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

Jason T. Kerwin
Nada Rostom
Olivier Sterck*

March 2025

Abstract

Economists often use balance tests to demonstrate that the treatment and control groups are comparable prior to an intervention. We show that typical implementations of balance tests have poor statistical properties. Pairwise t-tests leave it unclear how many rejections indicate overall imbalance. Omnibus tests of joint orthogonality, in which the treatment is regressed on all the baseline covariates, address this ambiguity but substantially over-reject the null hypothesis using the sampling-based p-values that are typical in the literature. This problem is exacerbated when the number of covariates is high compared to the number of observations. We examine the performance of alternative tests, and show that omnibus F-tests of joint orthogonality with randomization inference p-values have the correct size and reasonable power. We apply these tests to data from two prominent recent articles where standard F-tests indicate imbalance, and show that the study arms are actually balanced when appropriate tests are used. We provide code for implementing appropriate balance tests for RCTs.

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, Dean Eckles, Alan Griffith, Ethan Ligon, David McKenzie, and seminar participants at CSSS at the University of Washington and the CEGA Research Retreat at UC Berkeley 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 often used to demonstrate that treatment and control groups are comparable before treatment. Based on a review of all papers published in top five journals in economics between 2021 and 2023, we identified 69 randomized controlled trials (RCTs), out of which 62 (90%) discuss the results of balance tests (Table A.1 in Appendix).

Balance tests play a crucial role in establishing the comparability of the treatment and control groups in randomized controlled trials (RCTs) prior to any intervention. This is particularly 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. Kerwin and Thornton 2021, Gazeaud, Mvukiyehe, and Sterck 2023). Balance tests are especially relevant when administrators may have incentives 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) 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 (Siu, Sterck, and Rodgers 2023).

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 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

[1] See Mutz, Pemantle, and Pham (2019) for arguments against this practice.

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. The majority of papers (81%) report an omnibus F-test alongside pairwise t-tests.

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 the incorrect size: they substantially over-reject the null hypothesis. This problem is worst when many baseline variables are included in the test and when heteroskedasticity-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 distributed, robust omnibus F-tests of joint orthogonality reject the null hypotheses at $\alpha = 0.10$ approximately

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

50% of the time, instead of the expected 10%. The cause of this problem appears to be that, when there are many covariates, the asymptotic distribution of the F-statistic is actually normal rather than chi-squared (Anatolyev 2012).

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, as the uncertainty comes from the randomization process and not from sampling variation (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 value of our research, 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

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

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 assess both the statistical size and power of the tests. We also build on work that uses randomization inference for testing overall balance. Hansen and Bowers (2008) develop an overall balance test and show that it performs well when p-values are constructed using randomization inference. Their test is uncommon in economics. We show that the sampling inference-based F-tests that most economics papers actually use over-reject the null, but have the correct size and high power when randomization inference is used instead.

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 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 is suggested by McKenzie (2015). But there is no existing 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. 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. Our suggested alternative approach, which uses randomization inference instead, can help head off this issue.

Finally, this paper complements an extensive literature in econometrics on inference in models where there are many covariates. The most closely related previous paper on this topic is Anatolyev and Sølvsten (2023), which shows that tests of many restrictions (like the F-tests that we analyze) will over-reject the null, often quite badly, and propose a new test that performs better. We build on this work by showing that this problem is common in balance tests in economics, and that it can also 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 (DGP) to assess the size and power of different omnibus tests of joint orthogonality. We conduct all our simulations using Stata. Stata code for our simulations is available in the replication package for our

paper.

2.1 Test Size

In statistics, the size of a test refers to the probability of erroneously rejecting the null hypothesis – in other words, the likelihood of committing a Type I error. Scholars distinguish the nominal size of a test, which is the threshold set by the researcher for the maximum allowable probability of a Type I error, from its empirical size, which is the actual observed rate of Type I errors when the test is applied to a large number of datasets. Statisticians typically try to construct tests for which the empirical size is equal to the nominal size. If the empirical size of a test is less than its nominal size, the test is said to be more conservative, meaning the actual rate of Type I errors is lower than the pre-specified level. This is not problematic but may indicate that a more-powerful test is possible (Fisher and Robbins 2019). However, a statistical test should be avoided if its empirical size of a test is greater than its nominal size, as this means there is an increased likelihood of Type I error, i.e. false positives.

In balance tests, the null hypothesis is typically that the study arms have equal means. If two groups differ only because of random chance,[4] then a balance test will have correct size if its p-values are uniformly distributed, rejecting the null hypothesis at the x percent level in x percent of realizations. The test is conservative if p-values are skewed to the left, leading to a reduced risk of Type I error. By contrast, the test size is problematic if p-values are skewed to the right, leading to an over-rejection of the null hypothesis and increased likelihood of Type I error.

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 $\sim N(0, 1)$,[5] and an independent treatment that is randomly

[4]For example, due to sampling variation or assignment variation, which implies that the means are equal in expectation but different for any particular realizations of the sampling or randomization process.

[5]As a robustness check, we also considered a DGP with variables that are distributed following uniform,

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 100 clusters of equal size, the k variables are correlated within clusters (average intra-class correlation=0.2), and treatment is randomly assigned to half of the clusters (DGP 2). 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 variables k from 10 to 100 and the number of observations N from 200 to 5000.

The two other DGPs use original datasets from existing randomized controlled trials 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. First, we wanted to select 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. Then, we selected the individually randomized trial of 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. We selected the cluster randomized trial of 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 Garbiras-Díaz and Montenegro (2022), the authors study whether crowdsourcing the monitoring of elections is effective in combating electoral fraud. The intervention leverages a Facebook advertisement campaign aimed at encouraging citizens to report electoral irregularities during the 2019 mayoral elections in Colombia. A total of 698 municipalities (more than half the municipalities in the country) were randomized to one of three interventions or

chi-squared, or binary distributions. Results suggest the distribution of variables has only minor impact on test size.

a placebo control group, resulting in four main study arms. Citizens in the placebo control group received a basic message reminding them about the date of the coming elections. For citizens in the *"information"* study arm, the message also advertised the website where irregularities can be reported, together with a web link. Citizens in the *"call-to-action"* study arm received a message with a call to action to report irregularities and act against them. Citizens in the *"information + call-to-action"* study arm received both the message containing the link to report irregularities as well as the call to action.[6] In our simulations, we randomly re-assign the 698 municipalities into a "placebo" treatment arm or a control group and use balance tests with 33 baseline covariates to compare observations in both groups (DGP 3).

Auriol et al. (2020) examine the hypothesis that insurance can be a motive for religious donations, and that believers give money to churches hoping to receive insurance against future economic shocks. The authors conduct a two-stage lab-in-the-field experiment where they randomly assign participants to a formal and commercially available funeral insurance policy. Then, using a dictator game, researchers measure participants' willingness to contribute money to the church and two other charitable recipients.

Participants were recruited from different branches of a well-established Pentecostal church and grouped into sessions. Randomization was done at the session level. At the start of each session, one participant per group was invited to pick one out of three unmarked envelopes to assign the group to the "Insurance", "Insurance Information" or "No Insurance" group. The group of participants who received the paper with the insurance label received a funeral insurance policy offered by a leading micro insurer active in the Ghanaian market. Groups assigned to the "Insurance Information" received only information about the insurance policy. All groups were then invited to play the dictator game. Participants who received the "No Insurance" envelopes were not assigned to the policy and did not have any sort of discussion about insurance.

[6]In order to understand how candidates responded to the reporting campaign, the researchers additionally cross-randomized whether a candidate for mayor (or any of their staff) received a letter informing them about the reporting campaign. For simplicity and brevity, we do not consider this additional level of randomization here.

In our simulations, we randomly re-assign the clusters into three groups and use balance tests with 10 covariates to compare observations in two of the groups (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. In contrast, balance tests over-reject the null hypothesis if the distribution of p-values is skewed to the right and the cumulative distribution of p-values is above the 45 degree line for p-values below critical thresholds, and under-reject the null hypothesis if the opposite holds. When interpreting the results of simulations, we will often consider DGPs 1 and 3 together, as they are both individually randomized designs, and DGPs 3 and 4 together, as they are both clustered designs.

2.2 Statistical Power

The power of a test is the probability that the test correctly rejects the null hypothesis when a specific alternative hypothesis is true. The power of a test is equal to $1 - \beta$ where β is the probability of committing a type II error by wrongly failing to reject the null hypothesis.

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 higher 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.

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%).[7]

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. t-test difference 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

[7]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.

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.

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 ($\sqrt{\sigma_C^2 + \sigma_T^2}$). 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 be meaningful bias in estimated treatment effects as well. 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 creates additional "researcher degrees of freedom" (Simmons, Nelson, and Simonsohn 2011) to present balance tables as showing or not showing a problem, depending on the authors' preferences and audience pressures. Researchers may want to argue that there are no balance problems in order to make it easier to publish their papers. 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.[8] 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 percentage is 34% with a variance estimator assuming homoskedasiticty). 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 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 groupwise tests by considering baseline covariates jointly in a unique test yielding a unique statistic. Omnibus tests of joint orthogonality typically 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. Among the papers we reviewed, 32% report the results of one or more omnibus tests of joint orthogonality.

Researchers face different options when implementing omnibus tests of joint orthogonality 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

[8] 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.

randomization inference in Section 4.1). As the dependent variable in omnibus tests of joint orthogonality is either binary or categorical, heteroskedasticity is an issue when OLS is used to estimate a linear probability model (LPM). Statistical programs usually provide different heteroskedasticity-consistent (HC) variance estimators, including the HC1, HC2, and HC3 estimators (MacKinnon and White 1985).[9] 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.[10] Our key finding is that all versions of the omnibus 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 (incorrectly) 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.

[9](Angrist and Pischke 2008, Ch. 8) point out that these estimators can have poor finite-sample performance, and also that they themselves have a sampling distribution and can be smaller than homoskedastic standard errors. Moreover, HC1 standard errors are smaller than HO standard errors in expectation if the true error term is homoskedastic.

[10]The results using a probit model are very similar to those from the logit and hence are only reported in the Appendix (Figures A.4 and A.5).

3.2.1 Why do omnibus F-tests over-reject the null?

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 many 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. And 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.

Moreover, in our setting, the error terms are not actually heteroskedastic. Heteroskedastic errors are a general feature of LPMs, because extreme values of X run up against the boundary of the support of Y, approaching 1 or 0 respectively. But in our simulations the null hypothesis of zero treatment-control difference is true, and so the outcome (in our case T) is close to 50% for all observations.[11] This means that there is no heteroskedasticity problem to address. We conducted formal tests of the null of homoskedasticity in DGP 1, and found 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. Consistent with this, we find that our HC1 SEs are actually slightly smaller 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.

