimators deviate substantially from the theoretical distribution when the number of regressors is large relative to the sample size.

Consistent with this result, our simulations show that problems with heteroskedasticity-robust estimators are magnified when sample size is smaller and the number of covariates is larger. We also confirm Cattaneo, Jansson, and Newey's conjecture that this result extrapolates to clustered designs, which they do not analyze. Figures A.5 to A.7 in the Appendix represent the proportion of tests that are rejected at the 10% level for different versions of omnibus tests of joint orthogonality, as a function of the number of observations and covariates. By construction, 10% of tests should reject the null if the test size is correct, as the treatment is randomly assigned and hence independent of the baseline covariates in all four DGPs.

Under DGP 1, the size of omnibus F-tests of joint orthogonality assuming homoskedasticity is correct even when N is small and the number of covariates is large (Appendix Figure

Figure 1
Size of Omnibus Tests of Joint Orthogonality Using Sampling-based Inference



(a) DGP 1: 50 independent covariates $\sim$ N(0,1)

(b) DGP 2: clustered design

(c) DGP 3: Garbiras-Díaz and Montenegro (2022)

(d) DGP 4: Auriol et al. (2020)

Notes: Cumulative distributions of p-values with 500 simulations per figure. DGPs 1 and 3 are individually-randomized designs, while DGPs 2 and 4 are clustered designs. For each test, we estimate the proportion of p-values that are below a threshold t, for $t \in [0.01, 0.99]$ in steps of 0.01; curves show fractional polynomial fits. Figures A.5 and A.6 show how test size varies with the number of covariates and observations for DGP 1 and DGP 2 respectively.

A.5). The size of the other omnibus tests is correct only when N is very large (≈ 5000) or when the number of covariates is low ($n < 10$), but incorrect when the number of observations is small or moderate and the number of covariates is larger than 10. The HC2 and HC3 estimators perform no better, over-rejecting the null at comparable rates to HC1.

Under DGP 2 (the clustered design), test size problems emerge even for large datasets and a relatively small number of covariates if the number of clusters is fixed—at 100 in our simulations (Appendix Figure A.6). However, when we allow the number of clusters to increase proportionally with the sample size ($C = N/5$), the size problem decreases as we increase the sample size and—mechanically—the number of clusters (Appendix Figure A.7).

A plausible reason for the size problems in clustered designs is that the standard cluster-robust variance estimator is biased downwards in finite samples. Moreover, the asymptotics of the cluster-robust variance estimator are in the number of clusters rather than the number of observations (Cameron, Gelbach, and Miller 2008).[16]

Taken together, our simulations show that issues emerge when k is large compared to N in individually-randomized trials, and when k is large compared to the number of clusters in clustered RCTs. Overall, we conclude that the methods currently used by economists to assess covariate balance are generally inadequate. Pairwise and groupwise tests, while commonly used, rely on subjective assessments of multiple test results by researchers, introducing a degree of subjectivity and leaving room for interpretation. Although omnibus tests of joint orthogonality address these issues, they usually over-reject the null, both for simulated data and original data from existing RCTs.

4 Alternative methods

We examine three alternative approaches to assess balance: (1) omnibus tests of joint orthogonality with randomization inference, (2) sharpened q-values to adjust p-values from pairwise t-tests and control the false discovery rate, and (3) a Kolmogorov–Smirnov test. We first de-

[16]Thanks to Austin Nichols for raising the point that this is a likely driver of our results.

scribe the intuition for these three methods and then examine their properties in terms of test size and statistical power.

4.1 Omnibus test of joint orthogonality with randomization inference

Omnibus tests of joint orthogonality can be used with randomization inference instead of sampling-based inference, as proposed by Hansen and Bowers (2008) and Young (2019). Randomization inference involves comparing the observed test statistic with the theoretical distribution of the test, which is computed by re-estimating the test statistic for a random sample of all possible treatment assignment vectors.[17] In this paper, we use 500 random reassignments for each test. The intuition for using this approach is that the uncertainty in randomized experiments comes not from *sampling* variation but from *assignment* variation—differences across repetitions of the experiment in terms of which units are assigned to treatment versus control (Abadie et al. 2020). For a balance test, the sharp null hypothesis is that treatment assignment is orthogonal to the baseline covariates for every unit. Because the covariates are fixed and the randomly assigned "treatments" do not affect them, this sharp null holds in our simulations by design. We reject this null hypothesis if the observed test statistic is at the extreme of the estimated theoretical distribution—for example, beyond the 90^{th}, 95^{th}, or 99^{th} percentile.[18]

Randomization inference aligns well with the intuition of balance tests, which examine the uncertainty or variation resulting from the randomization process and not from sampling. Another advantage of randomization inference is that it only requires specifying the randomization protocol and does not require additionally specifying a probabilistic model of the error term, which typically involves unknown distributional or heteroskedasticity as-

[17] The exact theoretical distribution of the test can in theory be obtained by estimating the test statistic for all possible treatment assignment vectors. However, this is computationally infeasible for all but the smallest samples.

[18] We need not consider the lower tail of the probability distribution because F-statistics are weakly positive by construction.

sumptions. This makes randomization inference more robust to non-normality and violations of homoskedasticity (Young 2019).

Randomization inference tests for balance have the advantage of being very simple to implement in Stata using the `ritest` package (Heß 2017).[19] Below is example code for implementing such a test:

```
*set the seed
set seed 3134


*define the list of balance variables here
local list_x x1 x2 x3 x4 x5


*for individually-randomized experiments
 ritest T e(F),  reps(500)  : reg T 'list_x'


*for cluster-randomized experiments
ritest T e(F),  reps(500)  cluster(cluster_id): \\\
    reg T 'list_x' , cluster(cluster_id)
```

For stratified designs, strata fixed effects should be included in the balance regressions but excluded from the omnibus F-test—particularly when treatment probabilities vary across strata. This ensures that the test focuses on covariate balance within strata, rather than detecting variation driven mechanically by the stratification structure. In Stata, this can be implemented by combining the `ritest` command with `reghdfe`, which leverages the Frisch–Waugh–Lovell theorem to "absorb" high-dimensional fixed effects (such as strata indicators) without estimating their coefficients directly.

[19] It is important to be cautious with missing values and *if conditions* when using the command `ritest`, as it considers the entire dataset in memory during re-randomization—including observations excluded by the `if` condition. To avoid potential issues, Simon Heß recommends specifying the estimation sample and the observations excluded by the `if` condition as separate strata. You can then use `ritest`'s `strata()` option, which ensures that treatment is re-randomized only within strata.

Our paper focuses on OLS regressions and the associated omnibus F-tests, as these are more commonly used in economics. However, we note that chi-squared tests following logit and probit regressions exhibit similar size and power, and can therefore be used interchangeably. In the Appendix, we also assess the performance of the randomization-inference test proposed by Hansen and Bowers (2008), which is a close relative of Hotelling's T^2 test (Hotelling 1931) and combines baseline differences into a single statistic. This test is rarely used in applied economics. We show that the Hansen and Bowers test has correct size under the null (Appendix Figure A.18). Its statistical power is broadly similar to that of the randomization-inference F-test in individually randomized designs, but it has lower power in clustered designs with many covariates (Appendix Figures A.19 and A.20). For these reasons, we focus on the standard omnibus F-test with randomization inference in our main analyses.

4.2 Adjustments for multiple hypothesis testing

Pairwise t-tests and groupwise F-tests are problematic because multiple tests are used to test one hypothesis, which is that the treatment arms are balanced. This issue could in principle be addressed using methods that adjust p-values to account for multiple hypothesis testing.

Two main approaches to address multiple hypothesis testing have been proposed in the literature (Anderson 2008). A first group of corrections aim to control for the Family-Wise Error Rate (FWER), which is the probability of making at least one type I error among all the hypotheses being tested. The goal is to control this probability at a desired significance level. The second group of corrections aim to control for the False Discovery Rate (FDR), which is the expected proportion of false discoveries (type I errors) among the rejected hypotheses. If all null hypotheses are true, then FWER and FDR are equivalent (Anderson 2008, p. 1487). This equivalence is important because for balance tests, the null hypotheses are generally expected to be true.

Both FWER and FDR corrections can therefore be considered in the context of bal-

ance tests. In Appendix Figures A.3 and A.4, we compare Romano-Wolf stepdown p-values (Romano and Wolf 2005; Clarke, Romano, and Wolf 2020), which control the FWER, and sharpened q-values, which control the FDR (Anderson 2008). The figures show that both categories of corrections have similar statistical size and power for individually-randomized designs (DGPs 1 and 3) but only the FDR correction has an empirical test size below its nominal size for clustered designs (DGPs 2 and 4), while the FWER correction substantially over-rejects the null. A possible explanation for the poor performance of the FWER correction is that the Romano–Wolf step-down procedure relies on a bootstrap whose accuracy deteriorates when the number of clusters is small.

In what follows, we therefore focus on sharpened q-values. We reject the null hypothesis that treatment arms are balanced if the minimum sharpened q-value is below the nominal size (which is equivalent to rejecting the null if any q-value is less than the threshold).

4.3 Kolmogorov–Smirnov test

If a treatment is randomly assigned, then the p-values of pairwise t-tests should be uniformly distributed. This can be tested using a Kolmogorov–Smirnov (K–S) test, which is a non-parametric statistical test that can be used to compare a sample distribution with a known reference distribution.

In the context of balance tests, the null hypothesis of the K-S test is that the sample of p-values from pairwise t-tests comes from a uniform distribution. The K–S statistic quantifies the maximum vertical distance between the empirical distribution of p-values and a uniform distribution. A key limitation of the K–S test is that it has low statistical power, especially for small sample sizes (Razali and Wah 2011).

4.4 Test Size

We assess the size of these alternative balance tests in Figure 2, considering the four DGPs described in Section 2. For each method, we consider variance estimators assuming ho-

moskedasticity and either heteroskedasticity-robust or cluster-robust variance estimators, depending on whether the treatment is allocated to individuals or clusters. We consider individually-randomized designs (DGPs 1 and 3) and clustered designs (DGPs 2 and 4) separately.

For DGP 1, all of the tests we consider have an empirical size equal to or below the nominal size.[20] For DGP 3 however, only the omnibus F-tests with randomization inference and the minimum q-values from pairwise t-tests with a variance estimator assuming homoskedasticity have an empirical test size at or below the nominal size. We compare their statistical power in the next section. By contrast, minimum q-values with a HC1 variance estimator and both versions of the Kolmogorov–Smirnov test have incorrect sizes, over-rejecting the null hypothesis of balance.

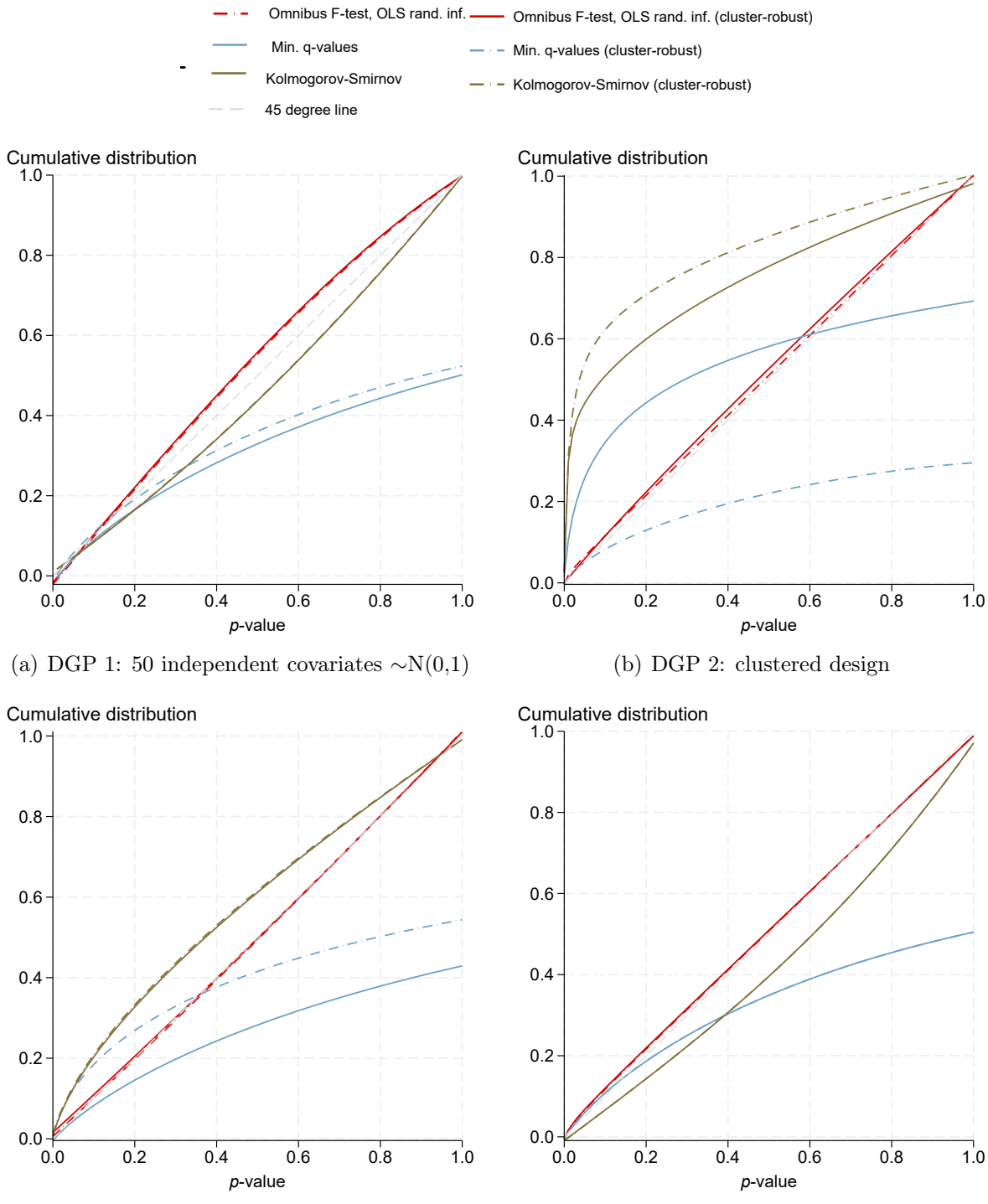
In DGPs that mimic a clustered RCT (DGPs 2 and 4), empirical size is equal to or below the nominal size only for the omnibus F-tests with randomization-based inference and minimum q-values from pairwise t-tests with cluster-robust variance estimators. We compare the statistical power of these tests in the next section. By contrast, minimum q-values with a variance estimator assuming homoskedasticity and both versions of the Kolmogorov–Smirnov tests tend to over-reject the null hypothesis.

These results are confirmed when varying both the sample size and the number of covariates (Appendix Figures A.8 and A.9). In all DGPs, minimum q-values appear to be conservative for high significance thresholds, indicating the tests tend to generate fewer type I errors than expected. This is not an issue *per se*, but it may indicate that a more powerful test could be designed (Fisher and Robbins 2019).

[20]The randomization inference curve closely follows the 45-degree line for p-values below 0.2, then rises slightly above it. This deviation results from the particular random seed used in all our simulations. Simulations with alternative seeds or additional repetitions (up to 2,000 as suggested by Young 2019) confirm that randomization inference yields correct test size.

Figure 2

Size of Omnibus Tests of Joint Orthogonality Using Randomization Inference, Kolmogorov–Smirnov Tests, and Minimum q-values



(a) DGP 1: 50 independent covariates ∼N(0,1)

(b) DGP 2: clustered design

(c) DGP 3: Garbiras-Díaz and Montenegro (2022)

(d) DGP 4: Auriol et al. (2020)

Notes: Cumulative distributions of p-values with 500 simulations per figure. DGPs 1 and 3 are individually-randomized designs, while DGPs 2 and 4 are clustered designs. For each test, we estimate the proportion of p-values that are below a threshold t, for $t \in [0.01, 0.99]$ in steps of 0.01; curves show fractional polynomial fits. Figures A.8 and A.9 show how test size varies with the number of covariates and observations for DGP 1 and DGP 2 respectively.

4.5 Statistical Power

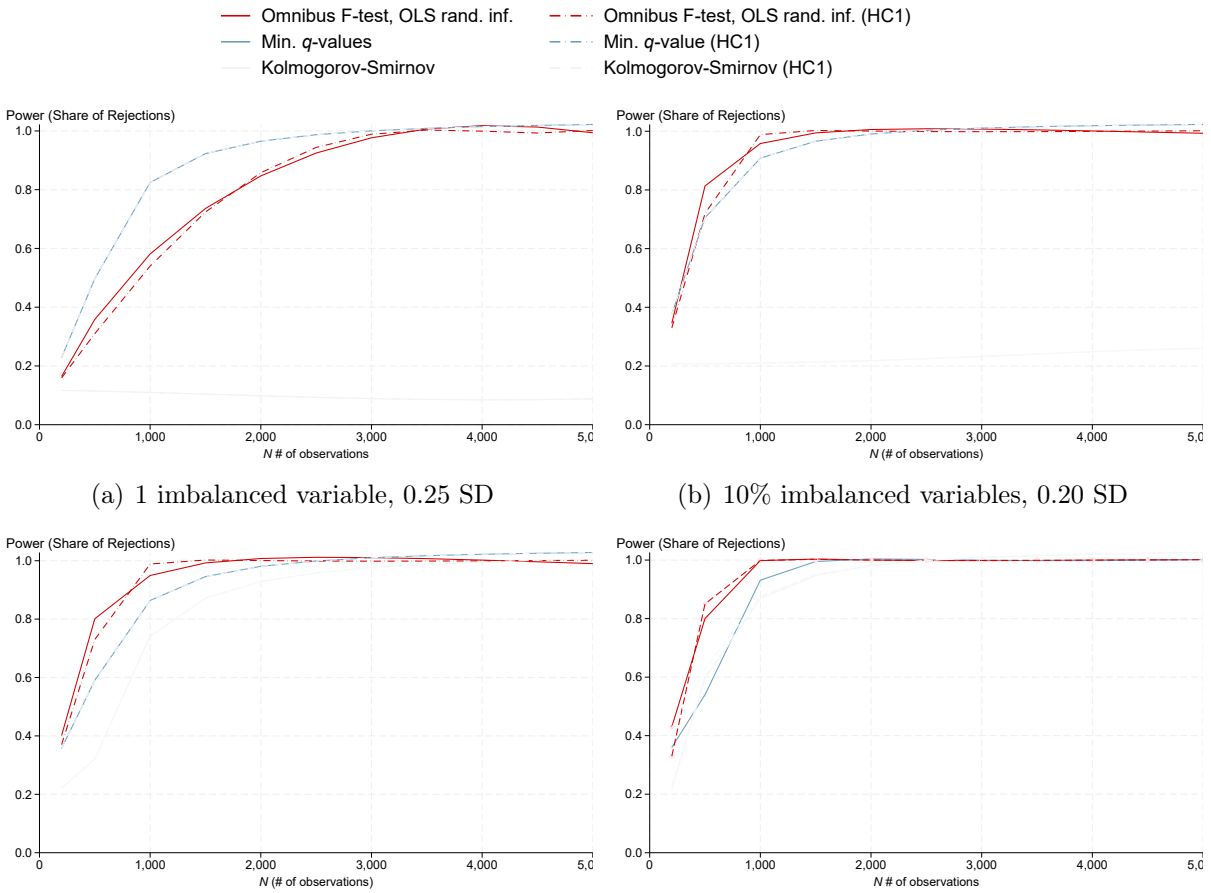
We assess the power of the balance tests in Figure 3 for DGP 1 and Figure 4 for DGP 2. We focus on balance tests that have correct sizes to avoid cluttering figures with misleading information.

For DGP 1, which assumes normally distributed variables and no clustering, we find that minimum q-values offer the best statistical power when only one variable is imbalanced, while the power of omnibus F-tests of joint orthogonality with randomization inference is intermediate.[21] Omnibus F-tests of joint orthogonality with randomization inference have higher power than minimum q-values when a larger number of imbalances, each of which is smaller in magnitude, are considered. This suggests that the two approaches might be complementary in individually-randomized RCTs. We obtain similar results when we vary the sample size and the number of covariates (Appendix Figure A.10).

