Asymptotically Robust Permutation-based Randomization Confidence Intervals for Parametric OLS Regression

Alwyn Young
London School of Economics
November 2023

Abstract

Randomization inference provides exact finite sample tests of sharp null hypotheses which fully specify the distribution of outcomes under counterfactual realizations of treatment, but the sharp null is often considered restrictive as it rules out unspecified heterogeneity in treatment response. However, a growing literature shows that tests based upon permutations of regressors using pivotal statistics can remain asymptotically valid when the assumption regarding the permutation invariance of the data generating process used to motivate them is actually false. For experiments where potential outcomes involve the permutation of regressors, these results show that permutation-based randomization inference, while providing exact tests of sharp nulls, can also have the same asymptotic validity as conventional tests of average treatment effects with unspecified heterogeneity and other forms of specification error in treatment response. This paper extends this work to the consideration of interactions between treatment variables and covariates, a common feature of published regressions, as well as issues in the construction of confidence intervals and testing of subsets of treatment effects.

---

*I am grateful to Guido Imbens, Taisuke Otsu and anonymous referees for helpful comments.

I. Introduction

Randomized experiments have achieved prominence as a key method in the credible identification of causal mechanisms in economics (Duflo et al 2007), and randomization inference has been advocated as a means of accurately testing null hypotheses regarding experimental outcomes (Athey & Imbens 2017, Young 2019). Randomization inference provides exact finite sample tests of sharp null hypotheses which fully specify the distribution of outcomes under counterfactual realizations of treatment, but its accuracy in testing average treatment effects with unspecified heterogeneity and other forms of specification error is less clear. In many cases, however, experimental potential outcomes involve the permutation of treatment regressors. Here a growing literature has shown that when based on asymptotically pivotal test statistics[1] permutation-based tests can share the asymptotic validity of conventional tests despite violations of the assumption regarding the permutation invariance of the data generating process used to motivate the test.[2] Janssen (1997) showed that studentized permutation tests of the equality of means across two samples are asymptotically exact even when the assumption of identical distributions motivating the permutation test is false, and Chung and Romano (2016) obtained similar results in a multi-parameter extension. Outside of studentization, prepivoting (Chung and Romano 2016) and martingale transforms (Chung and Olivares 2021) have also been found to give asymptotic accuracy of permutation tests in two sample problems. From the perspective of ordinary least squares (OLS) regression, these results concern binary treatment vs control comparisons, but in a major generalization DiCiccio and Romano (2017) show that studentized permutation tests are asymptotically equivalent to conventional heteroskedasticity robust Wald tests for testing treatment effects using iid treatment measures and ancilliary covariates defined on the real numbers. This establishes the asymptotic accuracy of permutation-based OLS randomization inference based upon a sharp null in the presence of heterogeneous treatment effects in iid environments with real experimental treatment measures as regressors.

This paper extends the DiCiccio and Romano result in several areas, with an eye to the regression environments encountered in published randomized experiments. First, it explicitly addresses interactions between treatment measures and covariates, both broadly defined on the reals, in published work a frequent characteristic[3] of often central importance[4], but not considered in DiCiccio and Romano. In such

[1]That is, those whose asymptotic distribution does not depend upon an unknown parameter.

[2]The precise meaning of the terms "randomization test" and "permutation test" is an object of disagreement with, for example, Edgington & Onghena (2007, p. 1) stating the former are a subclass of the latter, Lehman and Romano (2022, p. 831) stating the converse, and Hemerik & Goeman (2021) arguing neither is a subset of the other. This paper uses randomization inference and tests to refer to procedures motivated by a consideration of potential experimental outcomes, focusing on cases where these outcomes are the permutation of treatment regressors. An exact test based upon permutation of variables can be motivated without reference to an experimental procedure, e.g. via population sampling of exchangeable random variables, as is done in the papers cited above.

[3]Of the 53 published papers based upon randomized experiments surveyed in Young (2019), 34 interact treatment with non-treatment covariates in on average .46 of their estimating equations. Of the 39 of these with OLS regressions analyzed further below, 21 interact treatment with covariates in on average .40 of their OLS regressions.

[4]As examples: (i) Oster and Thornton (2011) examine the impact on school attendance of the provision of

cases a simple minded application of DiCiccio and Romano's methods, permuting the regressor made up of treatment interacted with a covariate, may allow for asymptotically accurate inference. Finite sample exactness, however, requires that the permuted regressor be independent of unpermuted ancilliary variables (DiCiccio & Romano 2017, p. 1216), a condition unlikely to be satisfied when the permuted regressor involves interaction with a non-treatment covariate. By focusing on potential experimental outcomes rather than permutation of regressors per se, i.e. by permuting treatment alone and creating new regressor values based upon its interaction with covariates that need not equal the permuted values of the original variable, this paper provides asymptotically robust permutation-based randomization inference methods for treatment-covariate interactions that remain finite sample exact in tests of sharp nulls.[5]

Second, the asymptotic accuracy of permutation-based inference is proven more generally for independently but not-necessarily identically distributed (inid) data while relaxing assumptions on the error term. Experimental data are often drawn from disparate regions[6] and are unlikely to be iid, as assumed by DiCiccio and Romano. When conducted in a single locale, experiments often apply treatment randomly to participants arriving at field or laboratory locations or participants contacted in home visits[7] and the characteristics of people who arrive at a location or are found at home are unlikely to be the same at different times of day or days of the week, producing inid data.[8] Furthermore, while DiCiccio and Romano assume that the mean of the error term is zero conditional on the regressors,[9] the analysis below merely assumes that the error term is uncorrelated with the regressors. This validates the use of the test in broader circumstances with specification error, such as the existence of non-linear effects that are

---

menstrual products to girls in Nepal in specifications that include the interaction of treatment with the pupil being on their period; (ii) Aker et al (2012) examine the impact on pupil outcomes of adult mobile phone education in Niger in diff-in-diff specifications with treatment interacted with dummies for post-experimental time periods.

[5]The permutation procedures used here equally motivate exact tests of sharp nulls for observational data whose distribution is exchangeable and independent of non-permuted covariates. Other frameworks, such as normal homoskedastic errors, also motivate exact tests. Lei & Bickel (2021) provide a quick survey of these as well as introducing a novel cyclic permutation test for exact inference given exchangeable errors. Tests based upon the permutation of OLS residuals or dependent variables, rather than regressors as in this paper, have been proposed (Anderson & Robinson 2001 provide a survey), but these are not finite sample exact and with heteroskedastic errors may not even be asymptotically valid (DiCiccio & Romano 2017).

[6]For example, Thornton (2008) investigated the demand for and effects of learning HIV status across north, central and south Malawi, which differ systematically in their ethnicity and religion.

[7]As an example of the former, Cai et al (2009) investigated saliency by randomly assigning restaurant arrivals in China to tables with different menu setups, while, as an example of the latter, Ashraf et al (2010) investigated the impact of prices on health product demand using door-to-door marketing in Zambia. Inid data are likely to arise even when prospective participants are assigned time slots, as often occurs in campus and field experiments, as the time slot a participant is actually able to attend will be a function of their characteristics.

[8]Although chronological information is rarely reported in experimental data, Cai et al (2009) give the time of day the restaurant bill was paid. With time denoting the 24 hour clock rescaled to 0 to 1, regressing the ln total restaurant bill on sine(π*time) I get a statistically significant negative value (t-stat -4.3 to -4.9 and -6.0 to -9.1, depending upon clustering, for the two public use data sets), as meals paid in the mid-night have the highest bills.

[9]In the case where there are non-permuted non-treatment covariates, as examined in this paper. When all regressors are permuted, they are able to make the stronger zero correlation assumption made here.

orthogonal to the linear effects estimated in the regression, where the zero conditional mean is violated but OLS remains consistent. These extensions broaden the applicability of the permutation procedures given here and (equivalently, when there are no treatment-covariate interactions) in DiCiccio and Romano. [10]

Third, this paper provides a novel approach to testing subsets of treatment measures. In randomized experiments multiple treatment measures are very rarely randomized and administered independently of each other.[11] Consequently, the counterfactual distribution of outcomes under a sharp null depends upon the null for all treatment measures. Thus, while permutation tests of individual coefficients are asymptotically identical to conventional tests, depending only upon the null for that treatment measure, in the finite sample they depend upon the null for all treatment measures in the regression. D'Haultfoeuille and Tuvaandorj (2022) address this issue by proposing that individual treatment measures be permuted within strata created by fixed values of other treatment, producing a valid subset of potential outcomes, an approach used in Young (2019) to provide alternative randomization p-values. In the finite sample this approach is not helpful in regressions with covariate interactions, as there are no possible permutations of treatment within values of treatment interacted with covariates or vice versa, and asymptotically has more limited validity than OLS or other forms of randomization inference, as it requires the covariance of errors and regressors to be zero in every stratum created by other treatment variables, which may be violated in the presence of heterogeneous treatment effects and correlated treatment measures. As an alternative, I propose calculating the maximum randomization inference p-value across all possible nulls for untested measures, producing a test that in both finite samples and asymptotically depends only upon the null for the desired subset. This approach can be applied in all environments, ensures control of the rejection probability below nominal level in tests of sharp nulls, retains the broad asymptotic validity of OLS, and in practice appears to provide 80 to 90% of the power of conventional inference and only slightly less than that found using other-treatment-stratification (where the latter can be applied).

Finally, this paper derives a number of practical results and techniques. The conditions needed for asymptotic accuracy in the very diverse settings found in empirical practice, where practitioners typically

[10]It has been brought to my attention that Zhao & Ding (2021), after the initial circulation of this paper in 2020, provide asymptotic results for specifications which include treatment-covariate interactions. Their paper (i) is limited to binary treatment, while this paper concerns treatment defined broadly on the reals; (ii) includes treatment-covariate interactions, following the suggestion of Lin (2013), but only as a means of improving power for inference on a single binary treatment and provides no results for inference on the treatment-covariate interactions themselves, while this paper provides procedures & results for testing both treatment regressors and their interactions with covariates; (iii) always tests the sharp null of no treatment effects using the coefficient on binary treatment alone without alerting readers to the fact (discussed below) that in finite samples this is actually a joint test of sharp nulls on both treatment and its interactions but asymptotically only a test of the coefficient on binary treatment, with consequent issues regarding finite sample null rejection probabilities and asymptotic power against alternatives, whereas this paper addresses the applied econometrician's desire in a multivariate regression model to maintain control of size and maximize power in separate tests of different elements of treatment.

[11]As noted below, of 2500+ treatment measures appearing in multi-treatment regressions in published papers, I find that only 94 measures were randomized independently of other treatment in the regression.

(see Young 2019) stratify and group treatment and use homoskedastic, heteroskedastic and clustered (at, below, above and across treatment groupings) covariance estimates are all covered (in the on-line appendix). While earlier work has focused on tests of zero average treatment effects alone, this paper shows how randomization based confidence intervals can be constructed analytically avoiding costly and potentially fruitless or inaccurate line searches (as these confidence intervals may be non-convex and of infinite width). Similarly, analytical techniques are developed to calculate maximum p-values across possible nulls for untested treatment measures along opposite rays of infinite length, thereby simplifying the search for maximum p-values in subset tests. Finally, different randomization inference techniques are applied and compared in a practical sample of 3000+ treatment measures appearing in 39 published papers, with results that confirm characteristics and patterns seen in Monte Carlos.

The proofs below merge results concerning the asymptotic distribution of permutation statistics of Wald & Wolfowitz (1944), Noether (1949) and Hoeffding (1951) with White's (1980) proof of the asymptotic accuracy of conventional inference using heteroskedasticity robust standard errors. Given White's assumptions on the moments of errors and regressors, conventional Wald statistics using heteroskedasticity robust variance estimates are asymptotically distributed chi-squared, and hence allow for accurate inference when evaluated using that distribution. With minimal additional assumptions, the permutation-based counterfactual distribution of these same Wald statistics is similarly asymptotically distributed chi-squared, even when the restrictions of the sharp null that underlie the calculation of the counterfactual distribution are false. Consequently, using the percentiles of the full permutation distribution to evaluate the conventional Wald statistic is asymptotically analogous to looking up chi-squared tables, and asymptotically yields identical p-values. Random sampling from the permutation distribution allows the calculation of randomization p-values and confidence intervals which are randomly weighted averages of those found using conventional tests at different levels. Consequently, in tests of true nulls, where the conventional coverage probability is asymptotically equal to a 45^{o} linear function of the nominal level, randomization confidence intervals based upon sampling from the permutation distribution have similarly accurate coverage probability. In tests of false nulls, where the conventional coverage probability of the false null is a convex function of nominal level, by Jensen's Inequality randomization confidence intervals have higher coverage probabilities and hence lower power, with the difference vanishing as the number of draws from the permutation distribution goes to infinity.

Within White's (1980) framework of independently but not necessarily identically distributed observations, the additional assumptions are: (1) treatment variables vary and are not colinear; (2) variables interacted with treatment appear separately as regressors in their own right; and (3) treatment, errors and interaction variables in combination have sufficiently high moments. These requirements are typically satisfied in experimental settings. Administered treatment measures vary, are not colinear with each other, and are drawn from distributions with moments of all order, satisfying (1) and (3). Variables interacted with treatment are usually entered separately in regressions to estimate their separate effects,

satisfying (2). White's framework is extended in the on-line appendix to allow for treatment applied to groupings of observations, correlations across observations of errors and regressors and the use of clustered standard errors (if the size of correlated observational groupings is bounded), as well as stratified treatment (if the first and second moments of treatment are asymptotically balanced across strata).[12]

Two key elements underlying the results are worth emphasizing. The use of Wald or studentized test statistics is crucial, because permutation of treatment variables breaks the correlation between the variance of the residual and the regressor introduced by treatment effect heterogeneity and specification error. Consequently, the asymptotic variance across permutations of coefficients is typically different than that of the conventional coefficient across the realizations of the data, so use of the distribution of permuted coefficients to evaluate the significance of the conventional coefficient estimate is usually inaccurate. Dividing permuted coefficient estimates by their standard error estimates corrects for the way in which permutation breaks the connection between regressors and the variance of residuals, producing a test statistic that is asymptotically distributed chi-squared. If the conventional test statistic is similarly distributed, i.e. the conventional OLS variance estimate is asymptotically correct, randomization inference asymptotically produces identical p-values. The key here is the use of a covariance matrix that asymptotically accurately estimates the differing correlation between regressors and residuals in the original data and the permutation distribution.

The inclusion of variables interacted with treatment separately in the regression plays a crucial role in ensuring asymptotic power identical to that of the conventional test. When the null underlying the counterfactual calculation of outcomes is false, the counterfactual data generating process has a mean bias which, absent assumption (2) above, results in permuted Wald statistics having an asymptotic non-central chi-squared distribution. When used to evaluate the experiment's test statistic, the higher tail probabilities of this distribution result in very low rejection rates. Including variables interacted with treatment as regressors in their own right partials out the bias, ensuring that Wald statistics on treatment effects are asymptotically distributed chi-squared, producing p-values and power identical to those of the conventional test.

The paper proceeds as follows: Section II lays out a general parametric linear regression model that encompasses the treatment-covariate interactions often found in published work in a specification that allows for linear treatment effect heterogeneity and non-linear specification error, stating the central theoretical results of the paper while further clarifying how they broaden DiCiccio and Romano's results and depend upon the given assumptions. Section III then uses Monte Carlos to illustrate the importance of White's moment conditions in conventional and randomization inference alike and the role assumption (2)

[12]The extension to stratification assumes that the clustered/robust (White 1980) standard error estimates are valid as treatment effects do not vary systematically by strata. Bugni, Canay & Shaikh (2018) show the need for a different standard error estimate when average treatment effects vary by strata and treatment balance is greater than that achieved by random sampling from a treatment distribution, proving that with test statistics based on this estimate randomization inference for binary treatment is again asymptotically robust to treatment heterogeneity.

plays in ensuring adequate randomization power, both in tests of heterogeneous and sharp treatment effects. Section IV addresses practical issues, such as algorithms for the construction of confidence intervals and differing methods for testing subsets of treatment effects in finite samples, highlighting where different approaches do or don't provide finite sample exact tests of sharp nulls or asymptotically valid tests of heterogeneous treatment effects. Monte Carlos show how other-treatment-stratification in subset testing provides finite sample exactness at the expense of narrower asymptotic validity in tests of heterogeneous treatment effects when treatment measures are correlated, while calculating maximum p-values over all possible nulls provides conservative control of the null rejection probability for sharp nulls while retaining the broader asymptotic validity of OLS. Section V compares randomization and conventional confidence intervals and p-values in a large practical sample of published papers, finding patterns which closely parallel those found in the Monte Carlos and, based on these, providing a practical summary and comparison of subset testing methods. The appendix below provides proofs of base results, while the on-line appendix lays out the extensions to grouped treatment, clustering and stratification. The programme *randcmdci*, available on the authors' website or in Stata through *ssc install*, calculates the parametric OLS randomization confidence intervals & p-values discussed in this paper for Stata users.

II. Permutation-based Randomization Inference in a Parametric Regression Model

This paper focuses on permutation-based randomization inference in a parametric regression model that encompasses the range of specifications typically encountered in applied work, namely:

$$(2.1)\quad \mathbf{y} = \mathbf{X_W}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon},$$

where $\mathbf{X_W} = \mathbf{X}\bullet\mathbf{W}$ and $\bullet$ denotes the row by row Kronecker or "face-splitting" product of two matrices, while $\mathbf{y}$ and $\boldsymbol{\varepsilon}$ are N x 1 vectors of outcomes and residuals, $\mathbf{Z}$ and $\boldsymbol{\gamma}$ the N x K matrix of covariates and K x 1 vector of associated parameters, $\mathbf{X}$ an N x P matrix of treatment variables, $\mathbf{W}$ an N x Q matrix of interaction covariates, and $\boldsymbol{\beta}$ the PQ x 1 vector of parameters of interest. This formulation allows for treatment entered simply into the regression ($\mathbf{W} = \mathbf{1}_N$, an N x 1 vector of 1s) or interacted with other non-treatment variables (the columns of $\mathbf{W}$), as is often the case. Treatment may be discrete or continuous and there are no restrictions on the elements of $\mathbf{X}$, $\mathbf{W}$ & $\mathbf{Z}$ other than that they are real numbers and satisfy moment conditions given further below. There may be heterogeneity or non-linear components to the impact of treatment regressors or covariates on the outcome and $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are interpreted as average linear effects,[13] with heterogeneous linear or non-linear aspects, if any such exist, implicitly included in the residuals. For example, with $\mathbf{x}'_{\mathbf{W}i}$ & $\mathbf{z}'_i$ denoting the i^{th} rows of $\mathbf{X_W}$ and $\mathbf{Z}$ and $f_i(\mathbf{x}'_{\mathbf{W}i}, \mathbf{z}'_i)$ some non-linear function, the data generating process for observation i might be

$$(2.2)\quad y_i = f_i(\mathbf{x}'_{\mathbf{W}i}, \mathbf{z}'_i) + \eta_i = \mathbf{x}'_{\mathbf{W}i}\boldsymbol{\beta} + \mathbf{z}'_i\boldsymbol{\gamma} + \varepsilon_i$$
$$\text{where } \varepsilon_i = \mathbf{x}'_{\mathbf{W}i}\boldsymbol{\beta}_i + \mathbf{z}'_i\boldsymbol{\gamma}_i + [f_i(\mathbf{x}'_{\mathbf{W}i}, \mathbf{z}'_i) - \mathbf{x}'_{\mathbf{W}i}(\boldsymbol{\beta} + \boldsymbol{\beta}_i) - \mathbf{z}'_i(\boldsymbol{\gamma} + \boldsymbol{\gamma}_i)] + \eta_i,$$

[13] Linear, that is, in the regressors, which themselves may be non-linear functions of an aspect of treatment. Thornton (2008), for example, randomly assigned monetary incentives to individuals in Malawi to learn their HIV results and used the incentive, its square, and an indicator if it is greater than 0 in many regression specifications.

where $\boldsymbol{\beta}_i$ & $\boldsymbol{\gamma}_i$ represent the observation specific heterogeneity in average linear effects, $f_i(\mathbf{x}'_{\mathbf{W}i}, \mathbf{z}'_i)$ - $\mathbf{x}'_{\mathbf{W}i}(\boldsymbol{\beta}+\boldsymbol{\beta}_i) - \mathbf{z}'_i(\boldsymbol{\gamma}+\boldsymbol{\gamma}_i)$ the non-linear components, and we assume that $E(\varepsilon_i(\mathbf{x}'_{\mathbf{W}i}, \mathbf{z}'_i)) = \mathbf{0}'_{PQ+K}$ (a $PQ+K$ row vector of 0s). This assumption allows for misspecification, with the understanding that we are trying to estimate population average linear effects while recognizing that there might be unspecified but orthogonal heterogeneity and non-linearity. Treatment is randomly applied at the observation level[14] and any of the $N!$ permutations of the rows of $\mathbf{X}$ is equally likely to have occurred, with the N x P matrix $\mathbf{T}$ used to represent one such outcome and $\mathbf{T}_\mathbf{W} = \mathbf{T}\bullet\mathbf{W}$ its interaction with covariates.

The dominant approach to inference in contemporary randomized experiments in the model described above uses conventional t- and Wald tests whose accuracy is validated by asymptotic results based upon characteristics of the data generating process, often motivated as representing random sampling from an infinite population. Within this framework, heterogeneity of variables and effects in the sample are viewed as an underlying characteristic of the population from which the experimental sample is drawn. Asymptotically heteroskedasticity or clustered robust covariance estimates are generally used although these have rejection probabilities that may deviate substantially from nominal levels in small samples, as seen for example in the Monte Carlos below.

An alternative approach to experimental inference, rooted in the history of the experimental literature, treats the allocation of experimental treatment $\mathbf{X}$ as the only source of stochastic variation, with the covariates $\mathbf{W}$ and $\mathbf{Z}$ and characteristics of $\boldsymbol{\varepsilon}$ unrelated to $\mathbf{X}$ taken as given. In particular, randomization tests of "sharp nulls", first advocated by Fisher (1935), use a precise specification of what outcomes would have been under counterfactual realizations of treatment to calculate the distribution of a test statistic. The test we examine in this paper is the standard one that assumes no treatment related heterogeneity or misspecification, so that $\mathbf{Z}\boldsymbol{\gamma}+\boldsymbol{\varepsilon} \mid \mathbf{X} = \mathbf{Z}\boldsymbol{\gamma}+\boldsymbol{\varepsilon} \mid \mathbf{T}$ for all alternative treatment allocations $\mathbf{T}$.[15] For the null hypothesis $\boldsymbol{\beta} = \boldsymbol{\beta}_0$, this allows the calculation of the counterfactual outcome under an alternative treatment allocation $\mathbf{T}$ as

$$(2.3)\quad \mathbf{y}(\mathbf{T}, \boldsymbol{\beta}_0) = \mathbf{y} + (\mathbf{T}_\mathbf{W} - \mathbf{X}_\mathbf{W})\boldsymbol{\beta}_0.$$

As the distribution of potential outcomes is uniform across the permutations $\mathbf{T}$ of $\mathbf{X}$, comparison of a test statistic for the experimental outcome $\tau(\mathbf{X},\boldsymbol{\beta}_0)$ with the percentiles of the distribution of $\tau(\mathbf{T},\boldsymbol{\beta}_0)$ across permutations $\mathbf{T}$ provides a finite sample exact test of the joint null that $\boldsymbol{\beta} = \boldsymbol{\beta}_0$ and $\mathbf{Z}\boldsymbol{\gamma}+\boldsymbol{\varepsilon} \mid \mathbf{X} = \mathbf{Z}\boldsymbol{\gamma}+\boldsymbol{\varepsilon} \mid \mathbf{T}$, regardless of the characteristics of $\mathbf{Z}$, $\mathbf{W}$ and $\boldsymbol{\varepsilon}$. Calculation of the entire distribution is not needed, as tests based on random sampling from that distribution are (under the joint null) equally exact.

As a concrete example, consider the Wald statistic for the conventional test that $\boldsymbol{\beta} = \boldsymbol{\beta}_0$, the test statistic used in this paper. The initial OLS coefficient and variance estimates $\hat{\boldsymbol{\beta}}$ & $\hat{\mathbf{V}}(\hat{\boldsymbol{\beta}})$ from the

[14]The on-line appendix examines the case of treatment applied to groupings of observations or stratified.

[15]In terms of (2.2), this implies that $\boldsymbol{\beta}_i = \mathbf{0}_{PQ}$ and f_i is a function of $\mathbf{z}_i$ alone. In principle, a sharp null could specify a precise set of heterogeneous effects (e.g. $\boldsymbol{\beta}_i$), but there is usually nothing to guide or discipline this choice. For completeness, one might also specify that $\mathbf{W} \mid \mathbf{X} = \mathbf{W} \mid \mathbf{T}$, but this is implicit in the statement that covariate values are taken as given.

regression of $\mathbf{y}$ on $(\mathbf{X_W},\mathbf{Z})$ are first used to calculate the Wald statistic $\tau(\mathbf{X},\boldsymbol{\beta}_0) = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)'\hat{\mathbf{V}}(\hat{\boldsymbol{\beta}})^{-1}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$. The treatment variables $\mathbf{X}$ are then permuted to $\mathbf{T}$, interacted with covariates to form $\mathbf{T_W} = \mathbf{T}\bullet\mathbf{W}$, and the regression of counterfactual outcome $\mathbf{y}(\mathbf{T},\boldsymbol{\beta}_0)$ on $(\mathbf{T_W},\mathbf{Z})$ is used to produce the coefficient and covariance estimates $\hat{\boldsymbol{\beta}}(\mathbf{T},\boldsymbol{\beta}_0)$ & $\hat{\mathbf{V}}(\hat{\boldsymbol{\beta}}(\mathbf{T},\boldsymbol{\beta}_0))$ and Wald statistic $\tau(\mathbf{T},\boldsymbol{\beta}_0) = (\hat{\boldsymbol{\beta}}(\mathbf{T},\boldsymbol{\beta}_0) - \boldsymbol{\beta}_0)'\hat{\mathbf{V}}(\hat{\boldsymbol{\beta}}(\mathbf{T},\boldsymbol{\beta}_0))^{-1}(\hat{\boldsymbol{\beta}}(\mathbf{T},\boldsymbol{\beta}_0) - \boldsymbol{\beta}_0)$. D random draws with replacement are made from the distribution of the permutations $\mathbf{T}$ of $\mathbf{X}$. With G and E denoting the number of times that $\tau(\mathbf{T},\boldsymbol{\beta}_0)$ is either greater than or equal to $\tau(\mathbf{X},\boldsymbol{\beta}_0)$, and u a random draw from the [0,1] uniform distribution, the randomization p-value given by

$$(2.4)\quad p_R = \frac{G + u(E+1)}{D+1},$$

is uniformly distributed under the sharp null.[16] D need not be "large", i.e. approximate the full distribution, although power increases with D, as shown below. However, while finite sample exactness is desirable, the sharp null and its assumption that $\mathbf{Z}\boldsymbol{\gamma}+\boldsymbol{\varepsilon} \mid \mathbf{X} = \mathbf{Z}\boldsymbol{\gamma}+\boldsymbol{\varepsilon} \mid \mathbf{T}$ is often seen as demanding.

This paper examines permutation-based randomization tests from the population sampling point of view, treating heterogeneity of regressors and parameters as part of the data generating process, which can be conceptualized as taking a random sample from an infinite population and then randomly allocating treatment across subjects. Given certain moment conditions, the use of the distribution of $\tau(\mathbf{T},\boldsymbol{\beta}_0)$ based on counterfactual outcomes $\mathbf{y}(\mathbf{T},\boldsymbol{\beta}_0)$ under the sharp null to evaluate the conventional Wald statistic $\tau(\mathbf{X},\boldsymbol{\beta}_0)$ is shown to be asymptotically identical to conventional inference when the latter is asymptotically accurate, even when the assumption $\mathbf{Z}\boldsymbol{\gamma}+\boldsymbol{\varepsilon} \mid \mathbf{X} = \mathbf{Z}\boldsymbol{\gamma}+\boldsymbol{\varepsilon} \mid \mathbf{T}$ underlying the calculation of $\mathbf{y}(\mathbf{T},\boldsymbol{\beta}_0)$ in (2.3) is in fact false. Thus, this permutation-based test has the same asymptotic validity of conventional tests of average linear treatment effects, while simultaneously providing, from the experimentalist perspective, exact finite sample tests of sharp nulls conditional on the given values of $\mathbf{W}$, $\mathbf{Z}$ and $\boldsymbol{\varepsilon}$.[17]

The main results of this paper (proven in the appendix below) are as follows:

(R1) Given White's (1980) assumptions W1 - W4 and the additional assumptions A1 - A3, all detailed below, for any $\boldsymbol{\beta}_0$ in a finite $\sqrt{N}$ neighbourhood of $\boldsymbol{\beta}$, i.e. such that $N(\boldsymbol{\beta} - \boldsymbol{\beta}_0)'(\boldsymbol{\beta} - \boldsymbol{\beta}_0) < \Delta$ (a constant) < ∞, the Wald statistic $\tau(\mathbf{T},\boldsymbol{\beta}_0)$ calculated using the heteroskedasticity robust covariance estimate is asymptotically distributed chi-squared with PQ degrees of freedom and in probability converges to the value for the true null $\boldsymbol{\beta}_0 = \boldsymbol{\beta}$

$$(2.5)\quad \tau(\mathbf{T},\boldsymbol{\beta}_0) \overset{d(\mathbf{T})|a.s.(\mathbf{X_W},\mathbf{Z},\boldsymbol{\varepsilon})}{\rightarrow} \chi^2_{PQ} \;\;\&\;\; \tau(\mathbf{T},\boldsymbol{\beta}_0) - \tau(\mathbf{T},\boldsymbol{\beta}) \overset{p(\mathbf{T})|a.s.(\mathbf{X_W},\mathbf{Z},\boldsymbol{\varepsilon})}{\rightarrow} 0,$$

where $\overset{d(\mathbf{T})|a.s.(\mathbf{X_W},\mathbf{Z},\boldsymbol{\varepsilon})}{\rightarrow}$ and $\overset{p(\mathbf{T})|a.s.(\mathbf{X_W},\mathbf{Z},\boldsymbol{\varepsilon})}{\rightarrow}$ denote convergence as $N \rightarrow \infty$ in distribution and probability across the permutations $\mathbf{T}$ of $\mathbf{X}$ almost surely given the realization of the data $(\mathbf{X_W},\mathbf{Z},\boldsymbol{\varepsilon})$.

(R2) Since locally $\tau(\mathbf{T},\boldsymbol{\beta}_0)$ ceases to be a function of $\boldsymbol{\beta}_0$, asymptotically in probability the 1 - α randomization confidence interval (*RCI*) based upon the set of nulls that are not rejected at the α level

[16] For proofs see, for example, Hoeffding (1952) or the on-line appendix of Young (2019).

[17] Thus, following the terminology of Abadie et al (2020) who usefully review these concepts, exact inference under design-based uncertainty provides asymptotically accurate inference under sampling based uncertainty.

using the p-value in (2.4) and D draws from the permutation distribution is given by

$$(2.6)\quad RCI(1-\alpha,D)=\{\boldsymbol{\beta}_0:(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)'\mathbf{V}(\hat{\boldsymbol{\beta}})^{-1}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)\le I_1\tau_{i^*}+(1-I_1)\tau_{i^*-1}\},$$

where i^* is the smallest integer greater than or equal to $(D+1)(1-\alpha)$, I_1 is the indicator function for the event $u\ge i^*-(D+1)(1-\alpha)$ with u uniformly distributed over [0,1], $\tau_1<...<\tau_D$ are the ordered values of $\tau(\mathbf{T},\boldsymbol{\beta}_0)$ and where if needed we define τ_0 as 0 and τ_{D+1} as ∞.