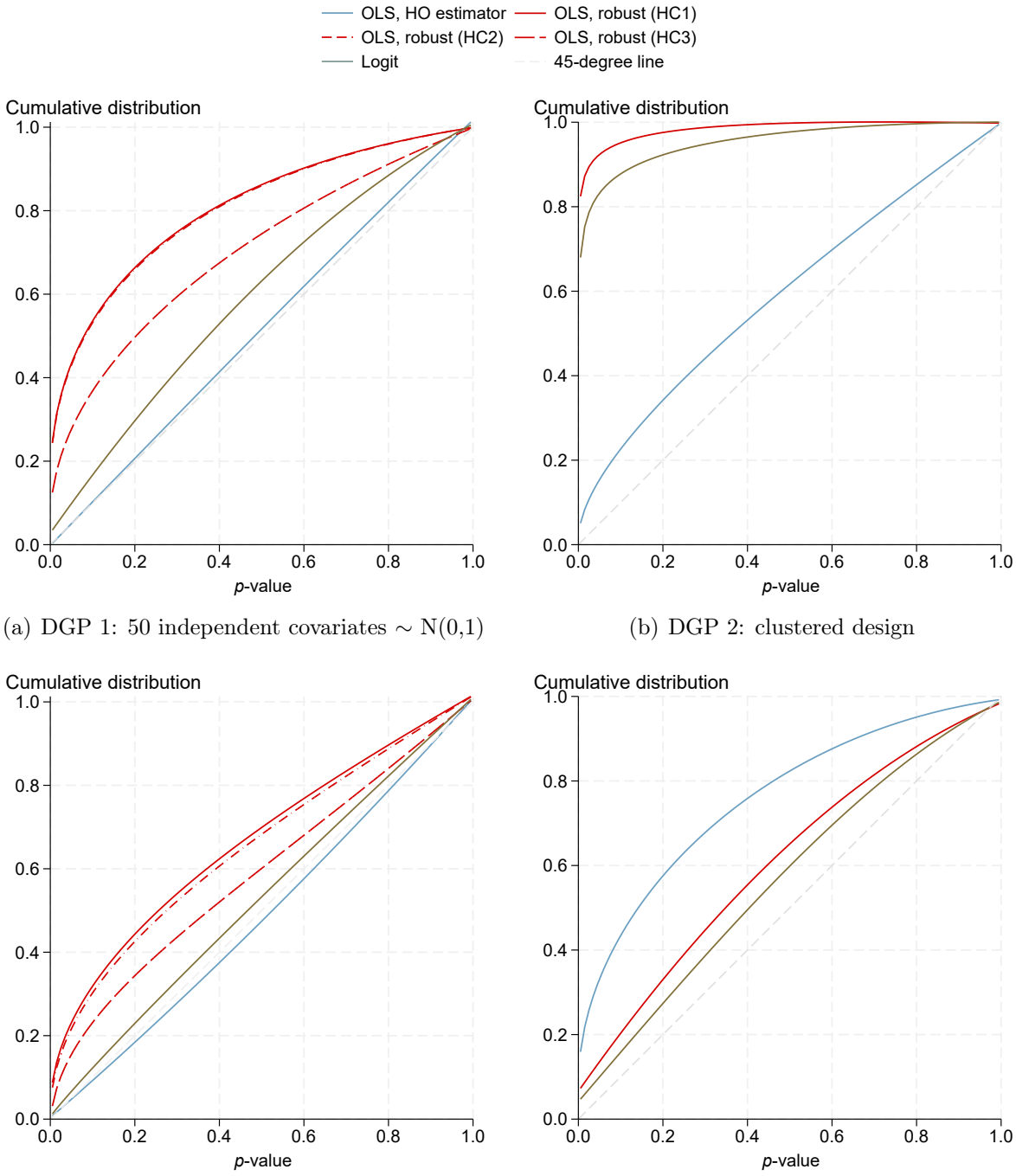
Because the HC1 SEs are only slightly smaller than the H0 SEs, the HC1 F-statistics are only slightly larger. This implies that the driver of the over-rejection problem is not the magnitude of the estimated F-statistic, but the reference distribution that it is being compared to. Indeed, Anatolyev (2012) shows that when there are many restrictions (i.e. k

[11] We thank Dean Eckles for pointing this out.

is an appreciable fraction of N), the F-statistic has an asymptotically *normal*, rather than chi-squared, distribution. While he shows that this leads to only modest over-rejections of the null, Anatolyev and Sølvsten (2023) shows that the over-rejection problem can be severe in simulated data, mirroring our findings.

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 from omnibus F- and chi-squared tests of joint orthogonality, using sampling-based inference, 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. Figures A.4 and A.5 show how test size varies with the number of covariates and sample size for DGP 1 and DGP 2 respectively.

These problems are magnified when sample size is smaller and the number of covariates is larger. Figures A.4 and A.5 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. 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.

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 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 first describe 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). 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.[12] 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). Since the randomized "treatments" do nothing, the sharp null hypothesis of a zero treatment effect for all observations is true by construction. 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.[13]

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 positive aspect of randomization inference is that it does not require specifying a model of the error term, which typically depends on a set of unknown parameters. This makes randomization inference more robust to certain non-normality or heteroskedasticity violations (Young 2019).

Randomization inference tests for balance have the advantage of being very simple to implement in Stata using the `ritest` package (Heß 2017). Below is example code for imple-

[12]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.