With DGP 2, we find that omnibus F-tests of joint orthogonality using randomization inference and cluster-robust standard errors have the highest power, while minimum q-values from pairwise t-tests with cluster-robust variance estimators largely fail to detect imbalance. The results are similar when we vary the sample size and the number of covariates (Appendix Figure A.11). For clustered RCTs, we conclude that omnibus F-tests of joint orthogonality with randomization inference are a valid tool to assess balance, achieving the correct test size and higher statistical power than other approaches.

[21] We do not show the results of Kolmogorov-Smirnov tests and minimum q-values with a HC1 variance estimator as these approaches have incorrect size in DGP 3. Figure A.10 shows that the statistical power of minimum q-values with a HC1 variance estimator is similar to that of the minimum q-values with a variance estimator assuming homoskedasticity. The statistical power of Kolmogorov-Smirnov tests is low.

Figure 3
Power of Omnibus Tests of Joint Orthogonality for DGP 1 Using
Randomization Inference, Kolmogorov–Smirnov Tests, and Minimum q-values



(a) 1 imbalanced variable, 0.25 SD

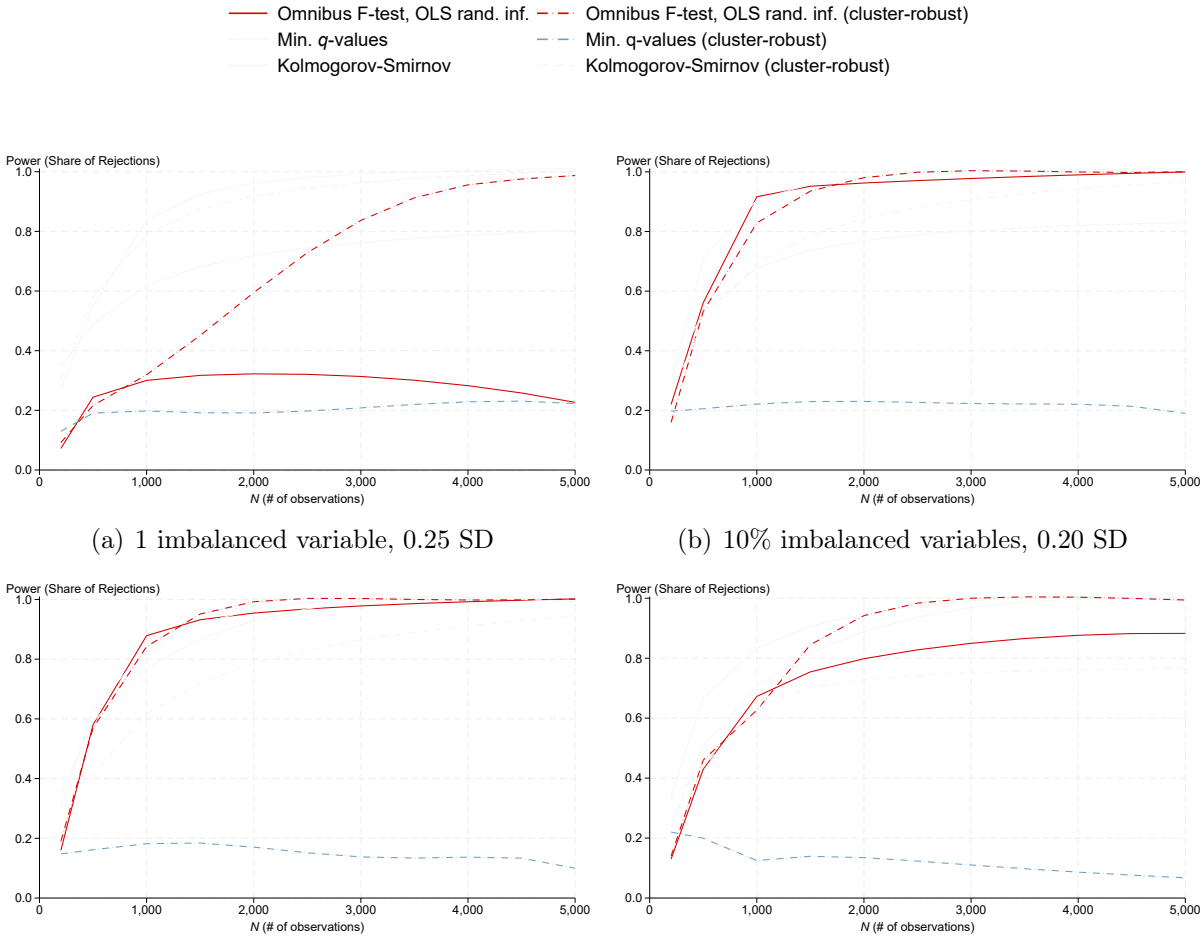
(b) 10% imbalanced variables, 0.20 SD

(c) 20% imbalanced variables, 0.15 SD

(d) 50% imbalanced variables, 0.10 SD

Notes: Power curves for DGP 1 with $k = 50$ and $N \in \{200, 500, 1000, 1500, 2000, 2500, 3000, 3500, 4000, 4500, 5000\}$. For each value of N, 100 simulated datasets are created and we compute the share of p-values or minimum q-values below 0.10; curves show fractional-polynomial fits. In Panel (a), one variable is made imbalanced by adding 0.25 to treated observations. In Panel (b), 10% of variables are made imbalanced by adding 0.2 to treated observations. In Panel (c), 20% of variables are made imbalanced by adding 0.15 to treated observations. In Panel (d), 50% of variables are imbalanced, by adding 0.1 to treated observations. Results for minimum q-values using HC1 SEs and K–S tests are not shown as these tests have incorrect size (see Figure 2). Figure A.10 shows how the power of the tests varies with the number of covariates and observations.

Figure 4
Power of Omnibus Tests of Joint Orthogonality for DGP 2
Using Randomization Inference, Kolmogorov–Smirnov Tests, and Minimum q-values



(a) 1 imbalanced variable, 0.25 SD

(b) 10% imbalanced variables, 0.20 SD

(c) 20% imbalanced variables, 0.15 SD

(d) 50% imbalanced variables, 0.10 SD

Notes: Power curves for DGP 2 with $k = 50$ and $N \in \{200, 500, 1000, 1500, 2000, 2500, 3000, 3500, 4000, 4500, 5000\}$. For each value of N, 100 simulated datasets are created and we compute the share of p-values or minimum q-values below 0.10; curves show fractional-polynomial fits. In Panel (a), one variable is made imbalanced by adding 0.25 to treated observations. In Panel (b), 10% of variables are made imbalanced by adding 0.2 to treated observations. In Panel (c), 20% of variables are made imbalanced by adding 0.15 to treated observations. In Panel (d), 50% of variables are imbalanced, by adding 0.1 to treated observations. Results for minimum q-values using HO SEs and K–S tests are not shown as these tests have incorrect size (see Figure 2). Figure A.11 shows how the power of the tests varies with the number of covariates and observations.

5 Multiple treatments and cross-randomization

Many randomized experiments involve more than one treatment, and thus more than two study arms. Out of the 69 RCTs that we identified in our literature review, 27 (39%) have more than one treatment. For these studies, we can examine balance between each treatment and the control group. That is the best course of action if we are concerned about the Leamer (1983) problem of unlucky random assignments. In that case, omnibus F-tests of joint orthogonality using randomization inference are the optimal approach.

For assessing the Heckman, Pinto, and Shaikh (2024) problem of violations of the randomization protocol, however, it is necessary to look for problems with *overall* balance. This means that we need to run combined omnibus tests of joint orthogonality across all study arms. Linear regression cannot handle this, but it is possible to implement a comparable test using multivariate analysis of variance and covariance (MANOVA) or multinomial logit with sampling-based or randomization inference. The minimum q-value approach and other multiple-hypotheses adjustments can also be easily adapted by considering the p-values from pairwise t-tests for each baseline covariate and each possible comparison between treatment arms. We consider individually-randomized designs (DGPs 1 and 3) and clustered designs (DGPs 2 and 4) separately. For DGPs 1 and 2, we consider four treatment arms with one quarter of observations randomly assigned to each. For DGPs 3 and 4, we consider four and three treatment arms, respectively, as in the original papers.

For the individually-randomized designs in DGPs 1 and 3, we show in Figure A.12 that all approaches have their empirical size equal or below the nominal size, except the multinomial logit with sampling-based inference, which over-rejects the null hypothesis.[22] We therefore omit this approach in power tests. Of the four remaining tests, the minimum q-value has the best statistical power to detect one large imbalance, while MANOVA and the multinomial logit with randomization inference have higher statistical power to detect multiple small

[22]The multinomial logit can fail to converge when the number of observations is low compared to the number of covariates (Figure A.15).

imbalances.

For DGPs 2 and 4, which are clustered designs, only MANOVA and multinomial logit with randomization inference have the correct test size, while minimum q-values are again conservative, under-rejecting the null (Figure A.12). We focus on these approaches in power tests and find that multinomial logit with randomization inference is, by far, the approach that has the best statistical power to detect single large or multiple small imbalances. We therefore recommend using this approach for clustered designs with multiple treatments. It is important to bear in mind, however, that this test is not high-powered: large sample sizes are needed to detect imbalances.

The Stata code to run this test parallels the code for F-tests above, but using the `mlogit` command instead of `reg`:

```
*for individiually-randomized experiments
ritest study_arm e(chi2),  reps(500): mlogit study_arm ‘list_x’

*for cluster-randomized experiments
ritest study_arm e(chi2),  reps(500)  cluster(cluster_id): \\\
    mlogit study_arm ‘list_x’ , cluster(cluster_id)
```

6 Revisiting existing papers

In this section, we reassess the balance of two RCTs whose results were recently published in top-five journals.

6.1 Garbiras et al. (2022)

To assess balance, Garbiras-Díaz and Montenegro (2022) report the results of pairwise t-tests, considering 33 baseline covariates (e.g. statistics on past reporting of irregularities,

socioeconomic covariates, political covariates, and region dummies). They use "vote counting" to conclude that municipalities are well balanced across treatment arms, reporting that *"Only 16 differences in means out of 264 comparisons in Table A2 are statistically significant at a 10 percent level or less."* The authors do not report the results of omnibus tests of joint orthogonality.

In Table 1, we report the results of the different tests of joint orthogonality discussed in our paper. With sampling-based inference and a heteroskedasticity-robust variance estimator, we find that two out of five tests are statistically significant at the 10% level. Two other p-values are just above 0.1, which could raise concerns. However, these results may be misleading: Section 3.2 concluded that such omnibus F-tests tend to over-reject the null hypothesis that groups have equal means. Indeed, if we use the randomization inference procedure that performs best in our simulations, we find that the p-values of omnibus F-tests of joint orthogonality are all well above 0.1. These results are consistent with our conclusion that these tests have correct test size when treatment is assigned at the individual level. We reach the same conclusion using minimum q-values.[23]

We also find that the control and treatment arms are well balanced when considering the multiple treatment arms together using MANOVA with sampling-based inference (p-value = 0.41), MANOVA with randomization inference (p-value = 0.41), multinomial logit with randomization inference (p-value = 0.48), and the minimum sharpened q-value from pairwise t-tests (minimum q-value = 1).

[23] All p-values of Kolmogorov-Smirnov tests are also above 0.1. However, we refrain from interpreting these results given the poor size and statistical power of this test.

Table 1
Replication of Existing Papers

Inference =	F-test p-value: Sampling-based HC1/Cluster	F-test p-value: Sampling-based H0	F-test p-value: Randomization HC1/Cluster	Min. q-value HC1/Cluster	K–S p-value HC1/Cluster
	(1)	(2)	(3)	(4)	(5)
Panel A: Replication of Garbiras-Díaz and Montenegro (2022)					
Any treatment vs. Control	0.088	0.287	0.278	1.000	0.737
Information vs. Control	0.155	0.366	0.562	0.413	0.595
Call to actions vs. Control	0.065	0.417	0.446	1.000	0.643
Info+call to action vs. Control	0.126	0.404	0.528	1.000	0.105
Any letter vs. No letter	0.428	0.631	0.664	1.000	0.349
Panel B: Replication of Auriol et al. (2020)					
Insurance vs. all other arms	0.988	0.967	0.996	1.000	0.005
Insurance info vs. all other arms	0.079	0.070	0.202	0.868	0.891
No insurance vs. all other arms	0.068	0.055	0.186	1.000	0.811

Notes: This table presents the results of different balance tests using the datasets from Garbiras-Díaz and Montenegro (2022) and Auriol et al. (2020). We rely on the same set of baseline covariates and present the same sets of comparisons as used by the authors in their original papers. For Auriol et al. (2020), we present the additional comparison "No insurance vs. all other arms".

6.2 Auriol et al. (2020)

Auriol et al. (2020) present a table of balance tests with twelve preregistered covariates in Table II of their paper.[24] They consider both pairwise t-tests and omnibus F-tests of joint orthogonality.[25] Out of the 24 p-values from pairwise t-tests, 4 are statistically significant at the 10% threshold (17%). A researcher using simple "vote counting" would conclude that the study arms are imbalanced. This is an example of the over-rejection problem we documented in Section 3.1. Moreover, their table also reports two p-values of omnibus F-tests of joint orthogonality—0.97 and 0.07—one of which is statistically significant at the 10% level.

In Table 1, we present the results of different omnibus tests of joint orthogonality. With sampling-based inference, two out of three omnibus F-tests of joint orthogonality are significant at the 10% level, both with a cluster-robust variance estimator and with a variance estimator assuming homoskedasiticty. Researchers using these tests may wrongly conclude that the RCT has a problem of imbalance. However, when we instead use randomization inference, all p-values are above 0.1. We obtain a similar conclusion with minimum q-values. Overall, these additional tests suggest there is no balance issue in the Auriol et al. (2020) RCT, contrary to the conclusions of conventional sampling-based inference.[26]

We also conclude that the control and treatment arms are well balanced when considering the multiple treatment arms together using multinomial logit with randomization inference (p-value = 0.28).

[24]The covariates are age, gender, total monthly income, a dummy of the employment status, three indicator variables reflecting ethnic group membership (Akan, Ewe, or Ga), and indicators for daily church attendance, praying multiple times per day, attending the revival week, and being recruited in the second wave.

[25]Pairwise t-tests are estimated for the twelve baseline covariates and two types of comparison: the "Insurance" treatment arm versus the "Insurance Information" treatment arm, and the "Information insurance" treatment arm versus the "No Insurance" treatment arm. For the two omnibus F-tests of joint orthogonality, the authors consider slightly different comparison groups, as the "Insurance" treatment arm is compared to the two other groups together in the first F-test, and the "Insurance Information" treatment arm is compared to the two other groups together in the second test. The F-tests are also estimated using a reduced list of baseline covariates, dropping the dummy variables identifying revival weeks and the second wave. We are able to exactly replicate the balance test results presented by the authors for both the pairwise t-tests and omnibus F-tests of joint orthogonality. In our analysis, we consider the same comparisons and variables as the authors used in their F-tests of joint orthogonality.

[26]We refrain from interpreting the results of Kolmogorov–Smirnov tests, which were shown to have poor test size and statistical power in Section 4.

7 Conclusion

Balance tests play a vital role in randomized experiments, especially when researchers lack full control over the randomization procedure. They allow researchers to verify whether randomization was implemented correctly and whether chance imbalances might undermine inference. They are even more important for natural experiments, where external factors fully determine treatment assignment. When used properly, balance tests can provide reassurance that treatment and control groups are comparable on observables—a key requirement for the internal validity of both RCTs and natural experiments.

However, the implementation of balance tests in economics often falls short. The most commonly used methods—pairwise t-tests combined with vote counting, or omnibus F-tests with sampling-based inference—tend to over-reject the null hypothesis of balance, especially when many covariates are included or sample sizes are modest. This can lead to incorrect conclusions about randomization failure even when treatment was truly assigned at random. As a result, researchers may avoid reporting balance test results that falsely suggest problems (Snyder and Zhuo 2024), or may even decide not to analyze or publish experimental findings at all due to misplaced concerns about the validity of the randomization (Miguel 2021; Franco, Malhotra, and Simonovits 2014).

We show that omnibus balance tests based on randomization inference perform substantially better than the conventional approaches that rely on sampling-based inference. Omnibus F-tests suffer from inflated Type I error rates, particularly when the number of covariates is large relative to the sample size, leading to frequent false rejections of the null of balance. In contrast, randomization inference yields tests with correct size and strong statistical power across for both individually- and cluster-randomized trials. Furthermore, the logic of randomization inference is better aligned with the purpose of balance tests: to assess whether observed differences across study arms could plausibly arise under the randomization procedure actually used, rather than under assumptions about asymptotic sampling distributions. This makes randomization inference not only statistically preferable, but also

conceptually appropriate for evaluating the success of random assignment.

Future work on balance tests should explore how these tests should best be used in practice. For example, what would happen if researchers abandoned all RCTs in which (correctly implemented) omnibus balance tests show an overall balance problem? And in particular, how does that vary with the true rate of randomization failures, and the bias caused by imperfect compliance with randomization protocols? If the null hypothesis holds, and all RCTs were in fact correctly run, then throwing out studies with imbalanced treatment allocations should not affect the expected value of the estimated treatment effect, but would mean that meta-analyses have higher standard errors due to relying on fewer studies. But if some experiments really are run incorrectly then throwing them out would *reduce* bias. Which pattern dominates is an empirical question, and one that should be informed by both careful simulations and engagement with the practitioners who actually implement these experiments in the field.

References

Abadie, Alberto, Susan Athey, Guido W. Imbens, and Jeffrey M. Wooldridge. 2020. "Sampling-based versus Design-Based Uncertainty in Regression Analysis." *Econometrica*, 88(1): 265–296.

Abebe, Girum, A Stefano Caria, Marcel Fafchamps, Paolo Falco, Simon Franklin, and Simon Quinn. 2021. "Anonymity or Distance? Job Search and Labour Market Exclusion in a Growing African City." *The Review of Economic Studies*, 88(3): 1279–1310.

Adhvaryu, Achyuta, Namrata Kala, and Anant Nyshadham. 2023. "Returns to On-the-job Soft Skills Training." *Journal of Political Economy*, 131(8): 2165–2208.

Afrouzi, Hassan, Spencer Y Kwon, Augustin Landier, Yueran Ma, and David Thesmar. 2023. "Overreaction in Expectations: Evidence and Theory." *The Quarterly Journal of Economics*, 138(3): 1713–1764.

Ainsworth, Robert, Rajeev Dehejia, Cristian Pop-Eleches, and Miguel Urquiola. 2023. "Why Do Households Leave School Value Added on the Table? the Roles of Information and Preferences." *American Economic Review*, 113(4): 1049–1082.

Alan, Sule, Ceren Baysan, Mert Gumren, and Elif Kubilay. 2021. "Building Social Cohesion in Ethnically Mixed Schools: An Intervention on Perspective Taking." *The Quarterly Journal of Economics*, 136(4): 2147–2194.

Alan, Sule, Gozde Corekcioglu, and Matthias Sutter. 2023. "Improving Workplace Climate in Large Corporations: a Clustered Randomized Intervention." *The Quarterly Journal of Economics*, 138(1): 151–203.

Alesina, Alberto, Armando Miano, and Stefanie Stantcheva. 2023. "Immigration and Redistribution." *The Review of Economic Studies*, 90(1): 1–39.

Allcott, Hunt, Joshua Kim, Dmitry Taubinsky, and Jonathan Zinman. 2022. "Are High-interest Loans Predatory? Theory and Evidence From Payday Lending." *The Review of Economic Studies*, 89(3): 1041–1084.

Anatolyev, Stanislav. 2012. "Inference in Regression Models with Many Regressors." *Journal of Econometrics*, 170(2): 368–382.

Anatolyev, Stanislav, and Mikkel Sølvsten. 2023. "Testing many Restrictions Under Heteroskedasticity." *Journal of Econometrics*, 236(1): 105473.