(R3) As White's assumptions ensure that for the conventional Wald statistic calculated using the heteroskedasticity robust covariance estimate

$$(2.7)\quad \tau(\mathbf{X},\boldsymbol{\beta}_0=\boldsymbol{\beta})\overset{d(\mathbf{X_W},\mathbf{Z},\boldsymbol{\varepsilon})}{\rightarrow}\chi^2_{PQ},$$

where $d(\mathbf{X_W},\mathbf{Z},\boldsymbol{\varepsilon})$ denotes convergence in distribution across the realizations of $(\mathbf{X_W},\mathbf{Z},\boldsymbol{\varepsilon})$, randomization confidence intervals based upon (2.6) asymptotically cover the true parameters with probability 1-α, i.e. with $\Pr\{a\}$ denoting the probability of event a,

$$(2.8)\quad \lim_{N\to\infty}\Pr\{\boldsymbol{\beta}\in RCI(1-\alpha,D)\}=1-\alpha.$$

(R4) Asymptotically the frequency with which the randomization confidence interval does not cover (rejects) false nulls $\boldsymbol{\beta}_0\neq\boldsymbol{\beta}$ in a finite $\sqrt{N}$ neighbourhood of $\boldsymbol{\beta}$, i.e. such that $N(\boldsymbol{\beta}-\boldsymbol{\beta}_0)'(\boldsymbol{\beta}-\boldsymbol{\beta}_0)<\Delta<\infty$, is strictly lower than that for the conventional confidence interval (*CCI*), but converges to the conventional level as D & $N\to\infty$,[18] i.e.

$$(2.9)\quad \lim_{N\to\infty}[\Pr\{\boldsymbol{\beta}_0(\neq\boldsymbol{\beta})\in RCI(1-\alpha,D)\}-\Pr\{\boldsymbol{\beta}_0(\neq\boldsymbol{\beta})\in CCI(1-\alpha)\}]>0$$
$$\lim_{D,N\to\infty}[\Pr\{\boldsymbol{\beta}_0(\neq\boldsymbol{\beta})\in RCI(1-\alpha,D)\}-\Pr\{\boldsymbol{\beta}_0(\neq\boldsymbol{\beta})\in CCI(1-\alpha)\}]=0.$$

(R5) As D & $N\to\infty$ the randomization p-value using the Wald statistic to test any null $\boldsymbol{\beta}_0$ in a finite $\sqrt{N}$ neighbourhood of $\boldsymbol{\beta}$ using D draws given the data, $p_R(\boldsymbol{\beta}_0,D|\,\mathbf{X_W},\mathbf{Z},\boldsymbol{\varepsilon})$, almost surely converges to the conventional p-value for the same null based upon the Wald statistic for the data, $p_C(\boldsymbol{\beta}_0|\mathbf{X_W},\mathbf{Z},\boldsymbol{\varepsilon})$, that is

$$(2.10)\quad \lim_{D,N\to\infty}[p_R(\boldsymbol{\beta}_0,D\,|\,\mathbf{X_W},\mathbf{Z},\boldsymbol{\varepsilon})-p_C(\boldsymbol{\beta}_0\,|\,\mathbf{X_W},\mathbf{Z},\boldsymbol{\varepsilon})]\overset{a.s.(\mathbf{X_W},\mathbf{Z},\boldsymbol{\varepsilon})}{=}0.$$

(R2) - (R5) apply equally to tests of subsets of coefficients, as shown in the appendix. The extension to errors that are homoskedastic or cluster correlated, treatment applied to groupings of observations and stratified treatment is given in the on-line appendix. The requirement that the null $\boldsymbol{\beta}_0$ remain root-N local to $\boldsymbol{\beta}$ implies drifting sequences such as $\boldsymbol{\beta}_0=\boldsymbol{\beta}+\boldsymbol{\delta}_0/N^{1/2}$. With additional conditions, (R1) can be changed to stating that the Wald statistic $\tau(\mathbf{T},\boldsymbol{\beta}_0)$ is asymptotically distributed chi-squared for *any* fixed $\boldsymbol{\beta}_0$, as shown in the on-line appendix. This paper focuses on drifting sequences as stated in (R1) above because (i) such sequences are necessary to ensure that p-values and coverage probabilities for incorrect nulls do not trivially converge to 0; (ii) the proof for fixed $\boldsymbol{\beta}_0$ in the extension to grouped treatment requires clustering of standard errors at levels of aggregation greater than or equal to treatment groupings, which is not always done in practice and is not necessary for (R1)'s extension to grouped treatment and

[18] We use $\lim_{D,N\to\infty}$ to denote the double-limit, i.e. for every $\Delta>0$ there exists an M such that for all D & $N>M$ the absolute value of the expression is less than Δ.

clustering;[19] and (iii) (R1)'s local convergence in probability of $\tau(\mathbf{T},\boldsymbol{\beta}_0)$ to $\tau(\mathbf{T},\boldsymbol{\beta})$, which is not true of fixed $\boldsymbol{\beta}_0$, allows the construction of subset tests that asymptotically are identical to a test based upon knowledge of the true parameter values for untested treatment effects, as described further in Section IV, which covers practical issues in calculating confidence intervals and implementing subset tests in finite samples.

Turning to the assumptions, combining the treatment and non-treatment regressors into more compact notation, our regression model can be described as

$$(2.11)\quad \mathbf{y} = \mathbf{Z}_+\boldsymbol{\gamma}_+ + \boldsymbol{\varepsilon} \quad \text{or (at the observation level)} \quad y_i = \mathbf{z}'_{+i}\boldsymbol{\gamma}_+ + \varepsilon_i,$$

where $\mathbf{Z}_+ = (\mathbf{X}_\mathbf{W}, \mathbf{Z})$ and $\boldsymbol{\gamma}'_+ = (\boldsymbol{\beta}', \boldsymbol{\gamma}')$ denote the full matrix of regressors and parameters and $\mathbf{z}'_{+i}$ the 1 x K_+ row vector representing the i^{th} row of $\mathbf{Z}_+$. In this notation, White (1980) proved that almost surely coefficient estimates converge on true parameter values and N times the heteroskedasticity robust covariance estimate to the covariance matrix of normally distributed $\sqrt{N}(\hat{\boldsymbol{\gamma}}_+ - \boldsymbol{\gamma}_+)$, thereby ensuring that Wald statistics are asymptotically distributed as chi-squared variables, using the following assumptions:

(W1) $(\mathbf{z}'_{+i}, \varepsilon_i)$ is a sequence of independent but not necessarily identically distributed random vectors such that $E(\mathbf{z}_{+i}\varepsilon_i) = \mathbf{0}_{K+}$.

(W2) There exist finite positive constants δ, Δ and γ such that (a) for all i, $E(|\varepsilon_i^2|^{1+\delta}) < \Delta$ and $E(|z_{+ij}z_{+ik}|^{1+\delta}) < \Delta$ for all j, $k = 1...K_+$; (b) $\mathbf{M}_N = \sum_{i=1}^N E(\mathbf{z}_{+i}\mathbf{z}'_{+i})/N$ is non-singular for all N sufficiently large, with determinant $\mathbf{M}_N > \gamma > 0$.

(W3) There exist finite positive constants δ, Δ and γ such that (a) for all i, $E(|\varepsilon_i^2 z_{+ij}z_{+ik}|^{1+\delta}) < \Delta$ for all j, $k = 1...K_+$; (b) $\mathbf{V}_N = \sum_{i=1}^N E(\varepsilon_i^2\mathbf{z}_{+i}\mathbf{z}'_{+i})/N$ is non-singular for all N sufficiently large, with determinant $\mathbf{V}_N > \gamma > 0$.

(W4) There exist finite positive constants δ and Δ such that for all i, $E(|z_{+ij}^2 z_{+ik}z_{+il}|^{1+\delta}) < \Delta$ for all j, k, l $= 1...K_+$ or, equivalently (by Hölder's Inequality), $E(|z_{+ij}^4|^{1+\delta}) < \Delta$ for all $j = 1...K_+$.

In addition to White's W1 - W4, we make use of three additional assumptions

(A1) There exists a finite positive constant γ such that $\mathbf{G}_N = \sum_{i=1}^N E(\mathbf{x}_i\mathbf{x}'_i)/N - \sum_{i=1}^N E(\mathbf{x}_i)/N \sum_{i=1}^N E(\mathbf{x}'_i)/N$ is non-singular for all N sufficiently large with determinant $\mathbf{G}_N > \gamma > 0$.

(A2) Either the matrix $\mathbf{W}$ is part of $\mathbf{Z}$, i.e. the interactions with treatment in $\mathbf{X}_\mathbf{W}$ are entered separately as covariates in the regression, or $\sum_{i=1}^N E(\mathbf{x}_i)/N \to \mathbf{0}_P$.

(A3) There exist finite positive constants θ, θ^* and Δ, with $\theta(1+2\theta^*) > 1$, such that for all i, $q = 1...Q$ and $p = 1...P$, $E(|w_{iq}^2\varepsilon_i^2|^{1+\theta}) < \Delta$ and $E(|x_{ip}^4|^{1+\theta^*}) < \Delta$.

Nothing in W1 - W4 and A1 - A3 requires that the data generating process behind $\mathbf{X}$ is such that all permutations $\mathbf{T}$ of $\mathbf{X}$ are actually equally likely. These assumptions guarantee that the distribution of the Wald statistic across row permutations $\mathbf{T}$ of $\mathbf{X}$ is asymptotically chi-squared, as is the conventional Wald statistic of the original regression. Using random draws from the distribution of $\tau(\mathbf{T},\boldsymbol{\beta}_0)$ to evaluate

[19] Provided the properties of ε_i are such that with the homoskedastic or clustered below or across treatment groupings covariance estimate the conventional Wald statistic is distributed chi-squared (see the on-line appendix).

$\tau(\mathbf{X},\boldsymbol{\beta}_0)$ asymptotically allows for accurate inference for the parameters associated with *any* set of regressors $\mathbf{X_W}$ that meet these conditions. It is also unnecessary, as one can more easily evaluate $\tau(\mathbf{X},\boldsymbol{\beta}_0)$ using tables of the chi-squared distribution. However, for data generating processes where all permutations **T** of **X** are actually equally probable, permutation-based randomization tests of sharp nulls are exact in finite samples. Results (R1) - (R5) then confirm that these tests have an asymptotic validity equal to those of conventional tests in tests of population average (heterogeneous) treatment effects when the restrictions of the sharp null are not satisfied. Since it is in this context that use of permutation based tests makes sense, the discussion in this paper focuses on randomized experiments where all permutations **T** of **X** are in fact equally likely.

The procedures, results and assumptions laid out above differ from those of DiCiccio and Romano (2017). DiCiccio & Romano show (2017, Theorem 3.3) that, given assumptions on moments and errors, heteroskedasticity robust Wald statistics arrived at by permuting regressors **X** in an OLS regression model with other covariates **Z** are asymptotically distributed chi-squared. The $\mathbf{X_W}$ in the model above could be taken as their **X** and permuted accordingly (i.e. permuting not merely the treatment, but the product of treatment with covariates). However, as DiCiccio & Romano note, when the **X** and **Z** in their model are not independent, as is most likely the case when their **X** represents the interaction of randomized treatment with participant covariates (our $\mathbf{X_W}$), the resulting test is not guaranteed to be finite sample exact, as the permutation distribution does not equal the sampling distribution of the original data. For the case of treatment interacted with covariates, which was not directly considered by DiCiccio & Romano, our procedure instead permutes treatment **X** to **T**, holding constant the matrix **W**, and calculates and uses $\mathbf{T_W} = \mathbf{T \bullet W}$ in the regression model. This allows the calculation under the sharp null of the counterfactual distribution of randomization outcomes, producing finite sample exact tests. Results (R1) - (R5) show that tests based upon this permutation distribution, which when covariate interactions are present is different than that considered by DiCiccio & Romano, are also asymptotically equivalent to the conventional test when the restrictions imposed by the sharp null are invalid. Practical interest in these results, however, stems from the fact that they are based upon a test that is otherwise finite sample exact for sharp nulls.

For the case where there are no interactions with covariates other than the constant term, $\mathbf{X_W} = \mathbf{X}$, the results above are identical to DiCiccio & Romano (2017), but the regression framework is more general. DiCiccio & Romano's proof requires that the regression include a constant term (the equivalent of A2), bounded fourth moments of iid regressors and errors, and that the conditional expectation of the error equal 0, i.e. $E(\varepsilon_i|\mathbf{z}_{+i}) = 0$. At the expense of requiring greater than fourth moments of regressors, W1-W4 and A1-A3 allow for independently but not identically distributed data and the weaker condition $E(\varepsilon_i\mathbf{z}_{+i}) = \mathbf{0}_{K+}$. As noted in the introduction, experimental data are often drawn from disparate regions and time periods across which participant characteristics are unlikely to be iid. As noted in example (2.2) above, the weaker condition $E(\varepsilon_i\mathbf{z}_{+i}) = \mathbf{0}_{K+}$ allows for specification error that is uncorrelated with the linear regressors and does not compromise the consistency of the OLS estimates of average linear treatment

effects. Finally, W1-W4 and A1-A3 provide some flexibility on moments, since when treatment measures are drawn from distributions with moments of all orders, in the DiCiccio and Romano model where **W** is simply $\mathbf{1}_N$ A3 can be satisfied without further restrictions beyond that given in W2a, i.e. the residuals need have only slightly greater than second moments.

In the context of the typical randomized experiment, the most challenging of the assumptions above is White's assumption W1, that the regressors are independent. In many experiments treatment is independently drawn from a fixed distribution, producing an iid regressor, but in some cases a preselected cumulative distribution function of treatment $F_{\mathbf{x}}$ is allocated to the sample, inducing correlation between the regressors.[20] In this case, we appeal to de Finetti's results regarding the asymptotic distribution of exchangeable random variables, including the reals (Hewitt and Savage 1955). In particular, de Finetti's Theorem implies that if the asymptotic cumulative distribution function of exchangeable random variables converges to a unique $F_{\mathbf{x}}$, then they are asymptotically iid with joint distribution equal to $F_{\mathbf{x}}$ raised to the N^{th} power (O'Neill 2009). The expectation of treatment measures in A1 - A3 allows for heterogeneity across observations, but this is only to indicate the generality of the permutation result. In situations where use of the permutation distribution has desirable finite sample characteristics, i.e. all permutations **T** of **X** are equally probable, the treatment variables are generally (at least asymptotically) iid. Many experiments apply common treatment to observational groupings, e.g. all individuals in a laboratory session, ensuring that observations are neither exchangeable nor asymptotically iid. Extension of the assumptions and results to allow treatment exchangeability only across observational groupings, as well as other cross observation correlations of regressors and residuals, is straightforward as long as one can define independent groupings of observations of bounded size and consider the asymptotics in terms of the number of such groupings going to infinity, as shown in the on-line appendix. The on-line appendix also shows that stratification of treatment is easily accommodated provided the first and second moments of treatment are asymptotically identical across strata.

The additional assumptions A1 - A3 are easily met by most experiments. A1 merely states that treatment measures vary across observations and are not perfectly collinear. Regarding A2, it would be unusual to seek the effect of the interaction of a covariate with treatment without wanting to know the direct effect of the covariate itself and in practice covariates that are interacted with treatment are almost always entered separately in the regression. The alternative assumption that the expectation of treatment is zero is unlikely to hold and is not used below. In the case where **W** is simply $\mathbf{1}_N$, an N vector of ones, A2 amounts to requiring that the regression include a constant term. If treatment measures are drawn from a distribution with bounded support and moments of all orders, as is usually the case, A3 (given A2

[20] As an example of the former, Robinson (2012) randomized weekly income shocks to married couples in Kenya by drawing (with replacement) one of 56 paper slips out of a bag. As an example of the latter, Galliani et al (2011) examined the impact of military service on crime using Argentina's annual national service lottery, which sequentially drew lottery balls numbered 1 through 1000 without replacement and assigned them to young males based upon the sequence of the last three digits of their national ID.

& W3) will be satisfied with $\theta(1+2\theta^{*}) > 1$.

The role A2 and A3 play in generating the results can be illustrated by considering the simplest example, that with a single treatment variable where the data generating process is $\mathbf{y} = \mathbf{x}\beta + \boldsymbol{\varepsilon}$. Under the null $\beta = \beta_0$, following each permutation $\mathbf{t}$ of treatment $\mathbf{x}$, the counterfactual $\mathbf{y}$ is taken to be $\mathbf{y}(\mathbf{t},\beta_0) = \mathbf{y} - \mathbf{x}\beta_0 + \mathbf{t}\beta_0 = \mathbf{x}(\beta-\beta_0) + \mathbf{t}\beta_0 + \boldsymbol{\varepsilon}$, producing coefficient estimates:

$$(2.12)\quad \hat{\beta}(\mathbf{t},\beta_0) = (\mathbf{t}'\mathbf{t})^{-1}\mathbf{t}'\mathbf{y}(\mathbf{t},\beta_0) \Rightarrow \sqrt{N}(\hat{\beta}(\mathbf{t},\beta_0) - \beta_0) = \left(\frac{\mathbf{t}'\mathbf{t}}{N}\right)^{-1}\frac{\mathbf{t}'\mathbf{x}}{N}\sqrt{N}(\beta-\beta_0) + \left(\frac{\mathbf{t}'\mathbf{t}}{N}\right)^{-1}\left(\frac{\mathbf{t}'\mathbf{1}_N}{N}\frac{\mathbf{1}_N'\boldsymbol{\varepsilon}}{\sqrt{N}} + \frac{\mathbf{t}'\mathbf{O}\boldsymbol{\varepsilon}}{\sqrt{N}}\right),$$

where $\mathbf{O} = \mathbf{I}_N - \mathbf{1}_N\mathbf{1}_N'/N$ is the centering matrix. The Wald & Wolfowitz (1944) theorem given in the appendix shows that the distribution across permutations of $\sqrt{N}$ times the correlation of a permuted variable and another sequence, a term such as $\mathbf{t}'\mathbf{O}\boldsymbol{\varepsilon}/\sqrt{N}$ in (2.12), converges to the normal if sufficiently high moments of the two sequences exist. This is similar to the restrictions on tail outcomes in central limit theorems which motivate White's (1980) bounds on higher moments, except more demanding, leading to the addition of A3 to White's assumptions. The mean of the product of a permuted sequence with a fixed sequence is easily shown (in the appendix) to converge in probability to the product of their means given much less stringent conditions, so that terms like $\mathbf{t}'\mathbf{x}/N$ converge to $(\mathbf{t}'\mathbf{1}_N/N)(\mathbf{x}'\mathbf{1}_N/N)$.

Examining (2.12), one sees that if the asymptotic mean of treatment is zero, as would be the case if the alternative assumption in A2 were true, then $\sqrt{N}(\hat{\beta}(\mathbf{t},\beta_0) - \beta_0)$ converges to a normal variable and when suitably normalized by a variance estimate takes on a chi-squared distribution, i.e. the asymptotic distribution of the Wald statistic for the original treatment allocation $\mathbf{x}$. When this condition does not hold, however, the coefficient estimates obtained by regressing counterfactual output $\mathbf{y}(\mathbf{t},\beta_0)$ on $\mathbf{t}$ have a bias induced by the $\sqrt{N}$ deviation of the null from the true parameter value, $\sqrt{N}(\beta-\beta_0)$, and the $\sqrt{N}$ deviation from zero of the finite sample mean of the error term, $\boldsymbol{\varepsilon}'\mathbf{1}_N/\sqrt{N}$. The associated Wald statistics are now distributed non-central chi-squared, do not match the asymptotic sampling distribution of the Wald statistics for treatment allocation $\mathbf{x}$, and hence cannot provide a suitable basis for population inference. The solution to this problem is to include the covariate interaction with treatment, which in this case is the regressor $\mathbf{1}_N$, in the regression, estimating the model $\mathbf{y} = \mathbf{x}\beta + \mathbf{1}_N\alpha + \boldsymbol{\varepsilon}$, even though the value of α is known to be zero. The coefficient estimate for permutation $\mathbf{t}$ is then given by

$$(2.13)\quad \hat{\beta}(\mathbf{t},\beta_0) = (\mathbf{t}'\mathbf{O}\mathbf{t})^{-1}\mathbf{t}'\mathbf{O}\mathbf{y}(\mathbf{t},\beta_0) \Rightarrow \sqrt{N}(\hat{\beta}(\mathbf{t},\beta_0) - \beta_0) = \left(\frac{\mathbf{t}'\mathbf{O}\mathbf{t}}{N}\right)^{-1}\frac{\mathbf{t}'\mathbf{O}\mathbf{x}}{N}\sqrt{N}(\beta-\beta_0) + \left(\frac{\mathbf{t}'\mathbf{O}\mathbf{t}}{N}\right)^{-1}\left(\frac{\mathbf{t}'\mathbf{O}\boldsymbol{\varepsilon}}{\sqrt{N}}\right),$$

and since $\mathbf{t}'\mathbf{O}\mathbf{x}/N = \mathbf{t}'\mathbf{O}\mathbf{O}\mathbf{x}/N$ converges in probability to $(\mathbf{t}'\mathbf{O}\mathbf{1}_N/N)(\mathbf{x}'\mathbf{O}\mathbf{1}_N/N) = 0*0,$ provided $\sqrt{N}(\beta-\beta_0)$ is finite the distribution of the associated Wald statistic will converge to the central chi-squared and for every realization of $\mathbf{t}$ will in probability be identical to that found setting $\beta_0 = \beta$. More generally, including all variables $\mathbf{W}$ interacted with the matrix of treatment $\mathbf{T}$ in the list of covariate regressors $\mathbf{Z}$, as stated in A2, solves these issues of bias and ensures an asymptotic chi-squared distribution & individual values identical to those found setting $\boldsymbol{\beta}_0 = \boldsymbol{\beta}$, as shown in the appendix below.

In (R1) and this paper in general, the emphasis is always on using the distribution of Wald statistics associated with permutations **T** and counterfactual outcomes $\mathbf{y}(\mathbf{T},\boldsymbol{\beta}_0)$ to evaluate the original Wald statistic based on **X** and **y**, rather than using the coefficient distribution to evaluate the original coefficient estimates. Permuting the original treatment **X** to **T** breaks any connection between the variance of the error term and the treatment regressors such as might exist because of heterogeneous treatment effects or specification error as described in (2.2) earlier. Consequently, the variance of the coefficient estimates across permutations of treatment brought about by the finite sample correlation between the permuted treatment vector and the errors, as in the last term of (2.13), is unlikely to be the same as the sampling variance of the original treatment vector **x**.[21] Normalizing by the permuted variance estimate, i.e. studentizing the test-statistic, corrects for how this variance differs between permuted treatment and the original population sample. When the sharp null is true, either the permuted coefficients or permuted Wald statistics provide a basis for accurate tests, but only the Wald test has asymptotic validity when $\mathbf{Z}\boldsymbol{\gamma}+\boldsymbol{\varepsilon} \mid \mathbf{X} \neq \mathbf{Z}\boldsymbol{\gamma}+\boldsymbol{\varepsilon} \mid \mathbf{T}$. This is illustrated in simulations below.

Finally, turning to the intuition for (R2) - (R5), from (R1) the limiting distribution across permutations of treatment of the Wald statistic is that of a chi-squared variable. As the number of draws from the randomization distribution goes to infinity, using the percentiles of this distribution to evaluate the conventional Wald statistic is asymptotically identical to using the chi-squared distribution to evaluate the conventional Wald statistic, i.e. produces the same p-values. This is the basis of (R5). Insofar as conventional tests of true nulls using the variance estimate are asymptotically exact, so will analogous tests using randomization inference, while the power of randomization inference in tests of locally false nulls will be identical, even if the sharp null that motivates the randomization test is false.

When using a finite number of randomization draws to evaluate the test statistic, the i^{th} order statistic of D draws from a chi-squared distribution is a random variable with expected value equal to the $i/(D+1)$ percentile of the chi-squared distribution. If $(1-\alpha)(D+1)$ is not an integer, we can create a random variable whose expected value equals the $1-\alpha$ percentile by probabilistically selecting the order statistic to be either the integer ceiling or floor of $(1-\alpha)(D+1)$, forming the basis for (R2). When the null is true, the asymptotic coverage probability of the conventional confidence interval, based upon the set of nulls such that the Wald statistic is less than the $1-\alpha$ percentile of the chi-squared distribution, equals $1-\alpha$, i.e., the conventional coverage probability is asymptotically a 45^o linear function of the nominal level. Hence, using a random variable whose mean is the chi-squared $1-\alpha$ percentile to determine the level of the confidence interval will have an expected coverage probability of $1-\alpha$, forming the basis for (R3). When testing false nulls, the probability the conventional confidence interval covers the false null is a convex function of the chi-squared $1-\alpha$ percentile. Hence, using a random variable whose mean is the chi-squared

[21] In (2.13), this variance can be shown to be the homoskedastic covariance estimate. This is not generally true, however, because when treatment is interacted with non-constant covariates **W**, as in $\mathbf{T}_\mathbf{W} = \mathbf{T}\bullet\mathbf{W}$, the coefficient estimates for permuted treatment continue to be affected by heteroskedasticity related to **W**.

1-α will by Jensen's Inequality yield a coverage probability greater than 1-α and lower power, unless the variance of the random variable goes to 0, as happens as $D \rightarrow \infty$. This forms the basis for (R4). The appendix below provides formal proofs.

III. Monte Carlos Illustrating the Importance of the Assumptions

This section uses Monte Carlos to highlight the role White's assumptions on moments, the additional randomization specific assumptions A2 and A3, and the sharp null hypothesis play in the finite sample and asymptotic performance of conventional and randomization inference. In Tables I and II below the Monte Carlo data generating process (*dgp*) is given by

$$(3.1)\ y_i = \beta x_i w_i + \varepsilon_i, \text{ with } \varepsilon_i = \beta_i x_i w_i + d|x_i w_i|^{½} + |w_i|^{½} + \eta_i,$$

$$iid\ \beta_i = U(-a,a)\ \&\ x_i = c + t(v), \text{ and } inid\ \eta_i = sin(i)*t(2.1)\ \&\ w_i = sin(i)*t(4.2),$$

where $U(-a,a)$ denotes the uniform distribution across $(-a,a)$, $t(v)$ the t distribution with v degrees of freedom which has all absolute moments up to v, and c and d constants with base values of 0 and 1 respectively. x_i is the iid treatment variable, w_i the inid sample characteristic interacted with treatment and β_i the heterogeneity in the linear treatment effect. The positive square root of $x_i w_i$ and w_i are included in the error term to illustrate how the DiCiccio and Romano (2017) assumption that $E(\varepsilon_i|\mathbf{z}_{+i}) = 0$ can be relaxed in favour of $E(\varepsilon_i \mathbf{z}_{+i}) = \mathbf{0}_{K+}$, thereby allowing the error term to incorporate non-linear specification error that is uncorrelated with the regressors.[22] Similarly, to illustrate the extension of their asymptotic results to inid data, the distributions of η_i and w_i follow a cyclical pattern dictated by the sine function. As noted above, experiments often apply treatment randomly to participants arriving at field or laboratory locations or participants contacted in home visits and these can create cyclically varying distributions.[23]

With both $x_i w_i$ and w_i included in the regressors $\mathbf{z}'_{+i}$ and the error term ε_i, $E(|\varepsilon_i^2|^{1+\delta})$, $E(|z_{+ij}^4|^{1+\delta})$ and $E(|z_{+ij} z_{+ik} \varepsilon_i^2|^{1+\delta})$ are uniformly bounded for some $\delta > 0$ provided v in $t(v)$ is greater than 4, thereby satisfying White's assumptions W1 - W4 earlier for asymptotically accurate heteroskedasticity robust conventional inference. In the tables below the degrees of freedom v for the treatment variable x_i varies between (a) 42.1, so that assumption A3 is met with $\theta(1+2\theta^*) > 1$;[24] (b) 4.21, so that $\theta(1+2\theta^*)$ is only greater than 0 but assumptions W1 - W4 are still met; and (c) .421, so that θ^* is not even positive and neither assumption A3 nor W2 - W4 are met. Assumption A2 is satisfied when the regression includes the interaction variable w_i or the *dgp* mean of x_i (c in equation 3.1) is 0. As noted above, the sharp null is that $\mathbf{Z\gamma+\varepsilon \mid X = Z\gamma+\varepsilon \mid T}$ for all alternative treatment allocations. In the context of (3.1), this requires that $\beta_i \equiv 0$ and $d = 0$, the latter illustrating how any misspecification of the sharp null for linear treatment effects is unallowable if finite sample exact inference is to be guaranteed.

---

[22] $E(\varepsilon_i \mathbf{z}_{+i}) = \mathbf{0}$ follows because the density of w_i is symmetric around 0 and would be true for any power of $|x_i w_i|$ and $|w_i|$ appearing in ε_i. I use the square root rather than greater powers to add non-linearity while not requiring the existence of additional moments beyond those already needed (as reviewed in the next paragraph).

[23] See the footnote on Cai et al (2009) in the introduction above.

[24] In terms of A3, for the *dgp* described above θ is just below min(.05,v/2-1), while θ^* is just below v/4-1.

Table I presents tests of true nulls of an average linear treatment effect β equal to zero, with sample sizes ranging from 20 to 20000 observations, 999 draws of **T** used in the calculation of randomization p-values and confidence intervals, and 10k iterations per *dgp*. In panel (a), which varies v, we see that conventional heteroskedasticity robust rejection probabilities show size distortions in small samples, but converge to nominal value when assumptions W1 - W4 are met. When $v = .421$, however, rejection probabilities show no tendency to converge to nominal value as the sample size increases. When v is sufficiently small the influence of individual observations on coefficient and standard error estimates does not decline with sample size and asymptotic theory simply does not apply.[25]

Turning to randomization inference, the topic of this paper, in panel (a) we see that with $v = 42.1$ or even 4.21, where $\theta(1+2\theta^*)$ is barely above 0, randomization p-values rapidly converge to the heteroskedasticity robust values, while the area covered by the intersection of the confidence intervals divided by the area of their union converges toward 1. While $\theta(1+2\theta^*) > 1$ is necessary to ensure that *all* moments of the randomization distribution converge to the normal, with $v = 4.21$ and $\theta(1+2\theta^*)$ merely > 0 the first four moments are still guaranteed to converge to those of the normal and this is enough, at least with this *dgp*, to produce rejection rates close to nominal value. In contrast, when $v = .421$ p-values and confidence intervals diverge with increases in sample size, and randomization inference shares the large size distortions of conventional inference. However, when the sharp null is true ($\beta_i \equiv 0$ and $d = 0$), indicated by the column with a *, randomization inference is exact even with $v = .421$ and delivers rejection probabilities that are within expected simulation variation from nominal value in all sample sizes. No matter how poorly behaved the data, in finite samples randomization inference provides exact tests of sharp nulls, while for well behaved variables with bounded moments of sufficiently high order it shares the same asymptotic validity as conventional robust inference in the presence of heterogeneous treatment effects. The sharp null, however, is more than a statement about a lack of heterogeneity in the tested linear treatment effect, as emphasized by the column marked with a #, where although $\beta_i \equiv 0$, $d = 1$ and the error term contains the specification error on the treatment effect. Randomization inference in this case is neither finite sample exact nor asymptotically accurate with misbehaving regressors.

Panel (b) of Table I varies the heterogeneity of treatment effects, increasing a in $U(-a,a)$ by orders of magnitude all the way up 5000, while keeping v at 42.1, so that assumptions W1-W4 and A1-A3 all hold. Relative to panel (a), greater heterogeneity and hence heteroskedasticity increase the size distortions of both conventional and randomization methods in the very smallest of samples, but these disappear as conventional rejection rates converge to nominal value and randomization p-values and confidence

[25]The leverage of an individual observation in the regression of **y** on **X** is usually defined as the diagonal element of the hat matrix $\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. For individual coefficient estimates, one can define a similar measure using the hat matrix for the partitioned regression, i.e. using the residuals of one regressor projected on the others. For the residuals of $x_i w_i$ regressed on w_i, in the regressions of panel (a) in Table I as the number of observations increases from 20 to 20k the average across 10k samples of the maximum leverage of an individual observation falls from .39 to .027 when $v = 42.1$, but actually rises from .78 to .85 when $v = .421$.

Table I: Conventional and Randomization Inference in Favourable and Unfavourable Conditions
(Monte Carlos: 10000 iterations per data generating process (*dgp*), following notation of (3.1))

		probability of rejecting true null for β at .05 level										joint distribution of p-values and confidence intervals for β									
		conventional robust inference					randomization inference					correlation between p-values					{CCI}∩{RCI}/{CCI}∪{RCI}				
(a) regression model $y_i = \beta x_i w_i + \gamma w_i + \varepsilon_i$; *dgp* of (3.1) with $c = 0$, $d = 1$, $a = ½$, by v																					
obs	v	42.1	4.21	.421	.421*	.421#	42.1	4.21	.421	.421*	.421#	42.1	4.21	.421	.421*	.421#	42.1	4.21	.421	.421*	.421#
20		.236	.264	.778	.427	.780	.072	.088	.576	.054	.536	.960	.962	.952	.847	.941	.579	.562	.261	.258	.255
200		.118	.154	.793	.557	.797	.060	.081	.795	.051	.672	.989	.990	.960	.789	.958	.831	.803	.622	.181	.310
2000		.069	.088	.787	.597	.789	.055	.067	.858	.050	.689	.995	.995	.919	.763	.968	.942	.926	.560	.119	.375
20000		.053	.059	.792	.598	.786	.052	.058	.884	.050	.719	.996	.997	.885	.756	.974	.971	.969	.478	.088	.443
(b) regression model $y_i = \beta x_i w_i + \gamma w_i + \varepsilon_i$; *dgp* of (3.1) with $c = 0$, $d = 1$, $v = 42.1$, by a																					
obs	a	5		50		500	5000	5		50	500	5000	5		50	500	5000	5		50	500	5000
20		.335		.353		.355	.355	.140		.157	.157	.157	.945		.944	.944	.944	.584		.588	.588	.588
200		.161		.163		.164	.164	.096		.099	.100	.100	.980		.979	.979	.979	.843		.844	.844	.844
2000		.079		.077		.077	.077	.065		.063	.063	.063	.990		.990	.990	.990	.942		.942	.942	.942
20000		.056		.055		.056	.056	.057		.056	.055	.055	.994		.994	.994	.994	.971		.971	.971	.971
(c) regression model $y_i = \beta x_i w_i + \varepsilon_i$; *dgp* of (3.1) with $v = 42.1$, $d = 1$, $a = ½$, by c																					
obs	c	0	10	100	100*	100#	0	10	100	100*	100#	0	10	100	100*	100#	0	10	100	100*	100#
20		.255	.218	.238	.158	.237	.072	.054	.049	.051	.050	.954	.193	.030	.017	.029	.515	.058	.006	.007	.006
200		.115	.099	.102	.075	.105	.058	.053	.050	.050	.048	.988	.191	.010	.023	.001	.825	.055	.005	.008	.006
2000		.068	.055	.057	.055	.059	.054	.045	.046	.050	.048	.995	.178	.019	.015	.011	.944	.035	.004	.007	.004
20000		.052	.048	.047	.051	.051	.052	.044	.048	.056	.053	.997	.203	.033	.011	.021	.972	.018	.002	.006	.003
(d) randomization inference using percentiles of coefficients (probability of rejecting true null for β at .05 level)																					
		model of panel (a) by v					model of panel (b) by a					model of panel (c) by c									
obs		42.1	4.21	.421	.421*	.421#	5	50	500	5000		0	10	100	100*	100#					
20		.072	.080	.459	.050	.387	.138	.165	.166	.165		.076	.054	.052	.052	.051					
200		.097	.147	.951	.053	.929	.231	.246	.247	.247		.102	.063	.053	.050	.047					
2000		.116	.226	.998	.052	.995	.250	.257	.257	.257		.116	.060	.049	.052	.050					
20000		.136	.291	1.00	.051	1.00	.257	.260	.261	.261		.137	.065	.049	.055	.054					

Notes: * = sharp null is true, $\beta_i \equiv 0$ & d=0; # = $\beta_i \equiv 0$, but sharp null is not true (d=1); {CCI}∩{RCI}/{CCI}U{RCI} = overlap divided by union of combined confidence intervals; randomization inference uses 999 draws of Wald statistics (panels a-c) or squared coefficients (panel d).

intervals converge to those of the conventional test. The results in panels (a) and (b) follow (R1) - (R3) and (R5) above. Panel (c) removes the interacted variable w_i from the regression, violating assumption A2 if the mean of x_i is not 0. The results for conventional inference and randomization inference when the mean c of x_i is 0 are much the same as before. However, when $c \neq 0$ and no version of A2 is satisfied, randomization p-values and confidence intervals do not converge to conventional values as the sample size increases. These results stem from the random bias and non-central chi-squared distribution highlighted earlier above.