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

menting 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 independently 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)
```

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 Familywise 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 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 as, with pairwise t-tests and groupwise F-tests and balance tests more generally, null hypotheses are expected to be true.

Both FWER and FDR corrections can therefore be considered in the context of balance tests. In Appendix Figures A.2 and A.3, 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.

In what follows, we therefore focus on sharpened q-values, which control the FDR. Controlling the FDR at level q implies imposing that the proportion of type I errors is below q. The basic method for this approach, from Benjamini and Hochberg (1995), is the following. Select a critical value for the test p_{crit}. Sort the p-values in increasing order and count them; call the total number M. Each has a rank, r, from 1 (the smallest) to M (the largest). Then, starting from the largest p-value, we test each one against $p_{crit} \times (r/M)$. So if there are 10 p-values then the largest is tested against 0.10, the second-largest against 0.09, and so forth. We stop when we get to the first rejection and reject all tests with smaller p-values.

The approach we use augments this method in three ways. First, we "sharpen" q_{crit} to improve statistical power while still controlling the FDR at the same rate, following Benjamini, Krieger, and Yekutieli (2006). Second, we use Anderson (2008)'s approach to compute not just whether a test was rejected at the q_{crit} level but the smallest q_{crit} for which the test would be rejected, which can be interpreted in the same way as a standard p-value.[14] Third, we use the minimum of all the sharpened q-values as a test for overall balance. When we apply the q-value procedure to the p-values of M pairwise t-tests, we obtain M test statistics. While this helps determine which variables are imbalanced, it does not provide a unique determination of whether there is overall imbalance. We do this by rejecting the null hypothesis

[14]One potential limitation of this approach is that it technically only works for independent tests (in Anderson's simulations it also works for positively dependent tests). Thus we may expect it to work better in DGP 1, where the covariates are independent, as compared with the other DGPs that do have a non-zero correlation structure.

that treatment arms are balanced if the minimum sharpened q-value is below a conventional significance threshold (usually 0.1 in economics). This 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 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 homoskedasticity 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. 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

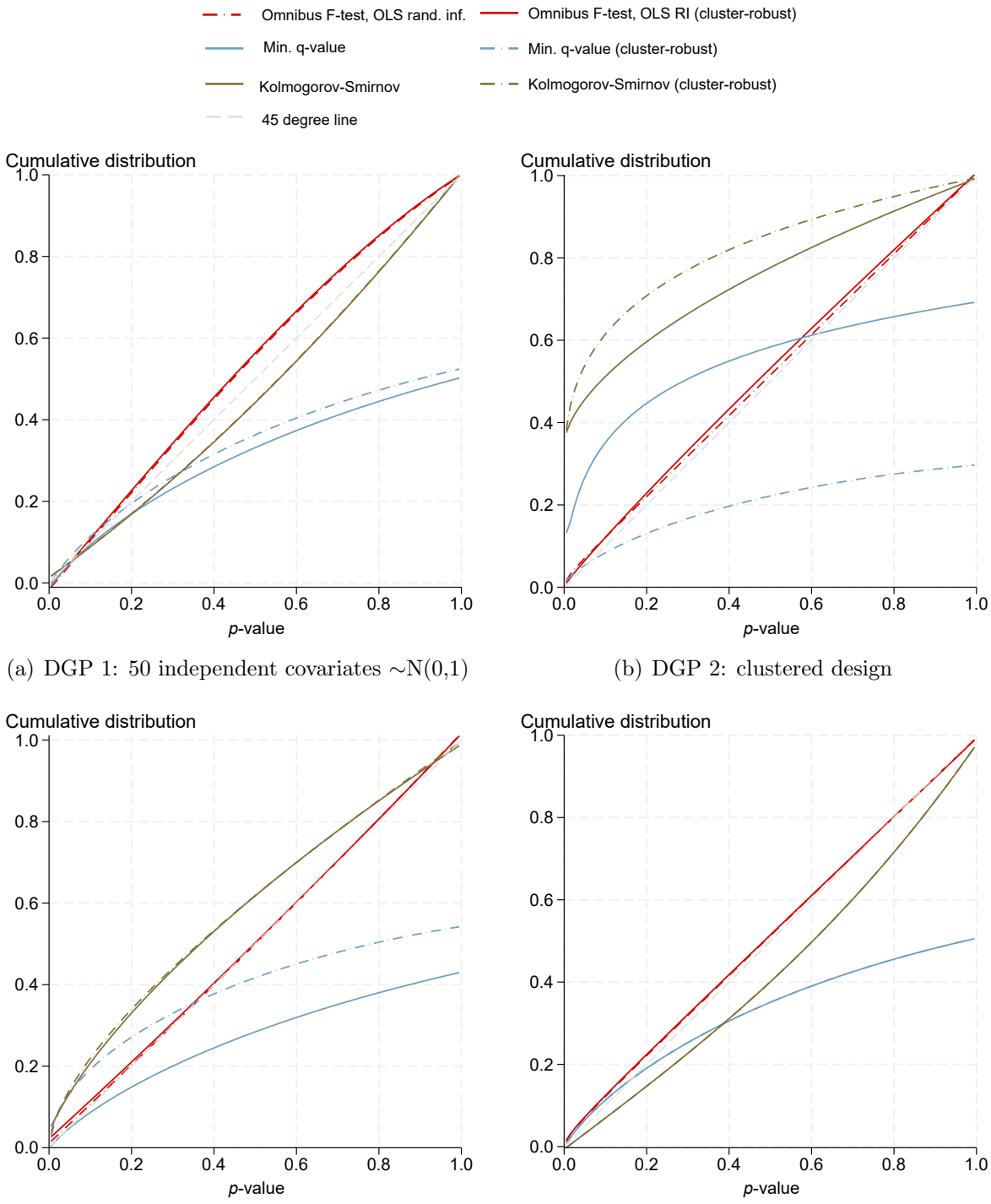
hypothesis of balance.

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 sample size and the number of covariates (Appendix Tables A.6 and A.7). 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).

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 from omnibus F- and chi-squared tests of joint orthogonality with 500 simulations per figure. DGPs 1 and 3 are individually-randomized designs, while DGPs 2 and 4 are clustered designs. Figures A.6 and A.7 show how test size varies with the number of covariates and sample size for DGP 1 and DGP 2 respectively.

4.5 Statistical Power

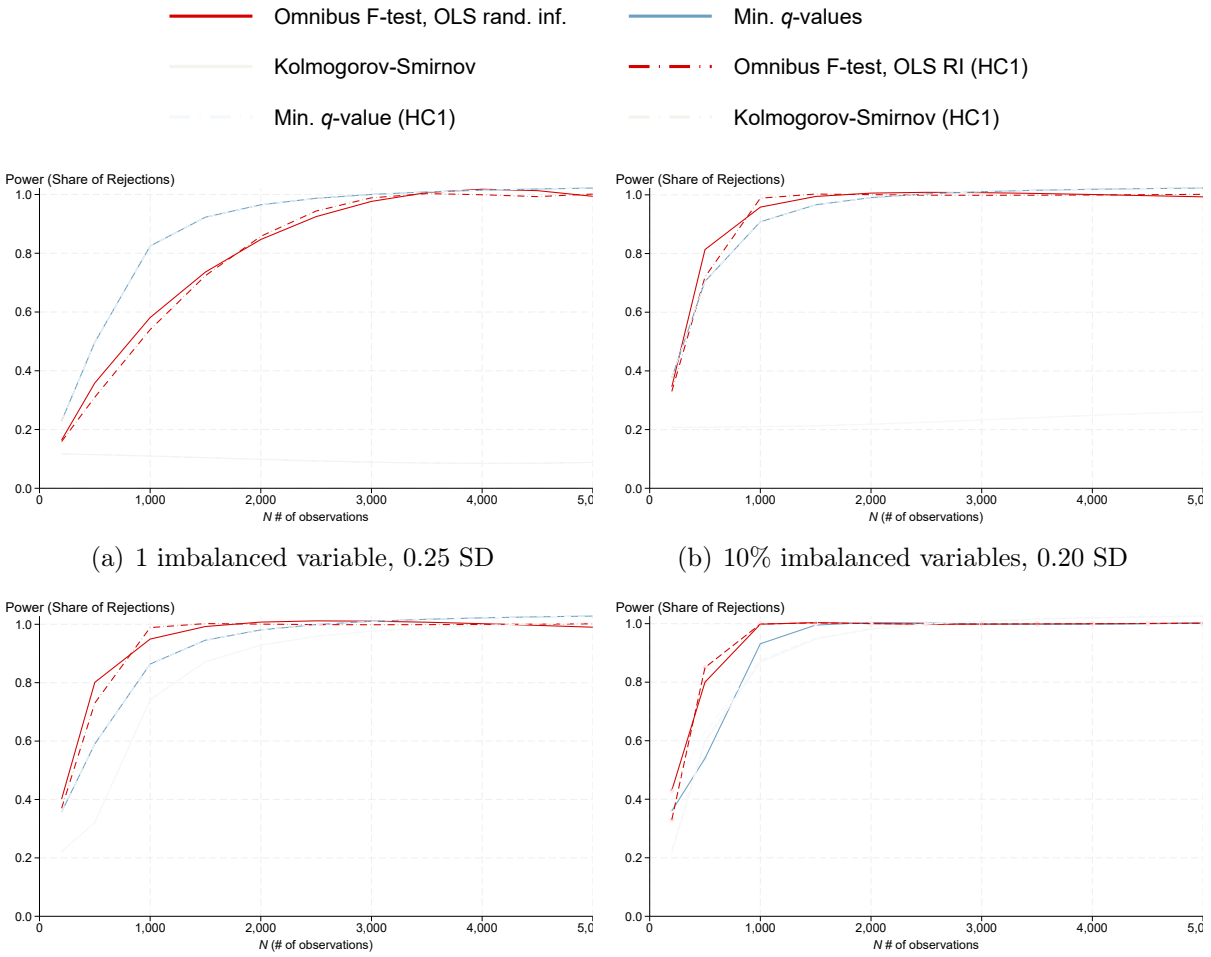
We assess the power of the balance tests in Figure 3 for DGP 1 and Figure 4 for DGP 2. We focus on balance tests that have correct sizes to avoid cluttering figures with misleading information.

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.[15] Omnibus F-tests of joint orthogonality with randomization inference have higher power than minimum q-values when a larger number of imbalances that each have a smaller magnitude are considered. This suggests that the two approaches might be complementary in individually randomized RCTs.

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 sample size and the number of covariates (Figures A.8 and A.9 in the Appendix). 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.

[15]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.8 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

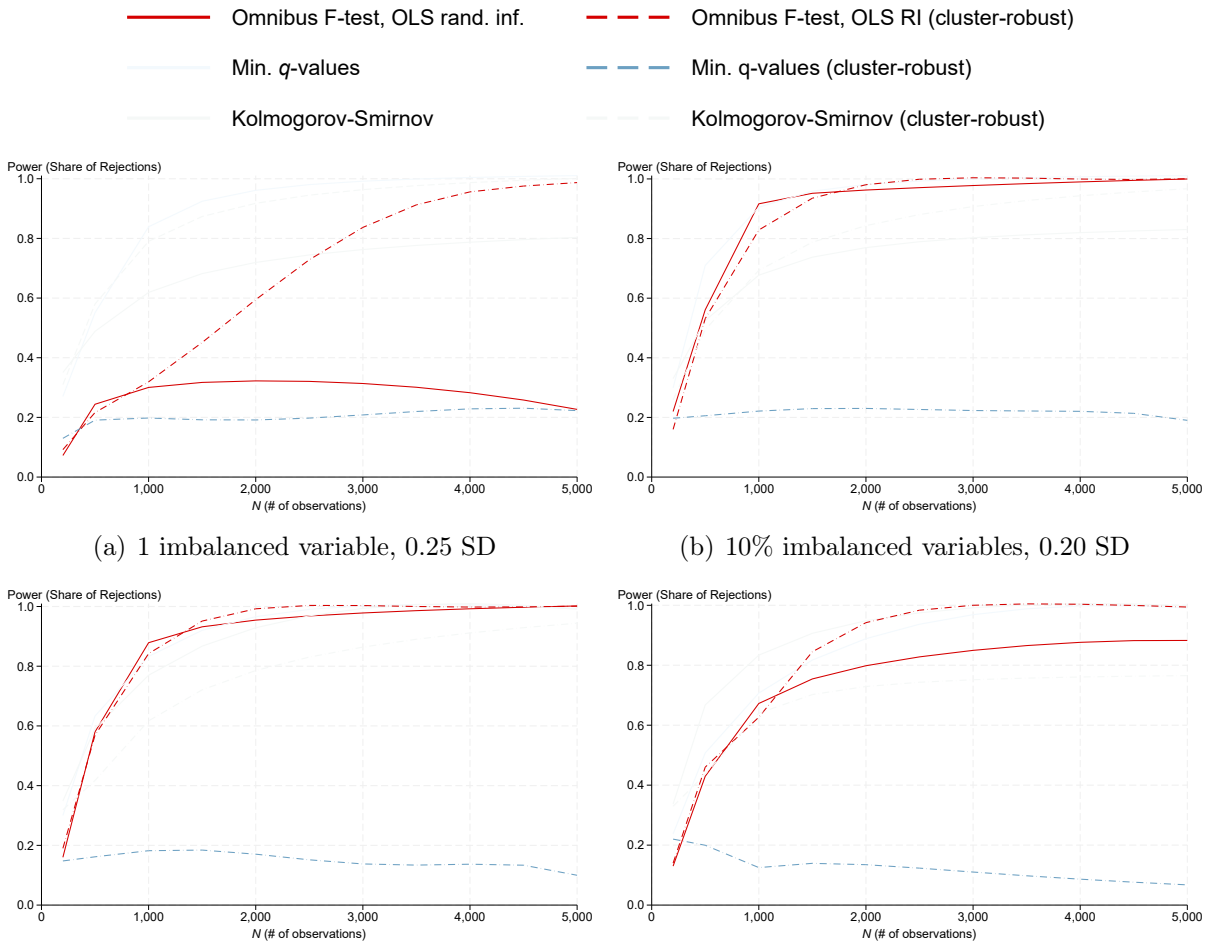