Anderson, Michael L. 2008. "Multiple inference and Gender Differences in the Effects of Early Intervention: A Reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects." *Journal of the American Statistical Association*, 103(484): 1481–1495.

Angelucci, Manuela, and Daniel Bennett. 2021. "Adverse Selection in the Marriage Market: HIV Testing and Marriage in Rural Malawi." *The Review of Economic Studies*, 88(5): 2119–2148.

Angrist, J.D., and J.S. Pischke. 2008. *Mostly harmless Econometrics: An Empiricist's Companion.* Princeton University Press.

Angrist, Joshua, David Autor, and Amanda Pallais. 2022. "Marginal Effects of Merit Aid for Low-income Students." *The Quarterly Journal of Economics*, 137(2): 1039–1090.

Angrist, Joshua D, Sydnee Caldwell, and Jonathan V Hall. 2021. "Uber Versus Taxi: a Driver's Eye View." *American Economic Journal: Applied Economics*, 13(3): 272–308.

Arteaga, Felipe, Adam J Kapor, Christopher A Neilson, and Seth D Zimmerman. 2022. "Smart Matching Platforms and Heterogeneous Beliefs in Centralized School Choice." *The Quarterly Journal of Economics*, 137(3): 1791–1848.

Athey, S., and G. W. Imbens. 2017. "The Econometrics of Randomized Experiments." In *Handbook of Economic Field Experiments*. Vol. 1, , ed. Esther Duflo and Abhijit V. Banerjee, 73–140. North-Holland. https://doi.org/10.1016/bs.hefe.2016.10.003.

Auriol, Emmanuelle, Julie Lassebie, Amma Panin, Eva Raiber, and Paul Seabright. 2020. "God Insures Those Who Pay? Formal Insurance and Religious Offerings in Ghana." *The Quarterly Journal of Economics*, 135(4): 1799–1848.

Aydin, Deniz. 2022. "Consumption Response to Credit Expansions: Evidence From Experimental Assignment of 45,307 Credit Lines." *American Economic Review*, 112(1): 1–40.

Bandiera, Oriana, Michael Carlos Best, Adnan Qadir Khan, and Andrea Prat. 2021. "The Allocation of Authority in Organizations: a Field Experiment with Bureaucrats." *The Quarterly Journal of Economics*, 136(4): 2195–2242.

Banerjee, Abhijit, Emily Breza, Arun G Chandrasekhar, and Benjamin Golub. 2023*a*. "When Less Is More: Experimental Evidence on Information Delivery During India's Demonetisation." *Review of Economic Studies*, rdad068.

Banerjee, Abhijit, Emily Breza, Arun G Chandrasekhar, Esther Duflo, Matthew O Jackson, and Cynthia Kinnan. 2023*b*. "Changes in Social Network Structure in Response to Exposure to Formal Credit Markets." *Review of Economic Studies*, forthcoming.

Barker, Nathan, Dean Karlan, Christopher Udry, and Kelsey Wright. 2024. "The Fading Treatment Effects of a Multifaceted Asset-transfer Program in Ethiopia." *American Economic Review: Insights*, 6(2): 277–294.

Baseler, Travis. 2023. "Hidden Income and the Perceived Returns to Migration." *American Economic Journal: Applied Economics*, 15(4): 321–352.

Battaglia, Marianna, Selim Gulesci, and Andreas Madestam. 2023. "Repayment flexibility and Risk Taking: Experimental Evidence from Credit Contracts." *Review of Economic Studies*.

Beaman, Lori, Ariel BenYishay, Jeremy Magruder, and Ahmed Mushfiq Mobarak. 2021. "Can Network Theory-based Targeting Increase Technology Adoption?" *American Economic Review*, 111(6): 1918–1943.

Beaman, Lori, Dean Karlan, Bram Thuysbaert, and Christopher Udry. 2023. "Selection Into Credit Markets: Evidence From Agriculture in Mali." *Econometrica*, 91(5): 1595–1627.

Bellemare, Charles, Marion Goussé, Guy Lacroix, and Steeve Marchand. 2023. "Physical Disability and Labor Market Discrimination: Evidence From a Video Résumé Field Experiment." *American Economic Journal: Applied Economics*, 15(4): 452–476.

Benjamini, Yoav, Abba M Krieger, and Daniel Yekutieli. 2006. "Adaptive Linear Step-up Procedures That Control the False Discovery Rate." *Biometrika*, 93(3): 491–507.

Bergman, Peter. 2021. "Parent-child Information Frictions and Human Capital Investment: Evidence From a Field Experiment." *Journal of Political Economy*, 129(1): 286–322.

Bessone, Pedro, Gautam Rao, Frank Schilbach, Heather Schofield, and Mattie Toma. 2021. "The Economic Consequences of Increasing Sleep Among the Urban Poor." *The Quarterly Journal of Economics*, 136(3): 1887–1941.

Brock, J Michelle, and Ralph De Haas. 2023. "Discriminatory Lending: Evidence From Bankers in the Lab." *American Economic Journal: Applied Economics*, 15(2): 31–68.

Brodeur, Abel, Mathias Lé, Marc Sangnier, and Yanos Zylberberg. 2016. "Star Wars: The Empirics Strike Back." *American Economic Journal: Applied Economics*, 8(1): 1–32.

Bruhn, Jesse, Kyle Greenberg, Matthew Gudgeon, Evan K Rose, and Yotam Shem-Tov. 2024. "The Effects of Combat Deployments on Veterans' Outcomes." *Journal of Political Economy*, 132(8): 2830–2879.

Bruhn, Miriam, and David McKenzie. 2009. "In Pursuit of Balance: Randomization in Practice in Development Field Experiments." *American Economic Journal: Applied Economics*, 1(4): 200–232.

Brune, Lasse, Eric Chyn, and Jason Kerwin. 2021. "Pay Me Later: Savings Constraints and The Demand for Deferred Payments." *American Economic Review*, 111(7): 2179–2212.

Buchmann, Nina, Erica Field, Rachel Glennerster, Shahana Nazneen, and Xiao Yu Wang. 2023. "A Signal to End Child Marriage: Theory and Experimental Evidence From Bangladesh." *American Economic Review*, 113(10): 2645–2688.

Byrne, David P, Leslie A Martin, and Jia Sheen Nah. 2022. "Price Discrimination by Negotiation: A Field Experiment in Retail Electricity." *The Quarterly Journal of Economics*, 137(4): 2499–2537.

Cai, Jing, and Shing-Yi Wang. 2022. "Improving Management Through Worker Evaluations: Evidence From Auto Manufacturing." *The Quarterly Journal of Economics*, 137(4): 2459–2497.

Cameron, A. Colin, Jonah B. Gelbach, and Douglas L. Miller. 2008. "Bootstrap-Based Improvements for Inference with Clustered Errors." *Review of Economics and Statistics*, 90(3): 414–427.

Carlana, Michela, Eliana La Ferrara, and Paolo Pinotti. 2022. "Goals and Gaps: Educational Careers of Immigrant Children." *Econometrica*, 90(1): 1–29.

Carrera, Mariana, Heather Royer, Mark Stehr, Justin Sydnor, and Dmitry Taubinsky. 2022. "Who Chooses Commitment? Evidence and Welfare Implications." *The Review of Economic Studies*, 89(3): 1205–1244.

Carter, Michael, Rachid Laajaj, and Dean Yang. 2021. "Subsidies and The African Green Revolution: Direct Effects and Social Network Spillovers of Randomized Input Subsidies in Mozambique." *American Economic Journal: Applied Economics*, 13(2): 206–229.

Casaburi, Lorenzo, and Tristan Reed. 2022. "Using Individual-level Randomized Treatment to Learn About Market Structure." *American Economic Journal: Applied Economics*, 14(4): 58–90.

Casey, Katherine, Rachel Glennerster, and Edward Miguel. 2012. "Reshaping institutions: Evidence on Aid Impacts Using a Preanalysis Plan." *The Quarterly Journal of Economics*, 127(4): 1755–1812.

Cattaneo, Matias D., Michael Jansson, and Whitney K. Newey. 2018. "Inference in linear Regression Models with Many Covariates and Heteroscedasticity." *Journal of the American Statistical Association*, 113(523): 1350–1361.

Chakravorty, Ujjayant, Manzoor H Dar, and Kyle Emerick. 2023. "Inefficient Water Pricing and Incentives for Conservation." *American Economic Journal: Applied Economics*, 15(1): 319–350.

Chen, Yiqun, Petra Persson, and Maria Polyakova. 2022. "The Roots of Health Inequality and the Value of Intrafamily Expertise." *American Economic Journal: Applied Economics*, 14(3): 185–223.

Christensen, Peter, and Christopher Timmins. 2023. "The Damages and Distortions From Discrimination in the Rental Housing Market." *The Quarterly Journal of Economics*, 138(4): 2505–2557.

Clarke, Damian, Joseph P Romano, and Michael Wolf. 2020. "The Romano–wolf Multiple-hypothesis Correction in Stata." *The Stata Journal*, 20(4): 812–843.

Cortés, Patricia, Jessica Pan, Laura Pilossoph, Ernesto Reuben, and Basit Zafar. 2023. "Gender Differences in Job Search and the Earnings Gap: Evidence From the Field and Lab." *The Quarterly Journal of Economics*, 138(4): 2069–2126.

Cullen, Zoë, and Ricardo Perez-Truglia. 2022. "How Much Does Your Boss Make? The Effects of Salary Comparisons." *Journal of Political Economy*, 130(3): 766–822.

Cullen, Zoë, Will Dobbie, and Mitchell Hoffman. 2023. "Increasing the Demand for Workers with a Criminal Record." *The Quarterly Journal of Economics*, 138(1): 103–150.

Dahl, Gordon B, Andreas Kotsadam, and Dan-Olof Rooth. 2021. "Does Integration Change Gender Attitudes? The Effect of Randomly Assigning Women to Traditionally Male Teams." *The Quarterly Journal of Economics*, 136(2): 987–1030.

Dal Bó, Ernesto, Frederico Finan, Nicholas Y Li, and Laura Schechter. 2021. "Information Technology and Government Decentralization: Experimental Evidence From Paraguay." *Econometrica*, 89(2): 677–701.

Davidson, Russell, James G MacKinnon, et al. 1993. *Estimation and inference in econometrics.* Vol. 63, Oxford New York.

De Janvry, Alain, Guojun He, Elisabeth Sadoulet, Shaoda Wang, and Qiong Zhang. 2023. "Subjective Performance Evaluation, Influence Activities, and Bureaucratic Work Behavior: Evidence From China." *American Economic Review*, 113(3): 766–799.

Depetris-Chauvin, Emilio, Ruben Durante, and Filipe Campante. 2020. "Building Nations through Shared Experiences: Evidence from African Football." *American Economic Review*, 110(5): 1572–1602.

Dhar, Diva, Tarun Jain, and Seema Jayachandran. 2022. "Reshaping Adolescents' Gender Attitudes: Evidence From a School-based Experiment in India." *American Economic Review*, 112(3): 899–927.

Diamond, Rebecca, Tim McQuade, and Franklin Qian. 2019. "The Effects of Rent Control Expansion on Tenants, Landlords, and Inequality: Evidence from San Francisco." *American Economic Review*, 109(9): 3365–3394.

Doerr, Annabelle, and Sarah Necker. 2021. "Collaborative Tax Evasion in the Provision of Services to Consumers: A Field Experiment." *American Economic Journal: Economic Policy*, 13(4): 185–216.

Dube, Oeindrila, Johannes Haushofer, Bilal Siddiqi, and Maarten Voors. 2021. "Building Resilient Health Systems: Experimental Evidence From Sierra Leone and The 2014 Ebola Outbreak." *The Quarterly Journal of Economics*, 1145: 1198.

Eble, Alex, Peter Boone, and Diana Elbourne. 2017. "On minimizing the Risk of Bias in Randomized Controlled Trials in Economics." *World Bank Economic Review*, 31(3): 687–707. Publisher: Oxford University Press.

Eckles, Dean. 2021. "Does the "Table 1 Fallacy" Apply If It Is Table S1 Instead?"

Fehr, Dietmar, Günther Fink, and B Kelsey Jack. 2022. "Poor and Rational: Decision-Making Under Scarcity." *Journal of Political Economy*, 130(11): 2862–2897.

Fisher, Thomas J, and Michael W Robbins. 2019. "A Cheap Trick to Improve the Power of a Conservative Hypothesis Test." *The American Statistician*, 73(3): 232–242.

Fowlie, Meredith, Catherine Wolfram, Patrick Baylis, C Anna Spurlock, Annika Todd-Blick, and Peter Cappers. 2021. "Default Effects and Follow-on Behaviour: Evidence From an Electricity Pricing Program." *The Review of Economic Studies*, 88(6): 2886–2934.

Franco, Annie, Neil Malhotra, and Gabor Simonovits. 2014. "Publication Bias in the Social Sciences: Unlocking the File Drawer." *Science*, 345(6203): 1502–1505.

Franklin, Simon, Clement Imbert, Girum Abebe, and Carolina Mejia-Mantilla. 2024. "Urban Public Works in Spatial Equilibrium: Experimental Evidence From Ethiopia." *American Economic Review*, 114(5): 1382–1414.

Freedman, David A. 2008. "On regression adjustments to experimental data." *Advances in Applied Mathematics*, 40(2): 180–193.

Garbiras-Díaz, Natalia, and Mateo Montenegro. 2022. "All Eyes on Them: a Field Experiment on Citizen Oversight and Electoral Integrity." *American Economic Review*, 112(8): 2631–2668.

Gazeaud, Jules, Eric Mvukiyehe, and Olivier Sterck. 2023. "Cash Transfers and Migration: Theory and Evidence From a Randomized Controlled Trial." *Review of Economics and Statistics*, 105(1): 143–157.

Gray-Lobe, Guthrie, Parag A Pathak, and Christopher R Walters. 2023. "The Long-Term Effects of Universal Preschool in Boston." *The Quarterly Journal of Economics*, 138(1): 363–411.

Guryan, Jonathan, Jens Ludwig, Monica P Bhatt, Philip J Cook, Jonathan MV Davis, Kenneth Dodge, George Farkas, Roland G Fryer Jr, Susan Mayer, Harold Pollack, et al. 2023. "Not Too Late: Improving Academic Outcomes Among Adolescents." *American Economic Review*, 113(3): 738–765.

Hanna, Rema, Esther Duflo, and Michael Greenstone. 2016. "Up in Smoke: the Influence of Household Behavior On The Long-run Impact of Improved Cooking Stoves." *American Economic Journal: Economic Policy*, 8(1): 80–114.

Hansen, Ben B, and Jake Bowers. 2008. "Covariate Balance in Simple, Stratified and Clustered Comparative Studies." *Statistical Science*, 219–236.

Hardy, Morgan, and Jamie McCasland. 2023. "Are Small Firms Labor Constrained? Experimental Evidence From Ghana." *American Economic Journal: Applied Economics*, 15(2): 253–284.

Hartman, Erin, and F Daniel Hidalgo. 2018. "An equivalence approach to balance and placebo tests." *American journal of political science*, 62(4): 1000–1013.

Heckman, James, Rodrigo Pinto, and Azeem M. Shaikh. 2024. "Dealing with Imperfect Randomization: Inference for The Highscope Perry Preschool Program." *Journal of Econometrics*, 105683.

Hedges, Larry V., and Ingram Olkin. 1980. "Vote-counting Methods in Research Synthesis." *Psychological Bulletin*, 88(2): 359–369.

Heß, Simon. 2017. "Randomization inference with Stata: A Guide and Software." *The Stata Journal: Promoting Communications on Statistics and Stata*, 17(3): 630–651.

Hotelling, Harold. 1931. "The Generalization of Student's Ratio." *The Annals of Mathematical Statistics*, 2(3): 360–378.

Hussam, Reshmaan, Atonu Rabbani, Giovanni Reggiani, and Natalia Rigol. 2022*a*. "Rational Habit Formation: Experimental Evidence From Handwashing in India." *American Economic Journal: Applied Economics*, 14(1): 1–41.

Hussam, Reshmaan, Erin M Kelley, Gregory Lane, and Fatima Zahra. 2022*b*. "The Psychosocial Value of Employment: Evidence From a Refugee Camp." *American Economic Review*, 112(11): 3694–3724.

Imbens, Guido W, and Donald B Rubin. 2015. *Causal Inference in Statistics, Social, and Biomedical Sciences.* Cambridge university press.

Jack, William, Michael Kremer, Joost De Laat, and Tavneet Suri. 2023. "Credit Access, Selection, and Incentives in a Market for Asset-collateralized Loans: Evidence From Kenya." *Review of Economic Studies*, 90(6): 3153–3185.

Jones, Maria, Florence Kondylis, John Loeser, and Jeremy Magruder. 2022. "Factor Market Failures and The Adoption of Irrigation in Rwanda." *American Economic Review*, 112(7): 2316–2352.

Kerwin, Jason T., and Rebecca L. Thornton. 2021. "Making The Grade: The Sensitivity of Education Program Effectiveness to Input Choices and Outcome Measures." *The Review of Economics and Statistics*, 103(2): 251–264.

Leamer, Edward E. 1983. "Let's take the Con Out of Econometrics." *The American Economic Review*, 73(1): 31–43. Publisher: American Economic Association.

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): 2213–2246.

Levy, Ro'ee. 2021. "Social Media, News Consumption, and Polarization: Evidence From a Field Experiment." *American Economic Review*, 111(3): 831–870.

Lin, Winston. 2013. "Agnostic notes on regression adjustments to experimental data: Reexamining Freedman's critique." *The Annals of Applied Statistics*, 295–318.

Lopez, Carolina, Anja Sautmann, and Simone Schaner. 2022. "Does Patient Demand Contribute to the Overuse of Prescription Drugs?" *American Economic Journal: Applied Economics*, 14(1): 225–260.

Lowe, Matt. 2021. "Types of Contact: A Field Experiment on Collaborative and Adversarial Caste Integration." *American Economic Review*, 111(6): 1807–1844.

Macchi, Elisa. 2023. "Worth Your Weight: Experimental Evidence on The Benefits of Obesity in Low-Income Countries." *American Economic Review*, 113(9): 2287–2322.

MacKinnon, James G, and Halbert White. 1985. "Some Heteroskedasticity-consistent Covariance Matrix Estimators with Improved Finite Sample Properties." *Journal of Econometrics*, 29(3): 305–325.

McGuirk, Eoin F, Nathaniel Hilger, and Nicholas Miller. 2023. "No Kin in the Game: Moral Hazard and War in the US Congress." *Journal of Political Economy*, 131(9): 2370–2401.

McKenzie, David. 2015. "Tools of The Trade: A Joint Test of Orthogonality When Testing for Balance." World Bank Development Impact Blog.

McKenzie, David, and Susana Puerto. 2021. "Growing Markets Through Business Training for Female Entrepreneurs: a Market-level Randomized Experiment in Kenya." *American Economic Journal: Applied Economics*, 13(2): 297–332.

Meghir, Costas, A Mushfiq Mobarak, Corina Mommaerts, and Melanie Morten. 2022. "Migration and Informal Insurance: Evidence From a Randomized Controlled Trial and a Structural Model." *The Review of Economic Studies*, 89(1): 452–480.

Miguel, Edward. 2021. "Evidence on Research Transparency in Economics." *Journal of Economic Perspectives*, 35(3): 193–214.

Mohanan, Manoj, Katherine Donato, Grant Miller, Yulya Truskinovsky, and Marcos Vera-Hernández. 2021. "Different strokes for Different Folks? Experimental Evidence on the Effectiveness of Input and Output Incentive Contracts for Health Care Providers with Varying Skills." *American Economic Journal: Applied Economics*, 13(4): 34–69.

Muralidharan, Karthik, Paul Niehaus, and Sandip Sukhtankar. 2023. "General Equilibrium Effects of (improving) Public Employment Programs: Experimental Evidence From India." *Econometrica*, 91(4): 1261–1295.

Muralidharan, Karthik, Paul Niehaus, Sandip Sukhtankar, and Jeffrey Weaver. 2021. "Improving Last-Mile Service Delivery Using Phone-based Monitoring." *American Economic Journal: Applied Economics*, 13(2): 52–82.