Any test statistic allows for exact randomization inference when the sharp null is true, but with heterogeneous treatment effects the distribution of coefficient estimates generated by treatment permutation is only guaranteed to converge to that of the original data if studentized by the appropriate variance estimate. Panel (d) of Table I illustrates this by moving away from Wald statistics and evaluating the significance of the estimated coefficient using the percentiles of the distribution of squared deviation of coefficient estimates from the null (in this case 0) generated by permuting treatment, i.e. comparing $\hat{\beta}(\mathbf{t}, \beta_0 = 0)^2$ to $\hat{\beta}^2$. When the sharp null is true randomization inference is finite sample exact, but otherwise there is no guarantee of asymptotic accuracy, even when the regressor has higher moments, as rejection rates based upon the distribution of coefficients generally worsen as the sample size increases.

Table II below evaluates relative power by adjusting the parameter β so that the power of conventional inference, when testing the false null $\beta = 0$, equals .25, .50 or .75, and then reporting randomization inference rejection probabilities for the same *dgp*.[26] The table focuses on results for $v = 42.1$ or 4.21, where conventional rejection probabilities of true nulls converge to nominal value (Table I). As shown in panel (a), randomization power is well below that of the conventional test in very small samples, where the larger size distortions of the conventional test (see Table I) produce greater power. However, in all sample sizes randomization power is increasing in the number of draws used to calculate the randomization p-value and converges to the conventional level in large samples with a large number of draws. This continues to be true in panel (b), when $v = 4.21$, A3 is not met, and only the first four moments of the randomization distribution are guaranteed to converge to the normal. Increasing the heterogeneity of effects by orders of magnitude (panels c and d) has no systematic effects. These results follow (R4) above. As shown in panel (e), when assumption A2 is violated, the regression is run without the interaction variable w_i, and the *dgp* for x_i has a non-zero mean, randomization inference performs extraordinarily poorly with power that hardly differs from the nominal level for true nulls. This is true even when the *dgp* conforms to that of the sharp null (panel f). As noted above, in this case the coefficient estimates of permuted treatment are asymptotically distributed thick tailed non-central chi-squared, producing wide confidence intervals. Assumption A2, that covariates **W** interacted with treatment **X** are included in the regressors **Z**, crucially guarantees randomization inference asymptotically identical power

[26]Since both positive and negative values of β can achieve the same conventional power, I compute randomization rejection rates for the two and report the average in the table.

Table II: Power of Randomization Inference by Power of Conventional Test for Tests of β & Number of Draws from the Permutation Distribution (Monte Carlos with 10000 iterations per *dgp*)

	conventional power = .25				conventional power = .50				conventional power = .75			
draws / *obs*	19	99	199	999	19	99	199	999	19	99	199	999
(a) regression model of Table Ia: $y_i = \beta x_i w_i + \gamma w_i + \varepsilon_i$, $x_i = t(42.1)$, $\beta_i = \beta + U(-½,½)$, $d = 1$												
20	.077	.080	.081	.082	.212	.239	.245	.246	.408	.463	.470	.475
200	.143	.158	.159	.161	.328	.370	.375	.380	.563	.630	.638	.643
2000	.197	.216	.220	.223	.406	.453	.459	.462	.651	.706	.716	.722
20000	.213	.237	.242	.246	.436	.484	.489	.495	.680	.732	.740	.744
(b) regression model of Table Ia: $y_i = \beta x_i w_i + \gamma w_i + \varepsilon_i$, $x_i = t(4.21)$, $\beta_i = \beta + U(-½,½)$, $d = 1$												
20	na	na	na	na	.215	.244	.248	.249	.412	.465	.472	.477
200	.138	.152	.155	.156	.320	.359	.365	.370	.553	.618	.627	.632
2000	.187	.204	.209	.211	.395	.440	.445	.450	.640	.701	.706	.711
20000	.207	.233	.237	.240	.434	.481	.486	.491	.679	.730	.738	.742
(c) regression model of Table Ib: $y_i = \beta x_i w_i + \gamma w_i + \varepsilon_i$, $x_i = t(42.1)$, $\beta_i = \beta + U(-50,50)$, $d = 1$												
20	na	na	na	na	.251	.277	.280	.282	.467	.517	.523	.530
200	.155	.169	.170	.171	.366	.401	.405	.408	.613	.661	.667	.671
2000	.200	.220	.221	.224	.417	.456	.460	.466	.669	.717	.723	.726
20000	.217	.243	.246	.248	.444	.486	.492	.500	.691	.739	.745	.750
(d) regression model of Table Ib: $y_i = \beta x_i w_i + \gamma w_i + \varepsilon_i$, $x_i = t(42.1)$, $\beta_i = \beta + U(-5000,5000)$, $d = 1$												
20	na	na	na	na	.251	.277	.279	.283	.468	.519	.524	.531
200	.156	.169	.170	.171	.368	.401	.406	.409	.613	.661	.666	.671
2000	.201	.220	.221	.225	.418	.456	.460	.466	.668	.716	.722	.725
20000	.216	.242	.246	.248	.445	.486	.491	.500	.691	.739	.745	.749
(e) regression model of Table Ic: $y_i = \beta x_i w_i + \varepsilon_i$, $x_i = 100 + t(42.1)$, $\beta_i = \beta + U(-½,½)$, $d = 1$												
20	.050	.050	.051	.051	.052	.054	.054	.054	.052	.054	.054	.055
200	.050	.050	.049	.050	.052	.050	.049	.050	.051	.050	.049	.050
2000	.048	.047	.047	.048	.048	.049	.048	.049	.048	.048	.048	.048
20000	.047	.049	.050	.050	.046	.048	.049	.049	.047	.047	.047	.049
(f) regression model of Table Ic: $y_i = \beta x_i w_i + \varepsilon_i$, $x_i = 100 + t(42.1)$, $\beta_i \equiv \beta$ & $d = 0$												
20	.052	.052	.053	.053	.052	.053	.054	.053	.053	.054	.054	.054
200	.052	.051	.050	.050	.053	.052	.051	.051	.053	.052	.052	.052
2000	.051	.051	.051	.051	.051	.052	.051	.052	.053	.053	.052	.054
20000	.052	.053	.054	.054	.053	.054	.053	.055	.054	.056	.055	.055

Note: β adjusted until the conventional rejection rate at the .05 level equals .25, .50 or .75; na - not applicable, conventional size distortions are so large that no null has a .05 rejection rate of .25; notation otherwise as in (3.1).

to conventional Wald tests of locally false heterogeneous *or* sharp nulls, as in results (R4) - (R5) above.

IV. Practical Issues: Finite Sample Algorithms and Testing Subsets of Coefficients

This section addresses algorithms for calculating randomization confidence intervals (RCI) and issues in testing subsets of coefficients. Asymptotically if assumptions W1-W4 & A1-A3 hold the Wald statistic for permuted treatment $\tau(\mathbf{T},\boldsymbol{\beta}_0)$ in probability does not depend upon the null $\boldsymbol{\beta}_0$ in a finite $\sqrt{N}$ neighbourhood of $\boldsymbol{\beta}$, so calculating the RCI merely involves sorting the values of $\tau(\mathbf{T})$ across D draws and using the order statistics to calculate a bound for $\tau(\mathbf{X},\boldsymbol{\beta}_0)$, as indicated by (2.6) in (R2) earlier. This defines the usual ellipsoid around $\hat{\boldsymbol{\beta}}$. In finite samples, $\tau(\mathbf{T},\boldsymbol{\beta}_0)$ generally does depend upon $\boldsymbol{\beta}_0$, but the calculation of the RCI remains straightforward, as whether using the heteroskedasticity robust or clustered robust

Table III: Non-Convexities and Asymmetries in Randomization Confidence Intervals for β
(Monte Carlos: 10000 iterations per data generating process based upon (3.1))

	(a) regression with w_i, specification of Table Ia $c = 0$, $d = 1$, $a = ½$, by v				(b) regression with w_i, specification of Table Ib $c = 0$, $d = 1$, $v = 42.1$, by a				(c) regression without w_i, specification of Table Ic $v = 42.1$, $d = 1$, $a = ½$, by c			
	42.1	4.21	.421	.421*	5	50	500	5000	0	10	100	100*
	(i) number of non-convex .95 RCI in 10000 regressions											
20	139	167	603	415	190	245	219	211	325	492	688	1025
200	9	18	513	59	25	30	24	21	47	539	846	1119
2000	0	2	201	8	0	1	1	1	2	907	1262	1303
20000	0	0	152	1	0	0	0	0	0	1247	1548	1579
	(ii) mean ratio of length of shorter side to longer side of convex cover of .95 RCI											
20	.751	.769	.753	.838	.726	.718	.717	.717	.736	.569	.573	.526
200	.904	.909	.876	.972	.892	.891	.891	.891	.901	.549	.539	.507
2000	.971	.969	.871	.992	.963	.963	.963	.963	.970	.545	.523	.504
20000	.986	.984	.830	.998	.983	.983	.983	.983	.986	.565	.513	.523

Notes: convex cover = smallest convex region covering all segments of RCI; * = sharp null is true, $\beta_i \equiv 0$ & $d = 0$; 20 ... 20000 = number of observations; notation otherwise as in Table I & (3.1).

covariance estimates, the equation $\tau(\mathbf{X},\boldsymbol{\beta}_0) = \tau(\mathbf{T},\boldsymbol{\beta}_0)$ defines a multivariate quartic equation in $\boldsymbol{\beta}_0$.[27]

In a test for estimating equations with a single treatment effect, β_j, the roots of $\tau(\mathbf{t},\beta_{0j}) = \tau(\mathbf{x},\beta_{0j})$ define regions on the real line where $\tau(\mathbf{t},\beta_{0j})$ is >, < or = to $\tau(\mathbf{x},\beta_{0j})$. To calculate the finite sample RCI, one simply takes D draws of $\mathbf{t}$, calculates the roots specific to each, orders them on the real line, and moving left to right keeps track of the number of $\tau(\mathbf{t},\beta_{0j})$ that are greater than, less than, or equal to $\tau(\mathbf{x},\beta_0)$. For a given draw from the uniform distribution u, one can calculate the (constant) p-value between any two adjacent roots and to the left and right of their extreme values, and in so doing find the RCI associated with any level 1-α. The finite sample RCI may have infinite width when there are few distinct potential realizations of $\mathbf{t}$ or draws D from the permutation distribution. It is also typically asymmetric and may be non-convex. For these reasons the analytic procedure just described (and further in the on-line appendix) is better than costly and potentially fruitless or inaccurate line searches.

Table III reports the number of non-convex RCI and the mean ratio of the shorter side to the longer side of the convex cover of the RCI for the simulations of Table I. As shown, when assumptions A2 and A3 are satisfied, with the interaction variable w_i in the regression and $v = 42.1$, non-convexities and asymmetries are rapidly eliminated as the sample size increases. When w_i is not included in the regression and the mean of x_i is not 0, i.e. both forms of assumption A2 are violated, asymmetries and non-convexities show no tendency to diminish with larger samples, even when the sharp null is true and there is no heterogeneity of treatment effects. Conditions W1-W4 and A1-A3 not only ensure an asymptotic equivalence with conventional inference in tests of heterogeneous treatment effects but also tend to produce conventional properties in finite sample exact randomization tests of sharp nulls.

In a joint test of multiple treatment effects, the boundaries in the space of $\boldsymbol{\beta}_0$ of $\tau(\mathbf{T},\boldsymbol{\beta}_0) = \tau(\mathbf{X},\boldsymbol{\beta}_0)$

[27]The on-line appendix lays out this equation and its use to calculate confidence intervals as described below.

can be calculated for each $\mathbf{T}$, and the intersections of these used to calculate a confidence interval for the joint test. This is obviously difficult and it is generally more straightforward to simply report the p-value for a given $\boldsymbol{\beta}_0$ by calculating the relative value of $\tau(\mathbf{T},\boldsymbol{\beta}_0)$ and $\tau(\mathbf{X},\boldsymbol{\beta}_0)$ for the D draws of $\mathbf{T}$. The same procedure of reporting the p-value for a specific multi-dimensional null, rather than calculating the whole confidence interval, is used in conventional tests, so this is not much of an issue.

More problematic is the fact that the Wald statistic of a permutation-based randomization test of a null for a *subset* of treatment coefficients depends upon the null for *all* treatment measures, as the counterfactual outcome $\mathbf{y}(\mathbf{T},\boldsymbol{\beta}_0)=(\mathbf{T}_\mathbf{W}-\mathbf{X}_\mathbf{W})\boldsymbol{\beta}_0$ generally depends upon all elements of $\boldsymbol{\beta}_0$. Let $\mathbf{P}$ denote a k x PQ matrix of zeros of full rank with a single one in each row, so that $\boldsymbol{\beta}_0^{\subset}=\mathbf{P}\boldsymbol{\beta}_0$, $\hat{\boldsymbol{\beta}}^{\subset}=\mathbf{P}\hat{\boldsymbol{\beta}}$, and $\hat{\boldsymbol{\beta}}^{\subset}(\mathbf{T},\boldsymbol{\beta}_0)=\mathbf{P}\hat{\boldsymbol{\beta}}(\mathbf{T},\boldsymbol{\beta}_0)$ denote k x 1 sub-vectors of the parameters and estimated coefficients, and $\hat{\mathbf{V}}(\hat{\boldsymbol{\beta}}^{\subset})=\mathbf{P}\hat{\mathbf{V}}(\hat{\boldsymbol{\beta}})\mathbf{P}'$ and $\hat{\mathbf{V}}(\hat{\boldsymbol{\beta}}^{\subset}(\mathbf{T},\boldsymbol{\beta}_0))=\mathbf{P}\hat{\mathbf{V}}(\hat{\boldsymbol{\beta}}(\mathbf{T},\boldsymbol{\beta}_0))\mathbf{P}'$ the k x k covariance estimates. The Wald statistic for the conventional test $\tau^{\subset}(\mathbf{X},\boldsymbol{\beta}_0^{\subset})=(\hat{\boldsymbol{\beta}}^{\subset}-\boldsymbol{\beta}_0^{\subset})'\hat{\mathbf{V}}(\hat{\boldsymbol{\beta}}^{\subset})(\hat{\boldsymbol{\beta}}^{\subset}-\boldsymbol{\beta}_0^{\subset})$ depends only on $\boldsymbol{\beta}_0^{\subset}$, but the equivalent Wald statistic for permuted treatment $\tau^{\subset}(\mathbf{T},\boldsymbol{\beta}_0)=(\hat{\boldsymbol{\beta}}^{\subset}(\mathbf{T},\boldsymbol{\beta}_0)-\boldsymbol{\beta}_0^{\subset})'\hat{\mathbf{V}}(\hat{\boldsymbol{\beta}}^{\subset}(\mathbf{T},\boldsymbol{\beta}_0))(\hat{\boldsymbol{\beta}}^{\subset}(\mathbf{T},\boldsymbol{\beta}_0)-\boldsymbol{\beta}_0^{\subset})$ in the finite sample depends upon all elements of $\boldsymbol{\beta}_0$.[28] Thus, in the finite sample a randomization test based upon a test statistic for a subset of coefficients remains a joint test of the null for all treatment measures, although this disappears asymptotically as the realizations of the test statistic become insensitive to nulls in a root-N neighborhood of $\boldsymbol{\beta}$ and the comparison of $\tau^{\subset}(\mathbf{T},\boldsymbol{\beta}_0)\rightarrow\tau^{\subset}(\mathbf{T},\boldsymbol{\beta})$ to $\tau^{\subset}(\mathbf{X},\boldsymbol{\beta}_0^{\subset})$ depends only $\boldsymbol{\beta}_0^{\subset}$.

One way to test a subset null without taking a stand on the null for other treatment measures is by permuting $\mathbf{X}_\mathbf{W}^{\subset}=\mathbf{P}\mathbf{X}_\mathbf{W}$ alone, calculating the counterfactual outcome $\mathbf{y}(\mathbf{T}_\mathbf{W}^{\subset},\boldsymbol{\beta}_0^{\subset})=\mathbf{y}+(\mathbf{T}_\mathbf{W}^{\subset}-\mathbf{X}_\mathbf{W}^{\subset})\boldsymbol{\beta}_0^{\subset}$ and treating the remaining treatment regressors as part of the matrix of covariates $\mathbf{Z}$. However, if the permutation of the treatment sub-vector alone is not consistent with the randomization protocol of the experiment, while results (R1) - (R5) still carry through for the test of the subset if conditions W1 - W4 and A1 - A3 hold, the randomization test will not carry the additional finite sample validity of randomization tests when the sharp null is true. Permutation of a treatment sub-vector is valid, for example, when treatment $\mathbf{X}_\mathbf{W}^{\subset}$ is independently applied, so that permutations of $\mathbf{X}_\mathbf{W}^{\subset}$ holding constant the remaining columns in $\mathbf{X}_\mathbf{W}$ constitute a valid subset of the universe of potential treatment realizations.[29] Unfortunately, in practice this is seldom the case, so while this method provides asymptotic validity, it only very rarely is guaranteed to be finite sample exact for tests of sharp nulls.

When multiple treatments are not separately randomized, as is more usual, one can still calculate a valid subset of potential outcomes by permuting treatment within strata defined by other treatments, e.g. examining the potential alternative realizations of $\mathbf{X}_\mathbf{W}^{\subset}$ within strata defined by the remaining columns of

[28]Similarly, while the confidence interval for an individual treatment effect β_j in a multi-treatment equation is easily calculated using the roots of $\tau(\mathbf{T},\beta_{0j})=\tau(\mathbf{X},\beta_{0j})$ following the method above, these roots and hence the confidence interval depend upon the maintained null for untested treatment measures as these determine $y(\mathbf{T},\boldsymbol{\beta}_0)$.

[29]For example, Cole et al (2013) investigated the demand for insurance in Gujarat by separately randomizing discounts. Permutation of these holding constant other treatment elements maps out a subset of potential outcomes.

$\mathbf{X_W}$. This is the procedure advocated by D'Haultfoeuille & Tuvaandorj (2022) in their extension of DiCiccio & Romano's (2017) analysis to cases where treatment measures are correlated with other covariates,[30] and was used earlier in Young (2019) to provide alternative randomization p-values in amenable equations. This method is not useful in regressions with covariate interactions, as the number of distinct other-treatment-stratified potential outcomes is generally the null set since treatment **t** generally cannot vary at all within strata defined by realizations of **t•w**,[31] and vice versa.[32] In these circumstances, the distribution is degenerate and while the p-value (based upon universal ties of all test statistics and a random draw from the uniform distribution in (2.4) earlier) is trivially exact, power is identical to size. Outside of this case, the conditional mean assumption $E(x_i\varepsilon_i|\mathbf{z}_i) = 0$ used by D'Haultfoeuille & Tuvaandorj to guarantee asymptotic validity may be problematic as $\mathbf{z}_i$ now contains other treatment variables that are correlated with x_i. As described earlier in (2.2), OLS parameters can be thought of as population average linear treatment effects, with misspecification in the form of orthogonal to average linear effects included in the error term. When x_i and $\mathbf{z}_i$ are correlated, average linear treatment effects can vary by strata so that even though $E(x_i\varepsilon_i) = 0$ holds in aggregate, $E(x_i\varepsilon_i|\mathbf{z}_i)$ may not equal 0 and this method is not asymptotically valid. Furthermore, if other-treatment-stratification restricts tested treatment variation for substantial shares of the sample, there may be a loss of power. In sum, while this method is finite sample exact for tests of sharp nulls in randomized experiments, it is often degenerate, and if not may lose power or be asymptotically invalid in the face of otherwise innocuous to OLS specification error.

A test that would remain exact and only depend upon the null for a subset of treatment measures while permuting all treatment regressors across the entire sample would be one that impractically always set the null on untested coefficients equal to their true values. A practical approximation of this test involves setting the null on untested coefficients equal to the estimated OLS coefficients. As stated in (R1) above, asymptotically for $\boldsymbol{\beta}_0$ in a finite root-N neighborhood of $\boldsymbol{\beta}$, $\tau(\mathbf{T},\boldsymbol{\beta}_0) = \tau(\mathbf{T},\boldsymbol{\beta})$. White's assumptions ensure that $\sqrt{N}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ is asymptotically normally distributed with a bounded covariance matrix, so that the $\sqrt{N}$ deviation of estimated values from $\boldsymbol{\beta}$ is bounded in probability and this test is asymptotically identical to setting the null on untested coefficients equal to their true values. Insofar as asymptotic properties carryover into the finite sample, this approach approximates the impractical finite sample exact test, while providing an asymptotically valid subset test of heterogeneous treatment effects

[30]D'Haultfoeuille & Tuvaandorj actually suggest permuting within strata defined by *all* other regressors, which in our case would include the non-treatment covariates **Z** as well. In practice covariates are often continuous, making that conditional permutation distribution degenerate, i.e. all strata contain only one observation. However, when treatment is randomly applied it can be taken as independent of non-treatment covariates and the approach simplified to permuting within strata defined by other treatment measures alone, as done in Young (2019) and below.

[31]Except for observations where **w** is 0.

[32]Holding **t** constant, **t•w** can only be varied by permuting **w**, which is not an alternative experimental outcome and cannot be justified on that basis nor, typically, on the grounds that its distribution is exchangeable. In any case, the inclusion of **w** as a separate regressor, as is usually done, renders this permutation distribution degenerate as well, as there is no way to vary **t•w** holding both **t** and **w** constant.

with the general validity conferred by the OLS assumption $E(x_i\varepsilon_i) = 0$. However, as sampling variation leads $\hat{\boldsymbol{\beta}}$ to vary from $\boldsymbol{\beta}$ it will not be truly finite sample exact and, as shown below, may perform poorly when White's assumptions do not hold.

Another approach to subset testing is to conservatively calculate the maximum p-value for the test of a subset $\boldsymbol{\beta}_0^{\subset}$ *across all possible nulls* for the remaining treatment effects in $\boldsymbol{\beta}_0$. While not exact, in a finite sample test of sharp treatment effects this will guarantee control of size, as the probability of rejecting a true null $\boldsymbol{\beta}_0^{\subset}$ must be less than or equal to the nominal level of the test. Searching for a maximum across an unbounded space can be problematic, but as shown in the on-line appendix the maximum p-value along opposite rays of infinite length can be calculated analytically. In this fashion, the problem is reduced, through the use of spherical coordinates, to one of searching in the bounded space of angles, each lying in $[0,\pi]$. When the regression contains two treatment effects and p-values for β_{0j} are calculated across all values of $\beta_{0\sim j}$, there are no angles to consider and the maximum can be calculated analytically.[33] This approach is all-purpose, serving across all regression types considered in this paper. It is however conservative, producing a loss of power, and this conservatism need not disappear asymptotically.[34] Power can be enhanced if bounds can be placed on the universe of true nulls for untested measures. If these bounds are based upon the conventional Wald statistic for nulls on the untested measures, given White's moment conditions in probability the search for a maximum p-value remains within a finite root-N neighborhood and asymptotically produces p-values identical to setting the null on untested measures equal to their true values. With liberal bounds, this provides protection against the finite sample variation of $\hat{\boldsymbol{\beta}}$ from $\boldsymbol{\beta}$ that bedevils the approach based upon setting untested nulls equal to estimated values described above, while ruling out theoretically unreasonable infinitely large values of parameters and assuring asymptotic p-values and power equal to that of the impractical test based upon knowledge of the parameter values for untested measures.

The Monte Carlos in Tables IV and V explore the issues described above. Table IV considers equations with covariate interactions, modifying the *dgp* of (3.1) to include the treatment vector itself as a source of variation:

$$(4.1)\ y_i = \beta_1 x_i + \beta_2 x_i w_i + \varepsilon_i, \text{ with } \varepsilon_i = \beta_{1i} x_i + \beta_{2i} x_i w_i + d|x_i|^{½} + d|x_i w_i|^{½} + |w_i|^{½} + \eta_i$$

$$iid\ x_i = t(v)\ \&\ \beta_{ji} = \beta_j + U(-a,a) \text{ for } j = 1 \text{ and } 2, \text{ and } inid\ \eta_i = sin(i)*t(2.1)\ \&\ w_i = sin(i)*t(4.2).$$

There are heterogeneous linear treatment effects for both x_i and $x_i w_i$ and to satisfy assumption A2 both $\mathbf{1}_N$ and $\mathbf{w}$ should be included as ancilliary regressors. The sharp null holds when $a = 0$ and $d = 0$. The table reports rejection rates of tests of correct nulls of the mean treatment effects β_1 and β_2. In the permute-one

[33]Techniques for finding the maximum for higher dimensional problems are reviewed in the section below.

[34]While, as noted earlier, the *distribution* of $\tau(\mathbf{T},\boldsymbol{\beta}_0)$ is asymptotically unaffected by *any* fixed value of $\boldsymbol{\beta}_0$ (not merely those in a root-N neighborhood), its individual realizations are, so by searching across $\boldsymbol{\beta}_0$ the p-value based on a finite number of draws from the distribution can be manipulated (as can be seen in the equations describing the calculation of the max p-value in the on-line appendix). Moreover, mathematically, convergence for a fixed value does not guarantee convergence of the maximum across all values.

Table IV: Subset Inference with Covariate Interactions by Method (true null rejection rates at .05 level)
Monte Carlos: 10000 iterations per data generating process based on (4.1)
regression model: $y_i = \beta_1 x_i + \beta_2 x_i w_i + \gamma + \delta w_i + \varepsilon_i$

	conventional robust		randomization inference									
			permute-one		$\beta_{0\sim j} = \hat{\beta}_{\sim j}$		$\beta_{0\sim j} = \beta_{\sim j}$		constrained max over $\beta_{0\sim j}$		max over $\beta_{0\sim j}$	
tests of:	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2
(a) $x_i = t(42.1)$, $d = 1$, $a = 0$												
20	.070	.145	.065	.059	.056	.048	.055	.052	.033	.028	.031	.026
200	.047	.086	.053	.058	.052	.058	.052	.058	.042	.045	.034	.035
2000	.047	.054	.051	.049	.051	.048	.051	.048	.047	.045	.036	.036
20000	.046	.053	.050	.053	.050	.053	.050	.053	.047	.052	.038	.040
(b) $x_i = t(42.1)$, $d = 1$, $a = 5000$												
20	.156	.302	.137	.147	.128	.129	.128	.136	.102	.098	.099	.096
200	.069	.164	.063	.102	.063	.102	.064	.102	.056	.089	.053	.082
2000	.052	.076	.051	.061	.052	.061	.051	.061	.045	.057	.040	.048
20000	.050	.053	.050	.052	.049	.052	.049	.052	.045	.051	.040	.039
(c) $x_i = t(42.1)$, $a = d = 0$ (sharp null is true)												
20	.070	.145	.065	.059	.056	.048	.055*	.052*	.033	.028	.031#	.026#
200	.047	.086	.053	.058	.052	.058	.052*	.058*	.042	.045	.034#	.035#
2000	.047	.054	.051	.049	.051	.048	.051*	.048*	.047	.045	.036#	.036#
20000	.046	.053	.050	.053	.050	.053	.050*	.053*	.047	.052	.038#	.040#
(d) $x_i = t(.421)$, $a = d = 0$ (sharp null is true)												
20	.238	.239	.051	.039	.072	.076	.048*	.049*	.034	.032	.030#	.025#
200	.368	.361	.053	.033	.094	.094	.050*	.049*	.051	.050	.046#	.044#
2000	.422	.428	.059	.031	.069	.073	.051*	.054*	.055	.058	.049#	.052#
20000	.429	.431	.064	.025	.057	.058	.047*	.048*	.050	.051	.047#	.047#

Notes: $\beta_{0\sim j}$ = null for coefficient not being tested, null for coefficient being tested always equals parameter of the *dgp*; * = method is exact; #= size guaranteed to be less than or equal to nominal level.

approach only the variable associated with the coefficient being tested is permuted, i.e. x_i is changed to t_i or $x_i w_i$ to $t_i w_i$, without changing the other. This is not a valid experimental outcome, and hence these tests need not be exact in tests of sharp nulls, but if W1-W4 and A1-A3 hold have the same asymptotic validity as the conventional test. In "$\beta_{0\sim j} = \hat{\beta}_{\sim j}$", following each permutation of treatment x_i to t_i both regressors are changed and the null on the coefficient that is not being tested is set equal to its estimated value. Again, given W1-W4 and A1-A3 this has the same asymptotic validity as the conventional test, but need not be exact for sharp nulls as $\beta_{0\sim j}$ does not equal the true value. In "$\beta_{0\sim j} = \beta_{\sim j}$" the untested null is set equal to its true value, providing a benchmark of an asymptotically valid and finite sample (sharp null) exact test. "max over $\beta_{0\sim j}$" reports rejection rates for the conservative test which calculates the maximum p-value across all possible untested nulls, providing conservative control over size in the finite sample for exact nulls and asymptotically for heterogeneous treatment effects. The "constrained max" restricts the search for a maximum p-value to those nulls for which the conventional p-value for the Wald statistic for $\beta_{0\sim j}$ is greater than 10^{-10}. When W1-W4 and A1-A3 hold, both the "constrained max" and "$\beta_{0\sim j} = \hat{\beta}_{\sim j}$" are

asymptotically identical to "$\beta_{0\sim j} = \beta_{\sim j}$". The method of other-treatment-stratified permutation is not examined here, as each observation would constitute a separate strata and the distributions are degenerate.

As seen in panels (a) - (c) of the table, when W1-W4 and A1-A3 hold, while small sample deviations from nominal rejection probability are much higher when there is a lot of regressor associated heteroskedasticity (a = 5000 in panel b), all approaches other than "max over $\beta_{0\sim j}$" asymptotically provide rejection probabilities equal to nominal value, whereas "max over $\beta_{0\sim j}$" remains conservative even in large samples,. When the sharp null is true, whether regressors satisfy assumptions W1-W4 & A1-A3 (panel c, $x_i = t(42.1)$) or not (panel d, $t(.421)$), setting "$\beta_{0\sim j} = \beta_{\sim j}$" provides an exact test (within the bounds of simulation variation) and "max over $\beta_{0\sim j}$" more practically implementable conservative control of size. Permute-one, which is not a valid counterfactual experimental outcome, does not provide an exact test when the sharp null is true, with rejection probabilities that actually deviate further from nominal value as the sample size grows in panel d. Results for both "$\beta_{0\sim j} = \hat{\beta}_{\sim j}$" and "constrained max" appear to converge to those of "$\beta_{0\sim j} = \beta_{\sim j}$", even in panel d where the sufficient moment conditions do not hold.[35] "$\beta_{0\sim j} = \hat{\beta}_{\sim j}$" is not, however, finite sample exact, as shown in panel d where with unruly regressors rejection rates are well above nominal value. The "constrained max" with the liberal bounds given above, however, controls size in finite samples while converging to the impractical test based upon "$\beta_{0\sim j} = \beta_{\sim j}$".

Table V below considers a Monte Carlo where other-treatment-stratified permutation is an option. The data generating process is given by

$$(4.2)\ y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i, \text{ with } \varepsilon_i = \beta_{1i} x_{1i} + \beta_{2i} x_{2i} + d_1 x_{1i}^2 + d_1 x_{2i}^2 + |w_i|^{½} + \eta_i$$

$$iid\ x_{ji} = U\{\text{-4,-3,-2,-1,1,2,3,4}\}\ \&\ \beta_{ji} = U(-a,a) + d_2 x_{\sim j},\ \text{and}\ inid\ \eta_i = sin(i)*t(v)\ \&\ w_i = sin(i)*t(4.2),$$

where $U\{\}$ denotes uniformly distributed on the given integers. I consider two scenarios: independent, where x_{1i} and x_{2i} are drawn independently, and correlated, where the two variables are constrained to have the same sign but are otherwise independently distributed across the integers. When $d_1 = 1$ (panel a) there is specification error in the form of quadratic treatment effects, whereas when $d_2 = 1$ (panel b) there is specification error in terms of treatment interactions. In both cases, $E(x_{ji}\varepsilon_i) = 0$ holds, while $E(x_{ji}\varepsilon_i|x_{\sim ji}) = 0$ when x_{1i} and x_{2i} are independent, but $E(x_{ji}\varepsilon_i|x_{\sim ji}) \neq 0$ when x_{1i} and x_{2i} are correlated. Other-treatment-stratified permutation is only guaranteed to be asymptotically valid when the x_{ji} are independent, whereas all other methods (provided $v > 2$ so that the error disturbance has enough moments) are guaranteed to have asymptotically accurate or conservative rejection rates. This is seen in panels (a) and (b) of the table, where with $v = 2.1$ rejection rates using all other methods other than "max over $\beta_{0\sim j}$" (which remains slightly conservative) rapidly converge to nominal value, while size distortions using other-treatment-stratified permutation actually increase with sample size when the regressors are correlated. This illustrates the more restrictive assumptions needed for asymptotic validity with this method. However,

[35]Table VII below gives more detailed evidence on convergence of individual p-values, but for the interim convergence to "$\beta_{0\sim j} = \beta_{\sim j}$" is discussed in terms of average rejection rates.