(b) 10% imbalanced variables, 0.20 SD

(c) 20% imbalanced variables, 0.15 SD

(d) 50% imbalanced variables, 0.10 SD

Notes: 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. To produce each figure, a total of 100 simulated datasets of N observations are generated, for each $N \in \{200, 500, 1000, 1500, 2000, 2500, 3000, 3500, 4000, 4500, 5000\}$. For each method, the light dots show the N-specific share of p-values or minimum q-values below 0.10, and the lines show fractional-polynomial predictions. Figure A.8 shows how the power of the tests varies with the number of covariates and sample size. Results for minimum q-values using HC1 SEs and K–S tests are hidden as these tests have the incorrect size (see Figure 2).

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: 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. To produce each figure, a total of 100 simulated datasets of N observations are generated, for each $N \in \{200, 500, 1000, 1500, 2000, 2500, 3000, 3500, 4000, 4500, 5000\}$. For each method, the light dots show the N-specific share of p-values or minimum q-values below 0.10, and the lines show fractional-polynomial predictions. Figure A.9 shows how the power of the tests varies with the number of covariates and sample size. Results for minimum q-values using HO SEs and K–S tests are hidden as these tests have the incorrect size (see Figure 2).

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 the individually randomized designs in DGPs 1 and 3, we show in Figure A.10 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.[16] 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 imbalances.

For DGPs 2 and 4, which are clustered designs, only MANOVA and multinomial logit

[16]The multinomial logit can fail to converge when the number of observations is low compared to the number of covariates (Figure A.13).

with randomization inference have the correct test size, while minimum q-values are again conservative, under-rejecting the null (Figure A.10). 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 independently 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.[17]

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).

[17]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 SI HC1/Cluster (1)	SI H0 (2)	RI HC1/Cluster (3)	Min. q-value HC1/Cluster (4)	K–S p-value HC1/Cluster (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)

In Table II of the main paper, Auriol et al. (2020) present a table of balance tests with twelve preregistered covariates.[18] They consider both pairwise t-tests and omnibus F-tests of joint orthogonality.[19] Out of the 24 p-values of 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. However, Section 3.1 showed that "vote counting" should be avoided as it leads to over-rejections of the null hypothesis.

In Table 1, we present the results of different joint tests of balance. 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.[20]

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).

[18]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.

[19]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.

[20]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

The use of balance tests in randomized experiments remains controversial. They can be powerful tools to demonstrate that randomized experiments were actually conducted as designed and did not lead to study arm allocations that may yield misleading estimates. However, some statisticians argue that they are unhelpful, falsely implying problems where there are none.

Our paper shows that this problem is even worse than previously recognized. The usual method of testing for overall balance yields far too many rejections of the null hypothesis, so that even experiments with no balance problems whatsoever appear to have serious issues. We show that randomization inference-based omnibus balance tests perform much better than the conventional approach using sampling-based inference. The latter has a high probability of Type I error with a large number of covariates, while the former has no such issue with test size, and it has high statistical power.

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 could cause treatment effect estimates to be biased on average. 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. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.3982/ECTA12675.

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*, rdad065.

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.