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.

Oh, Suanna. 2023. "Does Identity Affect Labor Supply?" *American Economic Review*, 113(8): 2055–2083.

Razali, Nornadiah Mohd, and Yap Bee Wah. 2011. "Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling Tests." *Journal of Statistical Modeling and Analytics*, 2(1): 21–33.

Romano, Joseph P, and Michael Wolf. 2005. "Stepwise Multiple Testing as Formalized Data Snooping." *Econometrica*, 73(4): 1237–1282.

Senn, Stephen. 1994. "Testing for Baseline Balance in Clinical Trials." *Statistics in Medicine*, 13(17): 1715–1726.

Sherry, Alexander D., Pavlos Msaouel, Zachary R. McCaw, Joseph Abi Jaoude, Eric J. Hsu, Ramez Kouzy, Roshal Patel, Yumeng Yang, Timothy A. Lin, Cullen M. Taniguchi, Claus Rödel, Emmanouil Fokas, Chad Tang, Clifton David Fuller, Bruce Minsky, Tomer Meirson, Ryan Sun, and Ethan B. Ludmir. 2023. "Prevalence and Implications of Significance Testing for Baseline Covariate Imbalance in Randomised Cancer Clinical Trials: The Table 1 Fallacy." *European Journal of Cancer*, 194. Publisher: Elsevier.

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): 1359–1366.

Snyder, Christopher M, and Ran Zhuo. 2024. "Examining Selection Pressures in the Publication Process through the Lens of Sniff Tests." *Review of Economics and Statistics*, 1–45.

Stephens Jr, Melvin, and Desmond Toohey. 2022. "The Impact of Health on Labor Market Outcomes: Evidence From a Large-scale Health Experiment." *American Economic Journal: Applied Economics*, 14(3): 367–399.

Tungodden, Jonas, and Alexander Willén. 2023. "When Parents Decide: Gender Differences in Competitiveness." *Journal of Political Economy*, 131(3): 751–801.

Wheeler, Laurel, Robert Garlick, Eric Johnson, Patrick Shaw, and Marissa Gargano. 2022. "Linkedin (to) Job Opportunities: Experimental Evidence From Job Readiness Training." *American Economic Journal: Applied Economics*, 14(2): 101–125.

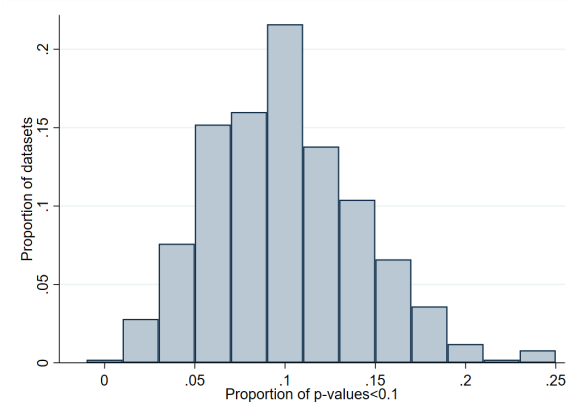
Young, Alwyn. 2019. "Channeling fisher: Randomization Tests and The Statistical Insignificance of Seemingly Significant Experimental Results." *The Quarterly Journal of Economics*, 134(2): 557–598.

Zárate, Román Andrés. 2023. "Uncovering Peer Effects in Social and Academic Skills." *American Economic Journal: Applied Economics*, 15(3): 35–79.

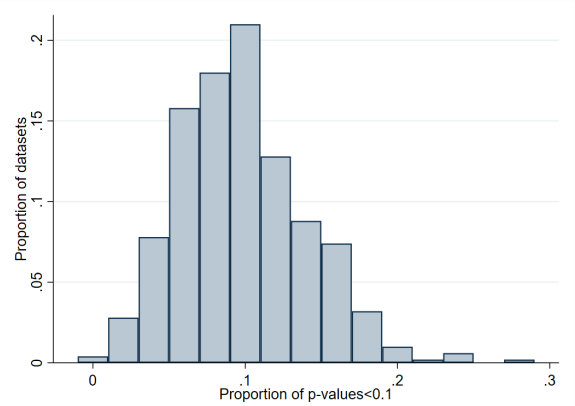
Zhao, Anqi, and Peng Ding. 2024. "No star is good news: A unified look at rerandomization based on p-values from covariate balance tests." *Journal of Econometrics*, 241(1): 105724.

Online Appendix

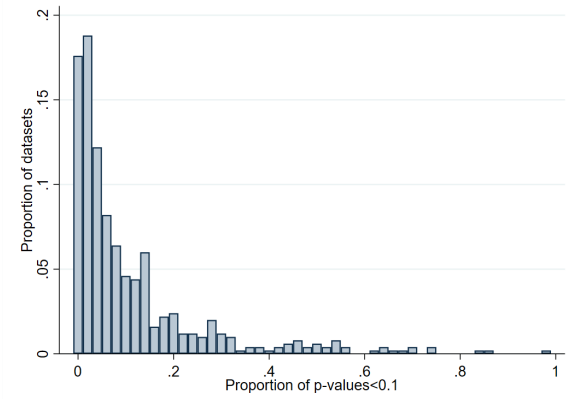
Figure A.1
Vote-Counting: Fraction of Pairwise t-tests with p-values Below 0.10



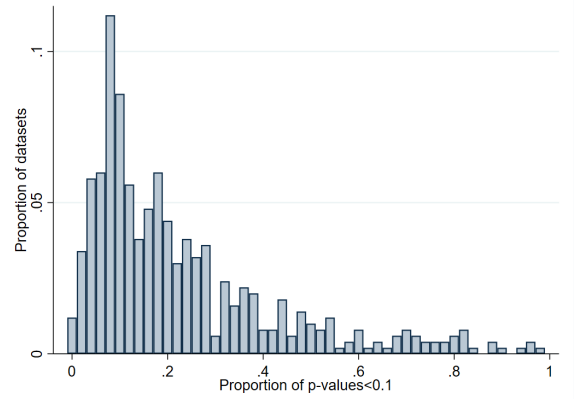
(a) Pairwise t-tests, OLS with HC1 variance estimator (DGP 1)



(b) Pairwise t-tests, OLS with HO variance estimator (DGP 1)



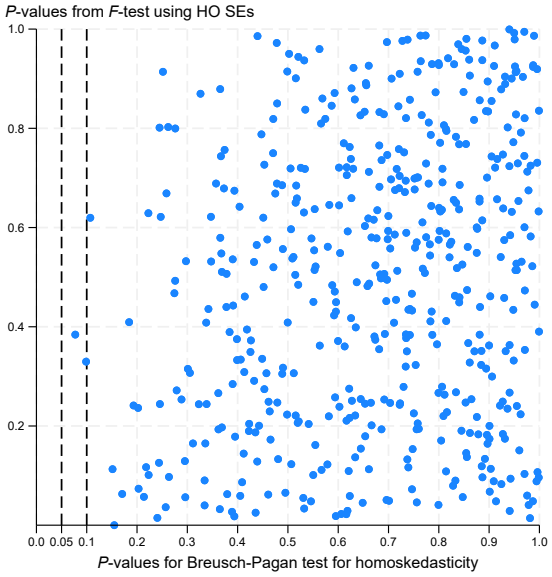
(c) Pairwise t-tests, OLS with clustered SEs (DGP 2)



(d) Pairwise t-tests, OLS with HO variance estimator (DGP 2)

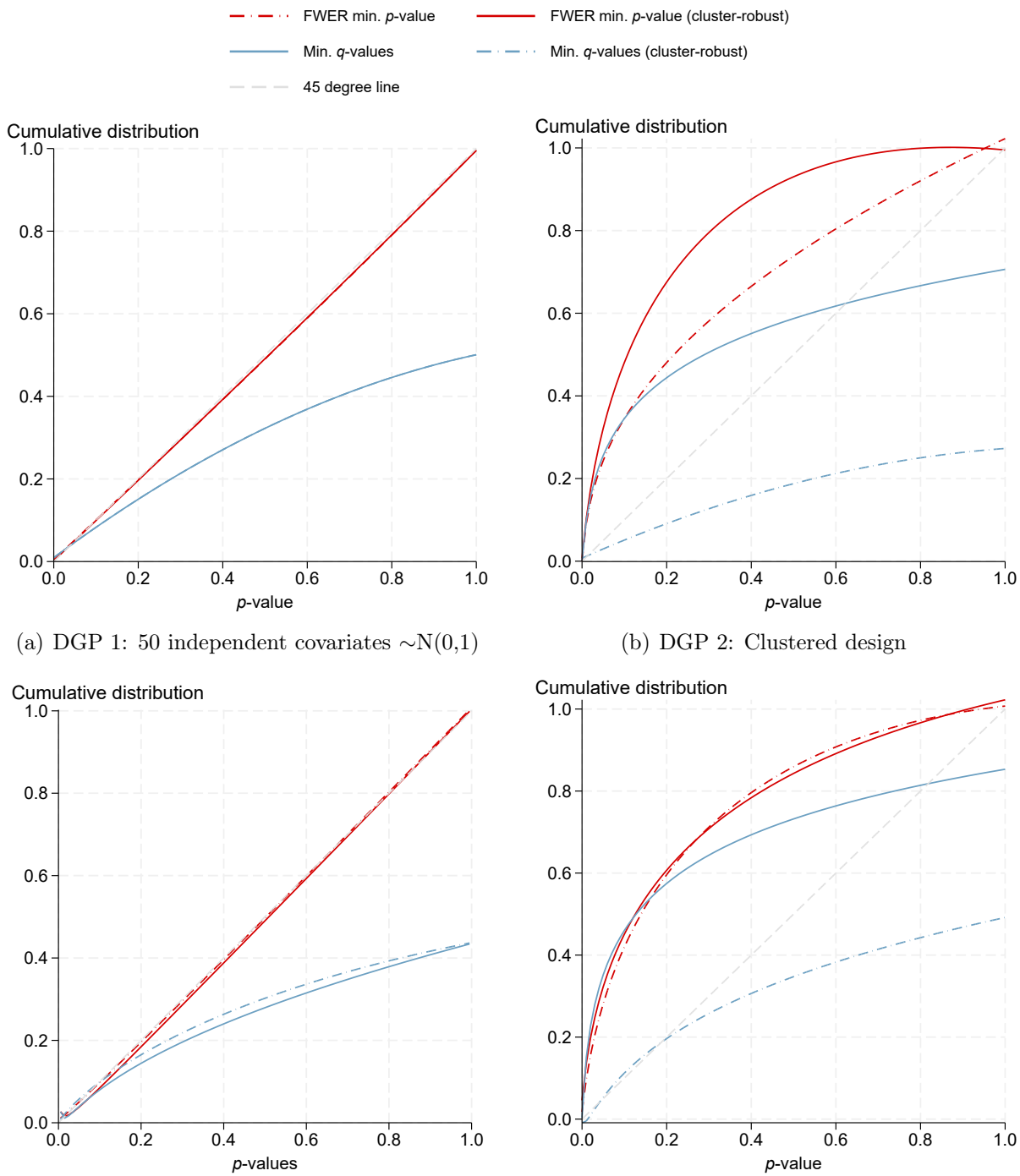
Notes: Each figure is based on 500 simulated datasets and shows the cumulative distribution of the share of p-values from pairwise t-test that are statistically significant at the 10% threshold. DGP 1 considers a data generating process with 500 observations, 50 independent variables that are normally distributed $\sim N(0,1)$, and an independent treatment randomly assigned to half of observations. DGP 2 considers a data generating process with 500 observations split in 100 clusters of equal size, 50 variables that are normally distributed, correlated within clusters (average coefficient of intra-cluster correlation = 0.2) and jointly correlated (average correlation = 0.2), and an independent treatment randomly assigned to half of the clusters. For each DGP, 500 datasets are generated, and for each dataset, we estimate the p-values of the 50 pairwise t-tests and calculate the share of p-values that are below 0.1. The four figures shows the distribution of these shares for DGPs 1 and 2 and for heteroskedasticity-robust and cluster-robust variance estimators as well as for a variance estimator assuming homoskedasticity.

Figure A.2
Breusch-Pagan Heteroskedasticity Test for DGP 1



Notes: The graph shows 500 simulated datasets based on DGP 1 with 500 observations and 50 independent variables that are normally distributed N∼(0,1), and an independent treatment randomly assigned to half of observations. Each point represents the intersection of the p-value from a F-test of joint orthogonality using HO standard errors and the p-value of from a Breusch-Pagan test of heteroskedasticity for one dataset.

Figure A.3
Size of Balance Tests Using FWER and FDR Multiple-Hypothesis Testing Adjustments



(a) DGP 1: 50 independent covariates ∼N(0,1)

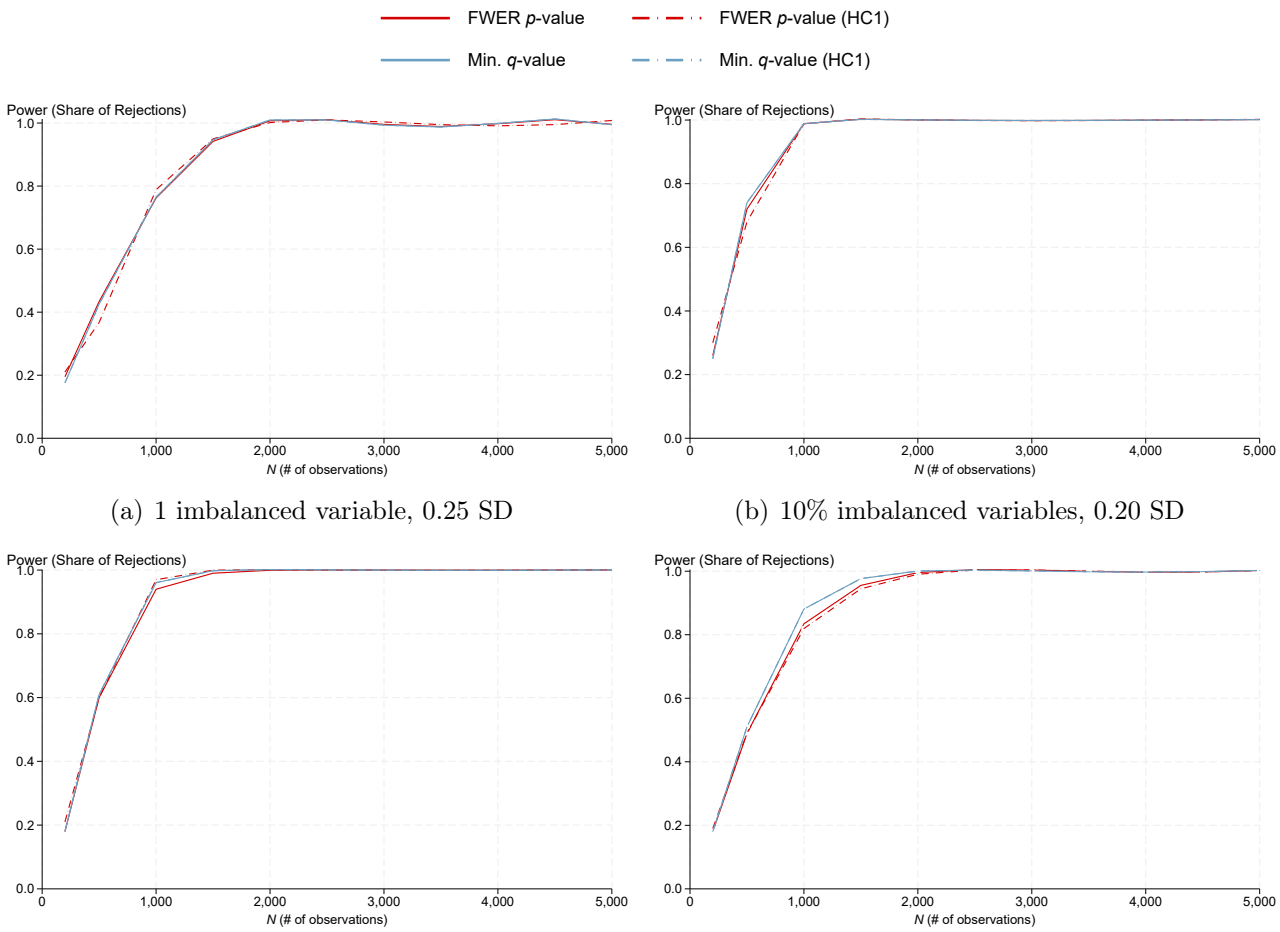
(b) DGP 2: Clustered design

(c) DGP 3: Garbiras-Díaz and Montenegro (2022)

(d) DGP 4: Auriol et al. (2020)

Notes: Cumulative distributions of minimum p- and q-values with 500 simulations per figure. DGPs 1 and 3 are individually-randomized designs, while DGPs 2 and 4 are clustered designs. For each test, we estimate the share of test statistics (out of 500) that are below a threshold t, for t ranging from 0.01 to 0.99 in steps of 0.01; curves show fractional polynomial fits.

Figure A.4
Power of Balance Tests Using FWER and FDR Multiple-Hypothesis Testing Adjustments for DGP 1



(a) 1 imbalanced variable, 0.25 SD

(b) 10% imbalanced variables, 0.20 SD

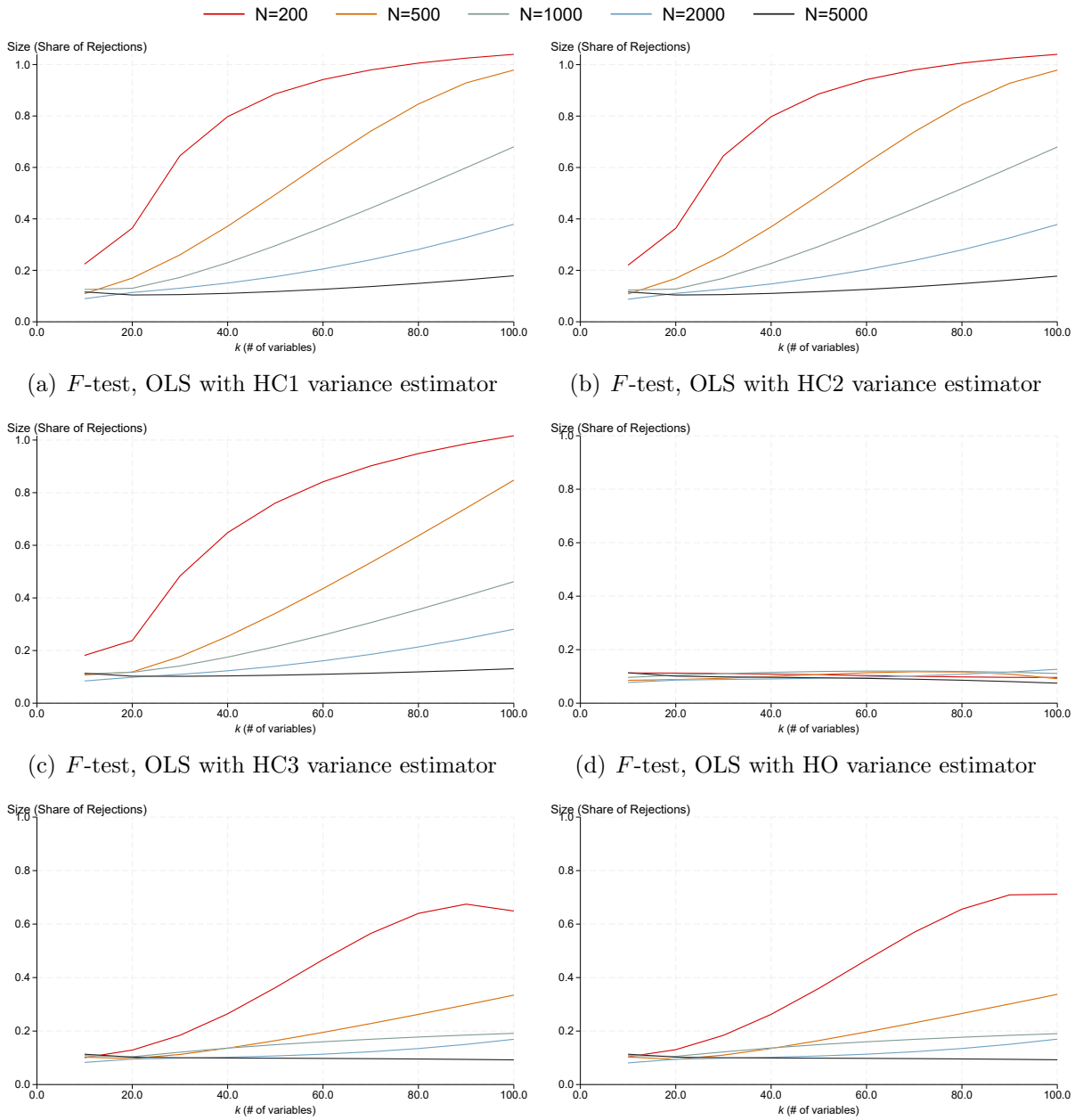
(c) 20% imbalanced variables, 0.15 SD

(d) 50% imbalanced variables, 0.10 SD

Notes: Power curves for DGP 1 with $k = 50$ and $N \in \{200, 500, 1000, 1500, 2000, 2500, 3000, 3500, 4000, 4500, 5000\}$. For each value of N, 100 simulated datasets are created and we compute the share of p-values or minimum q-values below 0.10; curves show fractional-polynomial fits. In Panel (a), one variable is made imbalanced by adding 0.25 to treated observations. In Panel (b), 10% of variables are made imbalanced by adding 0.2 to treated observations. In Panel (c), 20% of variables are made imbalanced by adding 0.15 to treated observations. In Panel (d), 50% of variables are imbalanced, by adding 0.1 to treated observations.