Table V: Subset Inference when Treatment Stratified Permutation is Non-Degenerate
(true null rejection rates of β_1 & β_2 at .05 level, 10k iterations per *dgp* (4.2), by correlation of regressors)
regression model: $y_i = \beta_1 x_i + \beta_2 x_i + \gamma + \delta w_i + \varepsilon_i$

	conventional robust		randomization inference											
			permute-one		$\beta_{0\sim j} = \hat{\beta}_{\sim j}$		$\beta_{0\sim j} = \beta_{\sim j}$		constrained max over $\beta_{0\sim j}$		max over $\beta_{0\sim j}$		$x_{\sim j}$ stratified	
	ind	corr	ind	corr	ind	corr	ind	corr	ind	corr	ind	corr	ind	corr
(a) specification error in the form of quadratic effects, $v = 2.1$, $a = ½$, $d_1 = 1$, $d_2 = 0$														
20	.072	.047	.054	.033	.055	.030	.055	.029	.043	.020	.040	.019	.058	.070
200	.050	.045	.049	.044	.050	.043	.049	.043	.043	.036	.039	.033	.054	.160
2000	.049	.049	.049	.049	.049	.049	.048	.049	.045	.045	.039	.039	.055	.166
20k	.052	.050	.051	.051	.052	.051	.052	.051	.050	.050	.041	.041	.057	.167
(b) specification error in the form of treatment interactions, $v = 2.1$, $a = ½$, $d_1 = 0$, $d_2 = 1$														
20	.107	.050	.088	.036	.090	.033	.087	.032	.076	.021	.072	.019	.079	.074
200	.053	.045	.052	.044	.053	.043	.053	.043	.046	.036	.041	.033	.050	.134
2000	.050	.049	.051	.049	.050	.049	.050	.049	.047	.045	.041	.039	.050	.142
20k	.048	.050	.048	.051	.048	.050	.048	.050	.046	.048	.039	.039	.049	.140
(c) sharp null is true, $v = 2.1$, $a = d_1 = d_2 = 0$														
20	.045	.054	.049*	.056	.044	.036	.048*	.052*	.024	.021	.021#	.019#	.051*	.051*
200	.044	.044	.050*	.050	.049	.049	.050*	.050*	.042	.037	.035#	.032#	.049*	.050*
2000	.045	.046	.049*	.051	.049	.051	.049*	.051*	.046	.046	.036#	.037#	.050*	.053*
20k	.045	.045	.048*	.049	.048	.049	.047*	.050*	.046	.047	.035#	.037#	.049*	.049*
(d) sharp null is true, $v = .21$, $a = d_1 = d_2 = 0$														
20	.002	.003	.051*	.075	.009	.003	.051*	.054*	.001	.000	.001#	.000#	.049*	.053*
200	.000	.000	.051*	.064	.016	.014	.050*	.052*	.002	.002	.000#	.000#	.051*	.050*
2000	.001	.000	.052*	.074	.027	.027	.052*	.051*	.009	.009	.000#	.000#	.051*	.052*
20k	.001	.001	.049*	.079	.031	.034	.049*	.053*	.017	.020	.001#	.000#	.049*	.050*

Notes: As the data generating process is symmetric, reported figures are averages for separate tests of the two coefficients. ind = x_{1i} and x_{2i} are independent; corr = x_{1i} and x_{2i} are constrained to be of the same sign and hence correlated; * = method is exact; #= method maintains control of size; otherwise as in Table IV.

when $a = d_1 = d_2 = 0$ and the sharp null is true (panels c and d) other-treatment-stratified permutation is always finite sample exact, whereas otherwise this is only true of the impractical "$\beta_{0\sim j} = \beta_{\sim j}$" and, when the x_{ji} are independent, permute-one. This is shown in panels (c), with $v = 2.1$, and (d), where $v = .21$ magnifies the deviations from nominal level of non-exact methods. "max over $\beta_{0\sim j}$" & "constrained max" provide control of rejection probabilities when the sharp null is true, albeit very conservatively in panel (d). This raises concerns over power, which is explored in Table VI which reports rejection rates when parameters are such that the conventional test has a .50 rejection frequency when the sharp null is true.[36] "max over $\beta_{0\sim j}$" has persistently lower rejection rates than other methods, although not as low as might be expected from Table V, while the "constrained max" has power that converges to that of "$\beta_{0\sim j} = \beta_{\sim j}$". As noted earlier, other-treatment-stratified permutation also has issues of power when inter-treatment

[36] As the conventional test has sustained distortions away from nominal level when the regressors or errors don't satisfy the assumptions needed for asymptotic conventional accuracy (Tables IV and V earlier), rejection probabilities for randomization inference do not necessarily converge to the .50 level in the table.

Table VI: Power When the Sharp Null is True
(randomization inference rejection rates for *dgp* (4.2) with .50 conventional rejection rates)

	permute-one		$\beta_{0\sim j}=\hat{\beta}_{\sim j}$		$\beta_{0\sim j}=\beta_{\sim j}$		constrained max over $\beta_{0\sim j}$		max over $\beta_{0\sim j}$		$x_{\sim j}$ stratified	
	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2
	(a) *dgp* of Table IV (c), $v = 42.1$											
20	.482	.298	.458	.272	.459*	.278*	.385	.210	.376#	.201#	na	na
200	.517	.427	.512	.425	.513*	.425*	.484	.392	.456#	.355#	na	na
2000	.512	.482	.512	.482	.512*	.483*	.499	.470	.467#	.436#	na	na
20000	.511	.503	.511	.503	.511*	.503*	.503	.498	.464#	.463#	na	na
	(b) *dgp* of Table IV (d), $v = .421$											
20	.272	.251	.307	.291	.301*	.283*	.246	.219	.200#	.166#	na	na
200	.200	.174	.265	.263	.248*	.248*	.214	.212	.204#	.203#	na	na
2000	.165	.123	.192	.186	.183*	.178*	.175	.169	.167#	.162#	na	na
20000	.150	.093	.163	.149	.160*	.146*	.156	.142	.151#	.136#	na	na
	ind	corr	ind	corr	ind	corr	ind	corr	ind	corr	ind	corr
	(c) *dgp* of Table V (c), $v = 2.1$											
20	.505*	.508	.494	.438	.503*	.458*	.417	.366	.398#	.356#	.433*	.302*
200	.519*	.517	.519	.514	.518*	.515*	.492	.480	.466#	.462#	.518*	.383*
2000	.512*	.513	.512	.513	.512*	.513*	.499	.498	.466#	.469#	.513*	.391*
20000	.509*	.509	.509	.509	.509*	.509*	.503	.501	.465#	.465#	.509*	.388*
	(d) *dgp* of Table V (d), $v = .21$											
20	.570*	.579	.532	.508	.566*	.553*	.491	.483	.486#	.481#	.565*	.543*
200	.616*	.611	.564	.543	.613*	.596*	.517	.509	.497#	.496#	.617*	.492*
2000	.638*	.634	.589	.563	.634*	.619*	.545	.526	.497#	.498#	.637*	.442*
20000	.646*	.647	.601	.580	.643*	.632*	.568	.547	.497#	.498#	.646*	.384*

Notes: Reported figures in panels (c) and (d), where treatment distributions are symmetric, are the average rejection probability of the two coefficients. β_1 and β_2 in (4.1) & (4.2) are adjusted so the conventional heteroskedasticity robust test of the null of 0 has a .50 rejection rate for each test. As there are two (+ and -) values for which this is true for each of two parameters, reported figures in each cell in panels a & b (c & d) are the average of two (four) rejection rates for two (four) different Monte Carlos with 10k iterations each. "na" = the subset permutation distribution is degenerate (i.e. only one outcome is possible) in these simulations. Otherwise as in Tables IV & V.

correlation restricts the range of outcomes observed in stratified permutations, with rejection rates perversely declining with sample size in panel (d) of Table VI.

We postpone a summary evaluation of subset testing methods to the next section, which applies randomization inference to a broad practical sample of published papers. Instead, Table VII examines the convergence of the individual p-values of other methods to those of the finite sample exact, broadly asymptotically valid, and totally impractical "$\beta_{0\sim j}=\beta_{\sim j}$" subset testing method by reporting the average absolute difference of p-values in testing true and false nulls in the realizations of the *dgps* of Tables IV - VI above. As can be seen, when the moment conditions hold, panels (a) and (c), the p-values of "constrained max" and especially "$\beta_{0\sim j}=\hat{\beta}_{\sim j}$" converge to those of "$\beta_{0\sim j}=\beta_{\sim j}$", while this is sometimes the case (panel b) and sometimes not (panel d) when the sufficient conditions do not hold. In contrast, in

Table VII: Absolute Difference of P-Values from those of $\beta_{0\sim j} = \beta_{\sim j}$ in Tests of True & False Nulls

	conventional robust		$\beta_{0\sim j} = \hat{\beta}_{\sim j}$		constrained max over $\beta_{0\sim j}$		max over $\beta_{0\sim j}$		other-treatment-stratified	
	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2
(a) *dgp* of Table IVa, IVb, & IVc: moment conditions hold, sharp & heterogeneous treatment effects										
20	.034	.094	.009	.014	.025	.036	.031	.039	na	na
200	.016	.045	.004	.003	.013	.012	.020	.018	na	na
2000	.012	.021	.002	.001	.008	.006	.017	.014	na	na
20000	.010	.014	.001	.001	.005	.003	.016	.015	na	na
(b) *dgp* of Table IVd: moment conditions don't hold, sharp treatment effects										
20	.118	.122	.032	.031	.038	.038	.043	.043	na	na
200	.178	.178	.020	.019	.014	.014	.017	.017	na	na
2000	.206	.209	.008	.008	.006	.006	.010	.009	na	na
20000	.217	.222	.004	.004	.003	.003	.008	.007	na	na
	ind	corr	ind	corr	ind	corr	ind	corr	ind	corr
(c) *dgp* of Table Va, Vb, & Vc: moment conditions hold, sharp & heterogeneous treatment effects										
20	.019	.025	.006	.008	.018	.021	.021	.022	.073	.138
200	.011	.012	.003	.003	.011	.012	.016	.016	.030	.095
2000	.010	.010	.002	.002	.006	.007	.016	.016	.029	.092
20000	.009	.009	.001	.001	.003	.004	.017	.016	.028	.091
(d) *dgp* of Table Vd: moment conditions don't hold, sharp treatment effects										
20	.112	.092	.045	.051	.046	.050	.047	.051	.096	.119
200	.173	.154	.065	.058	.058	.053	.062	.056	.032	.078
2000	.184	.176	.065	.061	.057	.055	.068	.063	.018	.085
20000	.186	.182	.056	.056	.049	.050	.070	.065	.014	.097

Notes: Tests of true and false nulls as in Tables IV, V and VI above. Reported figures are the average of average rejection rates for true nulls and false nulls. For false nulls, the average is across the two (positive & negative) parameter values yielding a .50 conventional rejection rate; for "ind" (independent) and "corr" (correlated), averages are calculated across the individual rejection rates for the two symmetric regressors. "na" = the subset permutation distribution is degenerate (i.e. only one outcome is possible) in these simulations.

large samples p-values for "max over $\beta_{0\sim j}$" generally remain substantively bounded away from "$\beta_{0\sim j} = \beta_{\sim j}$", confirming the persistent conservatism of this approach. Finally, while conventional robust p-values converge to those of "$\beta_{0\sim j} = \beta_{\sim j}$" when the moment conditions hold, those of other-treatment-stratification remain quite different, especially when treatment measures are correlated, in both small and large samples. This complicates the interpretation of results, as large differences between other-treatment-stratification and conventional and other forms of randomization inference may be due to failings of the latter, the more limited applicability of the former, or random chance in environments where both are valid. As shown below, in practice large absolute differences between conventional and other-treatment-stratification p-values arise frequently in environments where, based upon sample size and the influence of individual observations, conventional inference appears to have much of its asymptotic validity.

V. Randomization & Conventional Confidence Intervals in a Practical Sample

This section compares randomization and conventional confidence intervals in the sample of 53 published experimental papers examined in Young (2019), 44 of which contain reported results based on OLS regressions. I focus on treatment specifications that fit the framework used in this paper and the extensions in the on-line appendix, namely where (possibly stratified) permutations of treatment across observations or groupings of observations are counterfactual potential outcomes. This removes regressions where treatment is calculated from group characteristics or applied using multiple cross-cutting criteria in a fashion that does not allow for counterfactual permutation,[37] leaving 3213 treatment measures in 1066 OLS regressions in 39 papers (listed in the on-line appendix). The extension of this paper's results to stratification in the on-line appendix assumes that the number of strata is finite and the first and second moments of treatment are balanced across asymptotically non-negligible strata. Consequently, I also remove regressions where procedures ensure that the number of strata grows with the number of treatment groupings[38] or that treatment is not, at least in principle, asymptotically balanced across non-negligible strata.[39] Reduction in this fashion leaves 2944 treatment measures in 1013 regressions in 35 papers.[40] Results including all regressions dropped on the basis of stratification issues,

[37]As an example of the former, Duflo et al (2011) randomly assigned students to class sections in Kenyan schools and in some regressions the treatment variable is the average baseline test score of peers, permutations of which are not potential outcomes as individual's own scores contribute to the treatment measure for others but cannot contribute to the treatment measure for themselves. As an example of the latter, Dupas and Robinson (2013) randomized the provision of health savings technologies to members of credit associations in Kenya with individuals receiving multiple treatments depending upon which associations they belonged to. Permutations of these are not a potential outcome, as treatment received depends upon the configuration of individuals' memberships. In both cases, repeating the random treatment allocation procedure and recalculating regressors accordingly produces potential outcomes and allows for exact tests of sharp nulls, but these potential outcomes are not permutations of the baseline treatment regressors and hence do not fit within the framework of this paper.

[38]For example, Angrist & Lavy (2009) divided schools into treatment/control pairs on the basis of characteristics, so the number of strata equals the number of treatment groupings divided by 2.

[39]As examples: In Fong and Luttmer's (2009) investigation of racial group loyalty and charitable generosity whites and ethnic minorities were given different audiovisual treatment combinations. Regressions containing both strata are dropped from the analysis here. In contrast, although subjects in Robinson's (2012) income shocks experiment mentioned earlier did not receive an income shock in week 0, making that and the following 13 weeks separate strata with different mean treatments, as week 0 is in principle asymptotically negligible, and in practice accounts for only 5% of observations, I retain all those regressions.

[40]Bugni, Canay & Shaikh (2018, 2019) show that White's heteroskedasticity robust covariance estimate is conservative (i.e. too large) when stratified treatment balance is greater than that achieved by random sampling and there is heterogeneity of treatment effects across strata. For amenable equations in my sample (370 treatment effects in 168 equations in 7 papers), I have calculated the smallest standard error estimate implied by their theory (that mechanisms to balance treatment within strata ensure that root-N times the deviation of strata mean treatment from the target value has asymptotically zero variance), and find that on average it is only 3 percent smaller than White's standard error estimate (with strata fixed effects, which appear in some form in most of the equations in my sample; without strata fixed effects the difference rises to 6 percent) and no more than 10 percent smaller in 95 percent of cases. Since this standard error estimate mechanically equals White's when there is no heterogeneity of average treatment effects across strata, this result suggests that these concerns are generally not of substantive importance. (I say "amenable" because their theory concerns regressions without covariates other than strata fixed effects where treatment is measured by mutually exclusive dummies and applied to single observations. Only 2 such regressions

shown in the on-line appendix, are all but identical. In the case of a handful of regressions where authors do not include covariates interacted with treatment as separate regressors in their own right (assumption A2), I add these to the regression specification. Authors use a variety of standard error estimates, homoskedastic, heteroskedasticity robust, and clustered at, below, and across treatment groupings, as well as the occasional bootstrap or jackknife. The on-line appendix proves the asymptotic equivalence of conventional and randomization inference with clustering at levels other than the treatment grouping (when allowed by the cross-correlations in errors), but clustering decisions themselves often have substantive effects. To ease comparison, I homogenize methods by always calculating conventional and randomization test statistics using standard errors clustered at the treatment grouping level (equivalently, the heteroskedasticity robust standard error estimate when treatment is applied to individual observations).

Figure I graphs the ratio of the overlap to the union[41] of the .95 conventional and randomization confidence intervals for individual coefficients. In the case of equations with multiple treatment measures, as a baseline I set the randomization null on not-directly-tested measures equal to their estimated effects, reviewing results with alternative approaches later on. As the number of estimated treatment effects in a given paper ranges from a low of 4 to a median of 52 and an extraordinary high of 794, the figure reports results for individual treatment effects but also paper means, where each paper carries equal weight in the presentation. On the x-axes I place the $\log_{10}$ number of treatment groupings or the largest coefficient leverage share of any treatment grouping in the regression. While leverage is usually defined in terms of the diagonal elements of the hat matrix $(\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')$, for the case of a single coefficient one can define the same using the diagonal elements of the hat matrix for the partitioned regression, $\tilde{\mathbf{x}}(\tilde{\mathbf{x}}'\tilde{\mathbf{x}})^{-1}\tilde{\mathbf{x}}'$, where $\sim$ denotes the residuals from the projection on the remaining regressors. While maximum coefficient leverage generally falls with sample size, as noted earlier it does not do so when the tails of the distribution of the regressor are excessively thick, so that a finite number of observations retain a non-negligible influence on coefficient and standard error estimates, invalidating the assumptions of asymptotic theorems. Regressing the confidence interval overlaps or the absolute differences in p-values of randomization and conventional inference in the figure on the maximum leverage share of a single treatment grouping/cluster and the $\log_{10}$ number of clusters (in the on-line appendix), I find the leverage share produces higher R^2s and is more consistently statistically significant, and hence provides a better summary of conditions under which the two inference methods converge.

As seen in Figure I, in practical application randomization confidence intervals converge to their conventional counterparts as either the number of observations increases or, more clearly, the maximum

---

appear in my 1000+ practical sample. To broaden the comparison, I take all equations where treatment regimes are applied to single observations and can be recoded as mutually exclusive treatment effects and strip out all covariates other than strata fixed effects, producing the 168 regressions just mentioned.)

[41]In 33 (of 2992) cases the .95 randomization confidence interval is unbounded (due to there only being 4 or 5 treatment groups and hence a very small number of potential realizations) & the overlap ratio is zero. In an additional 3 cases the confidence interval is not convex and the convex cover is used to calculate the overlap.

Figure I: Overlap/Union of .95 Conventional and Randomization Confidence Intervals

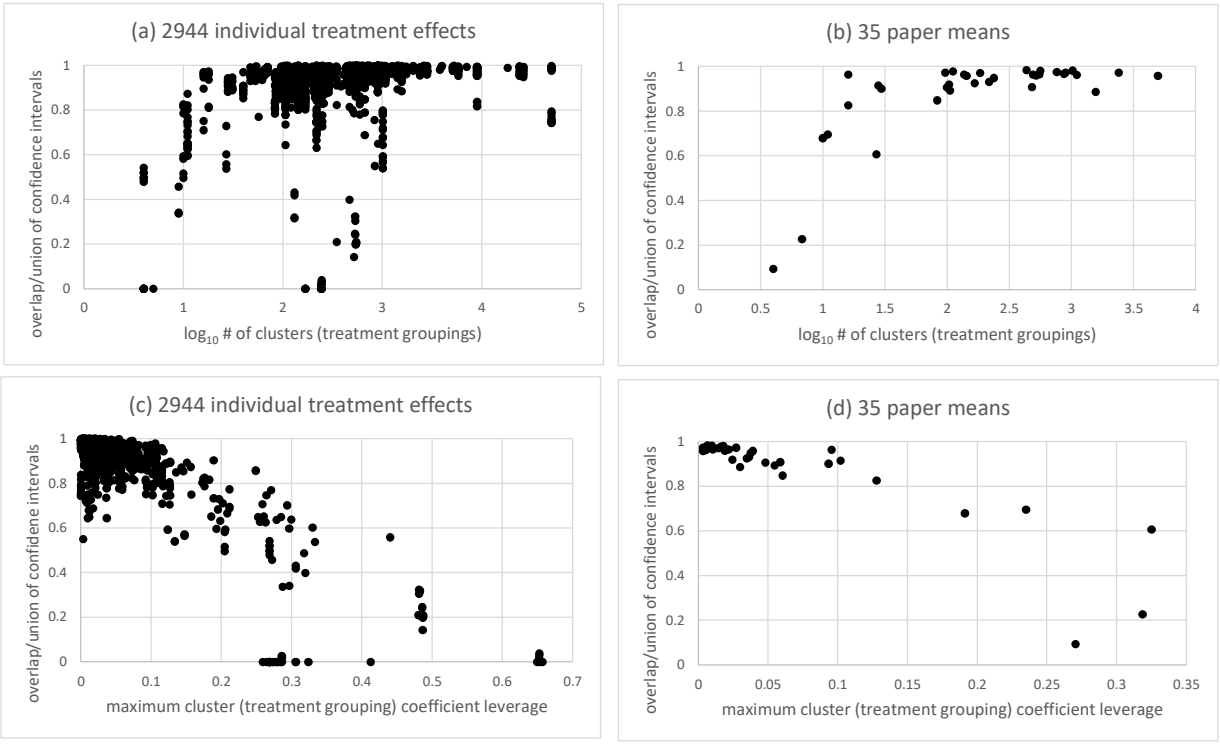
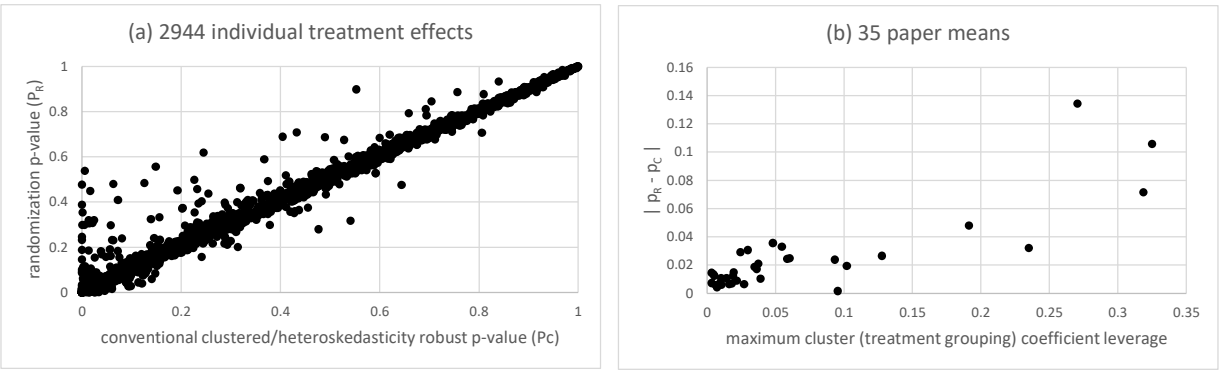


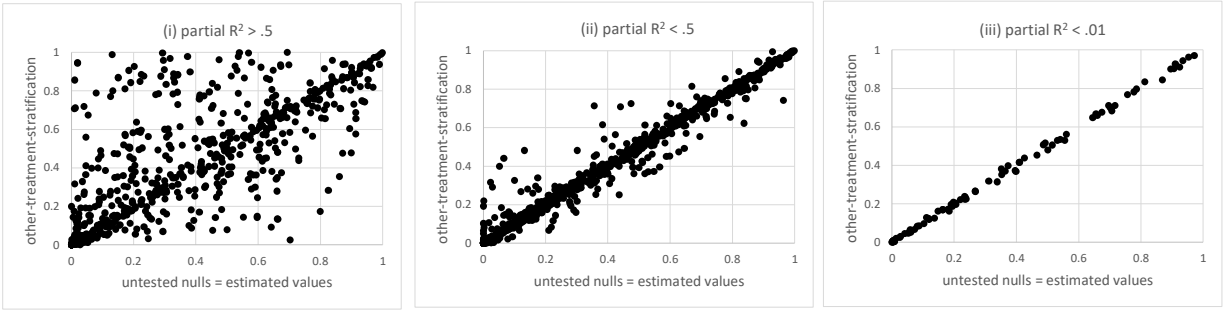
Figure II: Conventional vs Randomization P-Values



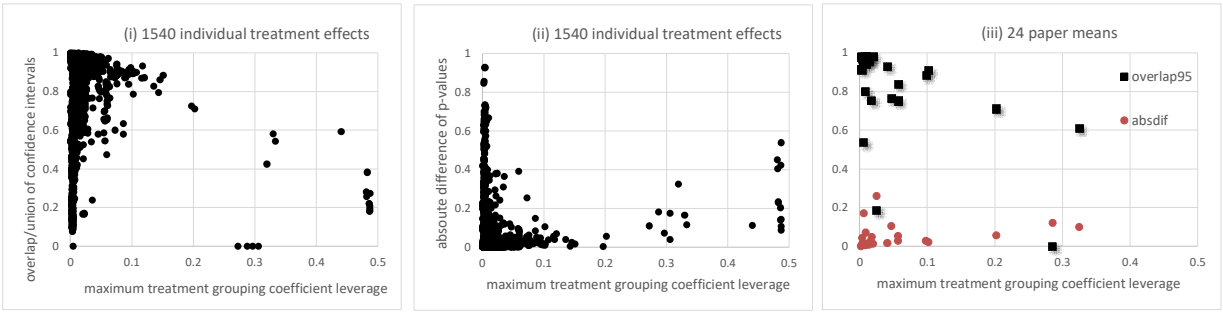
leverage share of a cluster/treatment grouping goes to zero. Figure II graphs the conventional and randomization p-values for individual treatment effects against each other, as well as the paper average of the absolute difference between the two against the paper average maximum cluster/treatment grouping leverage share. As can be seen, randomization p-values are often substantially larger than conventional estimates when the conventional p-value is low and where regression design is highly unbalanced, with one cluster exerting an unusually large influence, a factor that is known to produce size distortions in

Figure III: Randomization Inference using Other-Treatment-Stratification Compared with Other Methods

(a) 1540 individual treatment effect p-values compared to those with null on untested coefficients = to estimated values by R^2 of regression of permuted treatment on other-treatment-stratification dummies



(b) Confidence intervals and p-values of other-treatment-stratification compared to those of conventional inference by maximum coefficient leverage of a single cluster/treatment grouping



conventional tests.[42] Of the 756 treatment effects which are conventionally .05 significant, 70 are .05 randomization inference insignificant with an average maximum leverage share of .202, while of 707 treatment effects which are .05 randomization inference significant, 21 are conventionally .05 insignificant with an average maximum leverage share of .042. In contrast, for the 2853 treatment effects where the two methods agree on .05 significance or insignificance the average leverage share is merely .026.

Of the 2944 coefficients tested in Figures I and II, 2462 appear in equations with more than one treatment effect, and for these I set the randomization-inference null on untested treatment effects equal to estimated values ($\boldsymbol{\beta}_{0\sim j} = \hat{\boldsymbol{\beta}}_{\sim j}$). As noted earlier, an alternative approach to subset testing is to permute each treatment measure within the strata created by values of other treatment measures. Figure III implements this approach for the 1540 coefficients in 435 multi-treatment regressions in 24 papers where the resulting permutation distribution is not degenerate.[43] In the Monte Carlos above, it was seen that this

[42]In such circumstances, the robust standard error estimate is more volatile than indicated by standard degrees of freedom corrections (see Chesher 1989, Young 2016).

[43]As also noted earlier, D'Haultfoeuille & Tuvaandorj (2022) implement this method by stratifying on all covariates, but I take advantage of the fact that treatment is independent of non-treatment covariates and only stratify by other treatment measures (including treatment x covariate interactions) crossed with overall experimental strata, if such exist, so as to produce a valid subset of the potential outcomes of the experimental procedure. Because covariates often take on many values, implementing D'Haultfoeuille & Tuvaandorj's actual procedure produces non-degenerate distributions for only 667 coefficients (276 of which reside in regressions without non-treatment covariates). The results, in the on-line appendix, show more extreme versions of the same patterns, as the "strata" now account for even more of the variation in treatment.

approach gives rise to substantially different results than conventional and other forms of randomization inference when treatment measures are correlated. To this end, in panel (a), which compares the p-values found using this method to those found setting $\boldsymbol{\beta}_{0\sim j} = \hat{\boldsymbol{\beta}}_{\sim j}$, the results are divided according to whether the R^2 of the regression of the permuted treatment measure on other-treatment-stratification dummies is > .5, < .5 or < .01.[44] As seen, differences in the p-values delivered by the two methods are often very large when the R^2 is greater than .5, but virtually disappear when the R^2 is less than .01. These differences translate into similarly large differences with conventional p-values and confidence intervals and appear even when regressor leverage is minimal (panel b) and, given well-behaved errors, conventional inference is likely to have more of its asymptotic accuracy.[45] The correlation of treatment regressors is also strongly associated with the width of confidence intervals and the frequency with which treatment effects are found to be significant. For the 628 treatment measures where the R^2 of the regression on other-treatment dummies is greater than .5, the average ln ratio of the width of the other-treatment-stratified .95 confidence interval to that found using $\boldsymbol{\beta}_{0\sim j} = \hat{\boldsymbol{\beta}}_{\sim j}$ is .527[46] while 116 treatment measures are found to be .05 significant using $\boldsymbol{\beta}_{0\sim j} = \hat{\boldsymbol{\beta}}_{\sim j}$ but only 96 using other-treatment-stratification. In contrast, for the 906 treatment measures where the R^2 is less than .5, the average ln ratio of confidence intervals is only .017, while 247 and 238 treatment measures are found to be .05 significant using $\boldsymbol{\beta}_{0\sim j} = \hat{\boldsymbol{\beta}}_{\sim j}$ and other-treatment-stratification, respectively.

When individual treatment measures are independently randomized, permuting those measures alone without stratification by other treatment provides finite sample exact randomization tests of sharp nulls that are also asymptotically valid within the general framework of W1-W4, as noted earlier. In my sample, I find only 94 treatment measures in 51 multi-treatment equations in 4 papers that are independently randomized.[47] Figure IV graphs the permute-one randomization p-values of this method

---

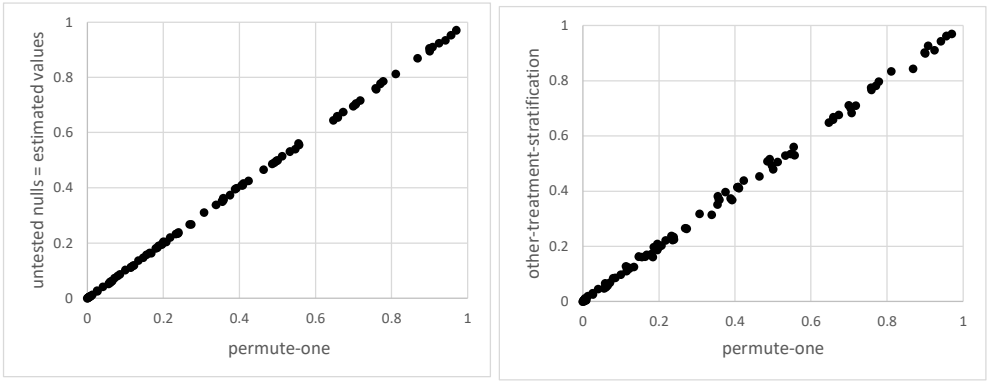
[44]Where the original experiment is stratified, the non-permuted treatment variables are crossed with the original stratification variable to create new strata so as to generate a subset of potential outcomes consistent with the experimental procedure. In such cases, the R^2 reported in the figure is the partial R^2 taking the regression on the original stratification dummies as the reduced model.

[45]I should note that the impact of the original experimental stratification on randomization p-values is minimal. The average R^2 of the regression of treatment measures on experimental strata dummies (for 2295 treatment measures in 21 papers which stratify) is .036 (including cases noted above removed on the grounds of unbalanced treatment across strata) and the correlation between the p-values found using randomization inference with and without the original stratification in such papers (in both cases setting untested nulls equal to estimated values) is .999, while the average overlap/union of confidence intervals is .980. In contrast, in Figure III the average partial R^2 of the regression of treatment measures on other-treatment strata is .433 and the correlation of the p-values using the two methods shown in panel (a) is .887 (.736 when the partial R^2 is > .5 and .508 when it is > .75).

[46]Excluding here and in the next sentence 6 cases where the other-treatment-stratified .95 confidence interval is unbounded.

[47]I determine these based upon methods described in the paper and the fact that the distribution of the treatment measure is largely invariant across strata defined by other treatment measures. Aside from the percentage discounts in Cole et al (2013) discussed above (which provide 4 measures in 4 regressions), I also find: (a) Beaman and MacGruder's (2012) investigation of social networks separately randomized the puzzle type given to participants (4 measures in 4 regressions); (b) Wisdom et al's (2010) investigation of information and healthy food choices

Figure IV: Randomization P-Values for Independently Randomized Treatment Effects by Method
(94 individual treatment effects)



against those found using $\boldsymbol{\beta}_{0\sim j} = \hat{\boldsymbol{\beta}}_{\sim j}$ or other-treatment-stratification. As can be seen, the $\boldsymbol{\beta}_{0\sim j} = \hat{\boldsymbol{\beta}}_{\sim j}$ p-values are virtually identical to permute-one, with an average absolute difference of .002. Other-treatment-stratification p-values vary somewhat more from permute-one, with an average absolute difference of .008, but are still quite close as the R^2 of the regression on other-treatment-stratification dummies is less than .01 in 88 of the 94 cases.