Benjamini, Yoav, and Yosef Hochberg. 1995. "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing." *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1): 289–300. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.2517-6161.1995.tb02031.x.

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, 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.

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. Publisher: ASA Website _eprint: https://doi.org/10.1080/01621459.2017.1328360.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

McKenzie, David. 2015. "Tools of the Trade: A Joint Test of Orthogonality When Testing for Balance."

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.

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.

Siu, Jade, Olivier Sterck, and Cory Rodgers. 2023. "The freedom to choose: Theory and quasi-experimental evidence on cash transfer restrictions." *Journal of Development Economics*, 161: 103027.

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.

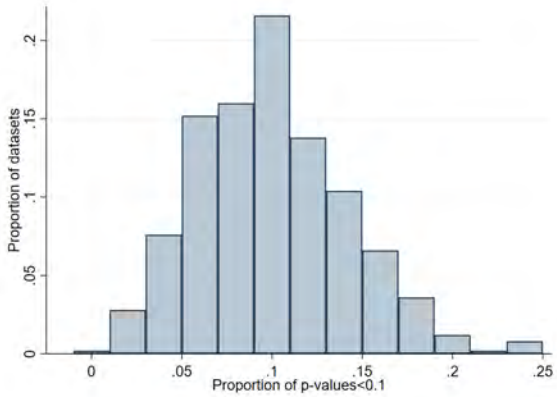
Tungodden, Jonas, and Alexander Willén. 2023. "When parents decide: Gender differences in competitiveness." *Journal of Political Economy*, 131(3): 751–801.

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.

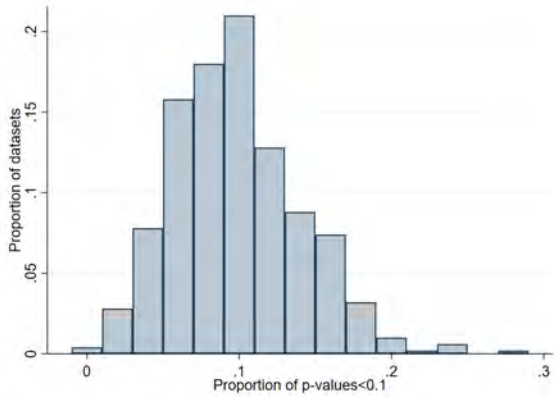
Young, Alwyn. 2019. "Channeling Fisher: Randomization Tests and the Statistical Insignificance of Seemingly Significant Experimental Results*." *The Quarterly Journal of Economics*, 134(2): 557–598.

Zárate, Román Andrés. 2023. "Uncovering peer effects in social and academic skills." *American Economic Journal: Applied Economics*, 15(3): 35–79.

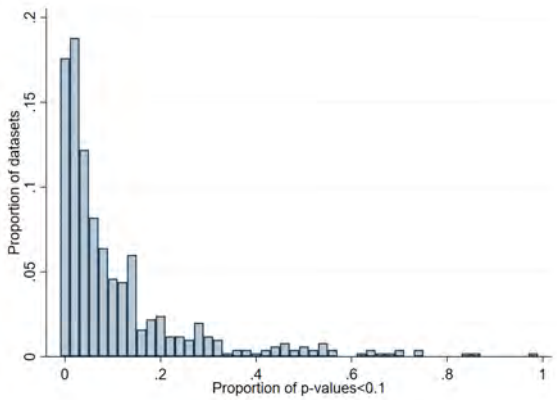
Online Appendix



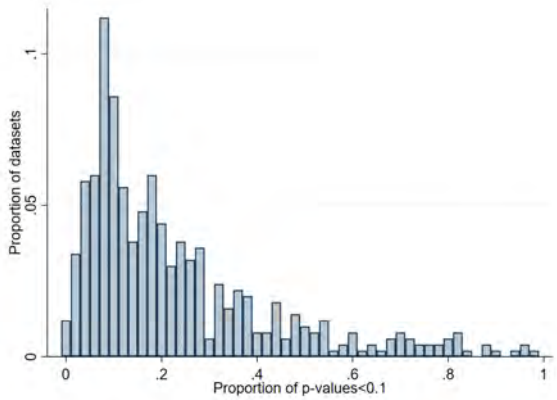
(e) Pairwise t-tests, OLS with HC1 variance estimator (DGP 1)



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



(g) Pairwise t-tests, OLS with cluster-robust variance estimator (DGP 2)



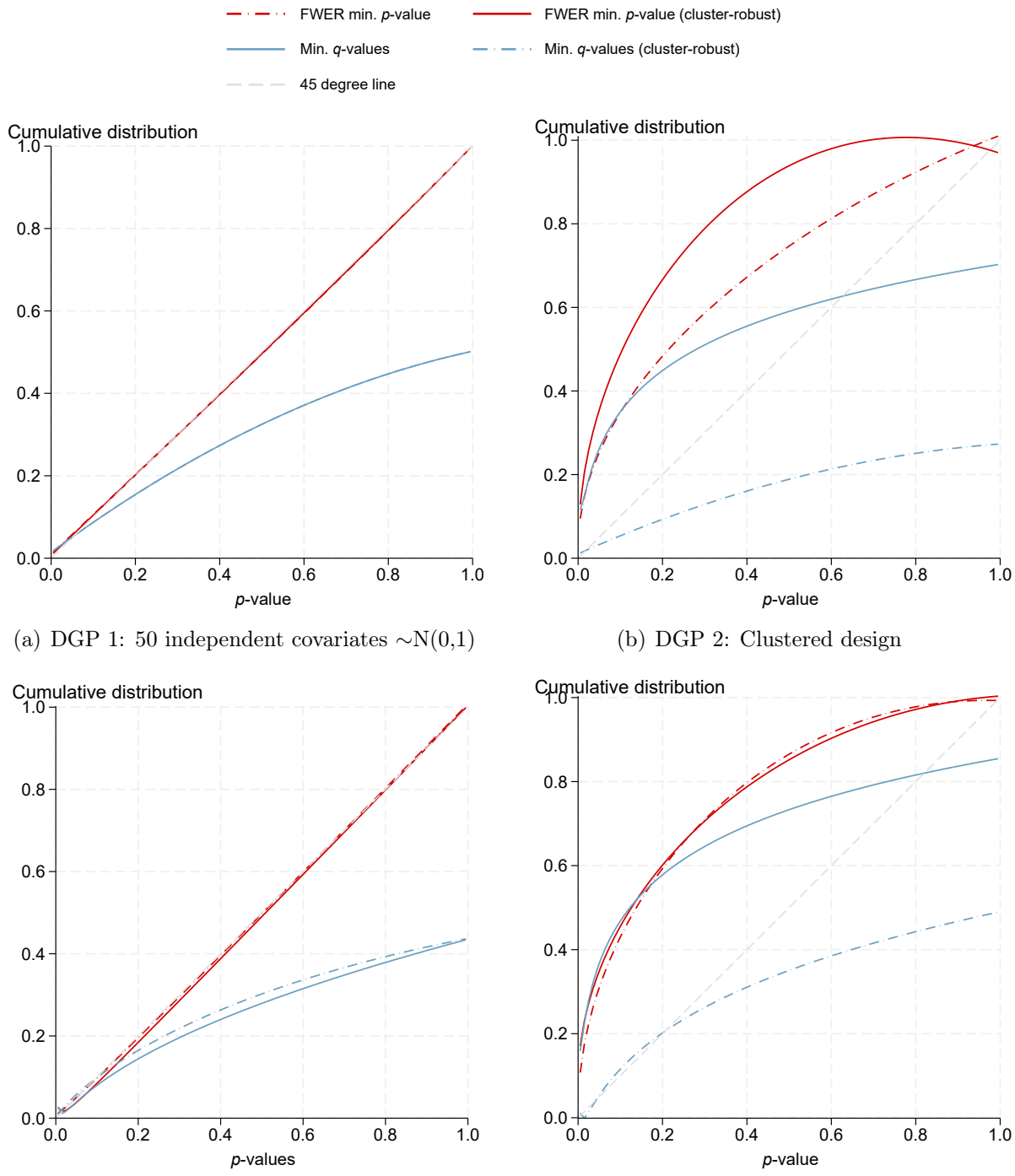
(h) Pairwise t-tests, OLS with HO variance estimator (DGP 2)

Figure A.1

Pairwise t-tests for DGP 1 and DGP 2: distribution of share of p-values below 0.10.

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 ∼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 and correlated within clusters (average coefficient of intra-cluster 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
Size of Balance Tests Using FWER and FDR Multiple-Hypothesis Testing Adjustments



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

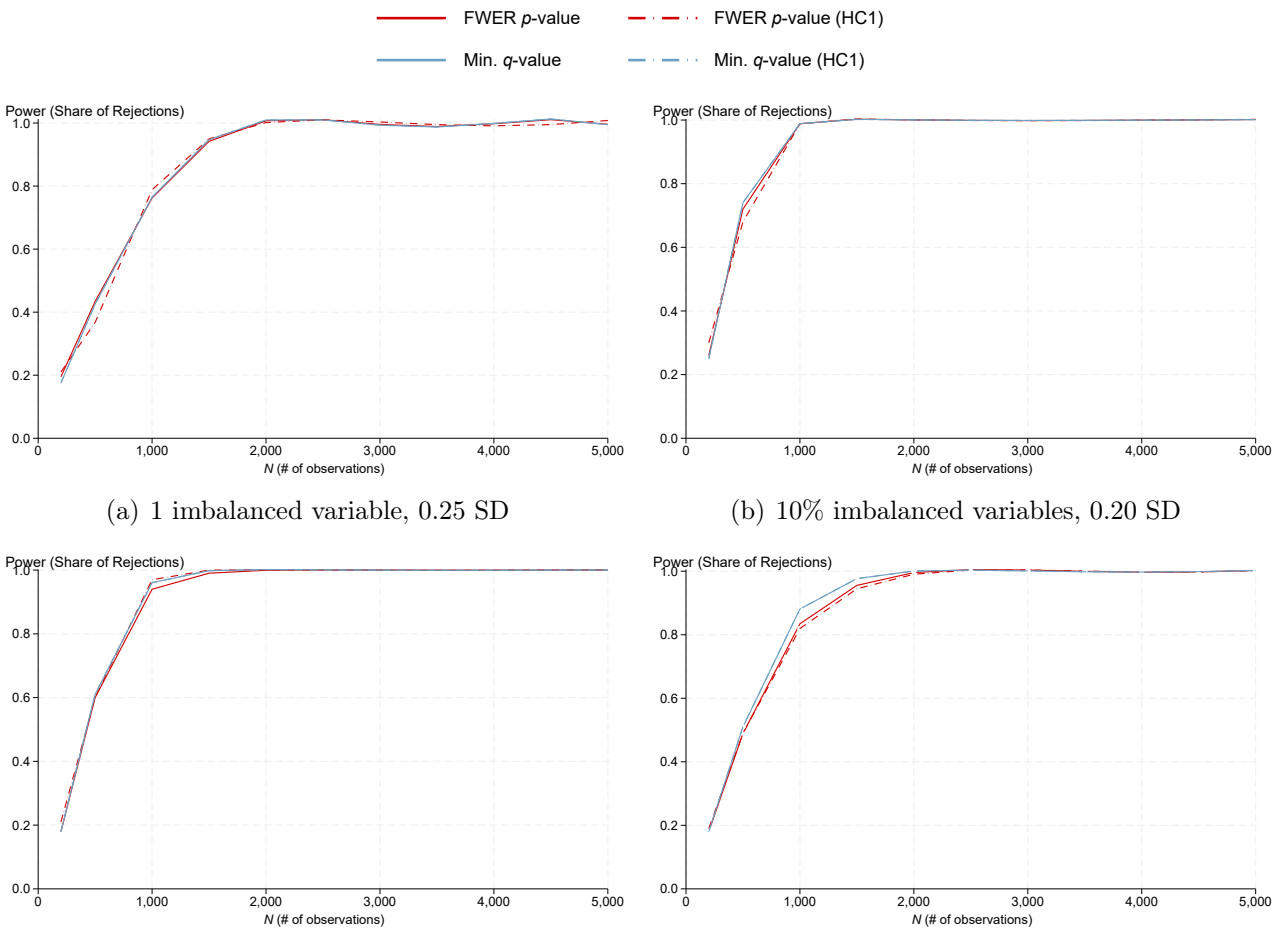
(b) DGP 2: Clustered design

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

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

Notes: Cumulative distributions of p-values from FWER minimum p-values and minimum q-values with 500 simulations per figure. DGPs 1 and 3 are individually-randomized designs, while DGPs 2 and 4 are clustered designs. Separately for each of the 500 simulated datasets and each method: (1) we use the Stata command rwolf2 to estimate k FWER-adjusted p-values and record the minimum of the p-values; and (2) we apply the method of Anderson (2008) to the k p-values from the pairwise t-tests to obtain k sharpened q-values and record the minimum of these sharpened q-values. For each method, 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.

Figure A.3
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

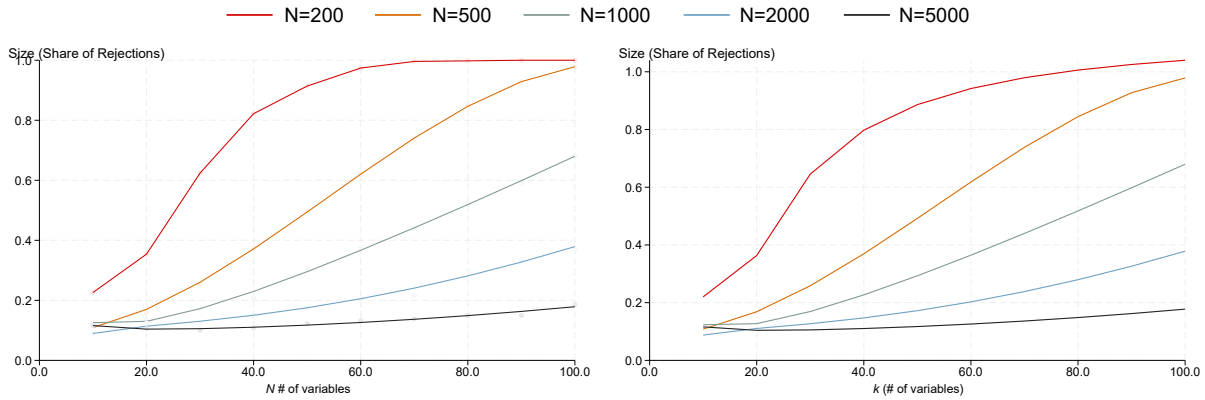
(c) 20% imbalanced variables, 0.15 SD

(d) 50% imbalanced variables, 0.10 SD

Notes: 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. To produce each figure, a total of 100 simulated datasets of N observations are generated, for each $N \in \{200, 500, 1000, 1500, 2000, 2500, 3000, 3500, 4000, 4500, 5000\}$. For each method, the light dots show the N-specific share of p-values or minimum q-values below 0.10, and the lines show fractional-polynomial predictions. Figure A.8 shows how the power of the tests varies with the number of covariates and sample size.

Figure A.4

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

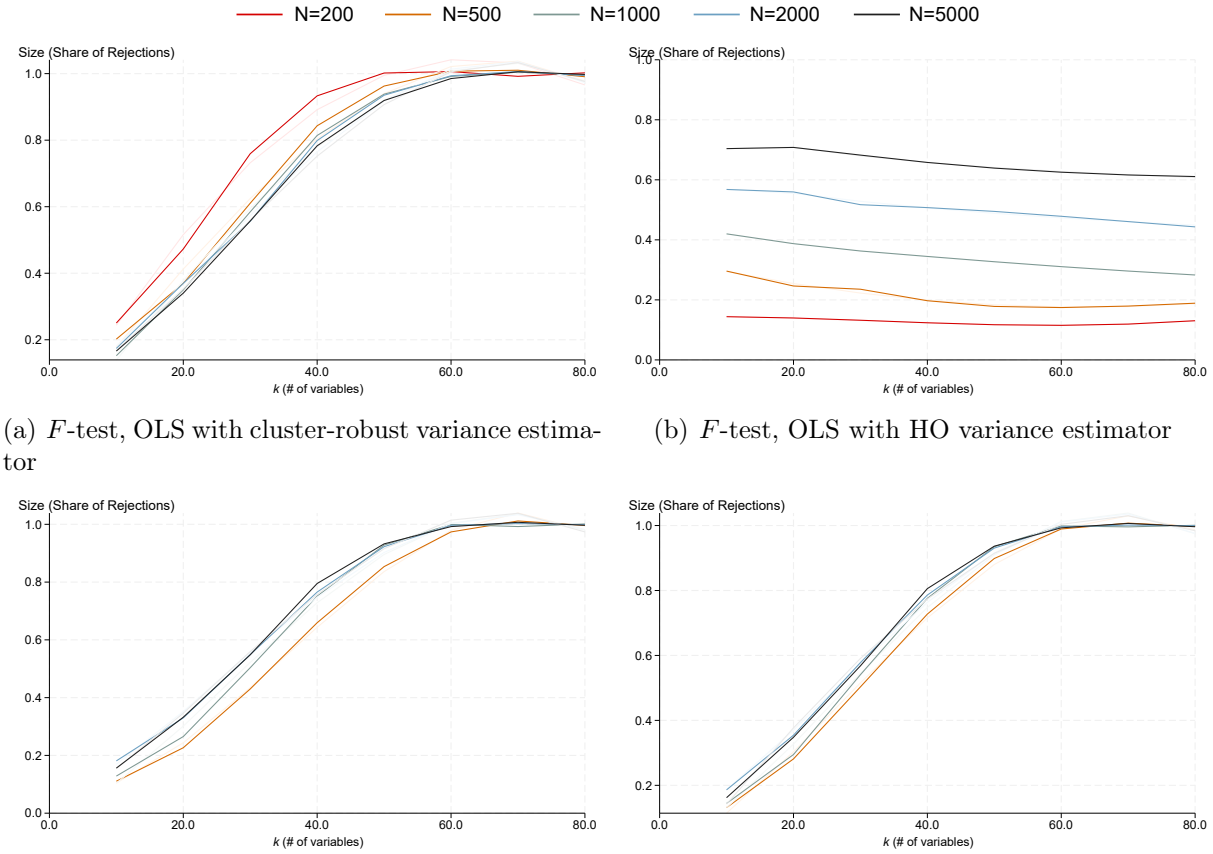
(e) Chi^2-test, Logit with HO variance estimator

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