Figure A.5

Size of Omnibus Tests of Joint Orthogonality for DGP 1 Using Sampling-based Inference



(a) F-test, OLS with HC1 variance estimator

(b) F-test, OLS with HC2 variance estimator

(c) F-test, OLS with HC3 variance estimator

(d) F-test, OLS with HO variance estimator

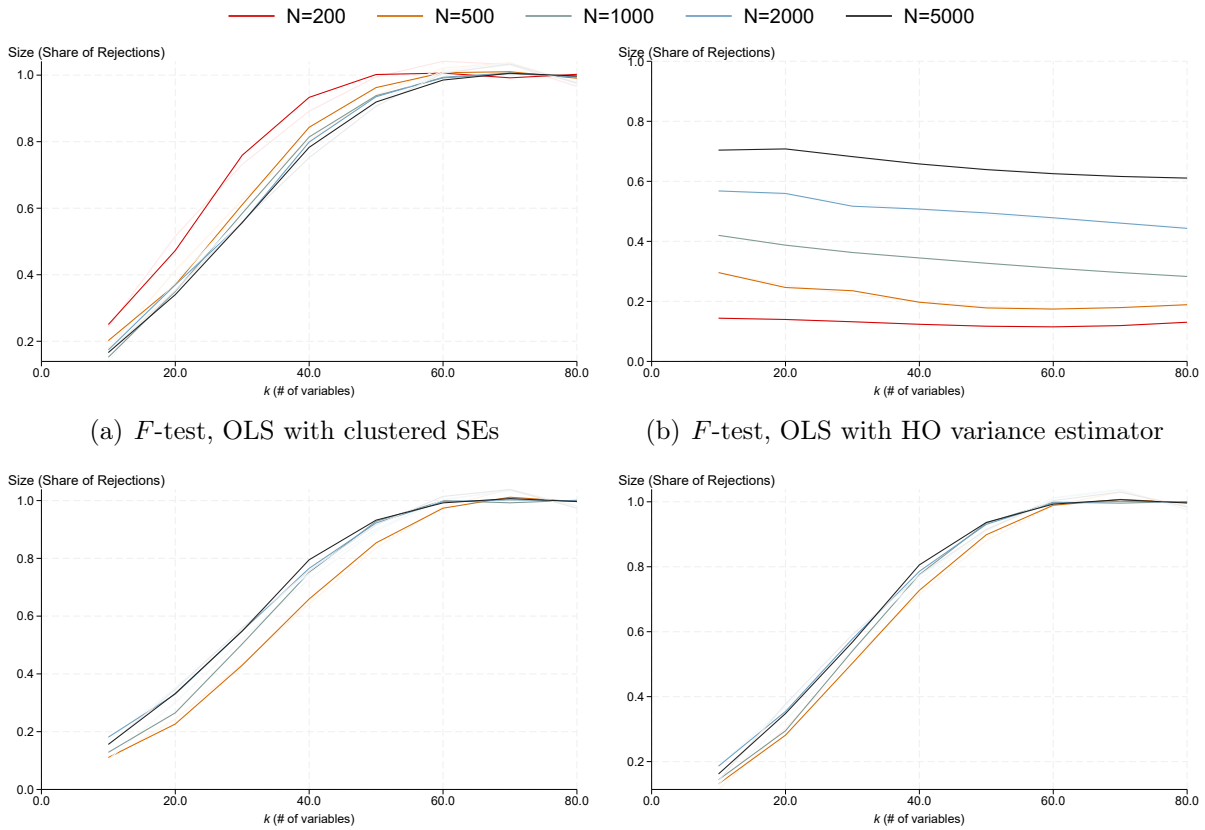
(e) Chi^2-test, Logit with HO variance estimator

(f) Chi^2-test, Probit with HO variance estimator

Notes: Size curves for DGP 1 with $k \in \{10, 20, 30, 40, 50, 60, 70, 80, 90, 100\}$ and $N \in \{200, 500, 1000, 2000, 5000\}$. For each combination of k and N, 500 simulated datasets are created and we compute the share of p-values below 0.10; curves show fractional-polynomial fits.

Figure A.6

Size of Omnibus Tests of Joint Orthogonality for DGP 2 Using Sampling-based Inference
(Holding # of Clusters Fixed, $C = 100$)



(a) F-test, OLS with clustered SEs

(b) F-test, OLS with HO variance estimator

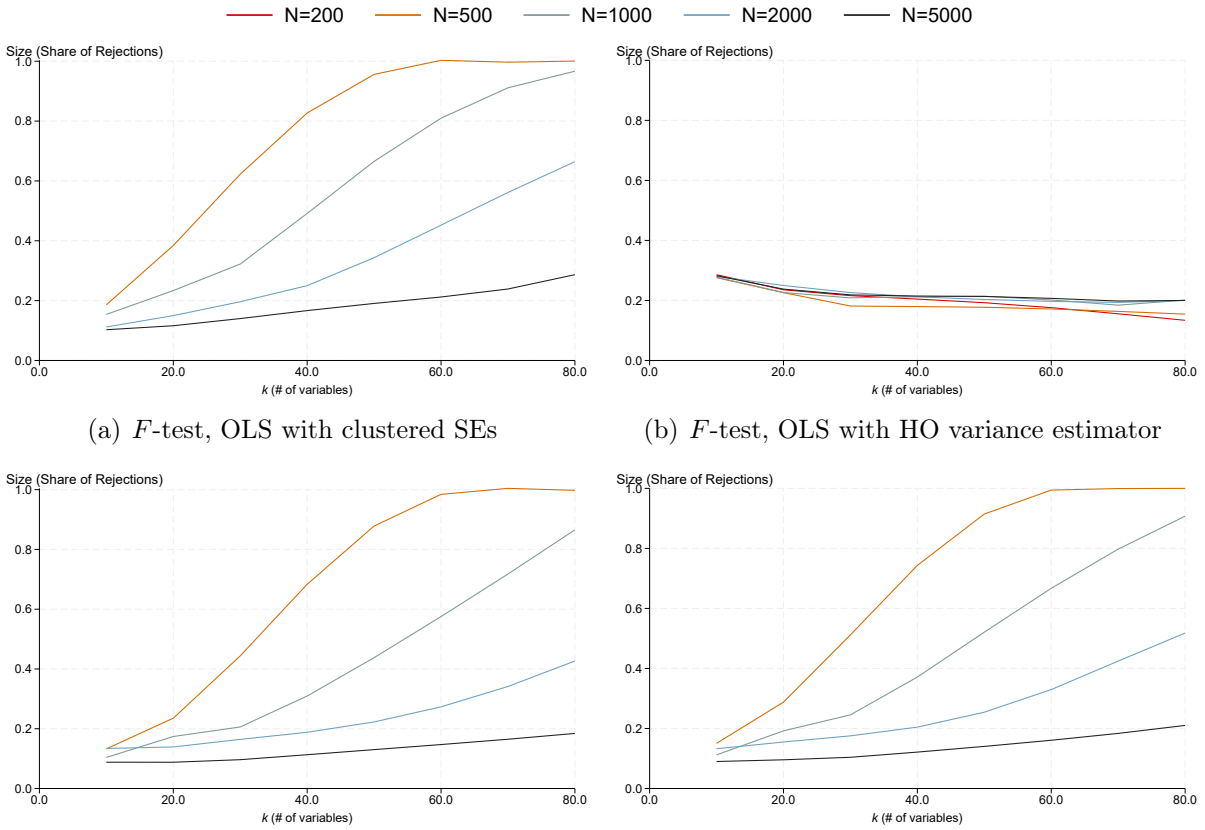
(c) F-test, Logit

(d) F-test, Probit

Notes: Size curves for DGP 2 with $k \in \{10, 20, 30, 40, 50, 60, 70, 80\}$, $N \in \{200, 500, 1000, 2000, 5000\}$, and 100 clusters. For each combination of k and N, 500 simulated datasets are created and we compute the share of p-values below 0.10; curves show fractional-polynomial fits.

Figure A.7

Size of Omnibus Tests of Joint Orthogonality for DGP 2 Using Sampling-based Inference
(Varying # of Clusters with Sample Size, $C = N/5$)



(a) F-test, OLS with clustered SEs

(b) F-test, OLS with HO variance estimator

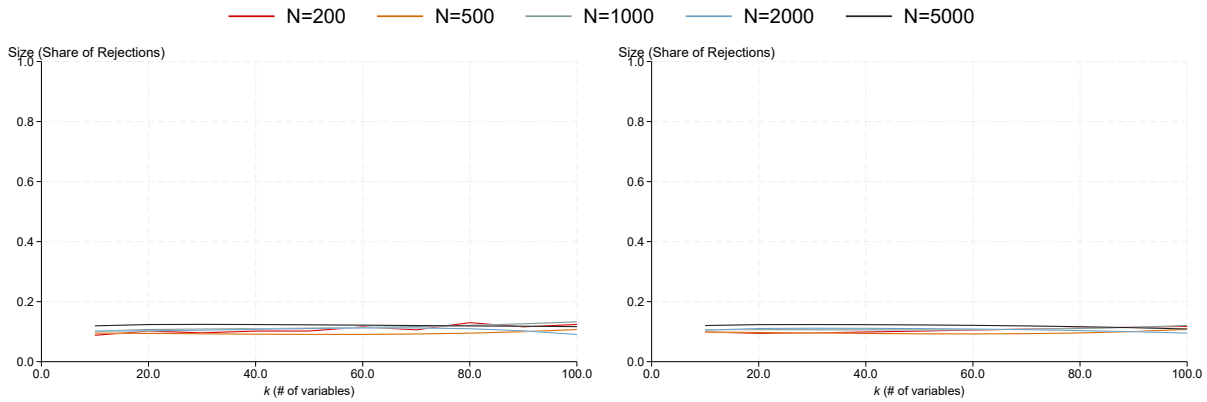
(c) F-test, Logit

(d) F-test, Probit

Notes: Size curves for DGP 2 with $k \in \{10, 20, 30, 40, 50, 60, 70, 80\}$, $N \in \{200, 500, 1000, 2000, 5000\}$, and $N/5$ clusters. For each combination of k and N, 500 simulated datasets are created and we compute the share of p-values below 0.10; curves show fractional-polynomial fits. Regressions with $N = 200$ and cluster-robust estimators cannot be estimated due to an insufficient number of clusters at k equal to and greater than 40.

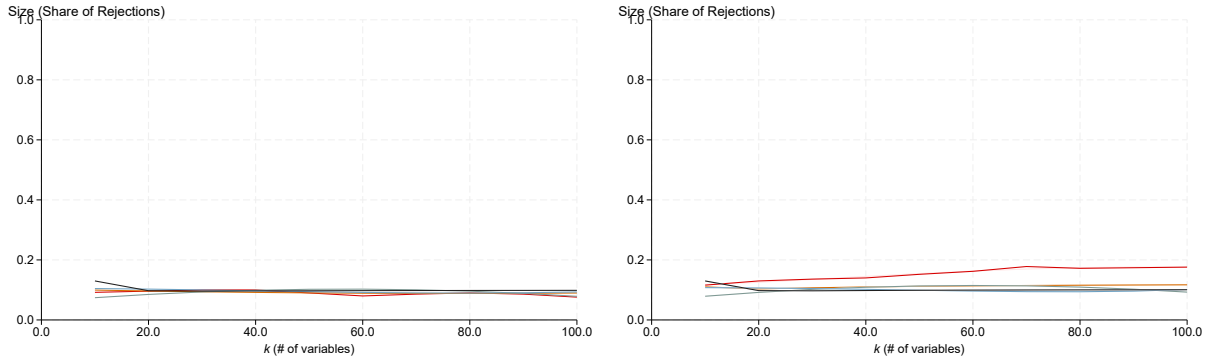
Figure A.8

Size of Omnibus Tests of Joint Orthogonality for DGP 1 Using Randomization Inference, Kolmogorov–Smirnov Tests, and Minimum q-values



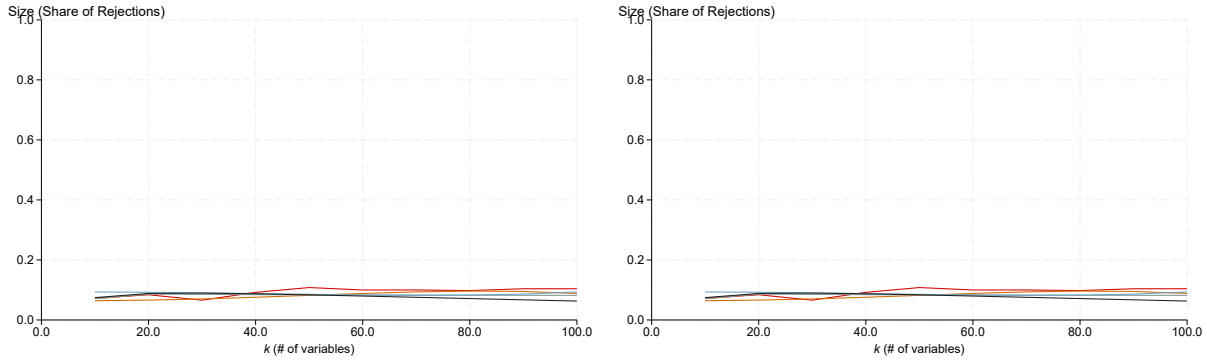
(a) F-test randomization inference, HO variance estimator

(b) F-test randomization inference, HC1 variance estimator



(c) q-values, HO variance estimator

(d) q-values, HC1 variance estimator



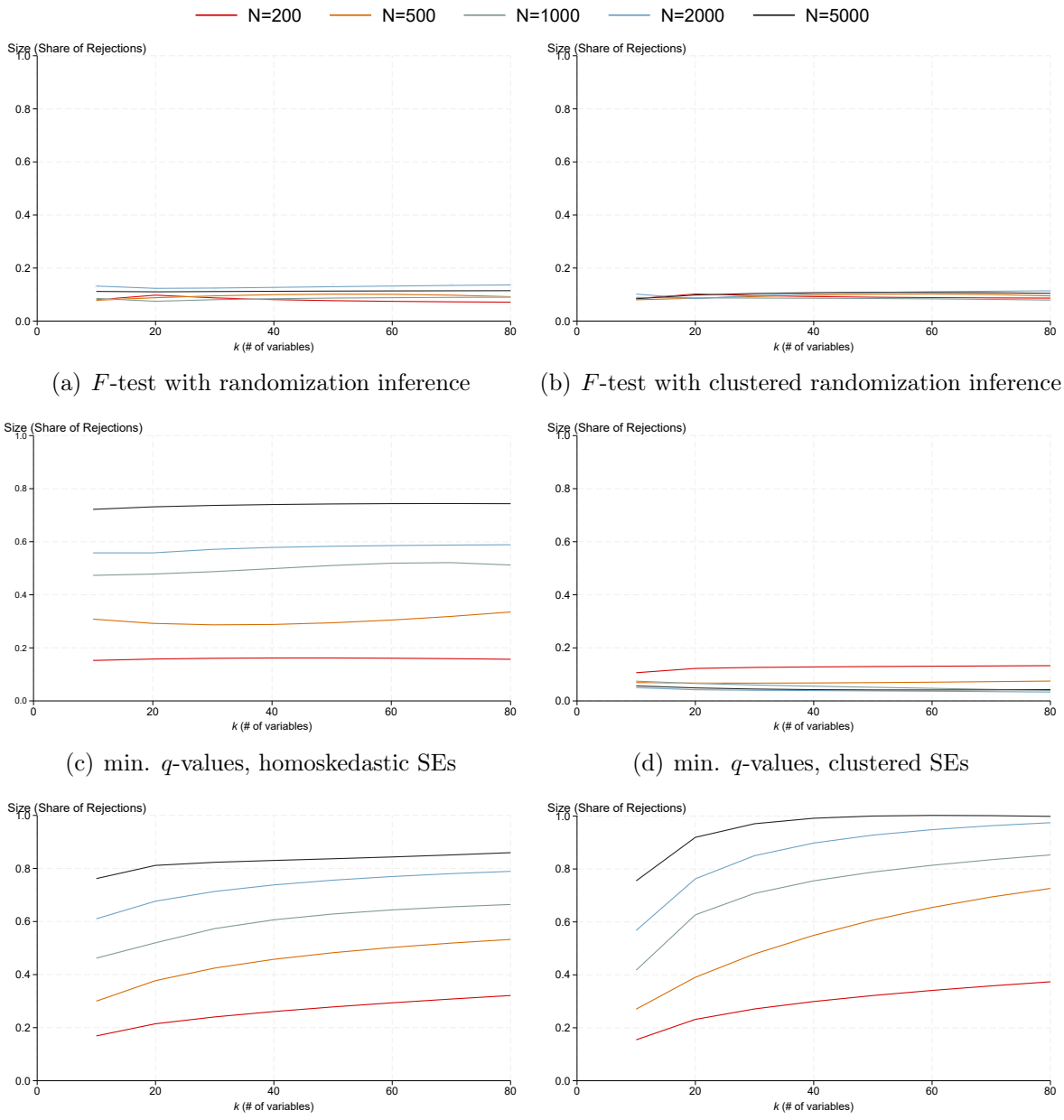
(e) Kolmogorov–Smirnov test, HO variance estimator

(f) Kolmogorov–Smirnov test, HC1 variance estimator

Notes: Size curves for DGP 1 with $k \in \{10, 20, 30, 40, 50, 60, 70, 80, 90, 100\}$ and $N \in \{200, 500, 1000, 2000, 5000\}$. For each combination of k and N, 500 simulated datasets are created and we compute the share of p-values or minimum q-values below 0.10; curves show fractional-polynomial fits.