The third approach to subset testing that provides control of size under sharp nulls discussed earlier above is that of calculating the maximum p-value across all possible nulls for untested measures. As was noted, for equations with two treatment measures, the maximum p-value across the one untested null can be calculated analytically. For equations with n > 2 treatment measures, with n - 1 untested nulls, the maximum p-value on unbounded opposite-rays can be calculated analytically. Thus, as shown in the on-line appendix, after transformation to spherical coordinates, the search for the maximum p-value in the unbounded $\mathbb{R}^{n-1}$ space of nulls for untested measures can be reduced to one of searching in the bounded n-2 dimensional square of spherical angles on $[0,\pi]^{n-2}$. This is still challenging, as the objective function is discrete and non-differentiable, involving the comparison of values of the test statistic under untested nulls for each permutation **T** of **X**. For n = 3 (n-2=1) I implement a simple grid search, dividing $[0,\pi]$ evenly into 100, 1000 and 10000 points, finding virtually no change between 1000 and 10000 points in the 330 such cases in my sample. For n > 3 (reaching as high as 18 in one paper) I implement a variety of search methods so as to evaluate their relative effectiveness:

(a) Grid search. Divide each angle evenly into m points such that the total number of points m^{n-2} is 10 to 15 thousand and calculate the maximum p-value on the opposite rays defined by each set of spherical angles. By n = 8 the grid is very sparse and results are clearly dominated by other techniques, so grid search is not used for n > 8.

---

appears to have separately randomized the calorie content and calorie recommendation information (34 measures in 17 regressions); (c) In Robinson's (2012) study of income shocks the marginal distribution of each partner's shock is largely unaffected by the realization of the other partner's shock (52 measures in 26 regressions).

(b) Random search. Draw a random vector from the uniform distribution on $[0,\pi]^{n-2}$ and calculate the maximum p-value on the opposite rays defined by those spherical angles. Call this max and set the counter C_R equal to 1. Draw another random vector. If max is exceeded, reset the value of max and C_R, else increase C_R by 1. Continue drawing from the uniform distribution on $[0,\pi]^{n-2}$ until the counter C_R hits 10000, i.e. no improvement on the current max is found in 10000 random draws.

(c) Nelder-Mead (1965) simplex method. Draw n-1 random vectors from the uniform distribution on $[0,\pi]^{n-2}$ and calculate the maximum on the opposite rays defined by each set of spherical angles. Implement the Nelder-Mead search algorithm across this simplex, recalculating the maximum across opposite rays for each point in $[0,\pi]^{n-2}$ considered, until the method converges. Call this max and set C_{NM} to 1. Draw a fresh set of n-1 vectors from $[0,\pi]^{n-2}$ and implement the Nelder-Mead search again. If max is exceeded, reset max and C_{NM}, else increase C_{NM} by 1. Continue creating simplexes and implementing Nelder-Mead until the counter C_{NM} hits 100 or 1000. I found little change in the maximum between $C_{NM} = 100$ and 1000. Of 253 coefficients in equations with $n > 4$ found to be .05 significant with $C_{NM} = 100$, only one is rendered .05 insignificant with $C_{NM} = 1000$.

(d) Greedy sequential grid search. Draw a random vector from the uniform distribution on $[0,\pi]^{n-2}$. Implement a sequential 100 point grid search on each of the n-2 dimensions (selected in random order), greedily setting the value for each dimension equal to that which maximizes the objective function in its 100 point unidimensional search (choosing randomly across values in case of ties). Continue until a full round of n-2 sequential grid searches fails to improve the value. Call this max and set C_{GR} equal to 1. Draw a new random vector and implement the sequential greedy grid search again. If the previous maximum is exceeded, reset max and C_{GR}, else increase C_{GR} by 1. Continue until C_{GR} equals 10. This method was found to be computationally more costly and less effective than Nelder-Mead with $C_{NM} = 100$, and hence higher values of C_{GR} were not implemented.

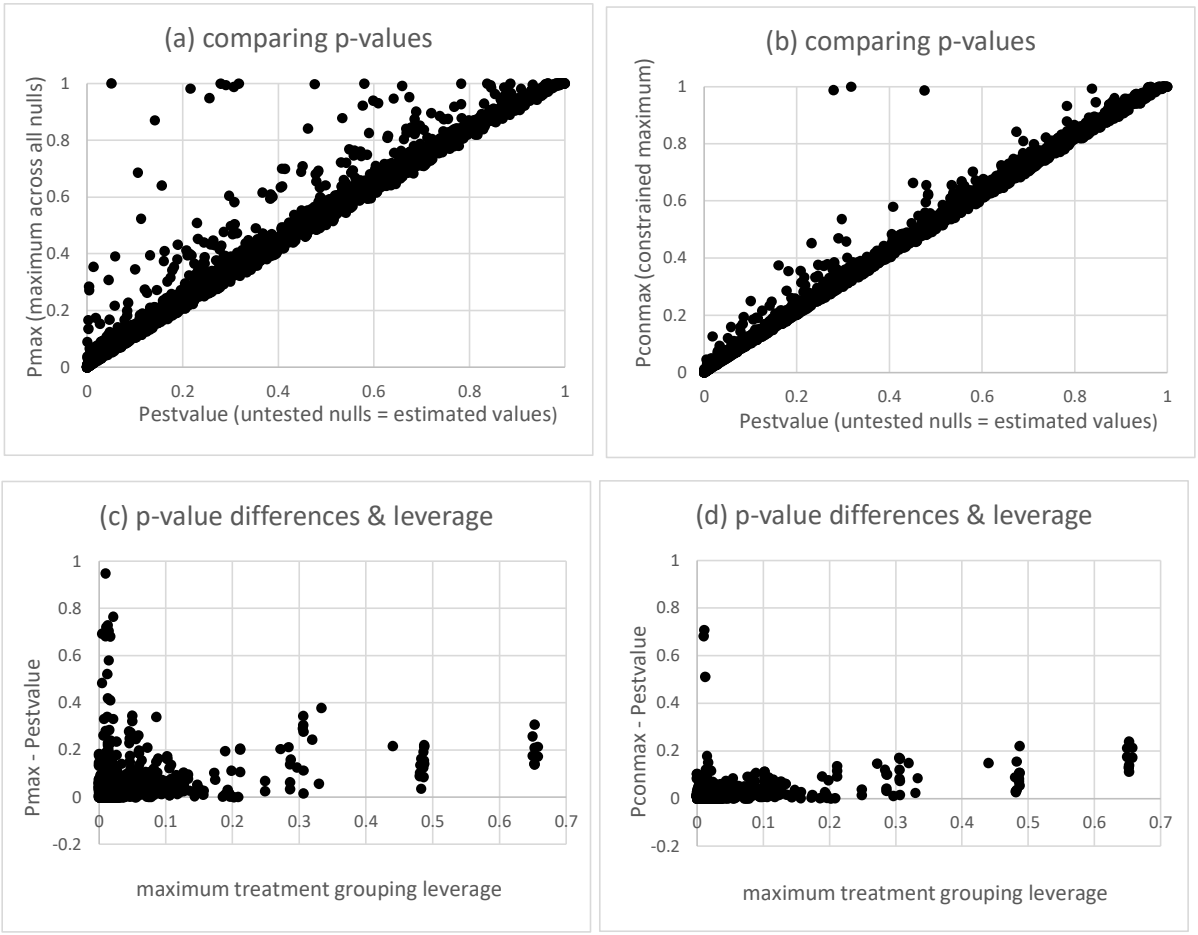
Of the above methods, I find Nelder-Mead search with $C_{NM} = 1000$ to be the most effective,[48] as it equals the maximum p-value across all methods 90 percent of the time, and differs from that maximum by an average of only .003 (.591 versus .594) in the remaining 10 percent of cases, where in no instance is a .05 significant Nelder-Mead result found to be .05 insignificant using other search techniques. Nelder-Mead with $C_{NM} = 100$ alone on average outperforms all other methods.[49]

Figure V below graphs the maximum randomization p-value across untested nulls found across all search methods described above against that found using the baseline approach setting $\boldsymbol{\beta}_{0\sim j} = \hat{\boldsymbol{\beta}}_{\sim j}$. As

[48]The actual number of Nelder-Mead searches used to find the maximum which is then not improved in 1000 subsequent tries is less than 1000 in .83 of cases and less than 2000 in .98 of cases.

[49]Average maximum p-values for individual coefficients in equations with $n > 4$ are .448 with random search until 10000 failures to improve, .455 with the computationally costly greedy sequential 100 point grid search until 10 failures to improve and .456 with comparatively rapid Nelder-Mead simplex search until 100 failures to improve. Nelder-Mead simplex search until 1000 failures to improve raises this average to .460 with, as noted above, almost no change for results found to be .05 significant with $C_{NM} = 100$.

Figure V: Maximum P-Value Across all Nulls for Untested Measures
Compared with Setting Untested Nulls Equal to Estimated Effects
(2462 individual treatment effects in 531 multi-treatment estimating equations)



shown in panel (a), the maximum p-value is in some cases very much higher, although these typically involve nulls on untested measures that are very far removed from estimated treatment effects. Panel (b) illustrates this by constraining the untested nulls to the region where the p-value of the conventional Wald statistic, $(\hat{\boldsymbol{\beta}}_{\sim j} - \boldsymbol{\beta}_{0\sim j})'\mathbf{V}(\hat{\boldsymbol{\beta}}_{\sim j})^{-1}(\hat{\boldsymbol{\beta}}_{\sim j} - \boldsymbol{\beta}_{0\sim j})$, is greater than 10^{-10}. Most of the extreme maximum p-values disappear. Regression analysis in the on-line appendix finds that differences between the "max across $\boldsymbol{\beta}_{0\sim j}$" (whether constrained or not) and $\boldsymbol{\beta}_{0\sim j} = \hat{\boldsymbol{\beta}}_{\sim j}$ p-values are strongly related to the maximum leverage of an individual treatment grouping. [50] Notwithstanding that, large differences do arise in cases with low maximum leverage, although these are largely eliminated by imposing bounds on the nulls, as illustrated

[50] Also significant, albeit less robustly so across specifications, are the number of untested treatment measures and the p-value and p-value squared found setting the null equal to estimated values (with p-values near one found to be, rather obviously, more difficult to raise, as also are p-values near zero).

in panels (c) and (d) of the figure.

In equations with more than one treatment regressor, $\boldsymbol{\beta}_{0\sim j} = \hat{\boldsymbol{\beta}}_{\sim j}$ finds 531 treatment measures to be statistically significant at the .05 level, but only 423 of the same are found to be so using the maximum p-value across all $\boldsymbol{\beta}_{0\sim j}$ nulls, suggesting relative power of .8 if failures to reject are taken to be Type II errors. Imposing the lax bound used in panel (b) raises the number of .05 significant effects to 469, or close to .9 of the level found setting nulls equal to estimated effects. Comparing results across the reduced set of coefficients that can be tested using other-treatment-stratification as well, the unconstrained maximum across $\boldsymbol{\beta}_{0\sim j}$ nulls finds .83 of $\boldsymbol{\beta}_{0\sim j} = \hat{\boldsymbol{\beta}}_{\sim j}$ .05 significant results to be .05 significant, while other-treatment-stratification finds .87 of these to be .05 significant.[51] Changes in p-values from those setting $\boldsymbol{\beta}_{0\sim j} = \hat{\boldsymbol{\beta}}_{\sim j}$ are much larger with other-treatment-stratification, averaging .069 across the panels of Figure III (a), while in the same sample the difference between the maximum across $\boldsymbol{\beta}_{0\sim j}$ nulls and $\boldsymbol{\beta}_{0\sim j} = \hat{\boldsymbol{\beta}}_{\sim j}$ is only .038. When other-treatment-stratification finds an otherwise $\boldsymbol{\beta}_{0\sim j} = \hat{\boldsymbol{\beta}}_{\sim j}$ .05 significant result to be .05 insignificant, only .33 of the other-treatment-stratification p-values lie below .1. In contrast, when the maximum across $\boldsymbol{\beta}_{0\sim j}$ nulls finds an otherwise $\boldsymbol{\beta}_{0\sim j} = \hat{\boldsymbol{\beta}}_{\sim j}$ .05 significant result to be .05 insignificant, fully .89 of the maximum p-values still lie below .1.

To summarize the results and discussion of this and the preceding section: Whenever treatment is independently applied, inference for a subset of treatment measures based upon permuting those treatment measures alone (permute-one) provides exact tests of sharp nulls while retaining broad asymptotic validity, making this procedure a natural choice. When treatment measures are correlated, other-treatment-stratification provides exact tests of sharp nulls for subsets of treatment measures, but may exhibit size distortions in both small and large samples when there is heterogeneity in average linear treatment effects across the strata defined by other treatment measures, i.e. it provides a biased test in environments where $E(x_{ji}\varepsilon_i|x_{\sim ji}) \neq 0$ but $E(x_{ji}\varepsilon_i) = 0$ and OLS remains consistent and conventional inference asymptotically accurate. When treatment measures are strongly correlated, as is often the case, it frequently produces very different p-values than conventional inference, even when regressor leverage is low and the latter is more likely to have some of its asymptotic validity, complicating the evaluation of results. Confidence intervals are also wider, producing lower power. Finally, in many circumstances, such as when treatment is interacted with covariates, this approach simply cannot be implemented, as the other-treatment stratified permutation distribution is degenerate.

Setting the null on untested measures equal to estimated values ($\boldsymbol{\beta}_{0\sim j} = \hat{\boldsymbol{\beta}}_{\sim j}$), while not finite sample exact, remains asymptotically valid in the broad environment where OLS is consistent and conventional inference asymptotically accurate and in practice produces results that converge to conventional values as the influence of individual observations goes to zero. When the data generating process allows for

[51]Overall, other-treatment-stratification achieves .92 of the 05 rejection rate found setting untested nulls equal to estimated values as it finds a number of otherwise insignificant results to be significant.

asymptotically valid tests of heterogeneous treatment effects, randomization inference is asymptotically locally insensitive to the nulls on untested treatment effects while $\hat{\boldsymbol{\beta}}_{\sim j}$ converges to the true parameter values. Not surprisingly, in these circumstance $\boldsymbol{\beta}_{0\sim j} = \hat{\boldsymbol{\beta}}_{\sim j}$ produces small sample results that are very close to exact methods such as permute-one (when that is valid) or setting the null on untested coefficients (impractically) equal to true values. This approach, however, is not exact, and may fail badly when the data generating process does not allow for asymptotically valid inference.

Calculating the maximum p-value across all $\boldsymbol{\beta}_{0\sim j}$ nulls for untested treatment measures provides control of rejection probabilities for tests of sharp nulls and hence insurance against some of the worst-cases scenarios that may be encountered when setting $\boldsymbol{\beta}_{0\sim j} = \hat{\boldsymbol{\beta}}_{\sim j}$. The loss of power in both small and large samples brought on by this conservative approach can be ameliorated if the range of plausible nulls for untested measures can be restricted. If this restriction keeps the universe of acceptable nulls root-*N* local to true parameter values, as it does when based upon the conventional Wald statistic if the latter is asymptotically valid, then both power and p-values asymptotically converge to those of conventional inference. Selection of very lax bounds allows for this while still providing, as seen in Monte Carlos, control of null rejection probabilities below nominal level with extreme data generating processes. The two approaches, $\boldsymbol{\beta}_{0\sim j} = \hat{\boldsymbol{\beta}}_{\sim j}$ and max over $\boldsymbol{\beta}_{0\sim j}$, can be implemented in all of the environments considered in this paper, including the common interaction of treatment with covariates.

When providing confidence intervals and p-values for individual coefficients in multi-treatment equations the Stata package *randcmdci* reports results using $\boldsymbol{\beta}_{0\sim j} = \hat{\boldsymbol{\beta}}_{\sim j}$ as a baseline. It also allows the user to call for the calculation of the maximum randomization p-value across $\boldsymbol{\beta}_{0\sim j}$ nulls, with or without bounds, so as to determine the sensitivity of individual coefficient results to nulls on untested measures.[52]

VII. Conclusion

Randomization inference provides researchers with a rare tool: tests of certain null hypotheses which are exact, regardless of sample size or the characteristics of regressors or errors. While the sharp null may seem restrictive, in some circumstances it is quite natural. A sharp null of 0, for example, is a test of complete and total irrelevance, a benchmark that most experimentalists would like to reject. This test remains exact in circumstances where, because of heavy tailed regressor or error distributions, conventional tests may have extraordinarily large size distortions, as seen in the tables above. However, consideration of confidence intervals, i.e. the range of non-zero population average linear treatment effects statistically consistent with experimental results, naturally suggests the possibility of heterogeneity in the response to experimental treatment. This paper builds on a literature that shows that permutation-based tests, despite mistakenly assuming the absence of interdependence between permuted treatment variables and errors such as would be induced by heterogeneous treatment effects and specification error, have an asymptotic validity for quite general treatment regressors equal to that of conventional tests.

[52]Users of *randcmdci* interested in applying other-treatment-stratification can do so by simply creating and specifying a stratum indicator based upon the values of other treatment measures.

This paper and its accompanying on-line appendix aim to provide results and techniques for the types of regression specifications found in applied experimental work. The above focuses on treatment applied at the observation level and interacted with non-treatment covariates, with population inference for not-identically-distributed data using heteroskedasticity robust covariance estimates. The on-line appendix generalizes the results to the remaining frameworks found in the published papers surveyed in Young (2019). If treatment is applied to groups of observations and the errors and other regressors are correlated across observations, the moment conditions in W1 - W4 and A1 - A3 can be defined in terms of groupings of observations. In this case, provided error correlations are such that conventional standard error estimates with clustering at, below, above or even across treatment groupings are asymptotically valid, Wald based randomization tests using the same covariance estimates are equally accurate if one adds the additional randomization assumption: (A4) the maximum size of an interdependent observational grouping is bounded. If treatment is stratified and hence only exchangeable and permuted within strata, within White's framework it suffices to add that: (A5) across strata treatment measures share the same asymptotic average first and second moments. These requirements are usually satisfied in experimental settings as potential correlations between observations are typically the result of locational groupings of bounded size, while experiments are stratified precisely in order to achieve treatment balance across strata. Without severely restrictive additional assumptions, randomization inference has an asymptotic validity equal to that of conventional population inference methods for the range of empirical specifications found in published experimental research, while providing finite sample exact tests of sharp nulls such as that of complete treatment irrelevance.

Bibliography

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.

Aker, Jenny C., Christopher Ksoll and Travis J. Lybbert (2012). "Can Mobile Phones Improve Learning? Evidence from a Field Experiment in Niger." *American Economic Journal: Applied Economics* 4 (4): 94-120.

Anderson, Marti J. and John Robinson (2001). "Permutation Tests for Linear Models." *Australian and New-Zealand Journal of Statistics* 43 (1): 75-88.

Angrist, Joshua, and Victor Lavy (2009). "The Effects of High Stakes High School Achievement Awards: Evidence from a Randomized Trial." *American Economic Review* 99 (4): 1384–1414.

Ashraf, Nava, James Berry, and Jesse M. Shapiro (2010). "Can Higher Prices Stimulate Product Use? Evidence from a Field Experiment in Zambia." *American Economic Review* 100 (5): 2383–2413.

Athey, Susan and Guido W. Imbens (2017). "The Econometrics of Randomized Experiments". In Handbook of Economic Field Experiments, Vol. 1, Esther Duflo and Abhijit Banerjee, editors, pp. 73-140. Amsterdam: North Holland Publishing Co.

Beaman, Lori and Jeremy Magruder (2012). "Who Gets the Job Referral? Evidence from a Social Networks Experiment." *American Economic Review* 102 (7): 3574–3593.

Bugni, Federico A., Ivan A. Canay, and Azeem M. Shaikh (2018). "Inference under Covariate-Adaptive Randomization." *Journal of the American Statistical Association* 113: 1784-1796.

Bugni, Federico A., Ivan A. Canay, and Azeem M. Shaikh (2019). “Inference under Covariate-Adaptive Randomization with Multiple Treatments.” *Quantitative Economics* 113: 1747-1785.

Cai, Hongbin, Yuyu Chen, and Hanming Fang (2009). “Observational Learning: Evidence from a Randomized Natural Field Experiment.” *American Economic Review* 99 (3): 864–882.

Chesher, Andrew. 1989. “Hajek Inequalities, Measures of Leverage and the Size of Heteroskedasticity Robust Wald Tests.” *Econometrica* 57 (4): 971-977.

Chung, Eun-yi and Mauricio Olivares (2021). "Permutation test for heterogeneous treatment effects with a nuisance parameter." *Journal of Econometrics* 225: 148-174.

Chung, Eun-yi and Joseph P. Romano (2016). "Multivariate and multiple permutation tests." *Journal of Econometrics* 193: 76-91.

Cole, Shawn, Xavier Giné, Jeremy Tobacman, Petia Topalova, Robert Townsend, and James Vickery. (2013). “Barriers to Household Risk Management: Evidence from India.” *American Economic Journal: Applied Economics* 5 (1): 104–135.

D'Haultfoeuille, Xavier and Purevdorj Tuvaandorj (2022). "A Robust Permutation Test for Subvector Inference in Linear Regressions." Manuscript, 2022.

DiCiccio, Cyrus J. and Joseph P. Romano (2017). "Robust Permutation Tests for Correlation and Regression Coefficients." *Journal of the American Statistical Association* 112: 1211-1220.

Duflo, Esther, Rachel Glennerster and Michael Kremer (2007). "Using Randomization in Development Economics Research: A Toolkit." In Handbook of Development Economics, Vol. 4, T. Paul Schultz and John A. Strauss, editors, pp. 3895-3962. Amsterdam: North Holland Publishing Co.

Duflo, Esther, Pascaline Dupas, and Michael Kremer (2011). “Peer Effects, Teacher Incentives, and the Impact of Tracking: Evidence from a Randomized Evaluation in Kenya.” *American Economic Review* 101 (5): 1739–1774.

Dupas, Pascaline and Jonathan Robinson (2013). “Savings Constraints and Microenterprise Development: Evidence from a Field Experiment in Kenya.” *American Economic Journal: Applied Economics* 5 (1): 163–192.

Edgington, Eugene S. and Patrick Onghena (2007). Randomization Tests, 4th edition. Boca Raton: Chapman & Hall, 2007.

Fisher, Ronald A. (1935). The Design of Experiments. Edinburgh: Oliver and Boyd, Ltd, 1935.

Fong, Christina M. and Erzo F. P. Luttmer (2009). “What Determines Giving to Hurricane Katrina Victims? Experimental Evidence on Racial Group Loyalty.” *American Economic Journal: Applied Economics* 1 (2): 64–87.

Galambos, Janos (1987). The Asymptotic Theory of Extreme Order Statistics, 2nd edition. Florida: Robert E. Krieger Publishing Co.

Galiani, Sebastian, Martín A. Rossi, and Ernesto Schargrodsky (2011). “Conscription and Crime: Evidence from the Argentine Draft Lottery.” *American Economic Journal: Applied Economics* 3 (2): 119–136.

Ghosh, M.N. (1950). "Convergence of Random Distribution Functions." *Bulletin of the Calcutta Mathematical Society* 42: 217-226.

Hemerik, Jesse and Jelle J. Goeman (2021). "Another Look at the Lady Tasting Tea and Differences Between Permutation Tests and Randomisation Tests." *International Statistical Review* 89 (2): 367-381.

Hewitt, Edwin and Leonard J. Savage (1955). "Symmetric Measures on Cartesian Products." *Transactions of the American Mathematical Society* 80: 470-501.

Hoeffding, Wassily (1951). "A Combinatorial Central Limit Theorem." *The Annals of Mathematical Statistics* 22 (4): 558-566.

Hoeffding, Wassily (1952). "The Large-Sample Power of Tests Based on Permutations of Observations." *The Annals of Mathematical Statistics* 23 (2): 169-192.

Janssen, Arnold (1997). "Studentized Permutation Tests for Non-iid Hypotheses and the Generalized Behrens-Fisher Problem." *Statistics & Probability Letters* 36: 9-21.

Jockel, Karl-Heinz (1986). "Finite Sample Properties and Asymptotic Efficiency of Monte Carlo Tests." *The Annals of Statistics* 14: 336-347.

Lehmann, E.L. & Joseph P. Romano (2022). Testing Statistical Hypotheses, 4th edition. Springer, 2022.

Lei, Lihua & Peter J. Bickel (2021). "An assumption-free exact test for fixed-design linear models with exchangeable errors." *Biometrika* 108 (2): 397-412.

Lin, Winston (2013). "Agnostic Notes on Regression Adjustments to Experimental Data: Reexamining Freedman's Critique". *The Annals of Applied Statistics* 7 (1): 259-318.

Nelder, J. A. and R. Mead (1965). "A simplex method for function minimization." *The Computer Journal* 7 (4): 308-313.

Noether, Gottfried E. (1949). "On a Theorem by Wald and Wolfowitz". *The Annals of Mathematical Statistics* 20 (3): 455-458.

O'Neill, Ben (2009). "Exchangeability, Correlation and Bayes' Effect." *International Statistical Review* 77 (2): 241-250.

Oster, Emily and Rebecca Thornton (2011). "Menstruation, Sanitary Products, and School Attendance: Evidence from a Randomized Evaluation." *American Economic Journal: Applied Economics* 3: 91-100.

Rao, C. Radhakrishna (1973). Linear Statistical Inference and its Applications. Second Edition. New York: John Wiley & Sons.

Robinson, Jonathan (2012). "Limited Insurance within the Household: Evidence from a Field Experiment in Kenya." *American Economic Journal: Applied Economics* 4 (4): 140–164.

Thornton, Rebecca L. (2008). "The Demand for, and Impact of, Learning HIV Status." *American Economic Review* 98 (5): 1829–1863.

Wald, Abraham. and Jacob Wolfowitz (1944). "Statistical Tests Based on Permutations of the Observations." *The Annals of Mathematical Statistics* 15 (4): 358-372.

White, Halbert (1980). "A Heteroskedasticity Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity." *Econometrica* 48 (4): 817-838.

White, Halbert (1984). Asymptotic Theory for Econometricians. London: Academic Press.

Wisdom, Jessica, Julie S. Downs, and George Loewenstein (2010). "Promoting Healthy Choices: Information versus Convenience." *American Economic Journal: Applied Economics* 2 (2): 164–178.

Young, Alwyn (2016). " Improved, Nearly Exact, Statistical Inference with Robust and Clustered Covariance Matrices using Effective Degrees of Freedom Corrections . Manuscript.

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

Zhao, Anqi and Peng Ding (2021). "Covariate-adjusted Fisher randomization tests for the average treatment effect." *Journal of Econometrics* 225: 278-294.

Table A1: Notation Used in Appendices A, B & C

(a) Sample means: $m(x_i) = N^{-1}\sum_{i=1}^{N} x_i$.

(b) $\mathbf{1}_N$ & $\mathbf{0}_N$ denote $Nx1$ vectors of 1s & 0s, $\mathbf{0}_{QxQ}$ a QxQ matrix of 0s & $\mathbf{I}_Q$ the QxQ identity matrix.

(c) Sample demeaned variables: $\widetilde{\mathbf{H}} = \mathbf{O}\mathbf{H}$, where $\mathbf{O} = \mathbf{I}_N - \mathbf{1}_N\mathbf{1}'_N/N$.

(d) Symmetric "square root" of a symmetric positive definite matrix: $\mathbf{A}^{-½}$ such that $\mathbf{A}^{-½}\mathbf{A}^{-½} = \mathbf{A}^{-1}$.

(e) Kronecker product $\otimes$ & face-splitting (or row by row Kronecker) product $\bullet$, with $\mathbf{T}_\mathbf{W} = \mathbf{T} \bullet \mathbf{W}$.

(f) As $N \to \infty$ converges almost surely across the distribution of the data $\mathbf{D} = (\mathbf{X},\mathbf{Z},\boldsymbol{\varepsilon})$: $\xrightarrow{a.s.}$.

(g) As $N \to \infty$ converges in probability or distribution across the permutations $\mathbf{T}$ of $\mathbf{X}$: $\xrightarrow{p}$ & $\xrightarrow{d}$.

(h) Expectation across row permutations $\mathbf{t}$ of $\mathbf{x}$: $E_\mathbf{t}$, as in $E_\mathbf{t}(t_i) = m(x_i)$.

(i) Counterfactual outcome, coefficient estimates and estimated residuals associated with permutation $\mathbf{T}$ of $\mathbf{X}$ under null $\boldsymbol{\beta}_0$: $\mathbf{y}(\mathbf{T},\boldsymbol{\beta}_0)$, $\hat{\boldsymbol{\beta}}(\mathbf{T},\boldsymbol{\beta}_0)$ & $\hat{\boldsymbol{\varepsilon}}(\mathbf{T},\boldsymbol{\beta}_0)$.

(j) Residual maker with respect to $\mathbf{Z}$: $\mathbf{M} = \mathbf{I}_N - \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'$.

(k) Rows, columns and elements of matrices: $\mathbf{x}'_{\mathbf{W}i}$, $\mathbf{x}_{\mathbf{W}j}$ and $x_{\mathbf{W}ij}$ are the i^{th} row, j^{th} column and ij^{th} element of $\mathbf{X}_\mathbf{W}$ (rows distinguished by use of subscript i, with columns referenced with j, k, l ...).

(l) Elements of matrices associated with particular columns of a $\bullet$ product: $t_{ip(j)}w_{iq(j)}$, where $p(j)$ and $q(j)$ denote the columns of $\mathbf{T}$ and $\mathbf{W}$ associated with the j^{th} column of $\mathbf{T}_\mathbf{W}$.

(m) Probability of the event $a \geq b$: $\Pr\{a \geq b\}$.

(n) $\mathbf{n}_{PQ}$ = the multivariate iid standard normal, indicated by $\mathbf{n}_{PQ} \sim \mathrm{N}(\mathbf{0}_{PQ},\mathbf{I}_{PQ})$.

Appendices

A. Foundational Theorems

The main result of this paper rests on a theorem first proven by Wald & Wolfowitz (1944) and later refined by Noether (1949) and Hoeffding (1951), regarding the asymptotic distribution of $\sqrt{N}$ times the correlation of a permuted sequence with another sequence:

Theorem I: Let $\mathbf{x}' = (x_1, \ldots, x_N)$ and $\mathbf{d}' = (d_1, \ldots, d_N)$ denote sequences of real numbers, not all equal, and $\mathbf{t}' = (t_1, \ldots, t_N)$ any of the $N!$ equally likely permutations of $\mathbf{x}$. Then as $N \to \infty$, the distribution across the realizations of $\mathbf{t}$ of the random variable

$$\text{(Ia)}\quad n(t_i,d_i) = \sum_{i=1}^{N} \frac{[t_i - m(t_i)][d_i - m(d_i)]}{\left(\sum_{i=1}^{N}\frac{[x_i - m(x_i)]^2}{N}\sum_{i=1}^{N}\frac{[d_i - m(d_i)]^2}{N}\right)^{½} N^{½}} \left[\text{where for } h = d,\ t \text{ or } x,\ m(h_i) = \sum_{i=1}^{N}\frac{h_i}{N}\right],$$

converges to that of the standard normal if for all integer $\tau > 2$

$$\text{(Ib)}\quad \lim_{N\to\infty} \frac{N^{\frac{\tau}{2}-1}\sum_{i=1}^{N}[x_i - m(x_i)]^\tau \sum_{i=1}^{N}[d_i - m(d_i)]^\tau}{\left(\sum_{i=1}^{N}[x_i - m(x_i)]^2\right)^{\tau/2}\left(\sum_{i=1}^{N}[d_i - m(d_i)]^2\right)^{\tau/2}} = 0.$$

The proof is based upon showing that the moments of $n(t_i,d_i)$ converge to those of the standard normal. A straightforward multivariate extension, proven in the on-line appendix, is that if $\mathbf{X}' = (\mathbf{x}_1,\ldots,\mathbf{x}_N)$ and

$\mathbf{D}' = (\mathbf{d}_1,...,\mathbf{d}_N)$ are sequences of vectors, $\mathbf{T}$ any of the row permutations of $\mathbf{X}$, $\mathbf{A}^{½}$ the "square root" of symmetric positive definite matrix $\mathbf{A}$,[53] $\mathbf{O} = \mathbf{I}_N - \mathbf{1}_N\mathbf{1}'_N/N$ the centering matrix, $\widetilde{\mathbf{H}} = \mathbf{OH}$, and $\otimes$ denotes the Kronecker product and $\bullet$ as above the row-by-row Kronecker product, then all the moments of

$$\text{(Ic)}\quad \mathbf{n}(\mathbf{t}_i,\mathbf{d}_i) = \left(\frac{\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}}}{N} \otimes \frac{\widetilde{\mathbf{D}}'\widetilde{\mathbf{D}}}{N}\right)^{-½} \frac{(\widetilde{\mathbf{T}} \bullet \widetilde{\mathbf{D}})'\mathbf{1}_N}{\sqrt{N}}$$

converge to those of the multivariate iid standard normal if (Ib) holds for all pairwise combinations of the elements of the columns of $\mathbf{X}$ and $\mathbf{D}$ and the matrices $\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}}/N$ and $\widetilde{\mathbf{D}}'\widetilde{\mathbf{D}}/N$ are bounded with determinant $> \gamma > 0$ for all N sufficiently large.