Notes: The figures show the results of simulations with a data generating process creating k independent variables that are normally distributed $\sim$N(0,1) (with $k \in \{10, 20, 30, 40, 50, 60, 70, 80, 90, 100\}$) and an independent treatment randomly assigned to half of observations. Simulated datasets of N observations are generated (with $N \in \{200, 500, 1000, 2000, 5000\}$). Each figure focuses on one type of balance test and is constructed as follows. For each combination of k and N, 500 simulated datasets are created and the light dots show the share of tests' p-values below 0.10. For each N, fractional-polynomial prediction lines are then used to predict the size of tests for $k \in [10, 100]$.

Figure A.5

Size of Omnibus Tests of Joint Orthogonality for DGP 2 Using Sampling-based Inference



(a) F-test, OLS with cluster-robust variance estimator

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

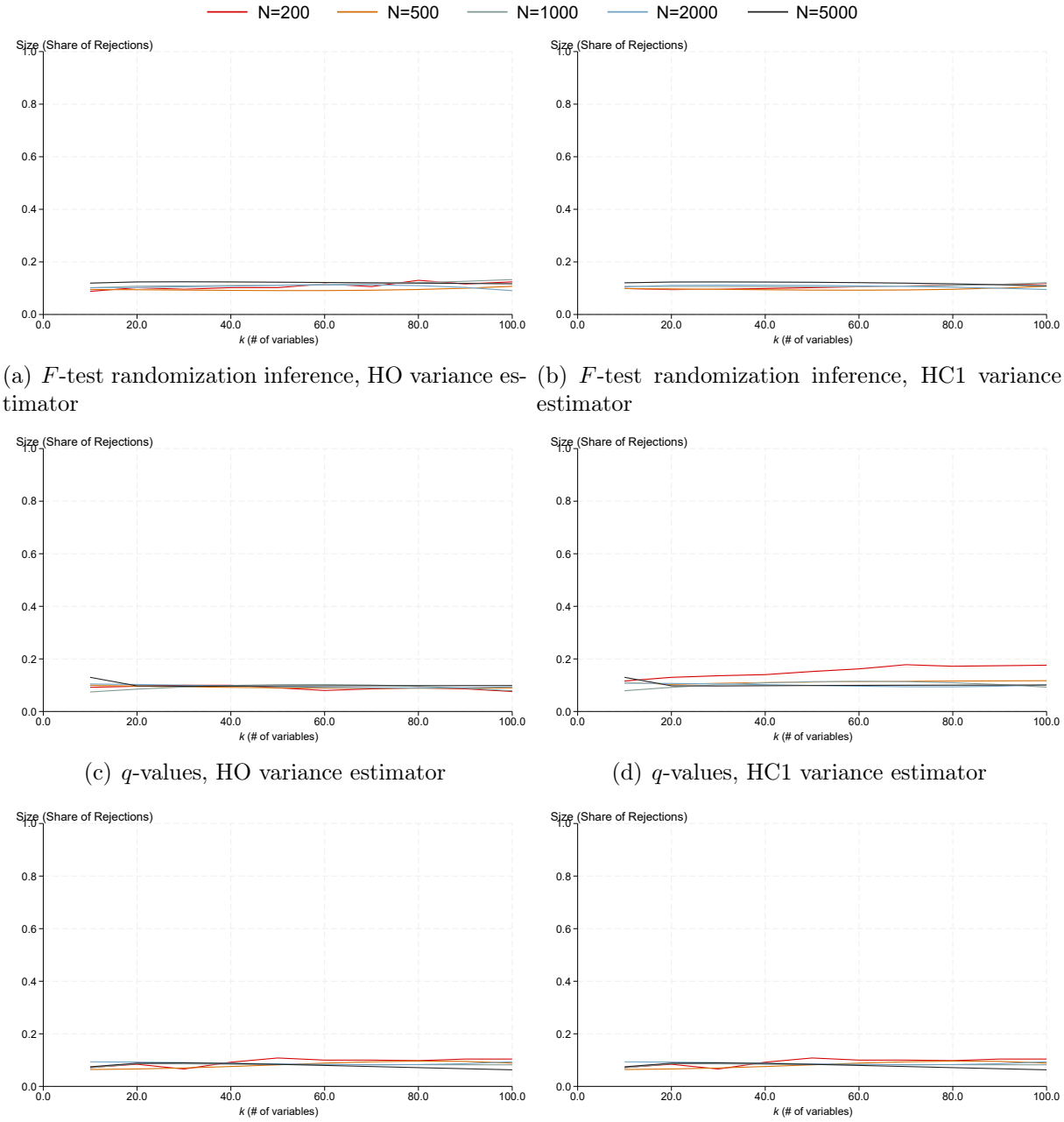
(c) F-test, Logit

(d) F-test, Probit

Notes: The figures show the results of simulations with a data generating process with N observations (with $N \in \{200, 500, 1000, 2000, 5000\}$) that are split into 100 clusters of equal size, and k variables that are normally distributed and correlated within clusters (with $n \in \{10, 20, 30, 40, 50, 60, 70, 80, 90, 100\}$, and an average intra-cluster correlation coefficient of 0.2). An independent treatment is equal to one in half of the clusters and zero otherwise. Each figure focuses on one type of balance test and is constructed as follows. For each combination of k and N, 500 simulated datasets are created and the light dots show the share of tests' p-values below 0.10. For each N, fractional-polynomial prediction lines are used to predict the size of tests for $n \in [10, 100]$.

Figure A.6

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

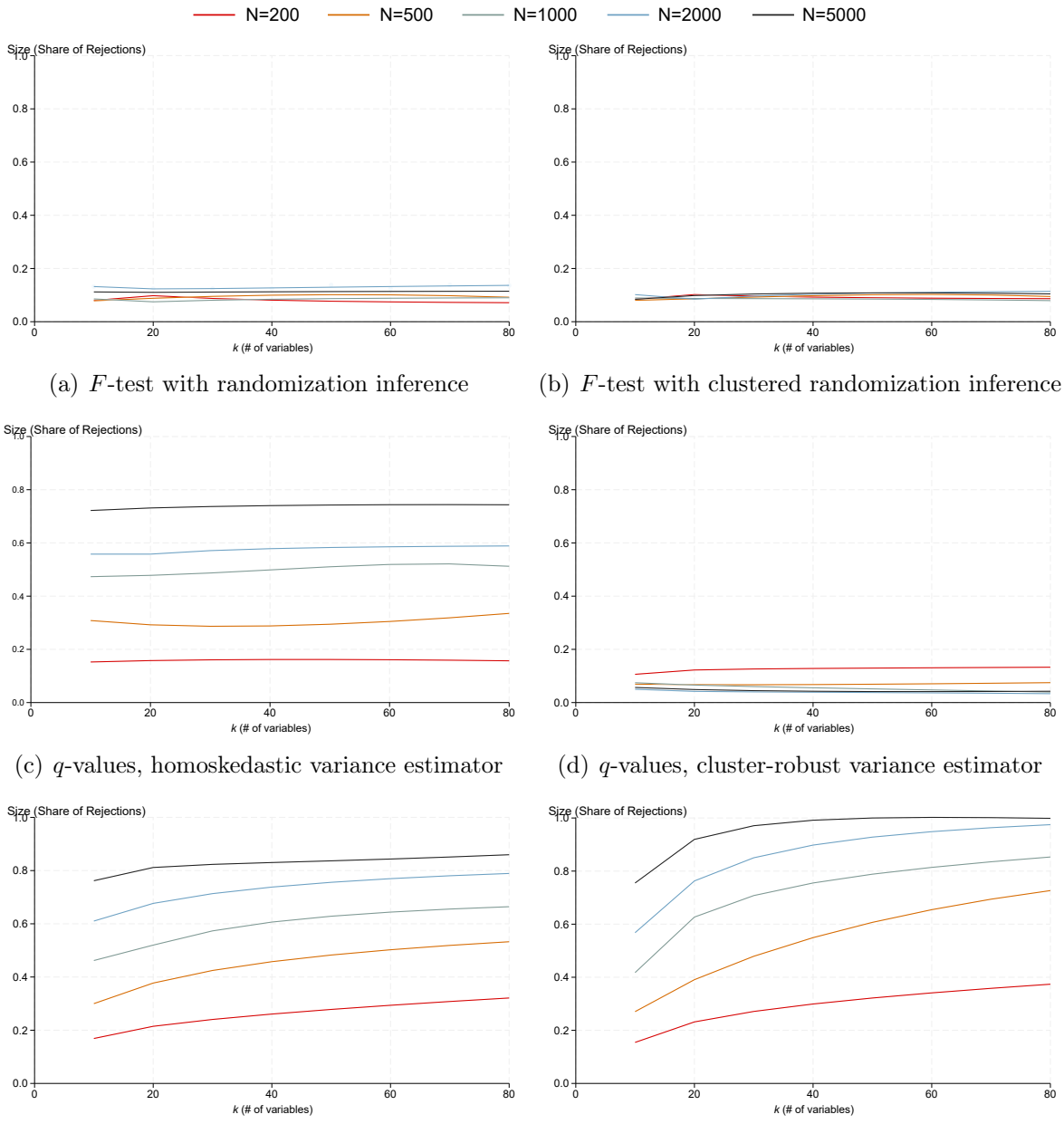
(e) Kolmogorov–Smirnov test, HO variance estimator

(f) Kolmogorov–Smirnov test, HC1 variance estimator

Notes: The figures show the results of simulations with a data-generating process that creates k independent variables that are normally distributed ~N(0,1) (with $k \in \{10, 20, 30, 40, 50, 60, 70, 80, 90, 100\}$) and an independent treatment randomly assigned to half of observations. Simulated datasets of N observations are generated (with $N \in \{200, 500, 1000, 2000, 5000\}$). For each combination of k and N, we create 500 simulated datasets and run pairwise t-tests for all k variables for each dataset. The Kolmogorov–Smirnov tests are run on the p-values of the pairwise t-tests. The minimum q-value is the minimum sharpened q-value obtained after applying the method of Anderson (2008) to the p-values from the pairwise t-tests. For each N, the light dots show the k-specific share of test statistics below 0.10 and the lines show fractional-polynomial predictions for $k \in [10, 100]$.

Figure A.7

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) q-values, homoskedastic variance estimator

(d) q-values, cluster-robust variance estimator

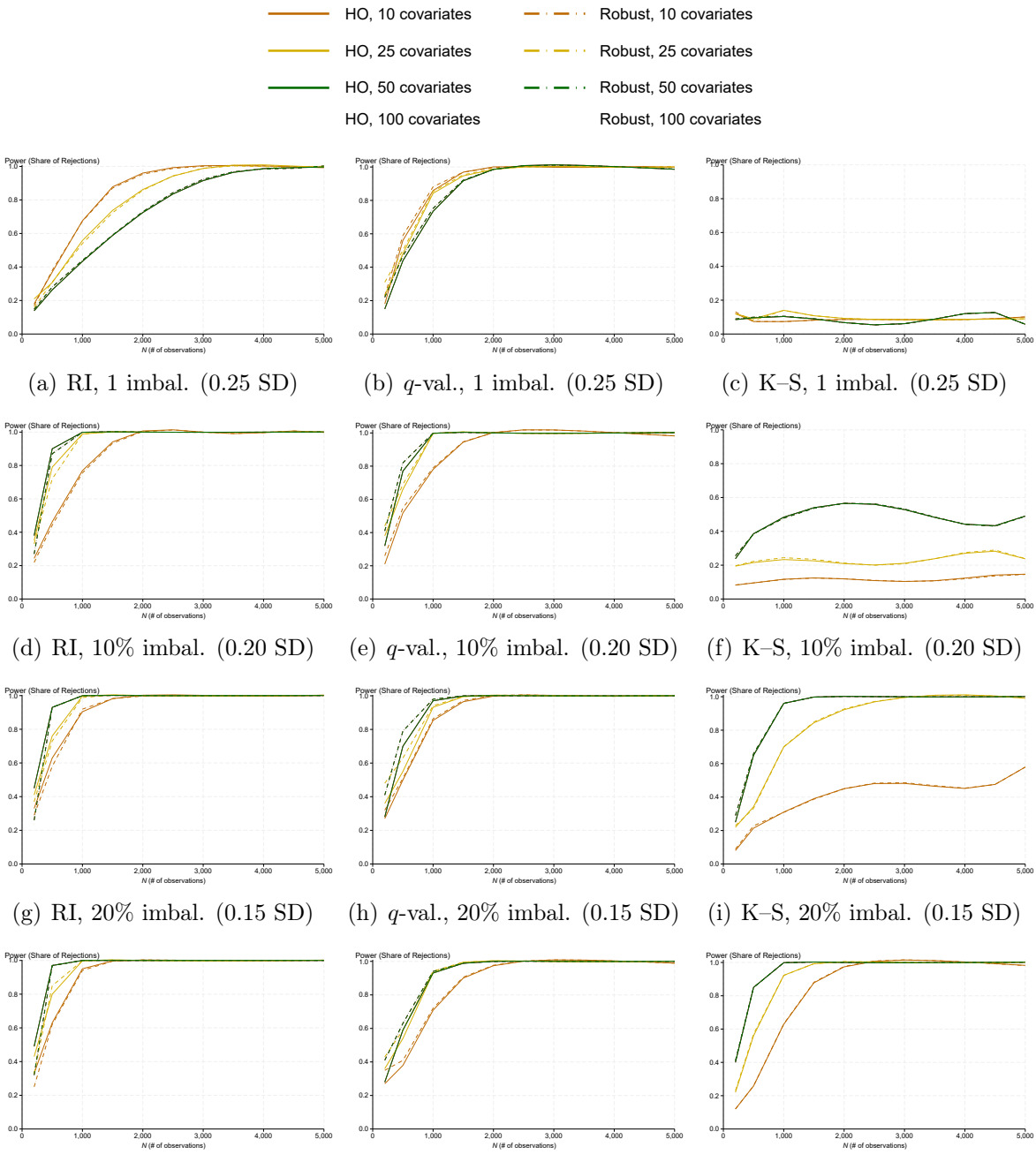
(e) K–S test, homoskedastic variance estimator

(f) K–S test, cluster-robust variance estimator

Notes: The figures show the results of simulations with a data-generating process with N observations (with $N \in \{200, 500, 1000, 2000, 5000\}$) that are split into 100 clusters of equal size, and k variables that are normally distributed and correlated within clusters (with $k \in \{10, 20, 30, 40, 50, 60, 70, 80, 90, 100\}$, and an average intra-cluster correlation coefficient of 0.2). An independent treatment is equal to one in half of the clusters and zero otherwise. For each combination of k and N, we create 500 simulated datasets and run pairwise t-tests for all k variables. The Kolmogorov–Smirnov tests are run on the p-values of the pairwise t-tests. The minimum q-value is the minimum sharpened q-value obtained after applying the method of Anderson (2008) to the p-values from the pairwise t-tests. The light dots show the share of p-values or minimum q-values below 0.10. For each N, fractional-polynomial prediction lines are used to predict the size of tests for $k \in [10, 100]$.

Figure A.8

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)

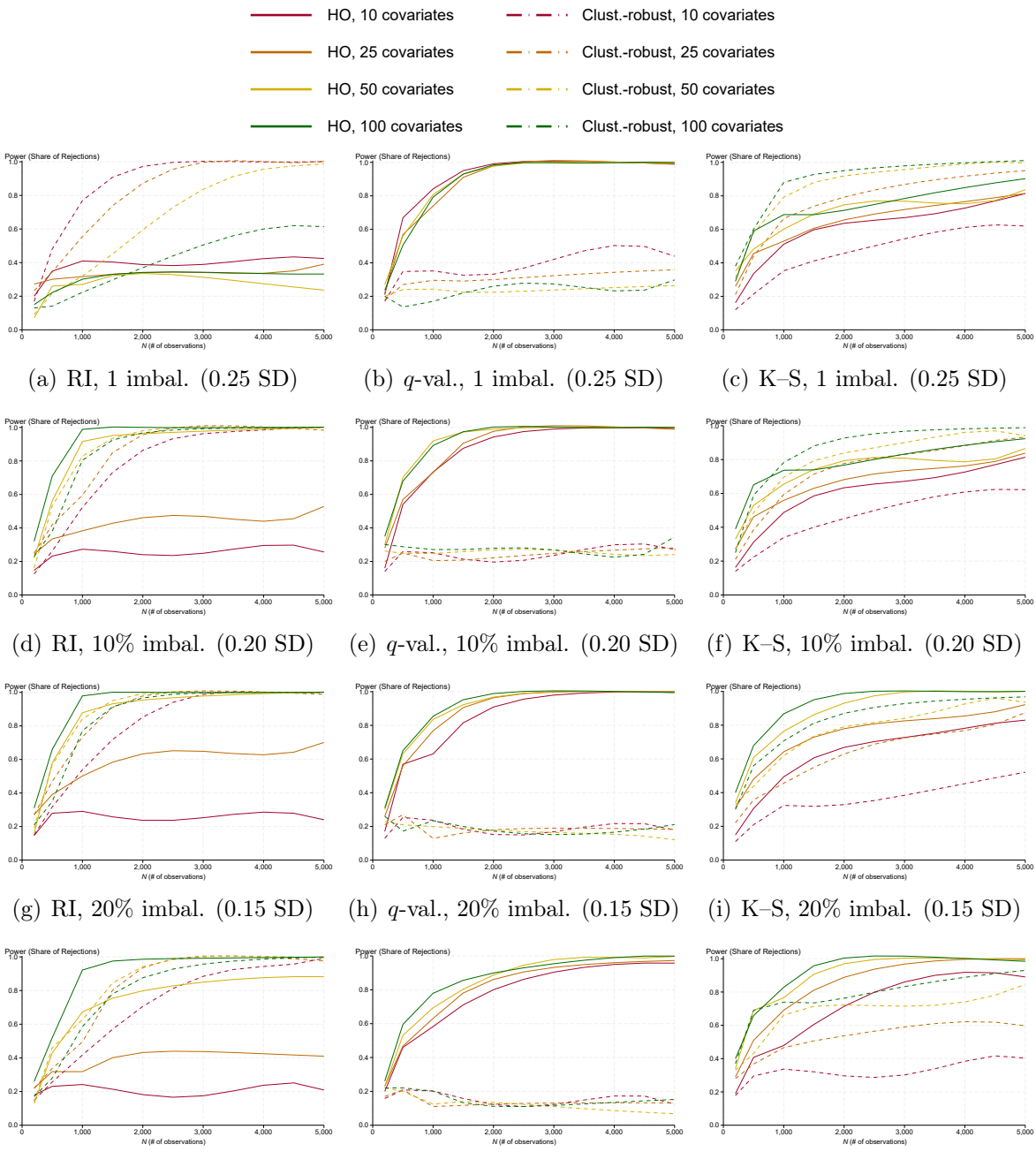
(g) RI, 20% imbal. (0.15 SD) (h) q-val., 20% imbal. (0.15 SD) (i) K–S, 20% imbal. (0.15 SD)

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

Notes: The figures show the results of simulations with $k \in \{10, 25, 50, 100\}$ independent $\sim$N(0,1) variables, and an independent treatment randomly assigned to half of observations. Imbalances are then created for a random subset of the k variables. 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 made imbalanced by adding 0.1 to treated observations. For each method and each k, the light dots show the N-specific share of p-values or minimum q-values below 0.10, and the lines show fractional-polynomial predictions for $N \in [200, 5000]$.