Figure A.9

Size of Omnibus Tests of Joint Orthogonality for DGP 2 Using Randomization Inference, Kolmogorov–Smirnov Tests, and Minimum q-values



(a) F-test with randomization inference

(b) F-test with clustered randomization inference

(c) min. q-values, homoskedastic SEs

(d) min. q-values, clustered SEs

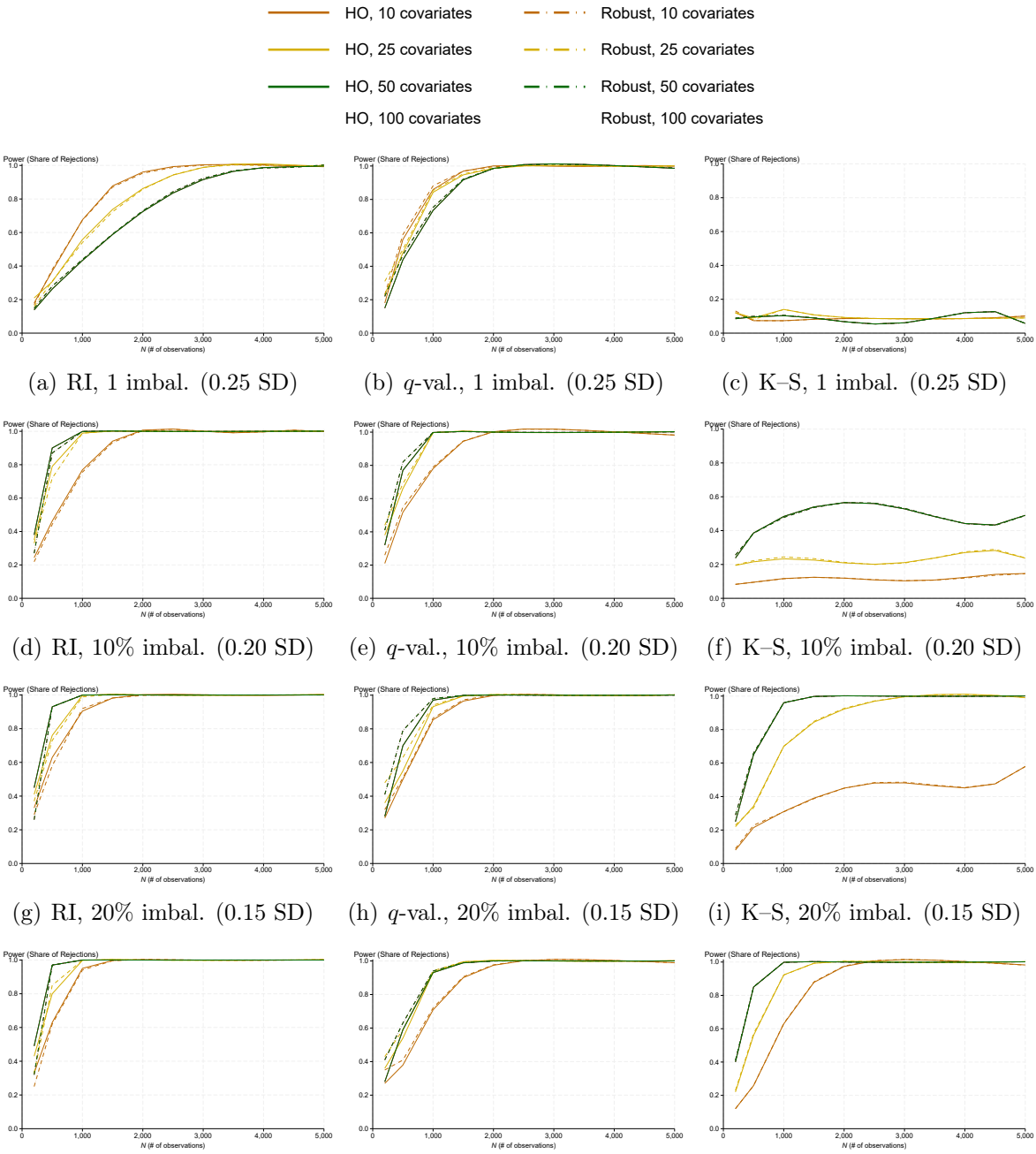
(e) K–S test, homoskedastic SEs

(f) K–S test, clustered SEs

Notes: Size curves for DGP 2 with $k \in \{10, 20, 30, 40, 50, 60, 70, 80\}$ and $N \in \{200, 500, 1000, 2000, 5000\}$. For each combination of k and N, 500 simulated datasets are created and we compute the share of p-values or minimum q-values below 0.10; curves show fractional-polynomial fits.

Figure A.10

Power of Omnibus Tests of Joint Orthogonality for DGP 1 Using Randomization inference, Kolmogorov–Smirnov Tests, and Minimum q-values



(a) RI, 1 imbal. (0.25 SD) (b) q-val., 1 imbal. (0.25 SD) (c) K–S, 1 imbal. (0.25 SD)

(d) RI, 10% imbal. (0.20 SD) (e) q-val., 10% imbal. (0.20 SD) (f) K–S, 10% imbal. (0.20 SD)

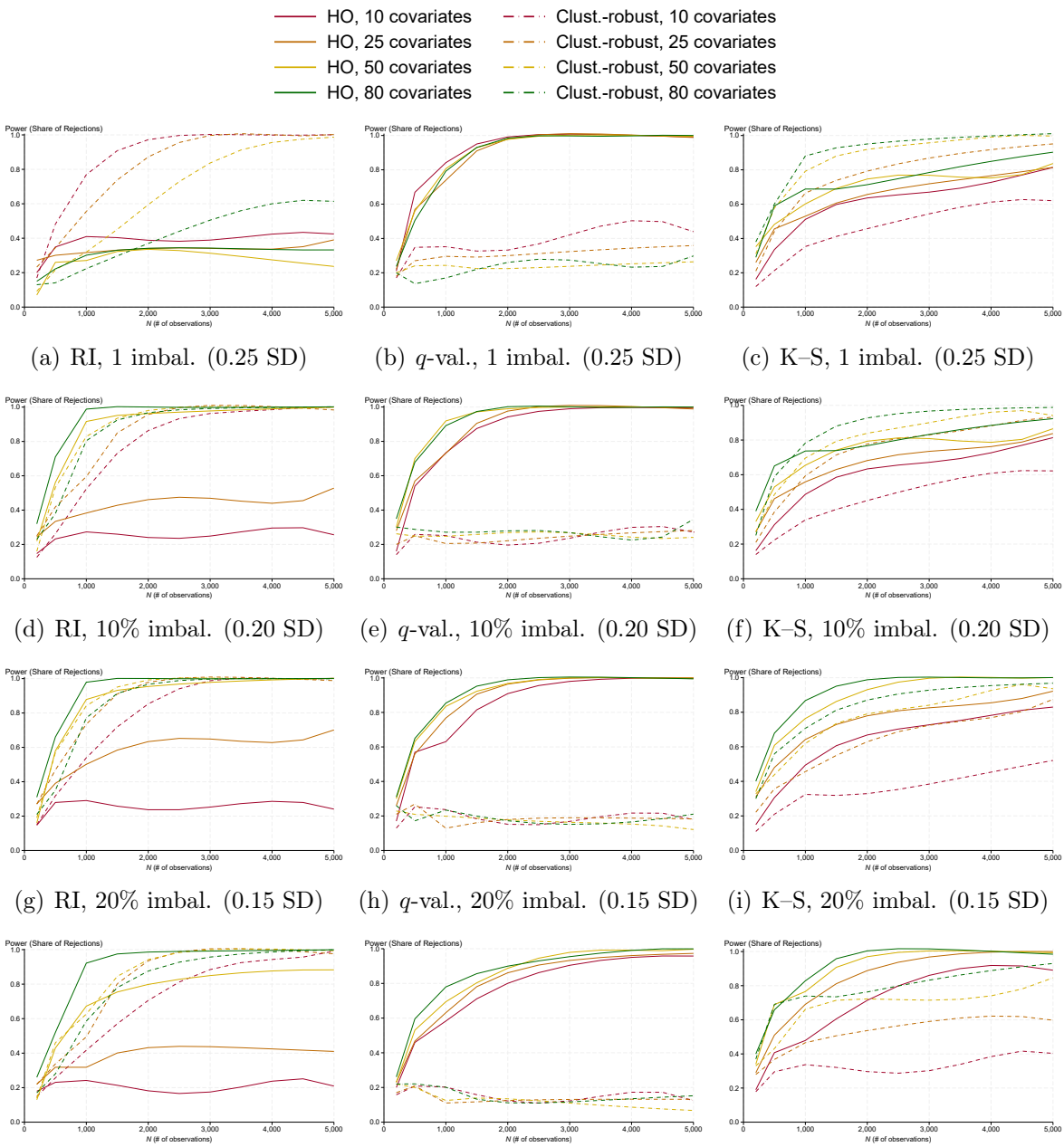
(g) RI, 20% imbal. (0.15 SD) (h) q-val., 20% imbal. (0.15 SD) (i) K–S, 20% imbal. (0.15 SD)

(j) RI, 50% imbal. (0.10 SD) (k) q-val., 50% imbal. (0.10 SD) (l) K–S, 50% imbal. (0.10 SD)

Notes: Power curves for DGP 1 with $k \in \{10, 25, 50, 100\}$ and $N \in \{200, 500, 1000, 1500, 2000, 2500, 3000, 3500, 4000, 4500, 5000\}$. For each method and combination of k and N, 100 simulated datasets are created and we compute the share of p-values or minimum q-values below 0.10; curves show fractional-polynomial fits. In Panels (a-c), one variable is made imbalanced by adding 0.25 to treated observations. In Panels (d-f), 10% of variables are made imbalanced by adding 0.2 to treated observations. In Panels (g-i), 20% of variables are made imbalanced by adding 0.15 to treated observations. In Panels (j-l) 50% of variables are imbalanced, by adding 0.1 to treated observations. RI = randomization inference. K–S = Kolmogorov-Smirnov Test.

Figure A.11

Power of Omnibus Tests of Joint Orthogonality for DGP 2 Using Randomization Inference, Kolmogorov–Smirnov Tests, and Minimum q-values



(a) RI, 1 imbal. (0.25 SD) (b) q-val., 1 imbal. (0.25 SD) (c) K–S, 1 imbal. (0.25 SD)

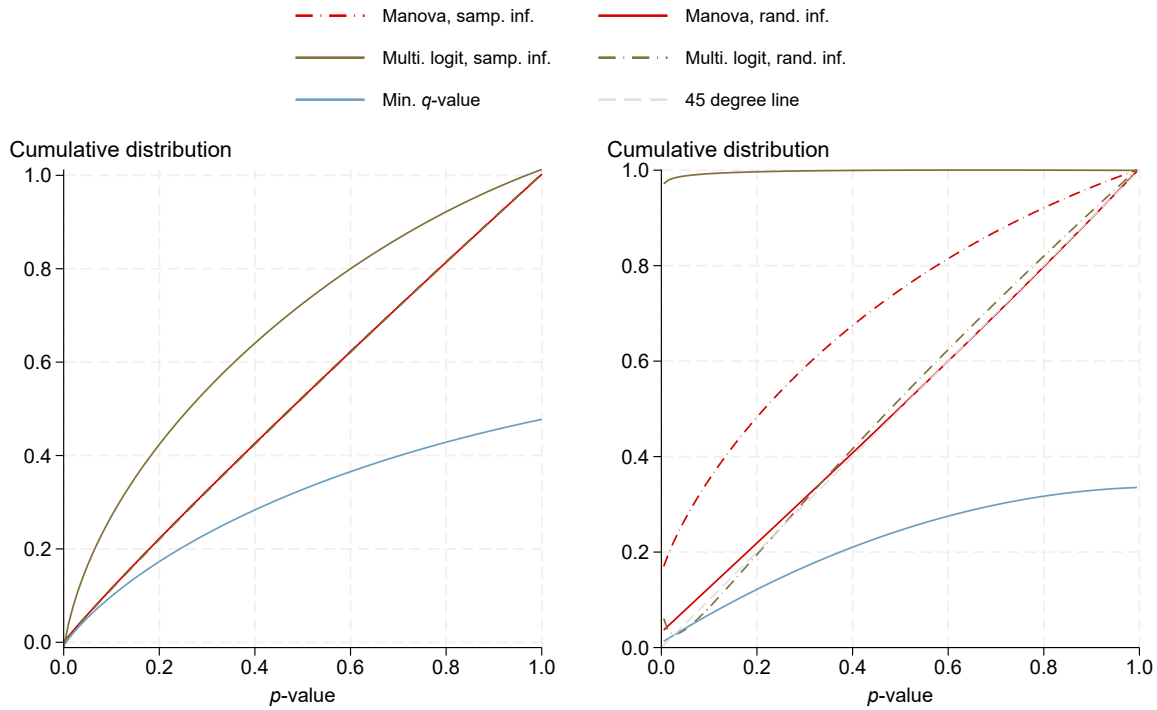
(d) RI, 10% imbal. (0.20 SD) (e) q-val., 10% imbal. (0.20 SD) (f) K–S, 10% imbal. (0.20 SD)

(g) RI, 20% imbal. (0.15 SD) (h) q-val., 20% imbal. (0.15 SD) (i) K–S, 20% imbal. (0.15 SD)

(j) RI, 50% imbal. (0.10 SD) (k) q-val., 50% imbal. (0.10 SD) (l) K–S, 50% imbal. (0.10 SD)

Note: Power curves for DGP 2 with $k \in 10, 25, 50, 80$ and $N \in \{200, 500, 1000, 1500, 2000, 2500, 3000, 3500, 4000, 4500, 5000\}$. For each value of N, 100 simulated datasets are created and we compute the share of p-values or minimum q-values below 0.10; curves show fractional-polynomial fits. In Panels (a-c), one variable is made imbalanced by adding 0.25 to treated observations. In Panels (d-f), 10% of variables are made imbalanced by adding 0.2 to treated observations. In Panels (g-i), 20% of variables are made imbalanced by adding 0.15 to treated observations. In Panels (j-l) 50% of variables are imbalanced, by adding 0.1 to treated observations. RI = randomization inference. K–S = Kolmogorov-Smirnov Test.

Figure A.12
Size of Omnibus Tests of Joint Orthogonality with Multiple Treatment Arms



(a) DGP 1: 50 independent covariates ∼ N(0,1)

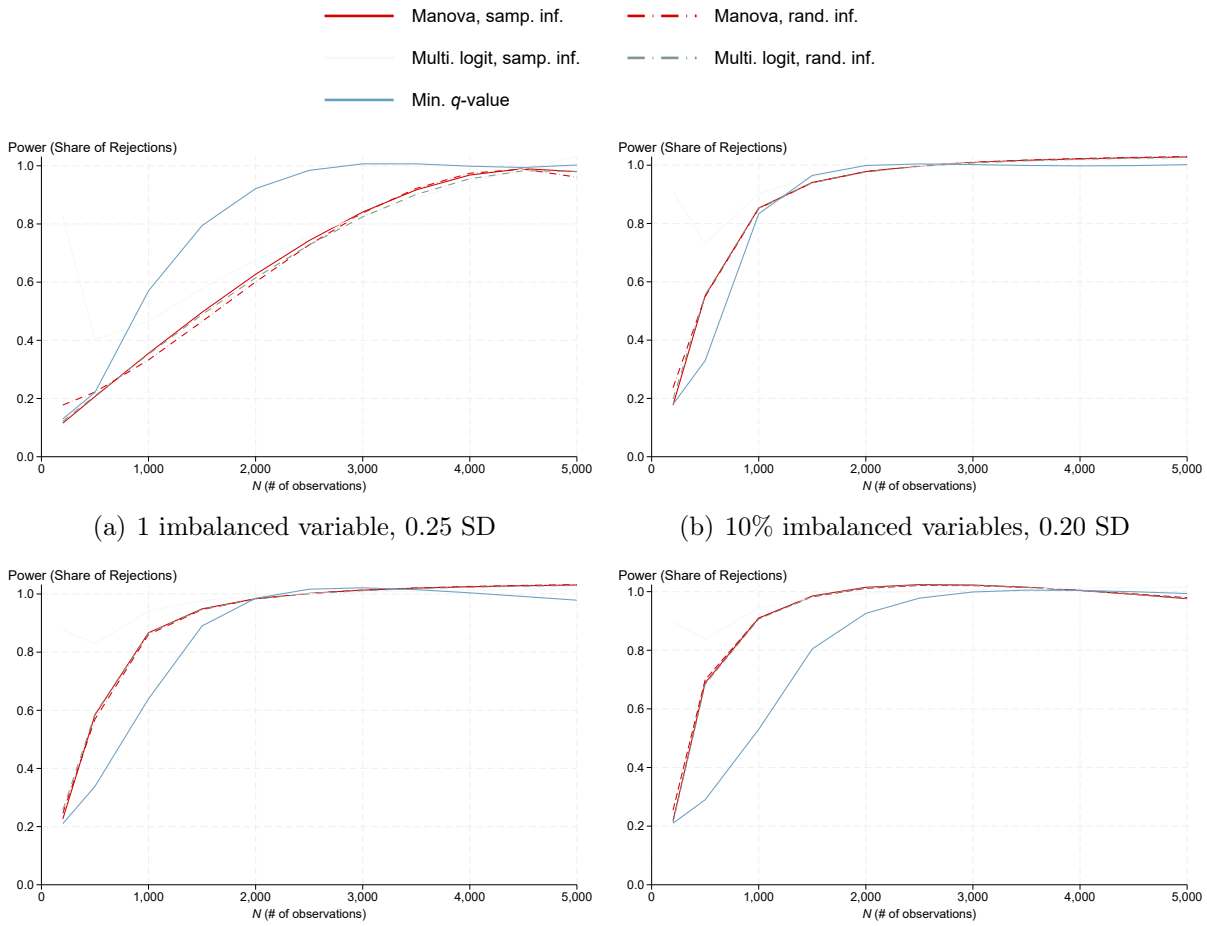
(b) DGP 2: clustered design

(c) DGP 3: data from Garbiras-Díaz and Montenegro (2022)

(d) DGP 4: data from Auriol et al. (2020)

Notes: Cumulative distributions of p-values with 500 simulations per figure. DGPs 1 and 3 are individually-randomized designs, while DGPs 2 and 4 are clustered designs. Observations are randomly assigned to four treatment arms for DGPs 1, 2, and 3, and three treatment arms for DGP 4. For each test, we estimate the proportion of p-values or minimum q-values (out of 500) that are below a threshold t, for t ranging from 0.01 to 0.99 in steps of 0.01; curves show fractional polynomial fits. Figures A.15 and A.16 show how test size vary with the number of covariates and observations for DGP 1 and DGP 2 respectively.