Theorem I is easily extended to a probabilistic environment by noting the following result due to Ghosh (1950) that allows us to translate the almost sure characteristics of an infinite number of moment conditions into an almost sure statement regarding a distribution:

> Theorem II: : If all the moments of the cumulative distribution function $F_N(x)$ converge almost surely (in probability) to those of $F(x)$ which possesses a density function and for which, with μ_{k+1} denoting the absolute moment of order k+1,
>
> $$\text{(IIa)}\quad \lim_{k\to\infty} \frac{\alpha^{k+2}\mu_{k+1}}{k+2!} = 0 \text{ for any given value of } \alpha,$$
>
> then $F_N(x)$ converges almost surely (in probability) to $F(x)$.

Condition (IIa) is, of course, true for the normal distribution. Hoeffding (1952) provides an alternate proof of convergence in probability without use of condition (IIa) to any $F(x)$ that is uniquely determined by its moments. By virtue of the Cramér-Wold device, Ghosh's Theorem covers the multivariate case given in (Ic) above, as for all $\boldsymbol{\lambda}$ such that $\boldsymbol{\lambda}'\boldsymbol{\lambda} = 1$, all moments of $\boldsymbol{\lambda}'\mathbf{n}(\mathbf{t}_i,\mathbf{d}_i)$ converge to those of the standard normal. Below, Theorems I and II and almost sure moment characteristics of the sequences $\mathbf{X}$, $\mathbf{Z}$ and $\boldsymbol{\varepsilon}$ guaranteed by White's assumptions W1 - W4 and the additional randomization assumptions A1 - A3 above are used to characterize the almost sure asymptotic distribution across row permutations $\mathbf{T}$ of $\mathbf{X}$ of products such as $\mathbf{T}'\mathbf{O}\boldsymbol{\varepsilon}/\sqrt{N}$.

A less demanding form of Theorem I provides weaker conditions under which the mean of products converges in probability across permutations to the product of means:

> Theorem III: Let $\mathbf{x}' = (x_1, ... , x_N)$ and $\mathbf{d}' = (d_1, ... , d_N)$ denote sequences of real numbers, possibly all equal, and $\mathbf{t}' = (t_1, ... , t_N)$ any of the $N!$ equally likely permutations of $\mathbf{x}$. Then as $N \to \infty$, across the permutations $\mathbf{t}$ of $\mathbf{x}$ the random variable
>
> $$\text{(IIIa)}\quad m(t_i d_i) - m(x_i)m(d_i) = \sum_{i=1}^{N} \frac{t_i d_i}{N} - \sum_{i=1}^{N} \frac{x_i}{N} \sum_{i=1}^{N} \frac{d_i}{N} \xrightarrow{p} 0,$$
>
> provided
>
> $$\text{(IIIb)}\quad \lim_{N\to\infty} \frac{\sum_{i=1}^{N} \frac{[x_i - m(x_i)]^2}{N} \sum_{i=1}^{N} \frac{[d_i - m(d_i)]^2}{N}}{N} = 0.$$

[53] With $\mathbf{E}$ the matrix of eigenvectors & $\boldsymbol{\Lambda}^{½}$ the diagonal matrix of square roots of the eigenvalues, $\mathbf{A}^{½} = \mathbf{E}\boldsymbol{\Lambda}^{½}\mathbf{E}'$.

If c_N is a sequence that converges to zero and the stronger condition

$$\text{(IIIc)} \sum_{i=1}^{N} \frac{[x_i - m(x_i)]^2}{N} \sum_{i=1}^{N} \frac{[d_i - m(d_i)]^2}{N} \text{ is asymptotically bounded}$$

holds, then across the permutations **t** of **x**

$$\text{(IIId)} \ \sqrt{N}[m(t_i d_i) - m(x_i)m(d_i)]c_N \xrightarrow{p} 0.$$

The proof is short and can be given here. If either the x_i or d_i are all identical ($x_i = x$ or $d_i = d$), then IIIa and IIId follow immediately as

$$\text{(A.1)} \ \sum_{i=1}^{N} \frac{t_i d_i}{N} = x \sum_{i=1}^{N} \frac{d_i}{N} = m(x_i)m(d_i) \ \text{ or } \ \sum_{i=1}^{N} \frac{t_i d_i}{N} = d \sum_{i=1}^{N} \frac{t_i}{N} = m(x_i)m(d_i).$$

Assuming this is not the case, we first use the symmetry and equal likelihood of permutations to calculate the expectation of t_i and products of t_i across the row permutations **t** of **x**:

$$\text{(A.2)} \ E_{\mathbf{t}}(t_i) = \sum_{i=1}^{N} \frac{x_i}{N} = m(x_i), \ E_{\mathbf{t}}(t_i^2) = \sum_{i=1}^{N} \frac{x_i^2}{N} = m(x_i^2)$$

$$\& \ E_{\mathbf{t}}(t_i t_{j \neq i}) = \sum_{i=1}^{N} \sum_{j=1}^{N} \frac{x_i x_j}{N(N-1)} - \sum_{i=1}^{N} \frac{x_i^2}{N(N-1)} = \frac{m(x_i)^2 N}{N-1} - \frac{m(x_i^2)}{N-1}.$$

Next, we calculate the mean and variance of $m(t_i d_i)$ - $m(x_i)m(d_i)$ across the realizations of **t**:

$$\text{(A.3)} \ E_{\mathbf{t}}[m(t_i d_i) - m(x_i)m(d_i)] = \sum_{i=1}^{N} \frac{E_{\mathbf{t}}(t_i)d_i}{N} - m(x_i)m(d_i) = 0,$$

$$E_{\mathbf{t}}[(m(t_i d_i) - m(x_i)m(d_i))^2] = E_{\mathbf{t}}[m(t_i d_i)^2] - m(x_i)^2 m(d_i)^2 = \sum_{i,j=1}^{N} \frac{E_{\mathbf{t}}(t_i t_j)d_i d_j}{N^2} + \sum_{i=1}^{N} \frac{E_{\mathbf{t}}(t_i^2)d_i^2}{N^2} - m(x_i)^2 m(d_i)^2$$

$$= \left(\frac{m(x_i)^2 N}{N-1} - \frac{m(x_i^2)}{N-1} \right) \left(m(d_i)^2 - \frac{m(d_i^2)}{N} \right) + m(x_i^2)\frac{m(d_i^2)}{N} - m(x_i)^2 m(d_i)^2$$

$$= \frac{[m(x_i^2) - m(x_i)^2][m(d_i^2) - m(d_i)^2]}{N-1},$$

where subscripted *i,j* denotes the summation across the two indices excluding ties between them. If the last line of (A.3) converges to 0 (condition IIIb), then across the permutations **t** of **x** $m(t_i d_i)$ - $m(x_i)m(d_i)$ converges in mean square and hence in probability to 0, as stated in IIIa. If $[m(x_i^2) - m(x_i)^2]$ * $[m(d_i^2) - m(d_i)^2]$ is bounded (condition IIIc), $\sqrt{N}[m(t_i d_i) - m(x_i)m(d_i)]$ is a mean zero random variable with bounded variance, so if $c_N \to 0$ then $\sqrt{N}[m(t_i d_i) - m(x_i)m(d_i)]c_N$ converges in probability to zero, as stated in the IIId.

Theorem III is used in proofs to make statements regarding the convergence in probability of means of products, such as $\mathbf{T}'\mathbf{Z}/N$. If the sequences **X** and **D** are such that conditions (IIIb) and (IIIc) hold almost surely and $c_N \xrightarrow{a.s.} 0$, we can speak of (IIIa) and (IIId) almost surely (across the sequences of the data) converging in probability (across the distribution generated by permutations **T** of **X** given the data). All references to almost sure convergence below refer to the sequences of the data, while all references to

convergence in probability and distribution are with respect to the permutations **T** of **X**.

B. Proof of (R1)

Following each permutation of treatment, the dependent variable is adjusted in accordance with the null and the realization **T** of treatment

$$\text{(B.1)}\quad \mathbf{y}(\mathbf{T},\boldsymbol{\beta}_0) = \mathbf{y} + (\mathbf{T}\bullet\mathbf{W} - \mathbf{X}\bullet\mathbf{W})\boldsymbol{\beta}_0 = \mathbf{X}_\mathbf{W}(\boldsymbol{\beta}-\boldsymbol{\beta}_0) + \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon} + \mathbf{T}_\mathbf{W}\boldsymbol{\beta}_0 .$$

With $\mathbf{M} = \mathbf{I} - \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'$ denoting the residual maker with respect to **Z**, using the fact that $\mathbf{MZ} = \mathbf{0}_{NxK}$ the estimated coefficients and residuals associated with **T** and $\boldsymbol{\beta}_0$ are seen to be

$$\text{(B.2)}\quad \hat{\boldsymbol{\beta}}(\mathbf{T},\boldsymbol{\beta}_0) \;=\; (\mathbf{T}_\mathbf{W}'\mathbf{M}\mathbf{T}_\mathbf{W})^{-1}\mathbf{T}_\mathbf{W}'\mathbf{M}\mathbf{y}(\mathbf{T},\boldsymbol{\beta}_0)$$

$$\Rightarrow\; \hat{\boldsymbol{\beta}}(\mathbf{T},\boldsymbol{\beta}_0) - \boldsymbol{\beta}_0 \;=\; (\mathbf{T}_\mathbf{W}'\mathbf{M}\mathbf{T}_\mathbf{W})^{-1}\mathbf{T}_\mathbf{W}'\mathbf{M}\mathbf{X}_\mathbf{W}(\boldsymbol{\beta}-\boldsymbol{\beta}_0) + (\mathbf{T}_\mathbf{W}'\mathbf{M}\mathbf{T}_\mathbf{W})^{-1}\mathbf{T}_\mathbf{W}'\mathbf{M}\boldsymbol{\varepsilon},$$

$$\text{(B.3)}\quad \hat{\boldsymbol{\varepsilon}}(\mathbf{T},\boldsymbol{\beta}_0) = \mathbf{M}\mathbf{y}(\mathbf{T},\boldsymbol{\beta}_0) - \mathbf{M}\mathbf{T}_\mathbf{W}\hat{\boldsymbol{\beta}}(\mathbf{T},\boldsymbol{\beta}_0) = \mathbf{M}\mathbf{X}_\mathbf{W}(\boldsymbol{\beta}-\boldsymbol{\beta}_0) + \mathbf{M}\boldsymbol{\varepsilon} - \mathbf{M}\mathbf{T}_\mathbf{W}(\hat{\boldsymbol{\beta}}(\mathbf{T},\boldsymbol{\beta}_0) - \boldsymbol{\beta}_0)$$

We begin by calculating the asymptotic distribution of the coefficients associated with permuted treatment and then establish the probability limits of the covariance estimates.

The following Lemma, proven in Appendix C below, will be useful:

Lemma 1: White's assumptions W1 - W4 and the additional A1 - A3 ensure that

(a) $\mathbf{Z}'\mathbf{Z}/N$, $\mathbf{W}'\mathbf{W}/N$, $\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}}/N$ & $\widetilde{\mathbf{W}}_\varepsilon'\widetilde{\mathbf{W}}_\varepsilon/N$ [where $\mathbf{W}_\varepsilon = \mathbf{W}\bullet\boldsymbol{\varepsilon}$] are all almost surely strictly positive definite with determinant $> \gamma > 0$ for all N sufficientlv large.

(b) $\mathbf{Z}'\boldsymbol{\varepsilon}/N \xrightarrow{a.s.} \mathbf{0}_K$, $\mathbf{X}_\mathbf{W}'\boldsymbol{\varepsilon}/N \xrightarrow{a.s.} \mathbf{0}_{PQ}$ & $\widetilde{\mathbf{W}}_\varepsilon'\widetilde{\mathbf{W}}_\varepsilon/N - \mathbf{W}_\varepsilon'\mathbf{W}_\varepsilon/N \xrightarrow{a.s.} \mathbf{0}_{QxQ}$.

(c) The sample means of the product of the elements of one through four columns of **X** or the elements of one, two or four columns of $\mathbf{D} = (\mathbf{X}_\mathbf{W},\mathbf{Z},\boldsymbol{\varepsilon})$ (no more than two of which are ε_i) are almost surely bounded, as are $(\mathbf{Z}'\mathbf{Z}/N)^{-1}$, $(\mathbf{W}'\mathbf{W}/N)^{-1}$, $(\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}}/N)^{-1}$ & $(\widetilde{\mathbf{W}}_\varepsilon'\widetilde{\mathbf{W}}_\varepsilon/N)^{-1}$.

(d) Condition IIIc of Theorem III almost surely holds for the mean of the product of the elements of one or two of the columns of **T** with the elements of two columns of $\mathbf{D} = (\mathbf{X}_\mathbf{W},\mathbf{Z},\boldsymbol{\varepsilon})$, no more than one of which is ε_i, so that with p and q denoting columns of **T** and j and k those of **D**

$$m(t_{ip}d_{ij}d_{ik}) - m(x_{ip})m(d_{ij}d_{ik}) \xrightarrow{p} 0\,,\; m(t_{ip}t_{iq}d_{ij}d_{ik}) - m(x_{ip}x_{iq})m(d_{ij}d_{ik}) \xrightarrow{p} 0 \;\&$$

$$\text{if } c_N \xrightarrow{a.s.} 0 \text{ then } \sqrt{N}[\,m(t_{ip}d_{ij}d_{ik}) - m(x_{ip})m(d_{ij}d_{ik})]c_N \xrightarrow{p} 0.$$

(e) x_{ip} & $w_{iq}\varepsilon_i$ almost surely satisfy condition Ib of Theorem I for all column pairs p of **X** and q of $\mathbf{W}_\varepsilon$, while $\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}}/N$ & $\widetilde{\mathbf{W}}_\varepsilon'\widetilde{\mathbf{W}}_\varepsilon/N$ are bounded with determinant $> \gamma > 0$ for all N sufficiently large, so that across the row permutations **T** of **X** we have

$$\left(\frac{\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}}}{N}\otimes\frac{\widetilde{\mathbf{W}}_\varepsilon'\widetilde{\mathbf{W}}_\varepsilon}{N}\right)^{-½}\frac{(\widetilde{\mathbf{T}}\bullet\widetilde{\mathbf{W}}_\varepsilon)'\mathbf{1}_N}{\sqrt{N}} \xrightarrow{d} \mathbf{n}_{PQ},\;\text{where } \mathbf{n}_{PQ} \sim \mathrm{N}(\mathbf{0}_{PQ},\mathbf{I}_{PQ}).$$

As elsewhere in this paper, almost sure limits are with respect to the data sequence $\mathbf{D} = (\mathbf{X}_\mathbf{W},\mathbf{Z},\boldsymbol{\varepsilon})$, while probability limits and limiting distributions are with respect to the probability distribution generated by the N! equally likely row permutations **T** of **X**.

(a) Asymptotic Distribution of Coefficient Estimates

Multiplying (B.2) by $\sqrt{N}$, we have

$$\text{(B.4)}\quad \sqrt{N}(\hat{\boldsymbol{\beta}}(\mathbf{T},\boldsymbol{\beta}_0)-\boldsymbol{\beta}_0) = \left(\frac{\mathbf{T}_{\mathbf{W}}'\mathbf{M}\mathbf{T}_{\mathbf{W}}}{N}\right)^{-1}\frac{\mathbf{T}_{\mathbf{W}}'\mathbf{M}\mathbf{X}_{\mathbf{W}}}{N}\mathbf{r}+\left(\frac{\mathbf{T}_{\mathbf{W}}'\mathbf{M}\mathbf{T}_{\mathbf{W}}}{N}\right)^{-1}\frac{\mathbf{T}_{\mathbf{W}}'\mathbf{M}\boldsymbol{\varepsilon}}{\sqrt{N}},$$

where $\mathbf{r}=\sqrt{N}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)$. Let $\mathbf{t}_{\mathbf{W}k}$ and $\mathbf{x}_{\mathbf{W}l}$ denote the k^{th} and l^{th} columns of $\mathbf{T}_{\mathbf{W}}$ and $\mathbf{X}_{\mathbf{W}}$, with the i^{th} elements of these vectors given by $t_{ip(k)}w_{iq(k)}$ and $x_{ip(l)}w_{iq(l)}$, where $p(j)$ and $q(j)$ denote the columns of $\mathbf{T}$ (or $\mathbf{X}$) and $\mathbf{W}$ associated with the j^{th} column of $\mathbf{T}_{\mathbf{W}}$ (or $\mathbf{X}_{\mathbf{W}}$). With this notation, we see that the kl^{th} element of $\mathbf{T}_{\mathbf{W}}'\mathbf{M}\mathbf{T}_{\mathbf{W}}/N$ can be expressed as

$$\text{(B.5)}\quad \frac{\mathbf{t}_{\mathbf{W}k}'\mathbf{M}\mathbf{t}_{\mathbf{W}l}}{N} = \frac{\mathbf{t}_{\mathbf{W}k}'[\mathbf{I}_N-\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}']\mathbf{t}_{\mathbf{W}l}}{N} = m(t_{ip(k)}t_{ip(l)}w_{iq(k)}w_{iq(l)})-\mathbf{m}(t_{ip(k)}w_{iq(k)}\mathbf{z}_i')\left(\frac{\mathbf{Z}'\mathbf{Z}}{N}\right)^{-1}\mathbf{m}(t_{ip(l)}w_{iq(l)}\mathbf{z}_i),$$

$$\text{so}\quad \frac{\mathbf{t}_{\mathbf{W}k}'\mathbf{M}\mathbf{t}_{\mathbf{W}l}}{N}-[m(x_{ip(k)}x_{ip(l)})-m(x_{ip(k)})m(x_{ip(l)})]m(w_{iq(k)}w_{iq(l)}) =$$

$$\underbrace{[m(t_{ip(k)}t_{ip(l)}w_{iq(k)}w_{iq(l)})-m(x_{ip(k)}x_{ip(l)})m(w_{iq(k)}w_{iq(l)})]}_{\xrightarrow{p}0\ \text{(Lemma 1d)}}+m(x_{ip(k)})m(x_{ip(l)})[m(w_{iq(k)}w_{iq(l)})-\underbrace{\mathbf{m}(w_{iq(k)}\mathbf{z}_i')\left(\frac{\mathbf{Z}'\mathbf{Z}}{N}\right)^{-1}\mathbf{m}(w_{iq(l)}\mathbf{z}_i)}_{\text{a.s.}=m(w_{iq(k)}w_{iq(l)})\ \text{for}\ N\ \text{sufficiently large (Lemma 1a)}}]$$

$$-\underbrace{[\mathbf{m}(t_{ip(k)}w_{iq(k)}\mathbf{z}_i')-m(x_{ip(k)})\mathbf{m}(w_{iq(k)}\mathbf{z}_i')]}_{\xrightarrow{p}\mathbf{0}_K'\ \text{(Lemma 1d)}}\underbrace{\left(\frac{\mathbf{Z}'\mathbf{Z}}{N}\right)^{-1}}_{\text{a.s. bounded (Lemma 1c)}}\underbrace{[\mathbf{m}(t_{ip(l)}w_{iq(l)}\mathbf{z}_i)-m(x_{ip(l)})\mathbf{m}(w_{iq(l)}\mathbf{z}_i)]}_{\xrightarrow{p}\mathbf{0}_K\ \text{(Lemma 1d)}}$$

$$-\underbrace{m(x_{ip(k)})\mathbf{m}(w_{iq(k)}\mathbf{z}_i')\left(\frac{\mathbf{Z}'\mathbf{Z}}{N}\right)^{-1}}_{\text{a.s. bounded (Lemma 1c)}}\underbrace{[\mathbf{m}(t_{ip(l)}w_{iq(l)}\mathbf{z}_i)-m(x_{ip(l)})\mathbf{m}(w_{iq(l)}\mathbf{z}_i)]}_{\xrightarrow{p}\mathbf{0}_K\ \text{(Lemma 1d)}}$$

$$-\underbrace{[\mathbf{m}(t_{ip(k)}w_{iq(k)}\mathbf{z}_i')-m(x_{ip(k)})\mathbf{m}(w_{iq(k)}\mathbf{z}_i')]}_{\xrightarrow{p}\mathbf{0}_K'\ \text{(Lemma 1d)}}\underbrace{\left(\frac{\mathbf{Z}'\mathbf{Z}}{N}\right)^{-1}m(x_{ip(l)})\mathbf{m}(w_{iq(l)}\mathbf{z}_i)}_{\text{a.s. bounded (Lemma 1c)}}\xrightarrow{p}0,$$

where $\mathbf{m}(t_{ip(k)}w_{iq(k)}\mathbf{z}_i')=(m(t_{ip(k)}w_{iq(k)}z_{i1}),\ldots,m(t_{ip(k)}w_{iq(k)}z_{iK}))$, and in the third line we use the fact that as $\mathbf{m}(w_{iq(k)},\mathbf{z}_i')=\mathbf{w}_{q(k)}'\mathbf{Z}/N$, where $\mathbf{w}_{q(k)}$ is the $q(k)^{th}$ column of $\mathbf{W}$ which is included in the covariates $\mathbf{Z}$ (assumption A2), so for all N sufficiently large that $\mathbf{Z}'\mathbf{Z}/N$ is guaranteed to be invertible $\mathbf{w}_{q(k)}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}$ is a row vector of zeros with a 1 in the column corresponding to the position of $\mathbf{w}_{q(k)}$ in $\mathbf{Z}$. Similarly, the kl^{th} element of $\mathbf{T}_{\mathbf{W}}'\mathbf{M}\mathbf{X}_{\mathbf{W}}/N$ can be expressed as

$$\text{(B.6)}\quad \frac{\mathbf{t}_{\mathbf{W}k}'\mathbf{M}\mathbf{x}_{\mathbf{W}l}}{N} = \frac{\mathbf{t}_{\mathbf{W}k}'[\mathbf{I}_N-\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}']\mathbf{x}_{\mathbf{W}l}}{N} = m(t_{ip(k)}w_{iq(k)}w_{iq(l)}x_{ip(l)})-\mathbf{m}(t_{ip(k)}w_{iq(k)}\mathbf{z}_i')\left(\frac{\mathbf{Z}'\mathbf{Z}}{N}\right)^{-1}\frac{\mathbf{Z}'\mathbf{x}_{\mathbf{W}l}}{N} =$$

$$\underbrace{[m(t_{ip(k)}w_{iq(k)}w_{iq(l)}x_{ip(l)})-m(x_{ip(k)})m(w_{iq(k)}w_{iq(l)}x_{ip(l)})]}_{\xrightarrow{p}0\ \text{(Lemma 1d)}}-\underbrace{[\mathbf{m}(t_{ip(k)}w_{iq(k)}\mathbf{z}_i')-m(x_{ip(k)})\mathbf{m}(w_{iq(k)}\mathbf{z}_i')]}_{\xrightarrow{p}\mathbf{0}_K'\ \text{(Lemma 1d)}}\underbrace{\left(\frac{\mathbf{Z}'\mathbf{Z}}{N}\right)^{-1}\frac{\mathbf{Z}'\mathbf{x}_{\mathbf{W}l}}{N}}_{\text{a.s. bounded (Lemma 1c)}}$$

$$+m(x_{ip(k)})m(w_{iq(k)}w_{iq(l)}x_{ip(l)})-m(x_{ip(k)})\quad\underbrace{\mathbf{m}(w_{iq(k)}\mathbf{z}_i')\left(\frac{\mathbf{Z}'\mathbf{Z}}{N}\right)^{-1}\frac{\mathbf{Z}'\mathbf{x}_{\mathbf{W}l}}{N}}_{\text{a.s.}=m(w_{iq(k)}w_{iq(l)}x_{ip(l)})\ \text{for}\ N\ \text{sufficiently large (Lemma 1a)}}\quad\xrightarrow{p}0,$$

where in the last line we again use the assumption that **W** is included in **Z**.

Combining these results, we have:

$$\text{(B.7)}\quad \frac{\mathbf{T}_{\mathbf{W}}'\mathbf{M}\mathbf{T}_{\mathbf{W}}}{N} - \frac{\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}}}{N}\otimes\frac{\mathbf{W}'\mathbf{W}}{N}\xrightarrow{p}\mathbf{0}_{PQ\,x\,PQ} \;\;\&\;\; \frac{\mathbf{T}_{\mathbf{W}}'\mathbf{M}\mathbf{X}_{\mathbf{W}}}{N}\xrightarrow{p}\mathbf{0}_{PQ\,x\,PQ}.$$

Finite values of $\mathbf{r}=\sqrt{N}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)$, multiplied by $\mathbf{T}_{\mathbf{W}}'\mathbf{M}\mathbf{X}_{\mathbf{W}}/N$, asymptotically have no influence in (B.4). As can be seen, assumption A2 that **W** is a part of **Z** (or the less plausible alternative that the mean of the x_{ip} are asymptotically zero) is key to this result.

The remaining part of (B.4) is the vector $\mathbf{T}_{\mathbf{W}}'\mathbf{M}\boldsymbol{\varepsilon}/\sqrt{N}$, the k^{th} term of which equals:

$$\text{(B.8)}\quad \frac{\mathbf{t}_{\mathbf{W}k}'\mathbf{M}\boldsymbol{\varepsilon}}{\sqrt{N}} = \sqrt{N}m(t_{ip(k)}w_{iq(k)}\varepsilon_i) - \sqrt{N}\mathbf{m}(t_{ip(k)}w_{iq(k)}\mathbf{z}_i')\left(\frac{\mathbf{Z}'\mathbf{Z}}{N}\right)^{-1}\frac{\mathbf{Z}'\boldsymbol{\varepsilon}}{N} = \underbrace{\sqrt{N}\begin{bmatrix} m(t_{ip(k)}w_{iq(k)}\varepsilon_i) - \\ m(x_{ip(k)})m(w_{iq(k)}\varepsilon_i)\end{bmatrix}}_{v_k}$$

$$\underbrace{-\sqrt{N}\begin{bmatrix}\mathbf{m}(t_{ip(k)}w_{iq(k)}\mathbf{z}_i') - \\ m(x_{ip(k)})\mathbf{m}(w_{iq(k)}\mathbf{z}_i')\end{bmatrix}\underbrace{\left(\frac{\mathbf{Z}'\mathbf{Z}}{N}\right)^{-1}\frac{\mathbf{Z}'\boldsymbol{\varepsilon}}{N}}_{\xrightarrow{a.s.}\mathbf{0}_K\text{ (Lemmas1b,1c)}}}_{\xrightarrow{p}0\text{ (Lemma1d)}} + \sqrt{N}\left[m(x_{ip(k)})m(w_{iq(k)}\varepsilon_i) - m(x_{ip(k)}) \quad \underbrace{\mathbf{m}(w_{iq(k)}\mathbf{z}_i')\left(\frac{\mathbf{Z}'\mathbf{Z}}{N}\right)^{-1}\frac{\mathbf{Z}'\boldsymbol{\varepsilon}}{N}}_{\text{a.s.}=m(w_{iq(k)}\varepsilon_i)\text{ for }N\text{ sufficiently large (Lemma1a)}}\right].$$

The only term that asymptotically is non-zero is v_k which, as $m(t_{ip(k)})=m(x_{ip(k)})$, equals the k^{th} element of $(\widetilde{\mathbf{T}}\bullet\widetilde{\mathbf{W}}_{\boldsymbol{\varepsilon}})'\mathbf{1}_N/\sqrt{N}$. Applying Lemma 1e we then see that

$$\text{(B.9)}\quad \left(\frac{\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}}}{N}\otimes\frac{\widetilde{\mathbf{W}}_{\boldsymbol{\varepsilon}}'\widetilde{\mathbf{W}}_{\boldsymbol{\varepsilon}}}{N}\right)^{-½}\frac{\mathbf{T}_{\mathbf{W}}'\mathbf{M}\boldsymbol{\varepsilon}}{\sqrt{N}}\xrightarrow{d}\mathbf{n}_{PQ},\ \text{where}\ \mathbf{n}_{PQ}\sim\mathrm{N}(\mathbf{0}_{PQ},\mathbf{I}_{PQ}),$$

$$\text{so ... (B.10)}\quad \overbrace{\left(\frac{\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}}}{N}\otimes\frac{\widetilde{\mathbf{W}}_{\boldsymbol{\varepsilon}}'\widetilde{\mathbf{W}}_{\boldsymbol{\varepsilon}}}{N}\right)^{-½}\left(\frac{\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}}}{N}\otimes\frac{\mathbf{W}'\mathbf{W}}{N}\right)}^{\text{almost surely bounded positive definite matrices (Lemmas1a,1c)}}\sqrt{N}(\hat{\boldsymbol{\beta}}(\mathbf{T},\boldsymbol{\beta}_0)-\boldsymbol{\beta}_0) =$$

$$\left(\frac{\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}}}{N}\otimes\frac{\widetilde{\mathbf{W}}_{\boldsymbol{\varepsilon}}'\widetilde{\mathbf{W}}_{\boldsymbol{\varepsilon}}}{N}\right)^{-½}\left(\frac{\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}}}{N}\otimes\frac{\mathbf{W}'\mathbf{W}}{N}\right)\underbrace{\left(\frac{\mathbf{T}_{\mathbf{W}}'\mathbf{M}\mathbf{T}_{\mathbf{W}}}{N}\right)^{-1}}_{\xrightarrow{p}\left(\frac{\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}}}{N}\otimes\frac{\mathbf{W}'\mathbf{W}}{N}\right)^{-1}}\underbrace{\frac{\mathbf{T}_{\mathbf{W}}'\mathbf{M}\mathbf{X}_{\mathbf{W}}}{N}}_{\xrightarrow{p}\mathbf{0}_{PQxPQ}}\underbrace{\mathbf{r}}_{\text{bounded}} +$$

$$\underbrace{\left(\frac{\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}}}{N}\otimes\frac{\widetilde{\mathbf{W}}_{\boldsymbol{\varepsilon}}'\widetilde{\mathbf{W}}_{\boldsymbol{\varepsilon}}}{N}\right)^{-½}\left(\frac{\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}}}{N}\otimes\frac{\mathbf{W}'\mathbf{W}}{N}\right)\left(\frac{\mathbf{T}_{\mathbf{W}}'\mathbf{M}\mathbf{T}_{\mathbf{W}}}{N}\right)^{-1}\left(\frac{\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}}}{N}\otimes\frac{\widetilde{\mathbf{W}}_{\boldsymbol{\varepsilon}}'\widetilde{\mathbf{W}}_{\boldsymbol{\varepsilon}}}{N}\right)^{½}}_{\xrightarrow{p}\mathbf{I}_{PQ}}\underbrace{\left(\frac{\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}}}{N}\otimes\frac{\widetilde{\mathbf{W}}_{\boldsymbol{\varepsilon}}'\widetilde{\mathbf{W}}_{\boldsymbol{\varepsilon}}}{N}\right)^{-½}\frac{\mathbf{T}_{\mathbf{W}}'\mathbf{M}\boldsymbol{\varepsilon}}{\sqrt{N}}}_{\xrightarrow{d}\mathbf{n}_{PQ}}\xrightarrow{d}\mathbf{n}_{PQ}.$$

(b) Probability Limit of the Heteroskedasticity Robust Covariance Estimate

For the heteroskedasticity robust covariance estimate we have

$$\text{(B.11)}\quad N\mathbf{V}_r(\hat{\boldsymbol{\beta}}(\mathbf{T},\boldsymbol{\beta}_0)) = \left(\frac{\mathbf{T}_{\mathbf{W}}'\mathbf{M}\mathbf{T}_{\mathbf{W}}}{N}\right)^{-1}\mathbf{A}\left(\frac{\mathbf{T}_{\mathbf{W}}'\mathbf{M}\mathbf{T}_{\mathbf{W}}}{N}\right)^{-1},\ \text{where}\ \mathbf{A}=\frac{(\mathbf{M}\mathbf{T}_{\mathbf{W}}\bullet\hat{\boldsymbol{\varepsilon}}(\mathbf{T},\boldsymbol{\beta}_0))'(\mathbf{M}\mathbf{T}_{\mathbf{W}}\bullet\hat{\boldsymbol{\varepsilon}}(\mathbf{T},\boldsymbol{\beta}_0))}{N}$$

with kl^{th} term given by

$$\text{(B.12)}\quad \mathbf{A}_{kl} = \frac{1}{N}\sum_{i=1}^{N}(t_{ip(k)}w_{iq(k)} - \sum_{a=1}^{K} z_{ia}\hat{\delta}_{ak})(t_{ip(l)}w_{iq(l)} - \sum_{b=1}^{K} z_{ib}\hat{\delta}_{bl})\hat{\varepsilon}_i(\mathbf{T},\boldsymbol{\beta}_0)^2,$$

$$\text{where } \hat{\varepsilon}_i(\mathbf{T},\boldsymbol{\beta}_0) = \varepsilon_i - \sum_{c=1}^{K} z_{ic}\hat{\eta}_c + \sum_{d=1}^{PQ}(x_{ip(d)}w_{iq(d)} - \sum_{e=1}^{K} z_{ie}\hat{\tau}_{ed})\frac{r_d}{\sqrt{N}} - \sum_{f=1}^{PQ}(t_{ip(f)}w_{iq(f)} - \sum_{g=1}^{K} z_{ig}\hat{\delta}_{gf})\frac{\hat{r}_f}{\sqrt{N}},$$

using the formula for $\hat{\boldsymbol{\varepsilon}}(\mathbf{T},\boldsymbol{\beta}_0)$ from (B.3) earlier with $\mathbf{r} = \sqrt{N}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)$, $\hat{\mathbf{r}} = \sqrt{N}(\hat{\boldsymbol{\beta}}(\mathbf{T},\boldsymbol{\beta}_0) - \boldsymbol{\beta}_0)$, $\hat{\boldsymbol{\delta}}_k = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{t}_{\mathbf{W}k}$, $\hat{\boldsymbol{\tau}}_k = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{x}_{\mathbf{W}k}$ and $\hat{\boldsymbol{\eta}} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\boldsymbol{\varepsilon}$. From Lemmas 1b and 1c we have $\hat{\boldsymbol{\eta}} \xrightarrow{a.s.} \mathbf{0}_K$ and from 1c know that $\hat{\boldsymbol{\tau}}_k$ is almost surely bounded. The plim of $\hat{\boldsymbol{\delta}}_k$ across the distribution of $\mathbf{T}$ is bounded as

$$\text{(B.13)}\quad \hat{\boldsymbol{\delta}}_k - \underbrace{m(x_{ip(k)})\left(\frac{\mathbf{Z}'\mathbf{Z}}{N}\right)^{-1}\frac{\mathbf{Z}'\mathbf{w}_{q(k)}}{N}}_{\text{a.s. bounded (Lemma 1c)}} = \underbrace{\left(\frac{\mathbf{Z}'\mathbf{Z}}{N}\right)^{-1}}_{\text{a.s. bounded (Lemma 1c)}} \underbrace{[\mathbf{m}(t_{ip(k)}w_{iq(k)}\mathbf{z}_i) - m(x_{ip(k)})\mathbf{m}(w_{iq(k)}\mathbf{z}_i)]}_{\xrightarrow{p}\mathbf{0}_K \text{ (Lemma 1d)}} \xrightarrow{p} 0.$$

As for all N sufficiently large $(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{w}_{q(k)}$ is a vector of zeros with a 1 in the row corresponding to the column position of $\mathbf{w}_{q(k)}$ in $\mathbf{Z}$ and $m(x_{ip(k)})$ is known to be bounded by Lemma 1c, plim $\hat{\delta}_{ak} = 0$ unless a is the column position of $\mathbf{w}_{q(k)}$ in $\mathbf{Z}$, in which case plim $\hat{\delta}_{ak} - m(x_{ip(k)}) = 0$. The elements of $\mathbf{r}$ are finite and of $\hat{\mathbf{r}}$ are asymptotically multivariate normal with almost surely bounded variance, so when divided by a positive power of N have a plim of zero.