Figure A.9

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)

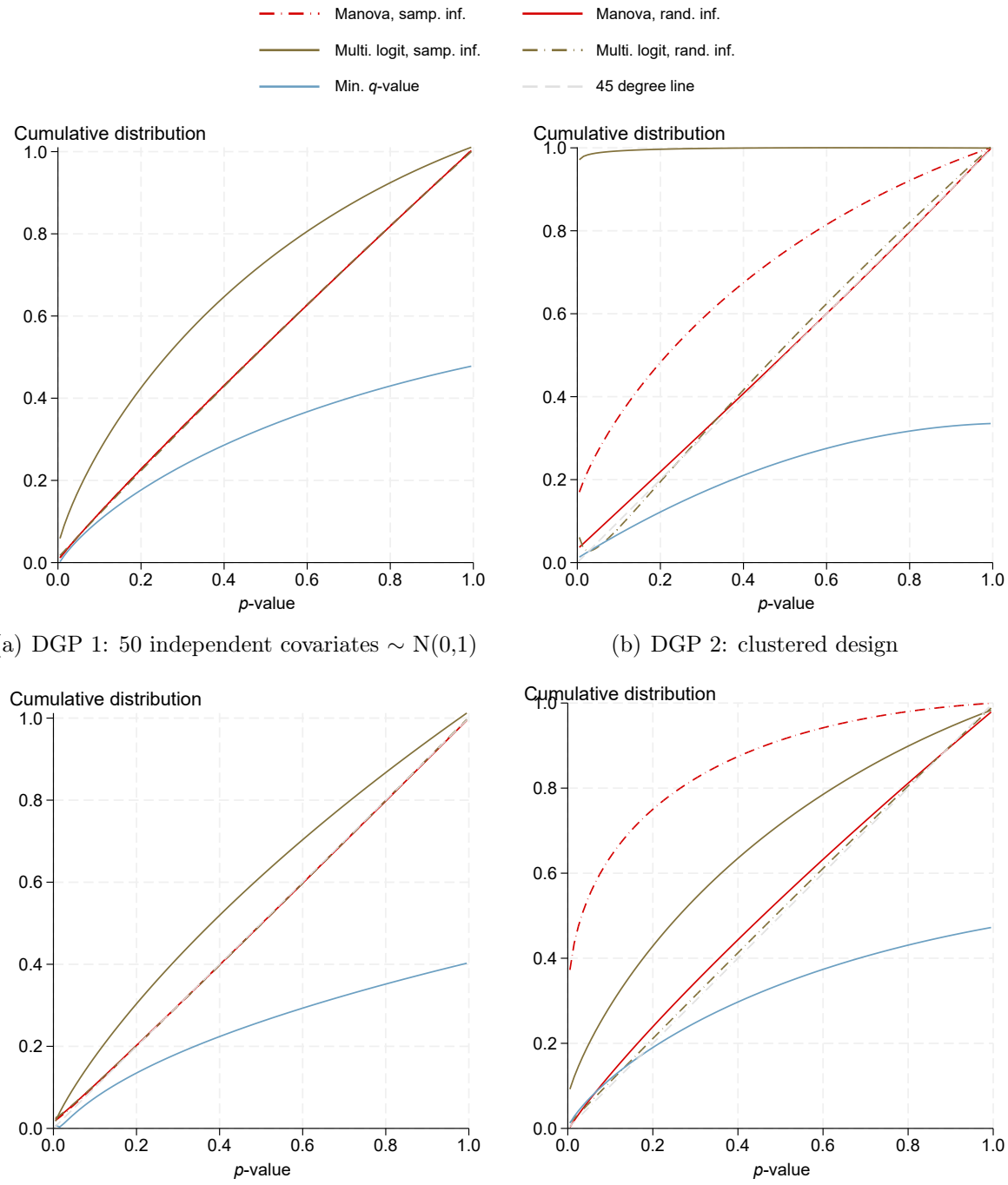
(d) RI, 10% imbal. (0.20 SD) (e) q-val., 10% imbal. (0.20 SD) (f) K–S, 10% imbal. (0.20 SD)

(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: The figures show the results of simulations with $N \in \{200, 500, 1000, 2000, 5000\}$ observations split into 100 clusters of equal size, and $k \in 10, 25, 50, 100$ normally distributed variables that are correlated within clusters (with an average intra-class correlation of 0.2). An independent treatment is equal to one in half of the clusters and zero otherwise. Imbalances are then created for a subset of the k variables. 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 made imbalanced, by adding 0.1 to treated observations. For each method and each k, the light dots show the N-specific share of p-values or minimum q-values below 0.10, and the lines show fractional-polynomial predictions for $N \in [200, 5000]$.

Figure A.10
Size of Omnibus Tests of Joint Orthogonality with Multiple Treatment Arms



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

(b) DGP 2: clustered design

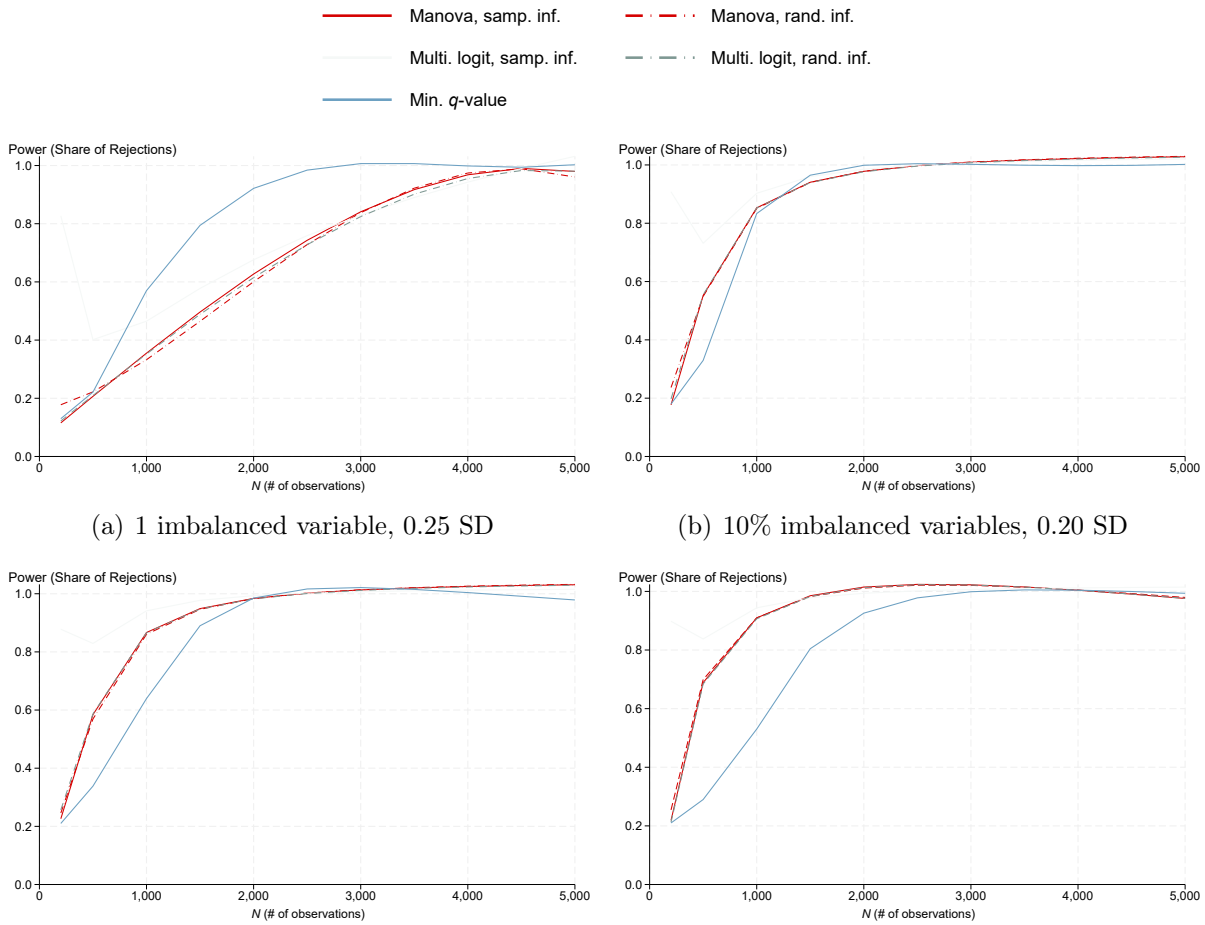
(c) DGP 3: data from Garbiras-Díaz and Montenegro (2022)

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

Notes: Cumulative distributions of p-values from omnibus F- and chi-squared tests of joint orthogonality, using sampling-based inference, 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 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. Figures A.13 and A.14 show how test size vary with the number of covariates and sample size for DGP 1 and DGP 2 respectively.

Figure A.11

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

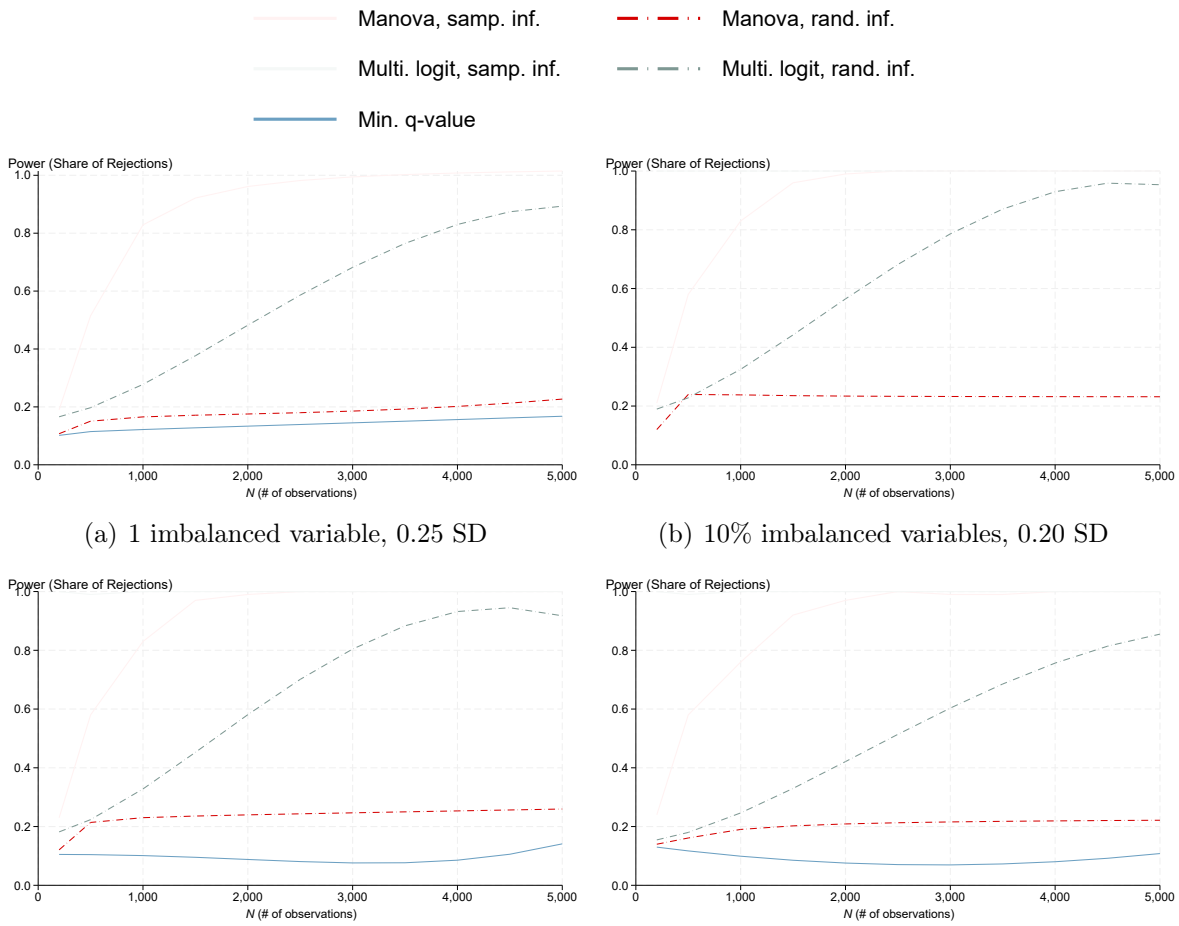
(c) 20% imbalanced variables, 0.15 SD

(d) 50% imbalanced variables, 0.10 SD

Notes: The figures show the results of simulations with a data generating process creating 50 independent variables that are normally distributed ∼N(0,1) and an independent treatment randomly assigned to half of observations. Imbalances are then created for a subset of the k variables and two treatment arms. 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. To produce each figure, a total of 100 simulated datasets of N observations are generated, for each $N \in \{200, 500, 1000, 1500, 2000, 2500, 3000, 3500, 4000, 4500, 5000\}$. We run pairwise t-tests for all k variables for each dataset and all possible comparisons between treatment arms. The minimum q-value is the minimum sharpened q-value obtained after applying the method of Anderson (2008) to the p-values from the pairwise t-tests. For each method, the light dots show the N-specific share of p-values or minimum q-values below 0.10, and the lines show fractional-polynomial predictions for $N \in [200, 5000]$. Results from multinomial logit with sampling-based inference are hidden as this approach has incorrect test size.

Figure A.12

Power of Omnibus Tests of Joint Orthogonality for DGP 2, with Multiple Treatment Arms



(a) 1 imbalanced variable, 0.25 SD

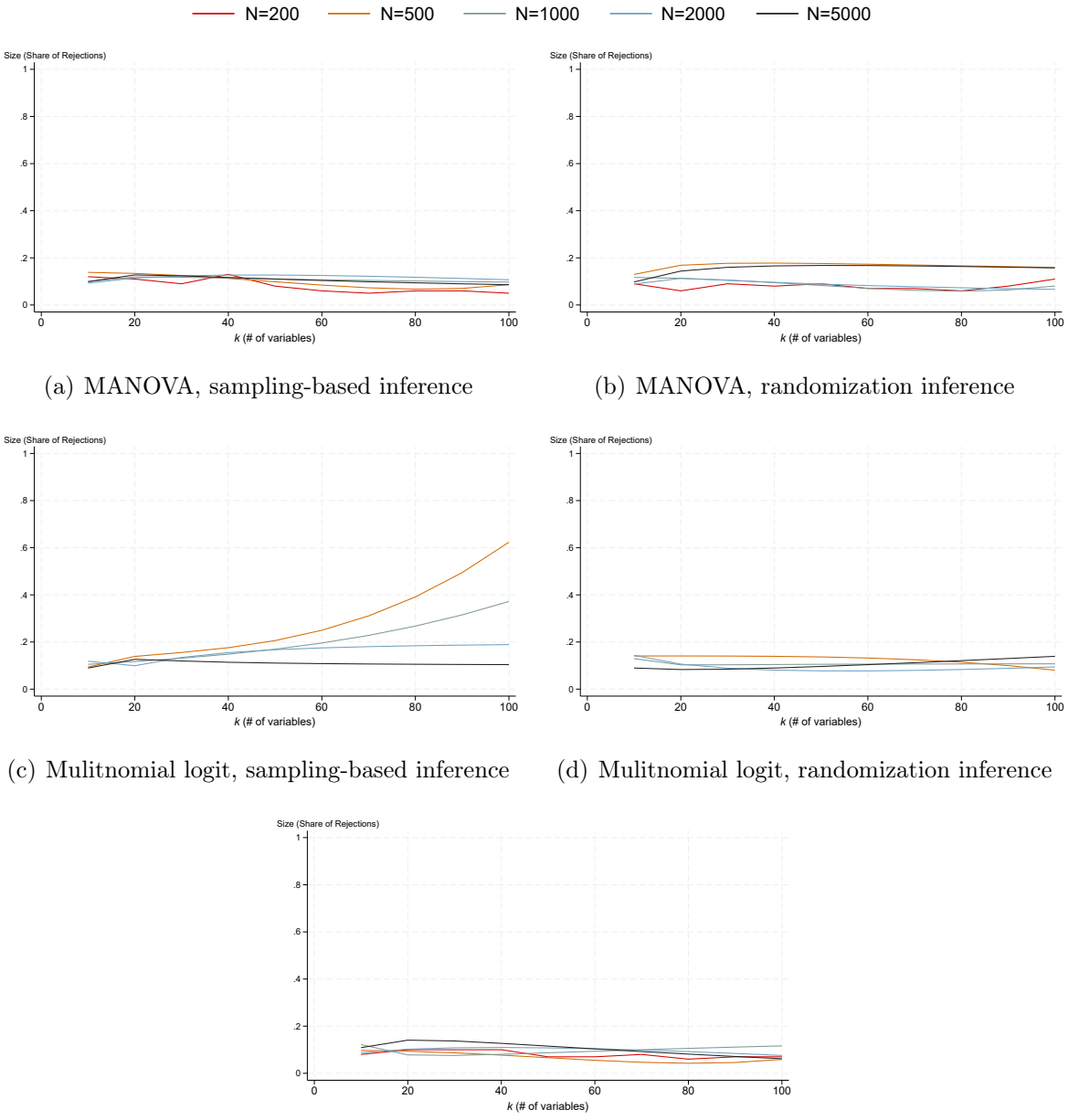
(b) 10% imbalanced variables, 0.20 SD

(c) 20% imbalanced variables, 0.15 SD

(d) 50% imbalanced variables, 0.10 SD

Notes: The figures show the results of simulations with a data generating process with N observations (with $N \in \{200, 500, 1000, 2000, 5000\}$) that are split into 100 clusters of equal size and 20 variables that are normally distributed and correlated within clusters (with an average intra-class correlation coefficient of 0.2). An independent treatment is equal to one in half of the clusters and zero otherwise. Imbalances are then created for a random subset of the k variables and two treatment arms. In Panel (a), one variable is imbalanced, by adding 0.25 to treated observations. In Panel (b), 10% of variables are imbalanced, by adding 0.2 to treated observations. In Panel (c), 20% of variables are imbalanced, by adding 0.15 to treated observations. In Panel (d), 50% of variables are imbalanced, by adding 0.1 to treated observations. To produce each figure, a total of 100 simulated datasets of N observations are generated, for each $N \in 200, 500, 1000, 1500, 2000, 2500, 3000, 3500, 4000, 4500, 5000$. For each N and each method, the light dots on each figure show the share of test statistics below 0.10. For each method, fractional-polynomial prediction lines are used to predict the power of tests for $N \in [200, 5000]$. Results from MANOVA and multinomial logit with sampling-based inference are hidden as these approaches have incorrect test size.