Figure A.13
Power of Omnibus Tests of Joint Orthogonality for DGP 1, with Multiple Treatment Arms



(a) 1 imbalanced variable, 0.25 SD

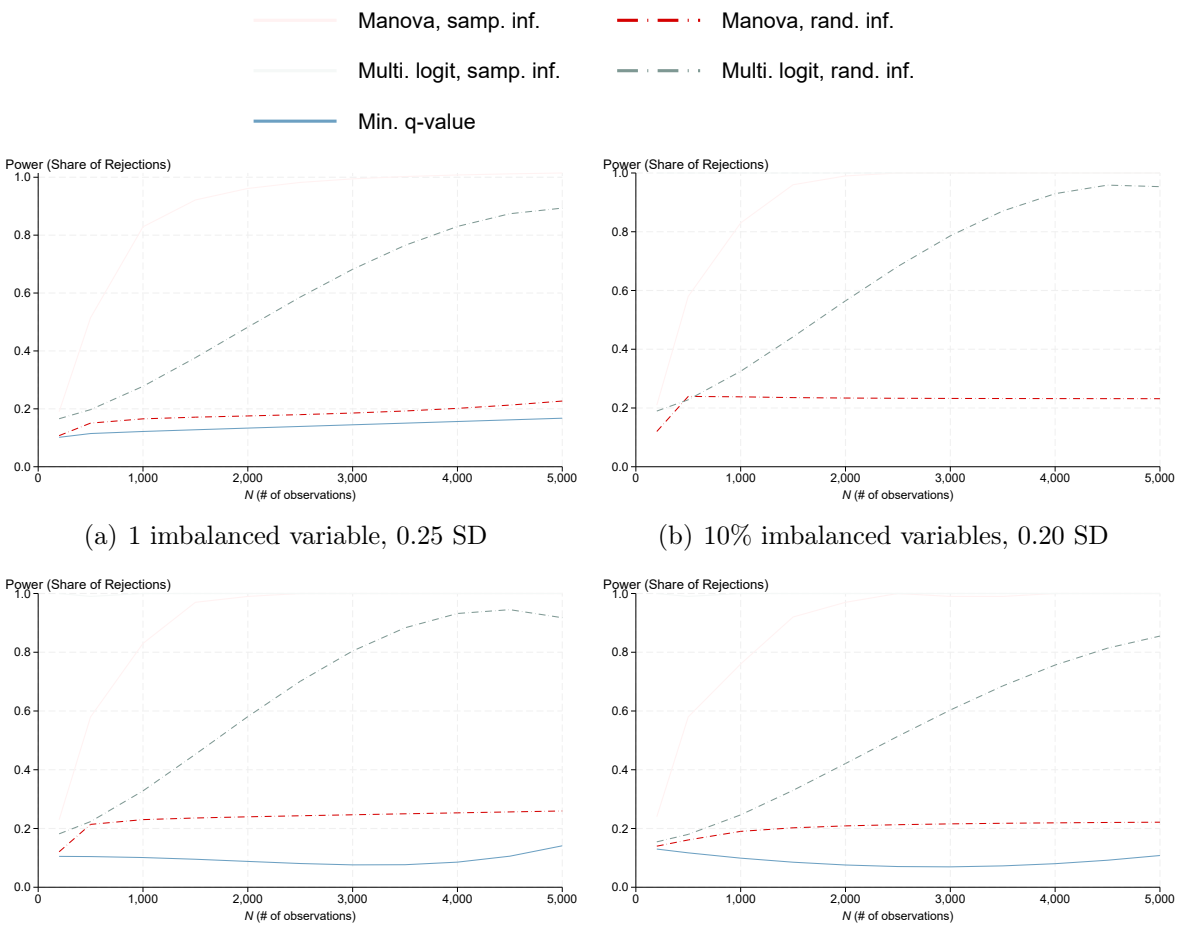
(b) 10% imbalanced variables, 0.20 SD

(c) 20% imbalanced variables, 0.15 SD

(d) 50% imbalanced variables, 0.10 SD

Notes: Power curves for DGP 1 with $k = 50$ and $N \in \{200, 500, 1000, 1500, 2000, 2500, 3000, 3500, 4000, 4500, 5000\}$. For each value of N, 100 simulated datasets are created and we compute the share of p-values or minimum q-values below 0.10; curves show fractional-polynomial fits. In Panel (a), one variable is made imbalanced by adding 0.25 to treated observations. In Panel (b), 10% of variables are made imbalanced by adding 0.2 to treated observations. In Panel (c), 20% of variables are made imbalanced by adding 0.15 to treated observations. In Panel (d), 50% of variables are imbalanced, by adding 0.1 to treated observations. Results from multinomial logit with sampling-based inference are not shown as this approach has incorrect test size (see Figure A.12).

Figure A.14
Power of Omnibus Tests of Joint Orthogonality for DGP 2, with Multiple Treatment Arms



(a) 1 imbalanced variable, 0.25 SD

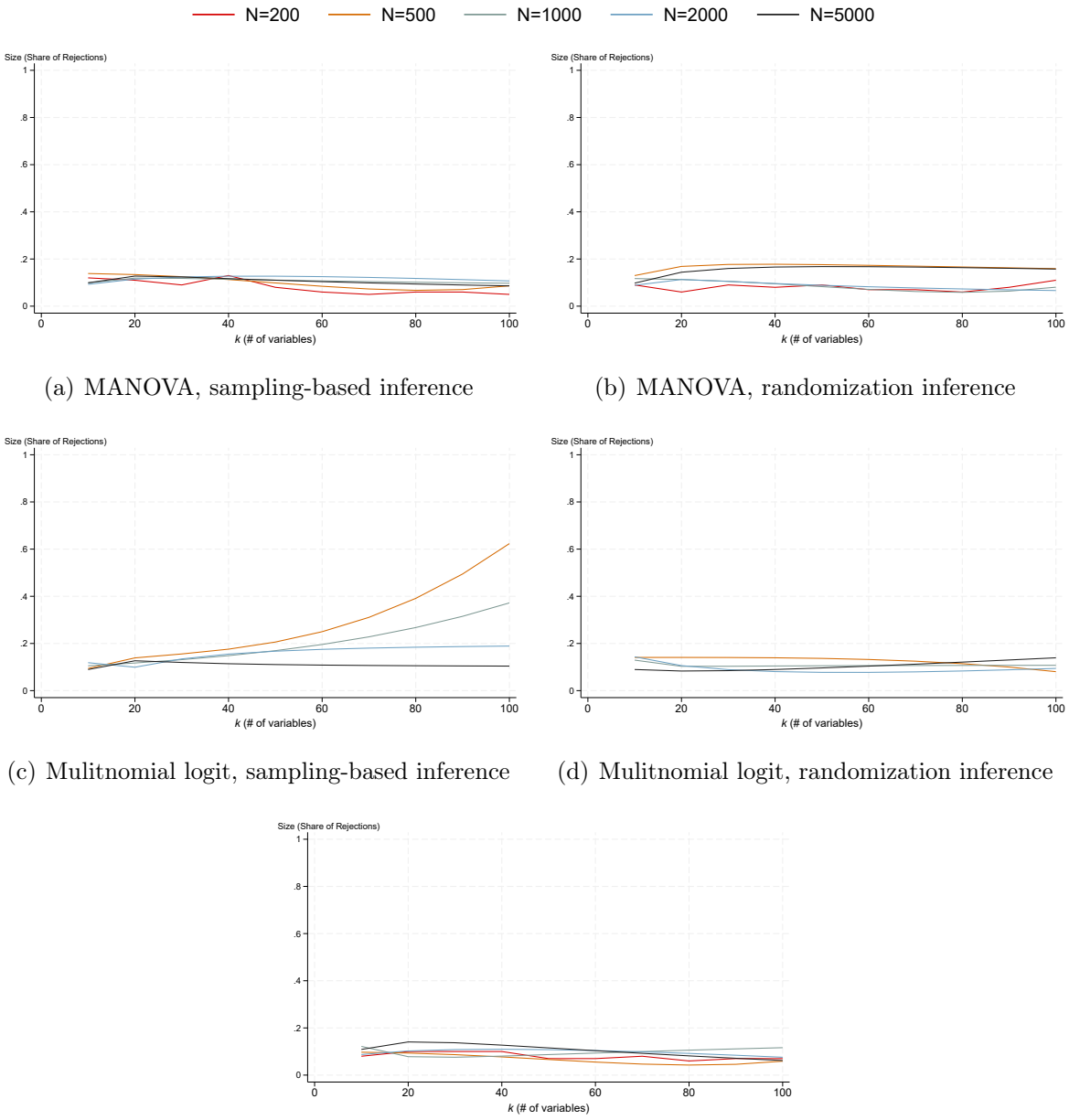
(b) 10% imbalanced variables, 0.20 SD

(c) 20% imbalanced variables, 0.15 SD

(d) 50% imbalanced variables, 0.10 SD

Notes: Power curves for DGP 2 with $k = 20$ and $N \in \{200, 500, 1000, 1500, 2000, 2500, 3000, 3500, 4000, 4500, 5000\}$. Clusters are randomly assigned to four treatment arms, with one quarter of observations in each. For each value of N, 100 simulated datasets are created and we compute the share of p-values or minimum q-values below 0.10; curves show fractional-polynomial fits. In Panel (a), one variable is made imbalanced by adding 0.25 to treated observations. In Panel (b), 10% of variables are made imbalanced by adding 0.2 to treated observations. In Panel (c), 20% of variables are made imbalanced by adding 0.15 to treated observations. In Panel (d), 50% of variables are imbalanced, by adding 0.1 to treated observations. Results from MANOVA and multinomial logit with sampling-based inference are not shown as these approaches have incorrect test size (see Figure A.12).

Figure A.15
Size of Omnibus Tests of Joint Orthogonality for DGP 1 with Multiple Treatment Arms, as a Function of Sample Size



(a) MANOVA, sampling-based inference

(b) MANOVA, randomization inference

(c) Mulitnomial logit, sampling-based inference

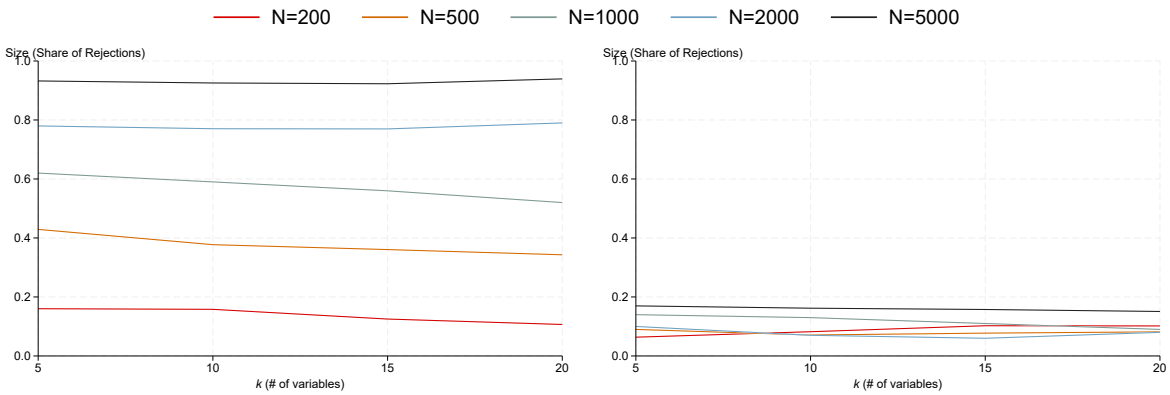
(d) Mulitnomial logit, randomization inference

(e) minimum q-value, HO variance estimator

Notes: Size curves for DGP 1 with $k \in \{10, 20, 30, 40, 50, 60, 70, 80, 90, 100\}$ and $N \in \{200, 500, 1000, 2000, 5000\}$ and treatments randomly assigned to four study arms with one quarter of the sample in each arm. For each combination of k and N, 500 simulated datasets are created and we compute the share of p-values or minimum q-values below 0.10; curves show fractional-polynomial fits. The multinomial logit did not converge for N=200.

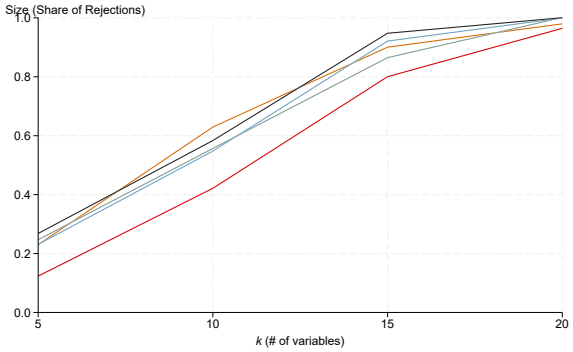
Figure A.16

Size of Omnibus Tests of Joint Orthogonality for DGP 2 with Multiple Treatment Arms, as a Function of Sample Size

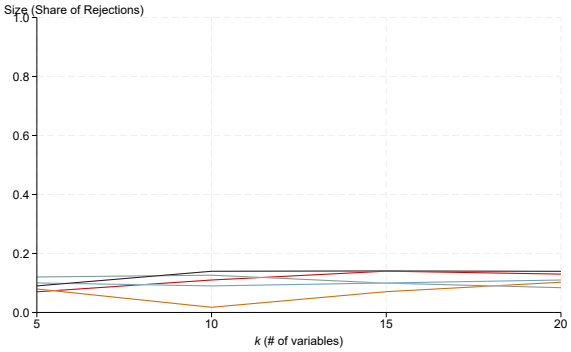


(a) MANOVA, sampling-based inference

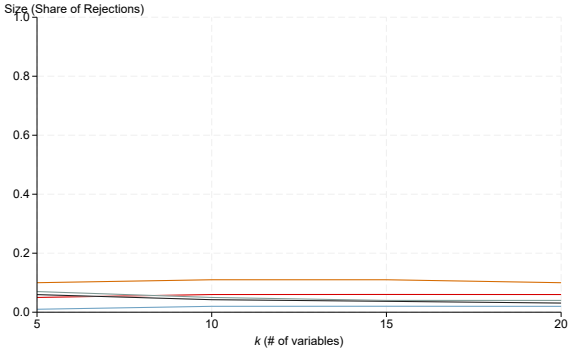
(b) MANOVA, randomization inference



(c) Mulitnomial logit, sampling-based inference and clustered SEs



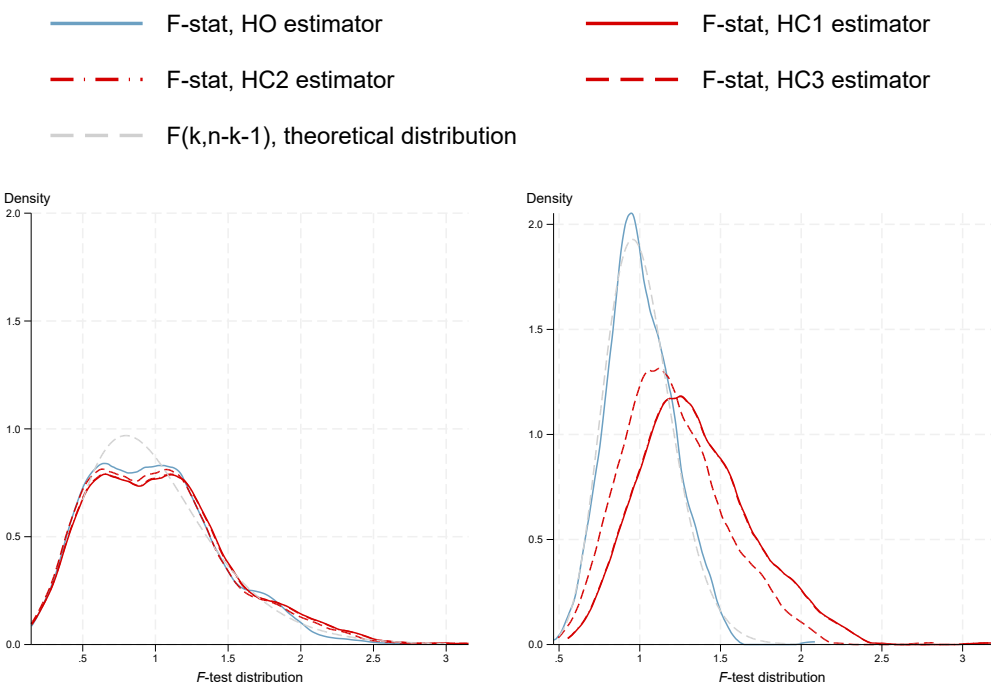
(d) Mulitnomial logit, randomization inference and clustered SEs



(e) minimum q-value, clustered SEs

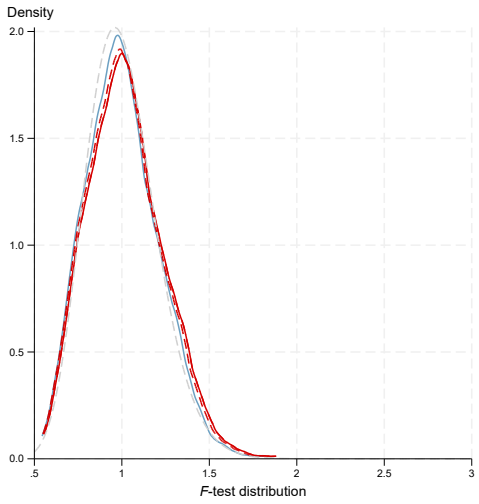
Notes: Size curves for DGP 2 with $k \in \{5, 10, 15, 20\}$ and $N \in \{200, 500, 1000, 2000, 5000\}$. Clusters are randomly assigned to four study arms, with one quarter of the clusters in each. For each combination of k and N, 500 simulated datasets are created and we compute the share of p-values or minimum q-values below 0.10; curves show fractional-polynomial fits.

Figure A.17
Empirical and Theoretical Distribution of F-statistics for DGP 1



(a) F-stat distribution, with N=500 and k=10

(b) F-stat distribution, with N=500 and k=50

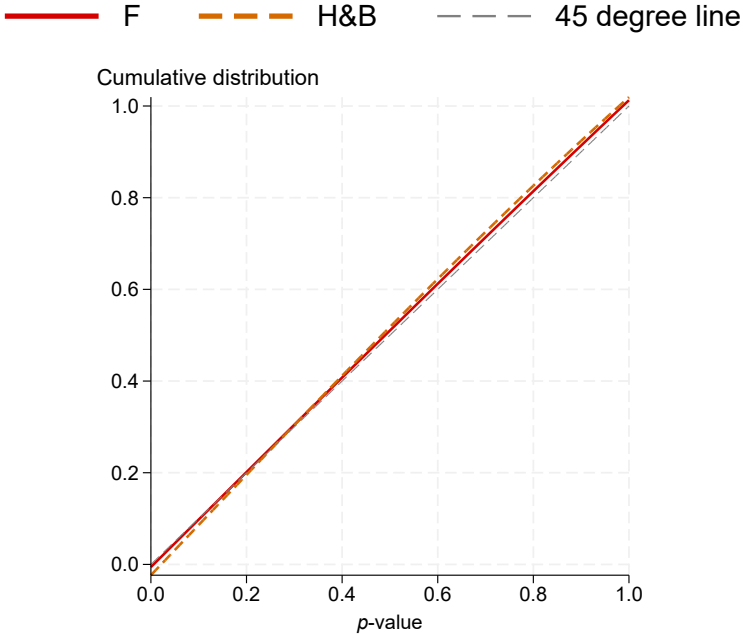


(c) F-stat distribution, with N=5000 and k=50

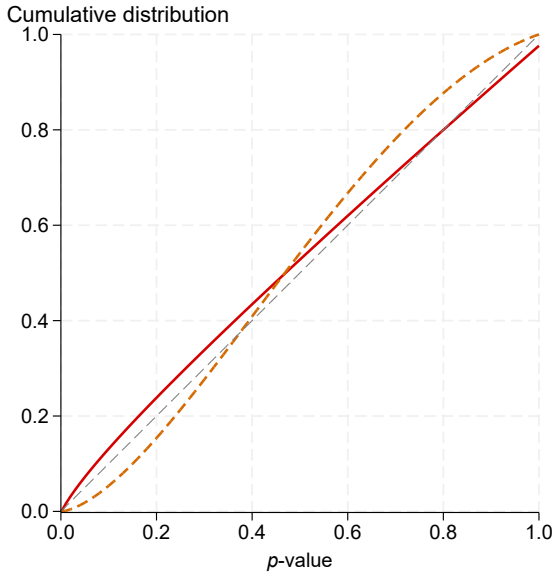
Notes: The figures show the empirical and theoretical distribution of F-statistics for DGP 1, varying the number of observations N and number of variables k.