When the terms in (B.12) are multiplied out, most involve a product with an element of $\mathbf{r}/\sqrt{N}$, $\hat{\mathbf{r}}/\sqrt{N}$, or $\hat{\boldsymbol{\eta}}$ that has a plim of zero, 0 to 4 parameters $\hat{\tau}$ and $\hat{\delta}$ with bounded probability limits, and the mean of the product of the elements of 0 to 4 columns of $\mathbf{T}$ and the elements of 4 columns of $\mathbf{D} = (\mathbf{X}_\mathbf{W},\mathbf{Z},\boldsymbol{\varepsilon})$ (no more than two of which are ε_i). The following lemma shows that the plim of all such terms is zero:

Lemma 2: White's assumptions W1 - W4 and the additional A1 - A3 ensure that for some a in (0,½) condition IIIb of Theorem III almost surely holds for the mean of the product of the elements of n = one through four columns of $\mathbf{T}$ divided by $N^{a\max(n-2,0)}$ with the elements of four columns of $\mathbf{D} = (\mathbf{X}_\mathbf{W},\mathbf{Z},\boldsymbol{\varepsilon})$, no more than two of which are ε_i, so that across the permutations $\mathbf{T}$ of $\mathbf{X}$

$$m(N^{-a\max(n-2,0)}(\prod_{o=1}^{n} t_{ip(o)})d_{ij}d_{ik}d_{il}d_{im}) - m(N^{-a\max(n-2,0)}\prod_{o=1}^{n} x_{ip(o)})m(d_{ij}d_{ik}d_{il}d_{im}) \xrightarrow{p} 0.$$

The sample means of the product of the elements of one through four columns of $\mathbf{X}$ or four columns of $\mathbf{D}$ are almost surely bounded (Lemma 1c), so the probability limit in the Lemma is bounded when n = 1 or 2 and 0 when n = 3 or 4. Consequently, in (B.12) every term that involves the product of an element of $\mathbf{r}/\sqrt{N}$, $\hat{\mathbf{r}}/\sqrt{N}$, or $\hat{\boldsymbol{\eta}}$ that has a plim of zero with the mean of the product of four columns of $\mathbf{D}$ with zero, one or two columns of $\mathbf{T}$ has a probability limit of zero. Every term in (B.12) that involves the product of n = three or four columns of $\mathbf{T}$ with four columns of $\mathbf{D}$ also includes at least n - 2 $\hat{\mathbf{r}}/\sqrt{N}$ terms which can be re-expressed as $(\hat{\mathbf{r}}/N^{½-a})(1/N^a)$ for some a in (0,½). The $1/N^a$ parts can be used to satisfy Lemma 2, while the $\hat{\mathbf{r}}/N^{½-a}$ part converges in probability to 0. Thus, all such terms also have a plim of 0.

The above only leaves terms in (B.12) that involve the product of two or less columns of $\mathbf{T}$ and do not include an element of $\mathbf{r}/\sqrt{N}$, $\hat{\mathbf{r}}/\sqrt{N}$, or $\hat{\boldsymbol{\eta}}$, namely

$$\text{(B.14)}\ \sum_{i=1}^{N}\frac{t_{ip(k)}w_{iq(k)}t_{ip(l)}w_{iq(l)}\varepsilon_i^2}{N}-\sum_{a=1}^{K}\hat{\delta}_{ak}\sum_{i=1}^{N}\frac{t_{ip(l)}w_{iq(l)}z_{ia}\varepsilon_i^2}{N}-\sum_{b=1}^{K}\hat{\delta}_{bl}\sum_{i=1}^{N}\frac{t_{ip(k)}w_{iq(k)}z_{ib}\varepsilon_i^2}{N}+\sum_{a=1}^{K}\sum_{b=1}^{K}\hat{\delta}_{ak}\hat{\delta}_{bl}\sum_{i=1}^{N}\frac{z_{ia}z_{ib}\varepsilon_i^2}{N}$$

$$=m(t_{ip(k)}t_{ip(l)}w_{iq(k)}w_{iq(l)}\varepsilon_i^2)-\sum_{a=1}^{K}\hat{\delta}_{ak}m(t_{ip(l)}w_{iq(l)}z_{ia}\varepsilon_i^2)-\sum_{b=1}^{K}\hat{\delta}_{bl}m(t_{ip(k)}w_{iq(k)}z_{ib}\varepsilon_i^2)+\sum_{a=1}^{K}\sum_{b=1}^{K}\hat{\delta}_{ak}\hat{\delta}_{bl}m(z_{ia}z_{ib}\varepsilon_i^2)$$

where $m(t_{ip(k)}t_{ip(l)}w_{iq(k)}w_{iq(l)}\varepsilon_i^2)-m(x_{ip(k)}x_{ip(l)})m(w_{iq(k)}w_{iq(l)}\varepsilon_i^2)\underbrace{\to}_{\text{Lemma 2}}0$ & $m(t_{ip(l)}w_{iq(l)}z_{ia}\varepsilon_i^2)-m(x_{ip(l)})m(w_{iq(l)}z_{ia}\varepsilon_i^2)\underbrace{\to}_{\text{Lemma 2}}0$,

$$\text{so } \mathbf{A}_{kl}-[m(x_{ip(k)}x_{ip(l)})-m(x_{ip(k)})m(x_{ip(l)})]m(w_{iq(k)}w_{iq(l)}\varepsilon_i^2)\xrightarrow{p}0,$$

where we use the boundedness of means of products of up to four terms (Lemma 1c) and the fact noted above that plim $\hat{\delta}_{ak}=0$ unless a is the column position of $\mathbf{w}_{q(k)}$ in $\mathbf{Z}$, in which case plim $\hat{\delta}_{ak}=m(x_{ip(k)})$ and $z_{ia}=w_{iq(k)}$. This allows us to state that

$$\text{(B.15)}\ \ \mathbf{A}-\frac{\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}}}{N}\otimes\frac{\mathbf{W}_{\boldsymbol{\varepsilon}}'\mathbf{W}_{\boldsymbol{\varepsilon}}}{N}\xrightarrow{p}\mathbf{0}_{PQxPQ}$$

and consequently for the heteroskedasticity robust covariance estimate we have

$$\text{(B.16)}\ N\mathbf{V}_r(\hat{\boldsymbol{\beta}}(\mathbf{T},\boldsymbol{\beta}_0))-\left(\frac{\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}}}{N}\otimes\frac{\mathbf{W}'\mathbf{W}}{N}\right)^{-1}\left(\frac{\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}}}{N}\otimes\frac{\mathbf{W}_{\boldsymbol{\varepsilon}}'\mathbf{W}_{\boldsymbol{\varepsilon}}}{N}\right)\left(\frac{\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}}}{N}\otimes\frac{\mathbf{W}'\mathbf{W}}{N}\right)^{-1}\xrightarrow{p}\mathbf{0}_{PQxPQ}\ ,$$

which from (B.10) and Lemma 1b is seen to be the asymptotic covariance matrix of normally distributed $\sqrt{N}(\hat{\boldsymbol{\beta}}(\mathbf{T},\boldsymbol{\beta}_0)-\boldsymbol{\beta}_0)$. This establishes that for bounded $\mathbf{r}=\sqrt{N}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)$ the distribution of the Wald statistic $\tau(\mathbf{T},\boldsymbol{\beta}_0)$ across permutations $\mathbf{T}$ converges to that of the chi-squared with PQ degrees of freedom. Moreover, every appearance of $\mathbf{r}$ in the equation for $\tau(\mathbf{T},\boldsymbol{\beta}_0)$ is multiplied by a term that in probability across $\mathbf{T}$ converges to 0, so that in probability $\tau(\mathbf{T},\boldsymbol{\beta}_0)$ converges to $\tau(\mathbf{T},\boldsymbol{\beta})$, as stated in (R1).

C. Proofs of Lemmas used in Appendix B

We make use below of corollaries to Markov's Law of Large Numbers and the Continuous Mapping Theorem given by White (1984) and to the Borel-Cantelli Lemma given by Galambos (1978):

Markov Corollary: Let d_i be a sequence of independent random variables such that $E(|d_i|^{1+\delta})<\Delta<\infty$ for some $\delta>0$ and all i. Then $m(d_i)-m((E(d_i))\xrightarrow{a.s.}0$.

Continuous Mapping Theorem Corollary: Let g: $\mathbb{R}^k\to\mathbb{R}^l$ be continuous on a compact set $C\subset\mathbb{R}^k$. Suppose that $b_N(\omega)$ and c_N are $kx1$ vectors such that $b_N(\omega)$ - $c_N\xrightarrow{a.s.}0$, and for all N sufficiently large, c_N is interior to C uniformly in N. Then $g(b_N(\omega))$ - $g(c_N)\xrightarrow{a.s.}0$.

Borel-Cantelli Corollary: Let d_1, d_2, ... be an infinite sequence of random variables, $F_j(d)$ the probability $d_j<d$, and u_N a non-decreasing sequence of real numbers such that for all j $\Pr\{d_j<\sup_N u_N\}=1$, then

$$\sum_{j=1}^{\infty}[1-F_j(u_j)]<\infty\Rightarrow\Pr\{\operatorname*{Max}_{i\le N}d_i\ge u_N\}\text{ infinitely often}=0.$$

Lemma 1a: By assumption W2 that $E(|z_{+ij}z_{+ik}|^{1+\delta})<\Delta$, and the Markov and Continuous Mapping Theorem Corollaries, $\mathbf{Z}_+'\mathbf{Z}_+/N$ converges almost surely to the matrix $\mathbf{M}_N$ in W2 and is non-singular with determinant $>\gamma>0$ for all N sufficiently large. Using Jensen's Inequality on $E(|z_{+ij}z_{+ik}|^{1+\delta})<\Delta$, its trace is

almost surely bounded from above by $\Delta^{1/(1+\delta)}K_+$. By the trace and determinant property of eigenvalues we then know its smallest eigenvalue is greater than $\lambda = \gamma/(\Delta^{1/(1+\delta)}K_+)\text{^}(K_+\text{-}1) > 0$. For a real symmetric matrix **V**, the min across all non-zero vectors **x** of the Rayleigh quotient $\mathbf{x}'\mathbf{V}\mathbf{x}/\mathbf{x}'\mathbf{x}$ equals the minimum eigenvalue of **V**. Since **Z** and **W** are part of $\mathbf{Z}_+$, it follows that the minimum eigenvalues of the sub-matrices $\mathbf{Z}'\mathbf{Z}/N$ and $\mathbf{W}'\mathbf{W}/N$ are greater than or equal to that of $\mathbf{Z}_+'\mathbf{Z}_+/N$. Consequently, for all N sufficiently large both matrices are almost surely positive definite with determinants $> \lambda^K$ and λ^Q, respectively.

By Jensen's Inequality the assumption $E(|x_{ip}^4|^{1+\theta^*}) < \Delta$ in A3 implies that $E(|x_{ip}^n|^{1+\theta^*}) < \Delta^{n/4}$ for n = 1...3, so by the Markov Corollary

$$\text{(C.1)}\quad \widetilde{\mathbf{X}}'\widetilde{\mathbf{X}}/N - \sum_{i=1}^{N} E(\mathbf{x}_i\mathbf{x}_i')/N + \sum_{i=1}^{N} E(\mathbf{x}_i)/N \sum_{i=1}^{N} E(\mathbf{x}_i')/N \xrightarrow{a.s.} \mathbf{0}_{PxP} \quad ,$$

which using A1 and the Continuous Mapping Theorem Corollary implies that the determinant of $\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}}/N$ is almost surely greater than $\gamma > 0$ for all N sufficiently large. By assumption W3 $E(|\varepsilon_i^2 z_{+ij} z_{+ik}|^{1+\delta}) < \Delta$ and, using Jensen's Inequality, $E(|\varepsilon_i z_{+ik}|^{1+\delta}) \le E(|\varepsilon_i^2 z_{+ik}^2|^{1+\delta})^{1/2} < \Delta^{1/2} < \infty$. Applying the Markov Corollary

$$\text{(C.2)}\quad \widetilde{\mathbf{W}}_{\boldsymbol{\varepsilon}}'\widetilde{\mathbf{W}}_{\boldsymbol{\varepsilon}}/N - \textstyle\sum_{i=1}^{N} E(\varepsilon_i^2 \mathbf{w}_i\mathbf{w}_i')/N \xrightarrow{a.s.} \mathbf{0}_{QxQ},$$

where we make use of assumption W1 $E(\varepsilon_i \mathbf{z}_{+i}) = \mathbf{0}_{K+}$, so that $m(E(\varepsilon_i \mathbf{w}_i)) = \mathbf{0}_Q$. As the minimum eigenvalue of $\mathbf{V}_N = \sum_{i=1}^{N} E(\varepsilon_i^2 \mathbf{z}_{+i}\mathbf{z}_{+i}')/N$ given in W3 is again greater than $\lambda = \gamma/(K_+ \Delta^{1/(1+\delta)})\text{^}(K_+\text{-}1)$ and $\mathbf{w}_i$ is part of $\mathbf{z}_{+i}$, following the argument above we can say that almost surely for all N sufficiently large $\widetilde{\mathbf{W}}_{\boldsymbol{\varepsilon}}'\widetilde{\mathbf{W}}_{\boldsymbol{\varepsilon}}/N$ is positive definite with determinant greater than $\lambda^Q > 0$.

Lemma 1b: Above we saw that $E(|\varepsilon_i z_{+ik}|^{1+\delta}) < \Delta^{1/2} < \infty$ and by W1 $E(\varepsilon_i \mathbf{z}_{+i}) = \mathbf{0}_{K+}$, so by the Markov Corollary $m(\varepsilon_i \mathbf{z}_{+i}) \xrightarrow{a.s.} \mathbf{0}_{K+}$, which establishes the first part as **Z** and $\mathbf{X}_\mathbf{W}$ are both part of $\mathbf{Z}_+$. The second part then follows as $\widetilde{\mathbf{W}}_{\boldsymbol{\varepsilon}}'\widetilde{\mathbf{W}}_{\boldsymbol{\varepsilon}}/N - \mathbf{W}_{\boldsymbol{\varepsilon}}'\mathbf{W}_{\boldsymbol{\varepsilon}}/N = -m(\varepsilon_i \mathbf{w}_i)m(\varepsilon_i \mathbf{w}_i)'$ and **W** is part of **Z**.

Lemma 1c: Regarding the sample mean of the product of the elements of one, two or four columns of $\mathbf{D} = (\mathbf{X}_\mathbf{W}, \mathbf{Z}, \boldsymbol{\varepsilon}) = (\mathbf{Z}_+, \boldsymbol{\varepsilon})$ (no more than two of which are ε_i) from W2 and W4 we know that for positive finite constants δ and Δ $E(|\varepsilon_i^2|^{1+\delta}) < \Delta$ and $E(|z_{+ik}^4|^{1+\delta}) < \Delta$. Jensen's Inequality then tells us that $E(|\varepsilon_i|^{1+\delta}) < \Delta^{1/2}$ and $E(|z_{+ik}^n|^{1+\delta}) < \Delta^{n/4}$ for n = 1, 2 or 3, and also that $E(|\varepsilon_i^n|) < \Delta^{n/2(1+\delta)}$ for n = 1 or 2 and $E(|z_{+ik}^n|) < \Delta^{n/4(1+\delta)}$ for n = 1 through 4. Applying the Markov Corollary, we have

$$\text{(C.3)}\quad m(\varepsilon_i^n) - m(E(\varepsilon_i^n)) \xrightarrow{a.s.} 0 \text{ (n = 1 or 2)} \quad \& \quad m(z_{+ik}^n) - m(E(z_{+ik}^n)) \xrightarrow{a.s.} 0 \text{ (n = 1, 2, 3 or 4)}$$

$$\text{which, as } |m(E(\varepsilon_i^n))| \le m(E(|\varepsilon_i^n|)) < \Delta^{n/2(1+\delta)} \quad \& \quad |m(E(z_{+ik}^n))| \le m(E(|z_{+ik}^n|)) < \Delta^{n/4(1+\delta)},$$

tells us that both $m(\varepsilon_i^n)$ & $m(z_{+ik}^n)$ are almost surely bounded. The proof of 1b showed that $m(\varepsilon_i z_{+ik})$ converges almost surely to zero. As by W3 $E(|\varepsilon_i^2 z_{+ij} z_{+ik}|^{1+\delta}) < \Delta$ by a similar application of Jensen's Inequality and the Markov Corollary the sample mean of the product of ε_i^2 with any two of the columns of $\mathbf{Z}_+$ is seen to be almost surely bounded. What remains is the product of two or four separate columns of $\mathbf{Z}_+$ or the product of ε_i with any three of the columns of $\mathbf{Z}_+$, but given the preceding results these can be bounded by repeatedly applying the Cauchy-Schwarz Inequality

$$(C.4)\ |m(z_{+ij}z_{+ik})| \le \sqrt{m(z_{+ij}^2)m(z_{+ik}^2)},\quad |m(z_{+ij}z_{+ik}z_{+il}z_{+im})| \le \sqrt[4]{m(z_{+ij}^4)m(z_{+ik}^4)m(z_{+il}^4)m(z_{+im}^4)}$$
$$\&\ |m(\varepsilon_i z_{+ij}z_{+ik}z_{+il})| \le \sqrt{m(\varepsilon_i^2 z_{+ij}^2)}\sqrt[4]{m(z_{+ik}^4)m(z_{+il}^4)},$$

as the right hand sides of these expressions have already been shown to be almost surely bounded.

Turning to the mean of the product of the elements of one through four columns of **X**, by Jensen's Inequality the assumption $E(|x_{ip}^4|^{1+\theta^*}) < \Delta$ in A3 implies that $E(|x_{ip}^n|^{1+\theta^*}) < \Delta^{n/4}$ and $E(|x_{ip}^n|) < \Delta^{n/4(1+\theta^*)}$ for n = 1, 2, 3 or 4, so by the Markov Corollary $m(x_{ip}^n) - m(E(x_{ip}^n)) \xrightarrow{a.s.} 0$ where $|m(E(x_{ip}^n))| \le m(E(|x_{ip}^n|)) < \Delta^{n/4(1+\theta^*)}$. The sample mean of products of 2, 3 or 4 different columns of **X** can then be bounded using the Cauchy-Schwarz inequality, as was done in (C.4) above. With regards to the matrix inverses in the Lemma, for an invertible positive definite matrix **A** the largest eigenvalue of $\mathbf{A}^{-1}$ is equal to the inverse of the smallest eigenvalue of **A**. Since from the proof of Lemma 1a, we know that each of the matrices in Lemma 1a is invertible with a smallest eigenvalue almost surely greater than some $\lambda > 0$ for all N sufficiently large, it follows that the elements of their inverses are all almost surely bounded.

Lemma 1d: We are considering the mean of the product of the elements of one or two of the columns of **T** with the elements of two columns of **D** = (**X**$_{\mathbf{W}}$,**Z**,**ε**) (no more than one of which is ε_i). To prove IIIc we need only show that the sample moments of the product of the elements of one, two or four columns of **T** and the product of two (at most one of which is ε_i) or four columns (of which possibly two are ε_i) of **D** are bounded. This has already been established in 1c.

Lemma 1e: Lemmas 1a and 1c already established that $\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}}/N$ & $\widetilde{\mathbf{W}}_{\boldsymbol{\varepsilon}}'\widetilde{\mathbf{W}}_{\boldsymbol{\varepsilon}}/N$ are almost surely bounded with determinant > $\gamma > 0$ for all N sufficiently large, so all that remains is condition Ib. Define $w_{iq\varepsilon} = w_{iq}\varepsilon_i$ and, as elsewhere, let superscripted ~ denote sample demeaned values. Our objective is to prove that for all integer $\tau > 2$ and all p and q

$$(C.5)\ N^{\frac{\tau}{2}-1}\sum_{i=1}^N \widetilde{x}_{ip}^{\tau}\sum_{i=1}^N \widetilde{w}_{iq\varepsilon}^{\tau} \Bigg/ \left(\sum_{i=1}^N \widetilde{x}_{ip}^2\right)^{\tau/2}\left(\sum_{i=1}^N \widetilde{w}_{iq\varepsilon}^2\right)^{\tau/2} \xrightarrow{a.s.} 0.$$

We begin by noting that:

$$(C.6)\ \left|\frac{N^{\frac{\tau}{2}-1}\sum_{i=1}^N \widetilde{x}_{ip}^{\tau}\sum_{i=1}^N \widetilde{w}_{iq\varepsilon}^{\tau}}{\left(\sum_{i=1}^N \widetilde{x}_{ip}^2\sum_{i=1}^N \widetilde{w}_{iq\varepsilon}^2\right)^{\tau/2}}\right| \le \left|\frac{N^{\frac{\tau}{2}-1}\left(\underset{i\le N}{\mathrm{Max}}\,\widetilde{x}_{ip}^2\ \underset{i\le N}{\mathrm{Max}}\,\widetilde{w}_{iq\varepsilon}^2\right)^{\frac{\tau}{2}-1}\sum_{i=1}^N \widetilde{x}_{ip}^2\sum_{i=1}^N \widetilde{w}_{iq\varepsilon}^2}{\left(\sum_{i=1}^N \widetilde{x}_{ip}^2\sum_{i=1}^N \widetilde{w}_{iq\varepsilon}^2\right)^{\tau/2}}\right| = \left|\frac{\underset{i\le N}{\mathrm{Max}}\,\widetilde{x}_{ip}^2\ \underset{i\le N}{\mathrm{Max}}\,\widetilde{w}_{iq\varepsilon}^2}{N\sum_{i=1}^N \frac{\widetilde{x}_{ip}^2}{N}\sum_{i=1}^N \frac{\widetilde{w}_{iq\varepsilon}^2}{N}}\right|^{\frac{\tau}{2}-1}.$$

From A3 we have bounds, $E(|w_{iq}^2\varepsilon_i^2|^{1+\theta}) < \Delta$ and $E(|x_{ip}^4|^{1+\theta^*}) < \Delta$ for some θ and $\theta^* > 0$ & $\Delta < \infty$, which by Jensen's Inequality imply that $E(|w_{iq}\varepsilon_i|^{1+\theta}) < \Delta^{½}$ & $E(|x_{ip}^n|^{1+\delta}) < \Delta^{n/4}$ for $n = 1$ or 2, as well as $E(w_{iq}^2\varepsilon_i^2) < \Delta^{1/(1+\theta)}$, $E(|w_{iq}\varepsilon_i|) < \Delta^{1/2(1+\theta)}$, $E(x_{ip}^2) < \Delta^{1/2(1+\theta^*)}$ & $E(|x_{ip}|) < \Delta^{1/4(1+\theta^*)}$. Consequently, with regards to the denominator above, by the Markov Corollary we know that for $d_i = x_{ip}$ or $w_{iq\varepsilon}$

$$(C.7)\ [m(d_i^2) - m(d_i)^2] - [m(E(d_i^2)) - m(E(d_i))^2] \xrightarrow{a.s.} 0.$$

For a K x K positive definite matrix **V** with determinant > $\gamma > 0$ and non-negative diagonal elements bounded from above by Δ', by the trace and determinant property of eigenvalues the smallest eigenvalue is

bounded from below by $\lambda(K)= \gamma/(K\Delta')^{(K-1)}$. By the Schur-Horn Theorem, the smallest diagonal element of a real symmetric matrix is greater than or equal to its smallest eigenvalue. Lemmas 1a and 1c showed that almost surely $\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}}/N$ & $\widetilde{\mathbf{W}}_{\boldsymbol{\varepsilon}}'\widetilde{\mathbf{W}}_{\boldsymbol{\varepsilon}}/N$ are bounded with determinants > γ for all N sufficiently large, so almost surely the denominator on the right of (C.6) is > $\lambda(P)\lambda(Q) > 0$ for all N sufficiently large.

Turning to the numerator, since

$$\text{(C.8)}\ \operatorname*{Max}_{i\le N}\widetilde{d}_i^{\,2} \le \operatorname*{Max}_{i\le N} d_i^2 + 2\sqrt{\operatorname*{Max}_{i\le N}(d_i^2)}\,|m(d_i)| + m(d_i)^2$$

and $m(w_{iq}\varepsilon_i)$ converges to zero while $m(x_{ip})$ is bounded (Lemma 1c), to prove (Ib) all that remains is to show that $\operatorname{Max}_{i\le N}x_{ip}^2\operatorname{Max}_{i\le N}w_{iq\varepsilon}^2/N \xrightarrow{a.s.} 0$. From assumption A3 we know that there exist finite positive constants θ, θ^* and Δ, with $\theta(1+2\theta^*) > 1$, such that $E(|x_i^4|^{1+\theta^*}) < \Delta$ and $E(|w_{iq\varepsilon}^2|^{1+\theta}) < \Delta$. Consequently, applying Markov's Inequality

$$\text{(C.9)}\ \sum_{N=1}^{\infty}\Pr\{x_{Np}^2 \ge N^a\} = \sum_{N=1}^{\infty}\Pr\{x_{Np}^4 \ge N^{2a}\} \le \sum_{N=1}^{\infty}\frac{\Delta}{N^{2a(1+\theta^*)}} < \infty \ \text{ if } \ 2a(1+\theta^*) > 1$$

$$\&\ \sum_{N=1}^{\infty}\Pr\{w_{Nq\varepsilon}^2 \ge N^b\} \le \sum_{N=1}^{\infty}\frac{\Delta}{N^{b(1+\theta)}} < \infty \ \text{ if } \ b(1+\theta) > 1.$$

Both conditions can be met with $a > 0$, $b > 0$ and $a + b < 1$ if $\theta(1+2\theta^*) > 1$ as

$$\text{(C.10)}\ 1 > a+b > \frac{1}{2(1+\theta^*)} + \frac{1}{1+\theta} = 1 - \frac{\theta(1+2\theta^*)-1}{2(1+\theta^*)(1+\theta)}$$

poses no contradiction. Applying the Borel-Cantelli Corollary, we see that $\operatorname{Max}_{i\le N}x_{ip}^2\operatorname{Max}_{i\le N}w_{iq\varepsilon}^2/N^{a+b}$ is almost surely bounded by 1, so $\operatorname{Max}_{i\le N}x_{ip}^2\operatorname{Max}_{i\le N}w_{iq\varepsilon}^2/N \xrightarrow{a.s.} 0$.

Lemma 2: The sample means of 1 through 4 columns of $\mathbf{X}$ and any 4 columns of $\mathbf{D} = (\mathbf{X_W},\mathbf{Z},\boldsymbol{\varepsilon})$ (no more than two of which are $\boldsymbol{\varepsilon}$) are known to be almost surely bounded (Lemma 1c), so to establish condition IIIb it suffices that there exists an a in (0,½) such that the following are almost surely bounded

$$\text{(C.11)}\ m\left(\frac{x_{ij}^2x_{ik}^2x_{il}^2}{N^{2a}}\right) \le \max_{i\le N}\frac{x_{ij}^2}{N^{2a}}m(x_{ik}^2x_{il}^2),\ m\left(\frac{x_{ij}^2x_{ik}^2x_{il}^2x_{im}^2}{N^{4a}}\right) \le \max_{i\le N}\frac{x_{ij}^2}{N^{2a}}\max_{i\le N}\frac{x_{im}^2}{N^{2a}}m(x_{ik}^2x_{il}^2)$$

$$\&\ \sum_{i=1}^{N}\frac{d_{ij}^2d_{ik}^2d_{il}^2d_{im}^2}{N^2} \le \max_{i\le N}\frac{d_{ij}^2d_{ik}^2}{N}m(d_{il}^2d_{im}^2)$$

where we select indices so that when two elements are ε_i, d_{ij} represents one and d_{il} the other. Lemma1c already established that $m(x_{ik}^2x_{il}^2)$ & $m(d_{il}^2d_{im}^2)$ are almost surely bounded and in the proof of 1e we saw that $\operatorname{Max}_{i\le N}x_{ij}^2/N^a$ is almost surely bounded for ½ > a > ½(1+θ^*)$^{-1}$. From (W3), (W4) & Hölder's Inequality we have $E(|z_{ij}^2\varepsilon_i^2|^{1+\delta}) < \Delta$ & $E(|z_{ij}^2z_{ik}^2|^{1+\delta}) < E(|z_{ij}^4|^{1+\delta})^{1/2}E(|z_{ik}^4|^{1+\delta})^{1/2} < \Delta$. Applying Markov's Inequality

$$\text{(C.12)}\ \sum_{N=1}^{\infty}\Pr\{d_{Nj}^2d_{Nk}^2 \ge N^b\} \le \sum_{N=1}^{\infty}\frac{\Delta}{N^{b(1+\delta)}} < \infty \ \text{ if } \ b(1+\delta) > 1.$$

As $\delta > 0$, we know that there exists a $b < 1$ such that (C.14) holds. so by the Borel-Cantelli Corollary $\max_{i\le N}d_{ij}^2d_{ik}^2/N^b$ is almost surely bounded and $\max_{i\le N}d_{ij}^2d_{ik}^2/N \xrightarrow{a.s.} 0$, which completes the proof.

Figure A1

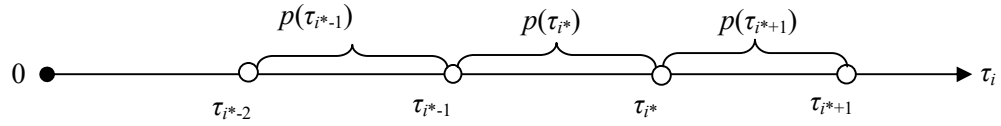


Table A2: Notation Used in Appendix D

(a) Coefficient & heteroskedasticity robust covariance estimates for the regression using $\mathbf{X}$, $\hat{\boldsymbol{\beta}}$ & $\hat{\mathbf{V}}(\hat{\boldsymbol{\beta}})$, and using permutation $\mathbf{T}$ of $\mathbf{X}$ under the null $\boldsymbol{\beta}_0$, $\hat{\boldsymbol{\beta}}(\mathbf{T},\boldsymbol{\beta}_0)$ & $\hat{\mathbf{V}}(\hat{\boldsymbol{\beta}}(\mathbf{T},\boldsymbol{\beta}_0))$.

(b) Subsets of parameters and the null: $\boldsymbol{\beta}^{\subseteq} = \mathbf{P}\boldsymbol{\beta}$ & $\boldsymbol{\beta}_0^{\subseteq} = \mathbf{P}\boldsymbol{\beta}_0$, where $\mathbf{P}$ is a k x PQ matrix of zeros of full rank with a single one in each row.

(c) Wald statistics for subset tests: $\tau^{\subseteq}$, as in $\tau^{\subseteq}(\mathbf{X},\boldsymbol{\beta}_0^{\subseteq}) = (\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)'\mathbf{P}'[\mathbf{P}\hat{\mathbf{V}}(\hat{\boldsymbol{\beta}})\mathbf{P}']^{-1}\mathbf{P}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)$ & $\tau^{\subseteq}(\mathbf{T},\boldsymbol{\beta}_0) = (\hat{\boldsymbol{\beta}}(\mathbf{T},\boldsymbol{\beta}_0)-\boldsymbol{\beta}_0)'\mathbf{P}'[\mathbf{P}\hat{\mathbf{V}}(\hat{\boldsymbol{\beta}}(\mathbf{T},\boldsymbol{\beta}_0))\mathbf{P}']^{-1}\mathbf{P}(\hat{\boldsymbol{\beta}}(\mathbf{T},\boldsymbol{\beta}_0)-\boldsymbol{\beta}_0)$.

(d) Ceilings and floors: i^* denotes the smallest integer greater than or equal to $(D+1)(1-\alpha)$ & $\underline{v}$ the greatest integer less than or equal to v.