Figure A.13
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

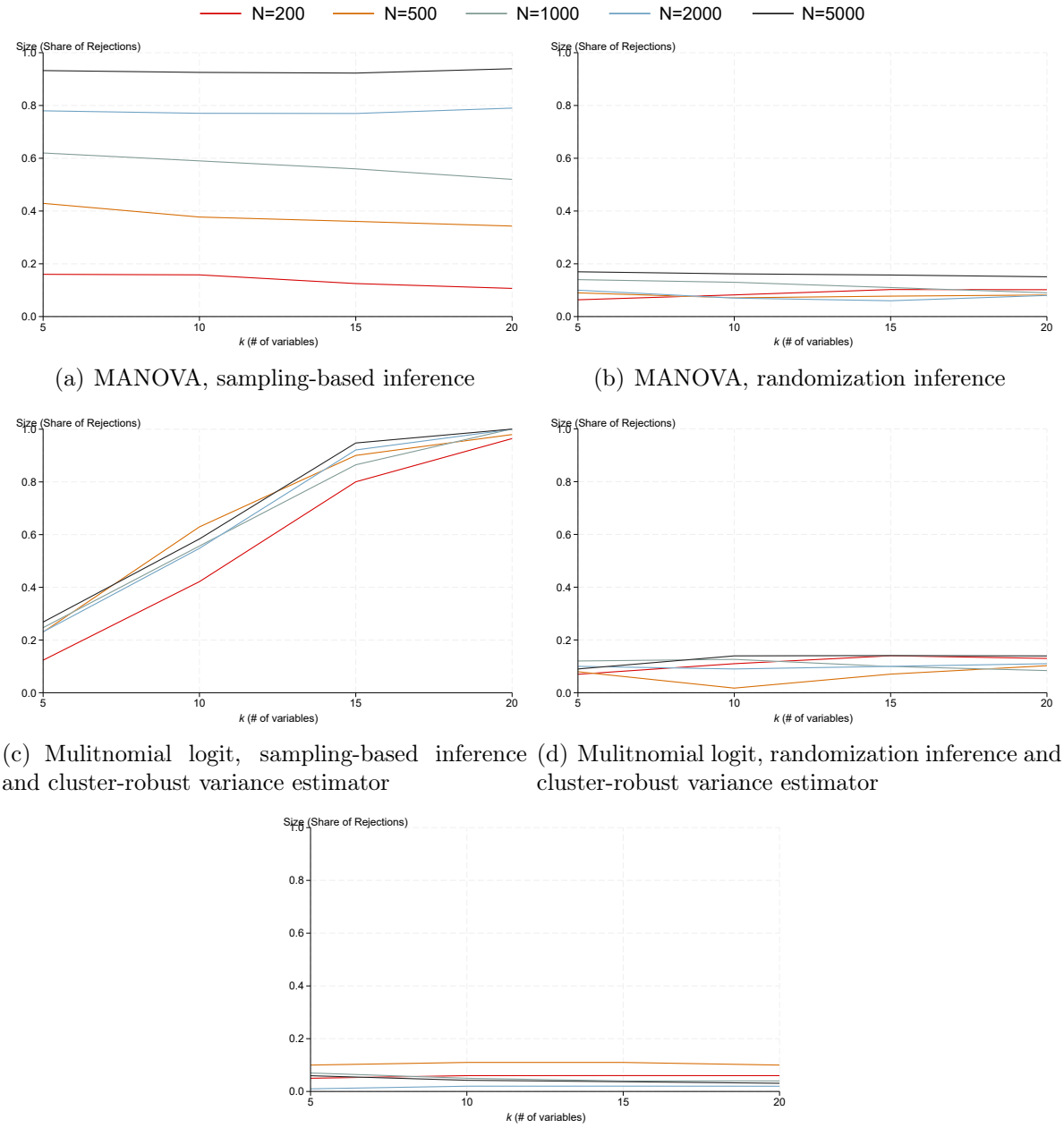
(d) Mulitnomial logit, randomization inference

(e) minimum q-value, HO variance estimator

Notes: The figures show the results of simulations with a data-generating process that creates k independent variables that are normally distributed ~N(0,1) (with $k \in \{10, 20, 30, 40, 50, 60, 70, 80, 90, 100\}$) and a categorical variable identifying four randomly-assigned treatment arms with one quarter of observations in each. Simulated datasets of N observations are generated (with $N \in \{200, 500, 1000, 2000, 5000\}$). For each combination of k and N, we create 500 simulated datasets and run pairwise t-tests for all k variables for each dataset. The minimum q-value is the minimum sharpened q-value obtained after applying the method of Anderson (2008) to the p-values from the pairwise t-tests. For each N, the light dots show the k-specific share of test statistics below 0.10 and the lines show fractional-polynomial predictions for $k \in [10, 100]$. The multinomial logit did not converge for N=200.

Figure A.14

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 cluster-robust variance estimator

(d) Mulitnomial logit, randomization inference and cluster-robust variance estimator

(e) minimum q-value, cluster-robust variance estimator

Notes: The figures show the results of simulations with a data-generating process with N observations (with $N \in \{200, 500, 1000, 2000, 5000\}$) that are split into 100 clusters of equal size, and k variables that are normally distributed and correlated within clusters (with $k \in \{5, 10, 15, 20\}$, and an average intra-cluster correlation coefficient of 0.2). A categorical variable identifies four randomly-assigned treatment arms with one quarter of observations in each. For each combination of k and N, we create 500 simulated datasets and run pairwise t-tests for all k variables. The minimum q-value is the minimum sharpened q-value obtained after applying the method of Anderson (2008) to the p-values from the pairwise t-tests. The light dots show the share of p-values or minimum q-values below 0.10. For each N, fractional-polynomial prediction lines are used to predict the size of tests for $k \in [5, 20]$.

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

	General Info.					Pairwise/Groupwise test				Joint Orthogonality	F-test		
Reference	Bal.	N	n	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	clu.-rob
Ainsworth et al. (2023)	Yes	2,629	12	2	Clu	Yes	No	No	SBI	Yes	OLS	SBI	clu.-rob
Guryan et al. (2023)	Yes	Exp. 1: 2,633	19	2	Ind	Yes	No	No	SBI	Yes	OLS	SBI	clu.-rob
		Exp. 2: 2,710									-	-	-
Macchi (2023)	Yes	Exp 1: 511	25	2	Ind	Yes	No	No	SB& RI	No	-	-	-
		Exp. 2: 238									-	-	-
Buchmann et al. (2023)	Yes	26,408	18	3	Clu	Yes	No	No	SBI	Yes	OLS	SBI	clu.-rob
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	clu.-rob
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	clu.-rob
Tungodden and Willén (2023)	No	-	-	-	-	-	-	-	-	-	-	-	-

	General Info.					Pairwise/Groupwise test				Joint Orthogonality F-test			
Reference	Bal.	N	n	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	-	-	-	-	-	-	-	-	-	-	-	-
Muralidharan, Niehaus, and Sukhtankar (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	clu.-rob
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	clu.-rob
Hussam et al. (2022*b*)	Yes	754	27	3	Ind	Yes	No	No	No	Yes	OLS	SBI	clu.-rob
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	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	-	-	-
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	-	-	-

	General Info.					Pairwise/Groupwise test				Joint Orthogonality F-test			
Reference	Bal.	N	n	T+C	Design	t	F	Nor.Diff	Inf.	J.F	Model	Inf.	VE
McKenzie and Puerto (2021)	Yes	3,537	14	3	Clu	Yes	No	No	SBI	Yes	OLS	SBI	clu.-rob
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	clu.-rob
Angrist, Caldwell, and Hall (2021)	Yes	1,031	11	3	Ind	Yes	No	No	SBI	Yes	ML	SBI	clu.-rob
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	No	Yes	OLS	SBI	HO
Dube et al. (2021)	Yes	504	17	3	Clu	Yes	No	No	No	Yes	ML	SBI	clu.-rob
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	-	-	-