Figure A.18

Size of omnibus F-tests and Hansen and Bowers (2008)'s test under randomization inference



(a) DGP 1: 50 independent covariates $\sim$ N(0,1)

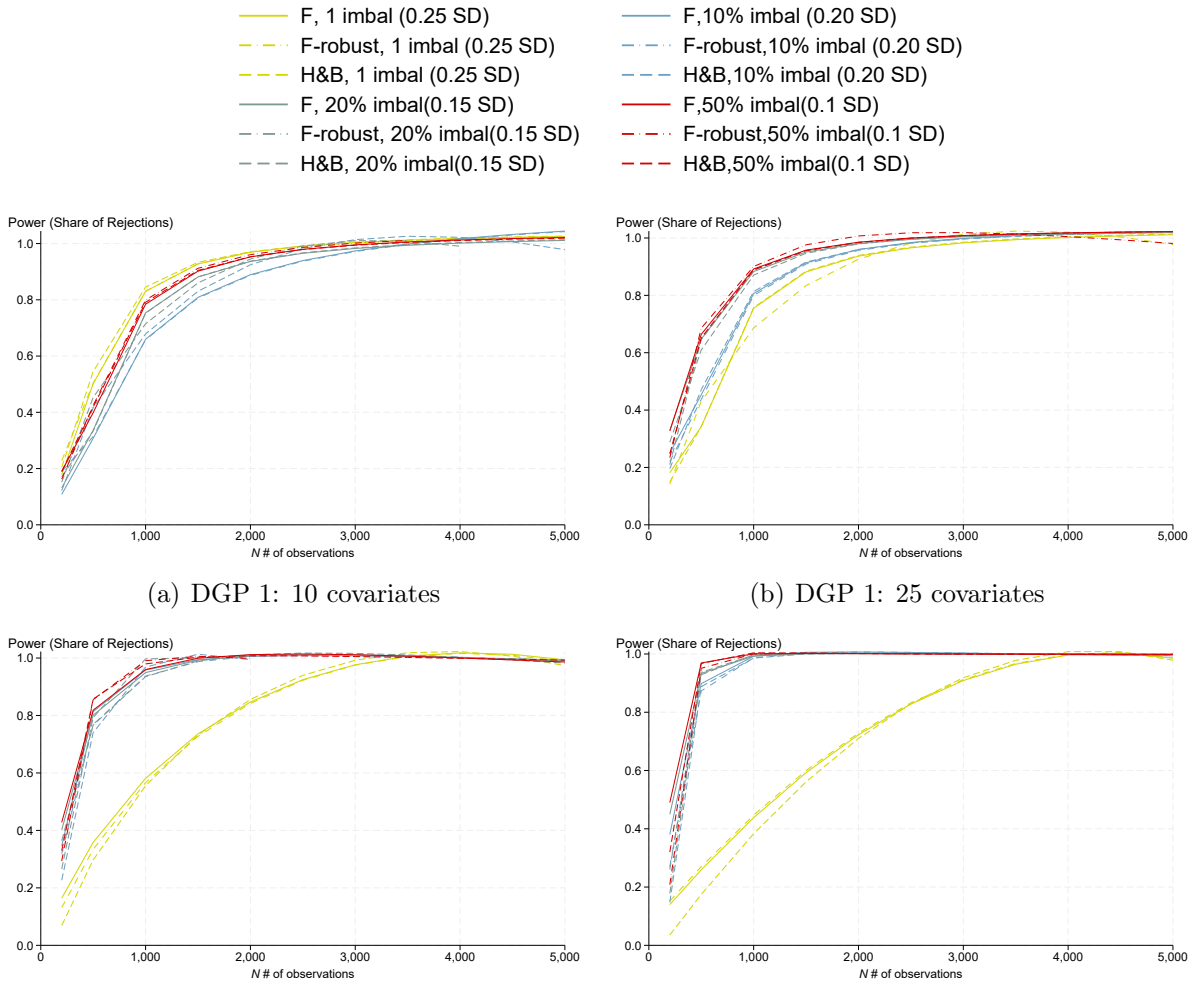


(b) DGP 2: clustered design

Notes: Notes: Cumulative distributions of p-values with 500 simulations per figure. DGP 1 is an individually-randomized design, while DGP 2 is a clustered design. For the clustered design, we specify the cluster option for the omnibus F-test with randomization inference and the test of Hansen and Bowers (2008). Curves show fractional polynomial fits.

Figure A.19

Power of omnibus F-tests and Hansen and Bowers (2008)'s test under randomization inference, for DGP 1



(a) DGP 1: 10 covariates

(b) DGP 1: 25 covariates

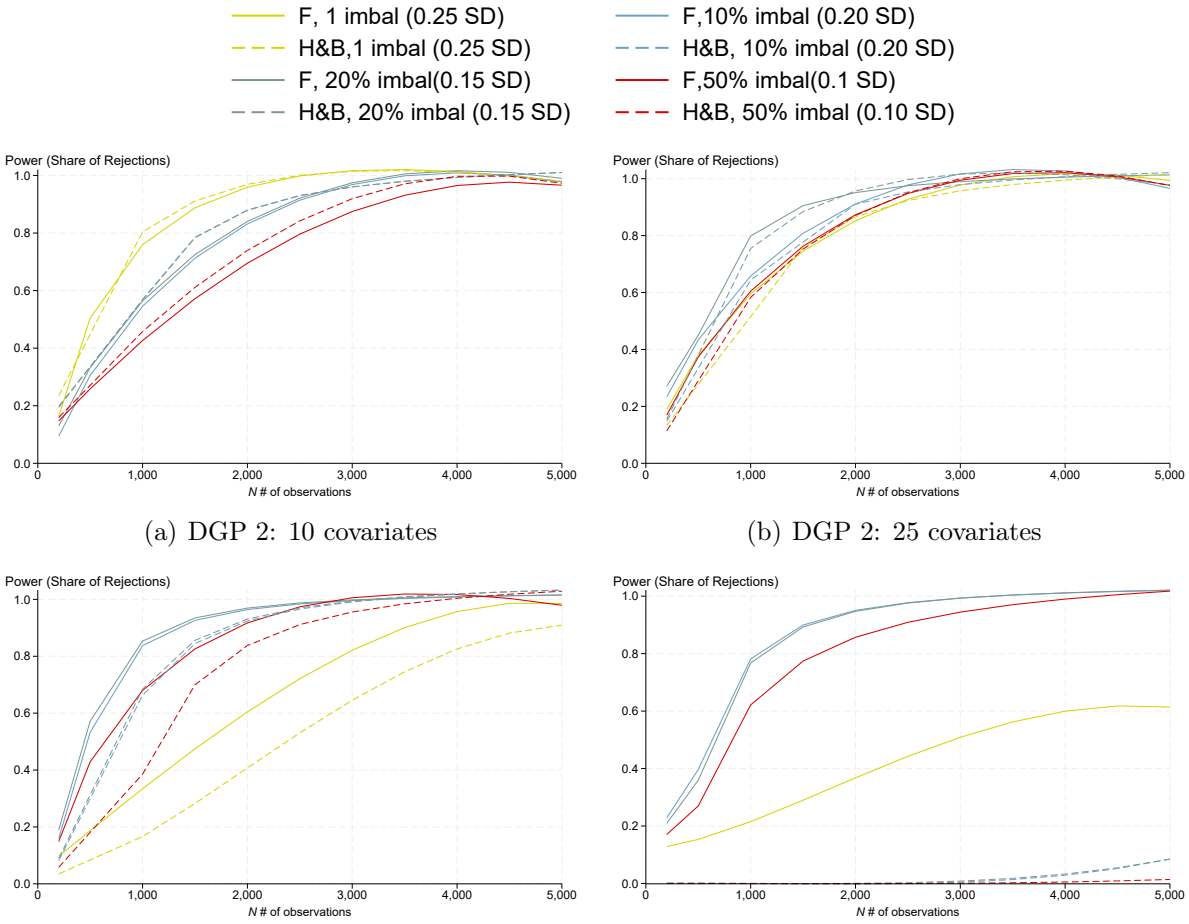
(c) DGP 1: 50 covariates

(d) DGP 1: 100 covariates

Notes: Power curves for DGP 1 with $k \in \{10, 25, 50, 100\}$ and $N \in \{200, 500, 1000, 1500, 2000, 2500, 3000, 3500, 4000, 4500, 5000\}$. For each method and combination of k and N, 100 simulated datasets. Curves show fractional-polynomial fits. Yellow lines show one variable made imbalanced by adding 0.25 to treated observations. Blue lines show 10% of variables are made imbalanced by adding 0.2 to treated observations. Grey lines show 20% of variables are made imbalanced by adding 0.15 to treated observations. Red lines show 50% of variables are imbalanced, by adding 0.1 to treated observations.

Figure A.20

Power of omnibus F-tests and Hansen and Bowers (2008)'s test under randomization inference, for DGP 2



(a) DGP 2: 10 covariates

(b) DGP 2: 25 covariates

(c) DGP 2: 50 covariates

(d) DGP 2: 80 covariates

Notes: Power curves for DGP 2 with $k \in \{10, 25, 50, 80\}$ and $N \in \{200, 500, 1000, 1500, 2000, 2500, 3000, 3500, 4000, 4500, 5000\}$. For each value of N, 100 simulated datasets are created and we compute the share of p-values ; curves show fractional-polynomial fits. Yellow lines show one variable made imbalanced by adding 0.25 to treated observations. Blue lines show 10% of variables made imbalanced by adding 0.2 to treated observations. Grey lines show 20% of variables made imbalanced by adding 0.15 to treated observations. Red lines show 50% of variables made imbalanced, by adding 0.1 to treated observations.

Table A.1
Literature Review of RCTs (2021–2023)

	General Info.					Pairwise/Groupwise test				Joint Orthogonality F-test			
Reference	Bal.	N	k	T+C	Design	t	F	Nor. Diff.	Inf.	J.F.	Model	Inf.	VE
De Janvry et al. (2023)	Yes	2,854	11	2	Clu	Yes	No	No	SBI	Yes	OLS	SBI	CRVE
Ainsworth et al. (2023)	Yes	2,629	12	2	Clu	Yes	No	No	SBI	Yes	OLS	SBI	CRVE
Guryan et al. (2023)	Yes	Exp. 1: 2,633	19	2	Ind	Yes	No	No	SBI	Yes	OLS	SBI	CRVE
		Exp. 2: 2,710									-	-	-
Macchi (2023)	Yes	Exp 1: 511	25	2	Ind	Yes	No	No	SBI&RI	No	-	-	-
		Exp. 2: 238									-	-	-
Buchmann et al. (2023)	Yes	26,408	18	3	Clu	Yes	No	No	SBI	Yes	OLS	SBI	CRVE
Brock and De Haas (2023)	Yes	2,054	6	3	Ind	No	No	No	SBI	Yes	OLS	SBI	HO
Hardy and McCasland (2023)	Yes	755	20	2	Ind	Yes	No	No	SBI	Yes	OLS	SBI	HO
Baseler (2023)	Yes	497	20	2	Ind	Yes	No	No	SBI	No	-	-	-
Zárate (2023)	Yes	6,147	23	3	Clu	Yes	No	No	SBI	Yes	ML	SBI	CRVE
Oh (2023)	Yes	630	14	2	Ind	Yes	No	No	SBI	No	-	-	-
Chakravorty, Dar, and Emerick (2023)	Yes	400	19	2	Clu	No	Yes	No	SBI	No	-	-	-
Afrouzi et al. (2023)	No										-	-	-
Alan, Corekcioglu, and Sutter (2023)	Yes	1,988	19	2	Clu	Yes	No	No	SBI	No	-	-	-
Cullen, Dobbie, and Hoffman (2023)	Yes	1,095	17	6	Ind	No	Yes	No	SBI	No	-	-	-
Gray-Lobe, Pathak, and Walters (2023)	Yes	4,125	14	2	Ind	Yes	No	No	SBI	No	-	-	-
Banerjee et al. (2023*b*)	Yes	7,511	42	2	Clu	Yes	No	No	SBI	No	-	-	-
Battaglia, Gulesci, and Madestam (2023)	Yes	2,717	31	2	Clu	Yes	No	Yes	SBI&RI	No	-	-	-
Alesina, Miano, and Stantcheva (2023)	Yes	22,506	8	3	Ind	No	No	No	SBI	No	-	-	-
Jack et al. (2023)	Yes	1,840	26	6	Clu	No	Yes	No	SBI	No	-	-	-
Christensen and Timmins (2023)	Yes	18,045	3	2	Clu	No	Yes	No	SBI	No	-	-	-
Beaman et al. (2023)	Yes	6,807	9	3	Clu	Yes	No	No	SBI	Yes	OLS	SBI	CRVE
Tungodden and Willén (2023)	No	-	-	-	-	-	-	-	-	-	-	-	-

Notes: Bal. refers to whether any balance tests were conducted in the paper. N is the number of observations and k is the number of covariates used to test balance. T+C is the total number of study arms in the paper. Design is the randomization method used: Clu means cluster-randomization while Ind means individual randomization. The t column indicates that the paper shows pairwise t-tests, the F column indicates that it shows groupwise F-tests, and the Nor. Diff. column indicates that it shows normalized differences. Inf. is the inference method used: SBI is sampling-based inference, RI is randomization inference, and ? means the inference approach is unspecified. J.F. indicates whether the paper shows omnibus F-tests of joint orthogonality. In the model column, OLS means ordinary least-squares regression and ML means multinomial logit. VE shows the variance estimator used to estimate standard errors: HO is the conventional homoskedastic SE estimator; HC1 is Eicker-Huber-White heteroskedasticity-robust SEs, and CRVE is the Cluster-Robust Variance Estimator.

	General Info.					Pairwise/Groupwise test				Joint Orthogonality F-test			
Reference	Bal.	N	k	T+C	Design	t	F	Nor. Diff.	Inf.	J.F.	Model	Inf.	VE
Muralidharan, Niehaus, and Sukhtankar (2023)	No	-	-	-	-		-	-	-	-	-	-	-
Adhvaryu, Kala, and Nyshadham (2023)	Yes	1,866	9	4	Clu	Yes	No	No	SBI	No	-	-	-
Cortés et al. (2023)	No	-	-	-	-	-	-	-	-	-	-	-	-
Banerjee et al. (2023*a*)	Yes	1,082	9	4	Clu	Yes	No	No	SBI	No	-	-	-
Bellemare et al. (2023)	Yes	2,021	15	4	Ind	Yes	No	No	SBI	No	-	-	-
Dhar, Jain, and Jayachandran (2022)	Yes	14,809	15	2	Clu	Yes	No	Yes	SBI	Yes	OLS	SBI	CRVE
Garbiras-Díaz and Montenegro (2022)	Yes	698	33	7	Ind	Yes	No	No	SBI&RI	No	-	-	-
Aydin (2022)	Yes	45,307	5	2	Ind	No	No	No	SBI	Yes	OLS	SBI	CRVE
Hussam et al. (2022*b*)	Yes	754	27	3	Ind	Yes	No	No	?	Yes	OLS	SBI	CRVE
Casaburi and Reed (2022)	Yes	1,079	12	2	Clu	Yes	No	No	SBI	No	-	-	-
Wheeler et al. (2022)	Yes	1,638	11	2	Clu	No	Yes	Yes	SBI	No	-	-	-
Chen, Persson, and Polyakova (2022)	Yes	743	17	2	Ind	Yes	No	No	SBI	No	-	-	-
Hussam et al. (2022*a*)	Yes	2,887	32	8	Clu	No	Yes	No	SBI	No	-	-	-
Lopez, Sautmann, and Schaner (2022)	Yes	2,055	27	3	Clu	No	Yes	No	SBI	No	-	-	-
Stephens Jr and Toohey (2022)	Yes	12,562	11	2	Ind	Yes	No	No	?	No	-	-	-
Angrist, Autor, and Pallais (2022)	Yes	8,190	15	2	Ind	Yes	Yes	No	SBI	No	-	-	-
Meghir et al. (2022)	No	-	-	-	-	-	-	-	-	-	-	-	-
Allcott et al. (2022)	Yes	1,177	14	2	Ind	Yes	Yes	No	SBI	Yes	OLS	SBI	HO
Carrera et al. (2022)	Yes	1,248	8	2	Ind	Yes	No	No	SBI	No	-	-	-
Byrne, Martin, and Nah (2022)	No	-	-	-	-	-	-	-	-	-	-	-	-
Cai and Wang (2022)	Yes	1,251	15	2	Clu	Yes	No	No	SBI	No	-	-	-
Arteaga et al. (2022)	Yes	2,050	5	3	Ind	No	Yes	No	SBI	No	-	-	-
Carlana, La Ferrara, and Pinotti (2022)	Yes	1,217	7	2	Clu	Yes	No	Yes	SBI	No	-	-	-
Fehr, Fink, and Jack (2022)	Yes	5,842	10	2	Ind	Yes	Yes	No	SBI	No	-	-	-
Cullen and Perez-Truglia (2022)	Yes	2,060	7	4	Clu	Yes	No	No	SBI	No	-	-	-

Notes: Bal. refers to whether any balance tests were conducted in the paper. N is the number of observations and k is the number of covariates used to test balance. T+C is the total number of study arms in the paper. Design is the randomization method used: Clu means cluster-randomization while Ind means individual randomization. The t column indicates that the paper shows pairwise t-tests, the F column indicates that it shows groupwise F-tests, and the Nor. Diff. column indicates that it shows normalized differences. Inf. is the inference method used: SBI is sampling-based inference, RI is randomization inference, and ? means the inference approach is unspecified. J.F. indicates whether the paper shows omnibus F-tests of joint orthogonality. In the model column, OLS means ordinary least-squares regression and ML means multinomial logit. VE shows the variance estimator used to estimate standard errors: HO is the conventional homoskedastic SE estimator; HC1 is Eicker-Huber-White heteroskedasticity-robust SEs, and CRVE is the Cluster-Robust Variance Estimator.

	General Info.					Pairwise/Groupwise test				Joint Orthogonality F-test			
Reference	Bal.	N	k	T+C	Design	t	F	Nor. Diff.	Inf.	J.F.	Model	Inf.	VE
Levy (2021)	Yes	37,494	21	3	Ind	No	No	No	SBI	Yes	OLS	SBI	-
Lowe (2021)	Yes	800	11	2	Clu	Yes	No	No	SBI	No	-	-	-
Brune, Chyn, and Kerwin (2021)	Yes	870	15	2	Ind	Yes	No	No	SBI	Yes	OLS	SBI	HC1
Leaver et al. (2021)	Yes	242	4	2	Ind	Yes	No	Yes	RI	No	-	-	-
McKenzie and Puerto (2021)	Yes	3,537	14	3	Clu	Yes	No	No	SBI	Yes	OLS	SBI	CRVE
Mohanan et al. (2021)	Yes	135	8	3	Clu	No	Yes	No	SBI	No	-	-	-
Muralidharan et al. (2021)	Yes	548	33	2	Clu	Yes	Yes	No	SBI	Yes	OLS	SBI	CRVE
Angrist, Caldwell, and Hall (2021)	Yes	1,031	11	3	Ind	Yes	No	No	SBI	Yes	ML	SBI	CRVE
Carter, Laajaj, and Yang (2021)	Yes	514	4	2	Ind	No	Yes	No	SBI	No	-	-	-
Dal Bó et al. (2021)	Yes	176	8	2	Clu	Yes	No	No	SBI	No	-	-	-
Bessone et al. (2021)	Yes	452	19	6	Ind	Yes	No	No	SBI	Yes	OLS	SBI	HO
Dahl, Kotsadam, and Rooth (2021)	Yes	781	9	2	Clu	No	No	No	?	Yes	OLS	SBI	HO
Dube et al. (2021)	Yes	504	17	3	Clu	Yes	No	No	?	Yes	ML	SBI	CRVE
Abebe et al. (2021)	Yes	3,049	33	3	Clu	Yes	Yes	No	SBI	Yes	OLS	SBI	-
Angelucci and Bennett (2021)	Yes	303	24	2	Clu	Yes	No	No	SBI	Yes	OLS	SBI	HO
Fowlie et al. (2021)	Yes	71,017	5	5	Clu	Yes	No	No	SBI	No	-	-	-
Doerr and Necker (2021)	Yes	2,543	1	7	Clu	No	Yes	No	SBI	No	-	-	-
Bandiera et al. (2021)	Yes	587	24	4	Clu	Yes	Yes	No	SBI&RI	No	-	-	-
Alan et al. (2021)	Yes	7,487	36	2	Clu	Yes	No	No	SBI	No	-	-	-
Bergman (2021)	Yes	306	12	2	Ind	Yes	No	No	SBI	No	-	-	-
Beaman et al. (2021)	Yes	14,300	12	4	Clu	No	Yes	No	SBI	No	-	-	-

Notes: Bal. refers to whether any balance tests were conducted in the paper. N is the number of observations and k is the number of covariates used to test balance. T+C is the total number of study arms in the paper. Design is the randomization method used: Clu means cluster-randomization while Ind means individual randomization. The t column indicates that the paper shows pairwise t-tests, the F column indicates that it shows groupwise F-tests, and the Nor. Diff. column indicates that it shows normalized differences. Inf. is the inference method used: SBI is sampling-based inference, RI is randomization inference, and ? means the inference approach is unspecified. J.F. indicates whether the paper shows omnibus F-tests of joint orthogonality. In the model column, OLS means ordinary least-squares regression and ML means multinomial logit. VE shows the variance estimator used to estimate standard errors: HO is the conventional homoskedastic SE estimator; HC1 is Eicker-Huber-White heteroskedasticity-robust SEs, and CRVE is the Cluster-Robust Variance Estimator.