(e) $F_{\mathbf{T}j}$ the probability $\tau^{\subseteq}(\mathbf{T},\boldsymbol{\beta}_0)$ is less than or equal to $\tau^{\subseteq}(\mathbf{T}_j,\boldsymbol{\beta}_0)$, $\mathbf{T}_j$ a draw from the universe of permutations $\mathbf{T}$ of $\mathbf{X}$, and $p_{\mathbf{T}j}$ the probability $\tau^{\subseteq}(\mathbf{T},\boldsymbol{\beta}_0)$ equals $\tau^{\subseteq}(\mathbf{T}_j,\boldsymbol{\beta}_0)$.

(f) Auxiliary test statistic $g(\mathbf{T}_j,\boldsymbol{\beta}_0) = F_{\mathbf{T}j} - u_{\mathbf{T}j}p_{\mathbf{T}j}$, where $u_{\mathbf{T}j}$ is a draw for each $\mathbf{T}_j$ from the uniform distribution on [0,1].

(g) $\Pr_{\mathbf{T}|\mathbf{D}}\{a < b\}$ = probability of the event $a < b$ across the permutation distribution $\mathbf{T}$ of $\mathbf{X}$ given the data $\mathbf{D} = (\mathbf{Z},\mathbf{X}_\mathbf{W},\boldsymbol{\varepsilon})$, and $\Pr_{\mathbf{D}}\{a < b\}$ = the probability of the event $a < b$ across the data $\mathbf{D}$.

(h) $P(I_1) = \Pr\{u > i^* - (D+1)(1-\alpha)\}$, where u is uniformly distributed on [0,1].

(i) Cumulative distribution functions of central & non-central chi-squared with k degrees of freedom and non-centrality parameter λ: $F\chi^2_k(x)$ & $F\chi^2_{k,\lambda}(x)$.

(j) Non-central coverage rates with central critical values: $G_{k,\lambda}(a) = F\chi^2_{k,\lambda}(x(a))$, x such that $F\chi^2_k(x) = a$.

(k) Expectation across permutations $\mathbf{T}$ of $\mathbf{X}$: $E_\mathbf{T}$.

D: Proof of (R2) - (R5)

We prove (R2) - (R5) generalized to cover subsets of coefficients. Let $\mathbf{P}$ denote a k x PQ matrix of zeros of full rank with a single one in each row and define $\tau^{\subseteq}(\mathbf{X},\boldsymbol{\beta}_0^{\subseteq}) = (\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)'\mathbf{P}'[\mathbf{P}\hat{\mathbf{V}}(\hat{\boldsymbol{\beta}})\mathbf{P}']^{-1}\mathbf{P}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)$, $\tau^{\subseteq}(\mathbf{T},\boldsymbol{\beta}_0) = (\hat{\boldsymbol{\beta}}(\mathbf{T},\boldsymbol{\beta}_0)-\boldsymbol{\beta}_0)'\mathbf{P}'[\mathbf{P}\hat{\mathbf{V}}(\hat{\boldsymbol{\beta}}(\mathbf{T},\boldsymbol{\beta}_0))\mathbf{P}']^{-1}\mathbf{P}(\hat{\boldsymbol{\beta}}(\mathbf{T},\boldsymbol{\beta}_0)-\boldsymbol{\beta}_0)$, $\boldsymbol{\beta}_0^{\subseteq} = \mathbf{P}\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}^{\subseteq} = \mathbf{P}\boldsymbol{\beta}$. While $\tau^{\subseteq}(\mathbf{X},\boldsymbol{\beta}_0^{\subseteq})$ is only a function of $\boldsymbol{\beta}_0^{\subseteq}$, in finite samples $\tau^{\subseteq}(\mathbf{T},\boldsymbol{\beta}_0)$ is a function of $\boldsymbol{\beta}_0$ as the full vector affects the calculation of counterfactual output $\mathbf{y}(\mathbf{T},\boldsymbol{\beta}_0)$. White's conditions W1 - W4 ensure that for a test of the true null $\tau^{\subseteq}(\mathbf{X},\boldsymbol{\beta}^{\subseteq})$ is asymptotically distributed chi-squared with k degrees of freedom. (R1) showed that in a finite $\sqrt{N}$ neighbourhood of $\boldsymbol{\beta}_0$ around $\boldsymbol{\beta}$ $\tau^{\subseteq}(\mathbf{T},\boldsymbol{\beta}_0)$ is asymptotically distributed chi-squared with k degrees of freedom and in probability equal to $\tau^{\subseteq}(\mathbf{T},\boldsymbol{\beta})$.

R2: As in the text define i^* as the smallest integer greater than or equal to $(D+1)(1-\alpha)$ and $\tau_1 < ... < \tau_D$ as the ordered values of $\tau^{\subseteq}(\mathbf{T},\boldsymbol{\beta})$, where if needed define τ_0 as 0 and τ_{D+1} as ∞. The logic behind (R2) is illustrated in Figure A1. For all $\boldsymbol{\beta}_0^{\subseteq}$ such that $\tau^{\subseteq}(\mathbf{X},\boldsymbol{\beta}_0^{\subseteq}) \in (\tau_{i-1},\tau_i)$, $D - i + 1$ of the $\tau^{\subseteq}(\mathbf{T},\boldsymbol{\beta})$ outcomes are greater than $\tau^{\subseteq}(\mathbf{X},\boldsymbol{\beta}_0^{\subseteq})$ and none equal, and hence the randomization p-value (2.4) for these nulls is

$[D\text{-}i\text{+}1\text{+}u]/(D\text{+}1)$, where u is a draw from the uniform distribution. The conditions for the p-values in each of the regions illustrated in the figure to be greater than some nominal level α are then[54]

$$\text{(D.1)}\quad p(\tau_{i^*-1}) = \frac{D - i^* + 2 + u}{D+1} > \alpha \;\Rightarrow\; u > i^* - (D+1)(1-\alpha) - 1 \text{ [always true]}$$

$$p(\tau_{i^*}) = \frac{D - i^* + 1 + u}{D+1} > \alpha \;\Rightarrow\; u > i^* - (D+1)(1-\alpha),\; p(\tau_{i^*+1}) = \frac{D - i^* + u}{D+1} > \alpha \;\Rightarrow\; u > i^* - (D+1)(1-\alpha) + 1 \begin{bmatrix} \text{never} \\ \text{true} \end{bmatrix}.$$

Given that i^* is the smallest integer greater than or equal to $(D\text{+}1)(1\text{-}\alpha)$, the p-value for the region below $\tau_{i^*\text{-}1}$ is always greater than α and the p-value for the region above τ_{i^*} is always less than α. Hence, the confidence interval is determined by τ_{i^*} if $u > i^* - (D\text{+}1)(1\text{-}\alpha)$, and by $\tau_{i^*\text{-}1}$ otherwise, as described in (R2).

R3 & R4: The proofs of (R3) & (R4) use ideas present in Jockel (1986) and Hoeffding (1952). We finesse complexities introduced by the discreteness of finite sample distributions by constructing an auxiliary test statistic whose distribution is always uniform. Let $\mathbf{\Omega}$ denote the universe of permutations $\mathbf{T}$ of $\mathbf{X}$, $\mathbf{T}_j$ one of those permutations, $F_{\mathbf{T}j}$ the probability $\tau^{\subseteq}(\mathbf{T},\boldsymbol{\beta}_0)$ is less than or equal to $\tau^{\subseteq}(\mathbf{T}_j,\boldsymbol{\beta}_0)$, and $p_{\mathbf{T}j}$ the probability $\tau^{\subseteq}(\mathbf{T},\boldsymbol{\beta}_0)$ equals $\tau^{\subseteq}(\mathbf{T}_j,\boldsymbol{\beta}_0)$ (> 0, as the number of permutations is finite). Define the new "test statistic" $g(\mathbf{T}_j,\boldsymbol{\beta}_0) = F_{\mathbf{T}j} - u_{\mathbf{T}j}p_{\mathbf{T}j}$, where $u_{\mathbf{T}j}$ is a draw for each $\mathbf{T}_j$ from the uniform distribution on [0,1]. $g(\mathbf{T}_j,\boldsymbol{\beta}_0)$ is by construction uniformly distributed across permutations $\mathbf{T}$ of $\mathbf{X}$ for all $\boldsymbol{\beta}_0$ in all sample sizes. Since $\mathbf{X}$ itself is one of the permutations of $\mathbf{X}$, we also have $g(\mathbf{X},\boldsymbol{\beta}_0)$, which with probability one lies in (0,1). Defining the p-value using (2.4), we will show that the probability of rejection at a level α using $g(\mathbf{T},\boldsymbol{\beta}_0)$ is always the same as using $\tau^{\subseteq}(\mathbf{T},\boldsymbol{\beta}_0)$. Consequently, if (R3) & (R4) apply for $g(\mathbf{T},\boldsymbol{\beta}_0)$, even though it is never actually observed, we know they apply for $\tau^{\subseteq}(\mathbf{T},\boldsymbol{\beta}_0)$ and the proof is complete.

We begin by noting that $\tau^{\subseteq}(\mathbf{T}_i,\boldsymbol{\beta}_0) < \tau^{\subseteq}(\mathbf{T}_j,\boldsymbol{\beta}_0)$ implies $g(\mathbf{T}_i,\boldsymbol{\beta}_0) < g(\mathbf{T}_j,\boldsymbol{\beta}_0)$, as with probability one $F_{\mathbf{T}j} - u_{\mathbf{T}j}p_{\mathbf{T}j} > F_{\mathbf{T}j} - p_{\mathbf{T}j} \geq F_{\mathbf{T}i} > F_{\mathbf{T}i} - u_{\mathbf{T}i}p_{\mathbf{T}i}$, since the probability $u_{\mathbf{T}j} = 1$ and $u_{\mathbf{T}i} = 0$ is zero. Let the draws $\mathbf{T}_1$, $\mathbf{T}_2$, ... $\mathbf{T}_D$ from $\mathbf{\Omega}$ contain G draws with values of $\tau^{\subseteq}(\mathbf{T},\boldsymbol{\beta}_0)$ strictly greater than $\tau^{\subseteq}(\mathbf{X},\boldsymbol{\beta}_0^{\subseteq})$ and E draws with values of $\tau^{\subseteq}(\mathbf{T},\boldsymbol{\beta}_0)$ equal to $\tau^{\subseteq}(\mathbf{X},\boldsymbol{\beta}_0^{\subseteq})$. Using (2.4), the p-value for $\tau^{\subseteq}(\mathbf{T},\boldsymbol{\beta}_0)$ is given by $[G + u(E\text{+}1)]/(D\text{+}1)$, while the p-value for $g(\mathbf{T},\boldsymbol{\beta}_0)$ is given by

$$\text{(D.2)}\quad \frac{G + E_G + u}{D+1},$$

where $E_G \leq E$ denotes the number of draws $\mathbf{T}_j$ with $\tau^{\subseteq}(\mathbf{T}_j,\boldsymbol{\beta}_0) = \tau^{\subseteq}(\mathbf{X},\boldsymbol{\beta}_0)$ which end up with $g(\mathbf{T}_j,\boldsymbol{\beta}_0) > g(\mathbf{X},\boldsymbol{\beta}_0)$ after the realization of the draws $u_{\mathbf{T}j}$ which determine $g(\mathbf{T}_j,\boldsymbol{\beta}_0)$. Select an α and let $\underline{\nu}$ denote the greatest integer less than or equal to ν. If $G \geq \underline{\alpha(D\text{+}1)}\text{+}1$ the p-value for both $\tau^{\subseteq}(\mathbf{T},\boldsymbol{\beta}_0)$ and $g(\mathbf{T},\boldsymbol{\beta}_0)$ is greater than α, i.e. neither test rejects, while if $G + E + 1 < \alpha(D\text{+}1)$ the p-value for both is less than α, i.e. both tests reject. Consequently, we need only concern ourselves with the case where $G \leq \underline{\alpha(D\text{+}1)}$ and $G + E + 1 \geq \alpha(D\text{+}1)$. In these circumstances, the test with $\tau^{\subseteq}(\mathbf{T},\boldsymbol{\beta}_0)$ rejects with probability $(\alpha(D\text{+}1)\text{-}G)/(E\text{+}1)$. Given the $u_{\mathbf{X}}$ which determined $g(\mathbf{X},\boldsymbol{\beta}_0) = F_{\mathbf{X}} - u_{\mathbf{X}}p_{\mathbf{X}}$, the probability a given element of the set

[54]Each calculation is for within the region bounded by two τ_i, as illustrated in the figure. The τ_i themselves, where the number of tied outcomes is 1, are of measure zero in the confidence interval.

of draws that have $\tau^{\subseteq}(\mathbf{T}_j,\boldsymbol{\beta}_0)=\tau^{\subseteq}(\mathbf{X},\boldsymbol{\beta}_0)$ ends up with $g(\mathbf{T}_j,\boldsymbol{\beta}_0) > g(\mathbf{X},\boldsymbol{\beta}_0)$ is $u_{\mathbf{X}}$. Say $G = \underline{\alpha(D+1)}$. For given $u_{\mathbf{X}}$, the probability (D.2) is less than or equal to α equals the probability $E_G = 0$ times the probability u in (D.2) is less than $\alpha(D+1)-G$, or:

$$(D.3)\ (1-u_{\mathbf{X}})^{E}(\alpha(D+1)-G).$$

Integrating across the uniform distribution of $u_{\mathbf{X}}$ yields

$$(D.4)\ \int_0^1 (1-u_{\mathbf{X}})^{E}(\alpha(D+1)-G)\,du_{\mathbf{X}} = (\alpha(D+1)-G)/(E+1),$$

which is the same probability as the test statistic using $\tau^{\subseteq}(\mathbf{T},\boldsymbol{\beta}_0)$. If $G \le \underline{\alpha(D+1)} - 1$, for given $u_{\mathbf{X}}$ the probability (D.2) is less than or equal to α is given by the probability $E_G \le \underline{\alpha(D+1)} - G - 1$ plus the probability $E_G = \underline{\alpha(D+1)} - G$ times the probability u in (D.2) is less than or equal to $\alpha(D+1)-\underline{\alpha(D+1)}$, or:

$$(D.5)\ \sum_{x=0}^{\underline{\alpha(D+1)}-G-1} \frac{E!}{x!(E-x)!}(1-u_{\mathbf{X}})^{E-x}u_{\mathbf{X}}^{x} + \frac{E!(1-u_{\mathbf{X}})^{E-\underline{\alpha(D+1)}+G}u_{\mathbf{X}}^{\underline{\alpha(D+1)}-G}[\alpha(D+1)-\underline{\alpha(D+1)}]}{(\underline{\alpha(D+1)}-G)!(E-\underline{\alpha(D+1)}+G)!}.$$

Again, integrating across the uniform distribution of $u_{\mathbf{X}}$

$$(D.6)\ \int_0^1 \left(\sum_{x=0}^{\underline{\alpha(D+1)}-G-1} \frac{E!}{x!(E-x)!}(1-u_{\mathbf{X}})^{E-x}u_{\mathbf{X}}^{x} + \frac{E!(1-u_{\mathbf{X}})^{E-\underline{\alpha(D+1)}+G}u_{\mathbf{X}}^{\underline{\alpha(D+1)}-G}[\alpha(D+1)-\underline{\alpha(D+1)}]}{(\underline{\alpha(D+1)}-G)!(E-\underline{\alpha(D+1)}+G)!} \right) du_{\mathbf{X}}$$

$$= \frac{\underline{\alpha(D+1)}-G}{E+1} + \frac{\alpha(D+1)-\underline{\alpha(D+1)}}{E+1} = \frac{\alpha(D+1)-G}{E+1},$$

which again is the same as in the case of the $\tau^{\subseteq}(\mathbf{T},\boldsymbol{\beta}_0)$ statistic. These examples cover all possible cases for G and E, so we see that for any null $\boldsymbol{\beta}_0$ and realized draws $\mathbf{T}_1$, $\mathbf{T}_2$, ... $\mathbf{T}_D$ from $\boldsymbol{\Omega}$, the p-value calculated using $\tau^{\subseteq}(\mathbf{T},\boldsymbol{\beta}_0)$ has the same probability of rejecting at level α as the p-value calculated using $g(\mathbf{T},\boldsymbol{\beta}_0)$.

We now focus on inference using g() alone. Let $g_1 < ... < g_D$ denote the ordered values of $g(\mathbf{T},\boldsymbol{\beta}_0)$ across D draws of $\mathbf{T}$ from $\boldsymbol{\Omega}$. By construction $g(\mathbf{T},\boldsymbol{\beta}_0)$ is uniformly distributed across draws $\mathbf{T}$ from $\boldsymbol{\Omega}$, so ties are a probability zero event, as are $g_1 = 0$ or $g_D = 1$. As before let i^* denote the smallest integer greater than or equal to $(D+1)(1-\alpha)$. If u in (D.2) is $> i^* - (D+1)(1-\alpha)$, the p-value is greater than α as long as $g(\mathbf{X},\boldsymbol{\beta}_0) < g_{i^*}$, whereas if u is $< i^* - (D+1)(1-\alpha)$ the p-value is greater than α as long as $g(\mathbf{X},\boldsymbol{\beta}_0) < g_{i^*-1}$. If we define $g_0 = 0$ & $g_{D+1} = 1$ these rules cover the case where i^* equals 1 or $D+1$.[55] Since $g(\mathbf{T},\boldsymbol{\beta}_0)$ is always uniformly distributed on [0,1], the probability $g_i = a$ is always $[D!/(D-i)!(i-1)!]a^{i-1}(1-a)^{D-i}$. With $\Pr_{\mathbf{T}|\mathbf{D}}$ denoting the probability across the permutation distribution $\mathbf{T}$ given the data $\mathbf{D}$, $g(\mathbf{X},\boldsymbol{\beta}_0)$ can be bounded:

$$(D.7)\ \Pr_{\mathbf{T}|\mathbf{D}}\{\tau^{\subseteq}(\mathbf{T},\boldsymbol{\beta}_0) < \tau^{\subseteq}(\mathbf{X},\boldsymbol{\beta}_0^{\subseteq})\} < g(\mathbf{X},\boldsymbol{\beta}_0) < \Pr_{\mathbf{T}|\mathbf{D}}\{\tau^{\subseteq}(\mathbf{T},\boldsymbol{\beta}_0) \le \tau^{\subseteq}(\mathbf{X},\boldsymbol{\beta}_0^{\subseteq})\}.$$

The distribution of $\tau^{\subseteq}(\mathbf{T},\boldsymbol{\beta}_0)$ across draws $\mathbf{T}$ given the data $\mathbf{D}$ (for $\boldsymbol{\beta}_0$ in a finite $\sqrt{N}$ neighbourhood of $\boldsymbol{\beta}$) almost surely converges to the chi-squared, as does that of $\tau^{\subseteq}(\mathbf{X},\boldsymbol{\beta}_0^{\subseteq})$ (for $\boldsymbol{\beta}_0 = \boldsymbol{\beta}$) across the data $\mathbf{D}$, and as

[55] If i^* equals $D+1$, $\alpha(D+1)$ is < 1. If $u > i^* - (D+1)(1-\alpha) = \alpha(D+1)$, the p-value is always greater than α, which is equivalent to saying $g(\mathbf{X},\boldsymbol{\beta}_0) < g_{i^*} = g_{D+1} = 1$, otherwise the p-value is greater than α if at least one of the D draws of $g(\mathbf{T},\boldsymbol{\beta}_0)$ is greater than $g(\mathbf{X},\boldsymbol{\beta}_0)$, i.e. $g(\mathbf{X},\boldsymbol{\beta}_0) < g_D$. If i^* equals 1, $\alpha(D+1)$ is $\ge D$. If $u > i^* - (D+1)(1-\alpha)$ and all D draws are greater than $g(\mathbf{X},\boldsymbol{\beta}_0)$, then the p-value is greater than α, equivalent to saying $g(\mathbf{X},\boldsymbol{\beta}_0) < g_1$, while if $u < i^* - (D+1)(1-\alpha)$ (which is only possible if $\alpha(D+1)$ is $> D$) it is impossible for the p-value to be greater than α, equivalent to saying $g(\mathbf{X},\boldsymbol{\beta}_0) < g_{i^*-1} = g_0 = 0$.

the chi-squared is continuous that convergence is uniform (Rao 1973, p. 120), so for each $\Delta > 0$ there exists an $N(\Delta)$ such that for all $N > N(\Delta)$ the two cumulative distributions are within $\Delta/2$ of the chi-squared with k degrees of freedom. Consequently, for $\boldsymbol{\beta}_0 = \boldsymbol{\beta}$ almost surely for all $N > N(\Delta)$

$$\text{(D.8) } F_{\chi 2}(\tau^{\subseteq}(\mathbf{X},\boldsymbol{\beta}_0^{\subseteq})) - \Delta/2 < \Pr_{\mathbf{T}|\mathbf{D}}\{\tau^{\subseteq}(\mathbf{T},\boldsymbol{\beta}_0) < \tau^{\subseteq}(\mathbf{X},\boldsymbol{\beta}_0^{\subseteq})\} < \Pr_{\mathbf{T}|\mathbf{D}}\{\tau^{\subseteq}(\mathbf{T},\boldsymbol{\beta}_0) \le \tau^{\subseteq}(\mathbf{X},\boldsymbol{\beta}_0^{\subseteq})\} < F_{\chi 2}(\tau^{\subseteq}(\mathbf{X},\boldsymbol{\beta}_0^{\subseteq})) + \Delta/2$$
$$\& \quad a - \Delta < \Pr_{\mathbf{D}}\{F_{\chi 2}(\tau^{\subseteq}(\mathbf{X},\boldsymbol{\beta}^{\subseteq})) \le a - \Delta/2\} \le \Pr_{\mathbf{D}}\{F_{\chi 2}(\tau^{\subseteq}(\mathbf{X},\boldsymbol{\beta}^{\subseteq})) \le a + \Delta/2\} < a + \Delta$$
$$\text{so using (D.7)} \quad a - \Delta < \Pr_{\mathbf{D}}\{g(\mathbf{X},\boldsymbol{\beta}_0) \le a\} < a + \Delta,$$

and we see that the cumulative distribution function of $g(\mathbf{X},\boldsymbol{\beta}_0=\boldsymbol{\beta})$ across $\mathbf{D}$ converges to that of the uniform on (0,1). As $N(\Delta)$ does not depend upon a, the convergence is uniform.

Using the above, we see that the probability across the data $\mathbf{D}$ that $g(\mathbf{X},\boldsymbol{\beta})$ will be less than the i^{th} order statistic g_i of D draws from the permutation distribution $\mathbf{T}$ of $\mathbf{X}$ converges to

$$\text{(D.9) } \Pr_{\mathbf{D}}\{g(\mathbf{X},\boldsymbol{\beta}) < g_i\} = \int_0^1 \Pr_{\mathbf{D}}\{g(\mathbf{X},\boldsymbol{\beta}) < a\} \frac{D!a^{i-1}(1-a)^{D-i}}{D-i!i-1!}\text{da} \underset{N\to\infty}{\to} \int_0^1 \frac{D!a^{i}(1-a)^{D-i}}{D-i!i-1!}\text{da} = \frac{i}{D+1},$$

where uniform convergence of $\Pr_{\mathbf{D}}\{g(\mathbf{X},\boldsymbol{\beta}) < a\} = \Pr_{\mathbf{D}}\{g(\mathbf{X},\boldsymbol{\beta}) \le a\}$, which is bounded continuous (as by construction $g(\mathbf{X},\boldsymbol{\beta})$ is continuous without mass points) and hence Riemann integrable, allows the limit of the Riemann integral. If i^* lies in $[2,D]$ the probability the p-value is greater than α then equals:

$$\text{(D.10) } [1 - i^* + (D+1)(1-\alpha)]\Pr_{\mathbf{D}}\{g(\mathbf{X},\boldsymbol{\beta}) < g_{i^*}\} + [i^* - (D+1)(1-\alpha)]\Pr_{\mathbf{D}}\{g(\mathbf{X},\boldsymbol{\beta}) < g_{i^*-1}\}$$
$$\underset{N\to\infty}{\to} [1 - i^* + (D+1)(1-\alpha)]\frac{i^*}{D+1} + [i^* - (D+1)(1-\alpha)]\frac{i^*-1}{D+1} = 1-\alpha.$$

For the special cases where i^* equals 1 or D+1, we have:

$$i^* = 1\text{: } \quad (D+1)(1-\alpha)\Pr_{\mathbf{D}}\{g(\mathbf{X},\boldsymbol{\beta}) < g_1\} + [1-(D+1)(1-\alpha)]\Pr_{\mathbf{D}}\{g(\mathbf{X},\boldsymbol{\beta}) < g_0\}$$
$$\underset{N\to\infty}{\to} (D+1)(1-\alpha)\frac{1}{D+1} + [1-(D+1)(1-\alpha)]*0 = 1-\alpha.$$
$$i^* = D+1\text{: } \quad [-D+(D+1)(1-\alpha)]\Pr_{\mathbf{D}}\{g(\mathbf{X},\boldsymbol{\beta}) < g_{D+1}\} + [D+1-(D+1)(1-\alpha)]\Pr_{\mathbf{D}}\{g(\mathbf{X},\boldsymbol{\beta}) < g_D\}$$
$$\underset{N\to\infty}{\to} [-\alpha D + (1-\alpha)]*1 + [(D+1)\alpha]\frac{D}{D+1} = 1-\alpha.$$

This establishes that the coverage probability of the true null of a test based upon g(), and hence one based upon $\tau^{\subseteq}$ as well, converges to 1-α, proving (R3).

Turning to (R4), for $\boldsymbol{\beta}_0^{\subseteq} \neq \boldsymbol{\beta}^{\subseteq}$ within a finite $\sqrt{N}$ neighbourhood of $\boldsymbol{\beta}$, $\tau^{\subseteq}(\mathbf{X},\boldsymbol{\beta}_0^{\subseteq})$ is asymptotically distributed non-central chi-squared with k degrees of freedom and non-centrality parameter $\lambda = \sqrt{N}(\boldsymbol{\beta}_0^{\subseteq} - \boldsymbol{\beta}^{\subseteq})'[N\hat{\mathbf{V}}(\hat{\boldsymbol{\beta}}^{\subseteq})]^{-1}(\boldsymbol{\beta}_0^{\subseteq} - \boldsymbol{\beta}^{\subseteq})\sqrt{N}$, which we denote as $\chi^2_{k,\lambda}$, and cumulative distribution function:

$$\text{(D.12) } F\chi^2_{k,\lambda}(x) = e^{-\lambda/2}\sum_{j=0}^{\infty}\frac{(\lambda/2)^j}{j!}\frac{\int_0^{x/2} t^{k/2+j-1}e^{-t}\text{dt}}{\Gamma(k/2+j)}.$$

The critical value x associated with the a^{th} percentile of the central chi-squared is determined by

$$\text{(D.13) } a = F\chi^2_k(x) = \Gamma(k/2)^{-1}\int_0^{x/2} t^{k/2-1}e^{-t}\text{dt}, \quad \text{with} \quad \frac{dx}{da} = 2\Gamma(k/2)(x/2)^{1-k/2}e^{x/2}.$$

Consequently, for

$$\text{(D.14)}\ \ G_{k,\lambda}(a) = F_{\chi^2_{k,\lambda}}(x(a)) = F_{\chi^2_{k,\lambda}}(F^{-1}_{\chi^2_k}(a)),\ \text{we have}$$

$$\frac{dG}{da} = e^{-\lambda/2}\sum_{j=0}^{\infty}\frac{(\lambda x/4)^j}{j!}\frac{\Gamma(k/2)}{\Gamma(k/2+j)} > 0\ \ \&\ \ \frac{d^2G}{da^2} = e^{-\lambda/2}\sum_{j=1}^{\infty}\frac{(\lambda x/4)^j}{x^* j-1!}\frac{\Gamma(k/2)}{\Gamma(k/2+j)}\frac{dx}{da} > 0,$$

that is, the conventional coverage probability of an incorrect null is convex in the nominal level. dG/da is bounded from above by $2e^{\lambda x(a)/4}$.

We prove pointwise convergence of the cumulative distribution of $g(\mathbf{X},\boldsymbol{\beta}_0\neq\boldsymbol{\beta})$ to $G_{k,\lambda}$ for $\boldsymbol{\beta}_0$ in a $\sqrt{N}$ neighbourhood of $\boldsymbol{\beta}$. For every a in (0,1) and $0 < \Delta/2 < \min(a,1\text{-}a)$, let $x^* = x(a+\Delta/4)$. There exists an $N(a, \Delta)$ such that for all $N > N(a, \Delta)$ the distribution of $\tau^{\subseteq}(\mathbf{T},\boldsymbol{\beta}_0)$ differs from the chi-squared by no more than $\Delta e^{-\lambda x^*/4}/4$ and that of $\tau^{\subseteq}(\mathbf{X},\boldsymbol{\beta}_0^{\subseteq})$ from the non-central chi-squared by no more than $\Delta/2$, so that we may modify (D.8) to read

$$\text{(D.15)}\begin{pmatrix} F_{\chi 2}(\tau^{\subseteq}(\mathbf{X},\boldsymbol{\beta}_0^{\subseteq})) \\ -\Delta e^{-\lambda x^*/4}/4 \end{pmatrix} < \Pr_{\mathbf{T}|\mathbf{D}}\begin{Bmatrix} \tau^{\subseteq}(\mathbf{T},\boldsymbol{\beta}_0) < \\ \tau^{\subseteq}(\mathbf{X},\boldsymbol{\beta}_0^{\subseteq}) \end{Bmatrix} < \Pr_{\mathbf{T}|\mathbf{D}}\begin{Bmatrix} \tau^{\subseteq}(\mathbf{T},\boldsymbol{\beta}_0) \leq \\ \tau^{\subseteq}(\mathbf{X},\boldsymbol{\beta}_0^{\subseteq}) \end{Bmatrix} < \begin{pmatrix} F_{\chi 2}(\tau^{\subseteq}(\mathbf{X},\boldsymbol{\beta}_0^{\subseteq})) \\ +\Delta e^{-\lambda x^*/4}/4 \end{pmatrix},$$

$$\begin{pmatrix} G_{k,\lambda}(a-\Delta e^{-\lambda x^*/4}/4) \\ -\Delta/2 \end{pmatrix} < \Pr_{\mathbf{D}}\begin{Bmatrix} F_{\chi 2}(\tau^{\subseteq}(\mathbf{X},\boldsymbol{\beta}_0^{\subseteq})) \leq \\ a-\Delta e^{-\lambda x^*/4}/4 \end{Bmatrix} < \Pr_{\mathbf{D}}\begin{Bmatrix} F_{\chi 2}(\tau^{\subseteq}(\mathbf{X},\boldsymbol{\beta}_0^{\subseteq})) \leq \\ a+\Delta e^{-\lambda x^*/4}/4\} \end{Bmatrix} < \begin{pmatrix} G_{k,\lambda}(a+\Delta e^{-\lambda x^*/4}/4) \\ +\Delta/2 \end{pmatrix},$$

$$\&\ (\text{as}\ \frac{dG}{da} < 2e^{-\lambda x/4}\ \&\ \frac{d^2G}{da^2} > 0)\ \ G_{k,\lambda}(a)-\Delta/2 < G_{k,\lambda}(a-\Delta e^{-\lambda x^*/4}/4) < G_{k,\lambda}(a+\Delta e^{-\lambda x^*/4}/4) < G_{k,\lambda}(a)+\Delta/2,$$

$$\text{so using (D.7)}\quad G_{k,\lambda}(a)-\Delta < \Pr_{\mathbf{D}}\{g(\mathbf{X},\boldsymbol{\beta}_0)\leq a\} < G_{k,\lambda}(a)+\Delta,$$

proving pointwise convergence of $\Pr_{\mathbf{D}}\{g(\mathbf{X},\boldsymbol{\beta}_0\neq\boldsymbol{\beta})\leq a\}$ to $G_{k,\lambda}(a)$. For $\boldsymbol{\beta}_0 \neq \boldsymbol{\beta}$ we then have:

$$\text{(D.16)}\ \Pr_{\mathbf{D}}\{g(\mathbf{X},\boldsymbol{\beta}_0) < g_i\} = \int_0^1 \Pr_{\mathbf{D}}\{g(\mathbf{X},\boldsymbol{\beta}_0) < a\}\frac{D!a^{i-1}(1-a)^{D-i}}{D-i!i-1!}\mathrm{da} \underset{N\to\infty}{\to} \int_0^1 G_{k,\lambda}(a)\frac{D!a^{i-1}(1-a)^{D-i}}{D-i!i-1!}\mathrm{da} > G_{k,\lambda}\left(\frac{i}{D+1}\right),$$

where the limit of the integral follows from Arzelà's Dominated Convergence Theorem and the fact that the functions in the sequence are bounded and continuous (hence Riemann integrable). The inequality follows from Jensen's Inequality and the fact that

$$\text{(D.17)}\ \ \int_0^1 \frac{D!a^{i-1}(1-a)^{D-i}}{D-i!i-1!}\mathrm{da} = 1\ \ \&\ \ \int_0^1 a\frac{D!a^{i-1}(1-a)^{D-i}}{D-i!i-1!}\mathrm{da} = \frac{i}{D+1},$$

so the integral on the right hand side can be considered the expectation across the random variable a of $G_{k,\lambda}(a)$. Letting $P(I_1)$ denote the probability $u > i^*$ - $(D+1)(1-\alpha)$ and the p-value is based on the i^* order statistic, the probability the incorrect null lies in the 1-α confidence interval is then seen to be

$$\text{(D.18)}\ P(I_1)\Pr_{\mathbf{D}}\{g(\mathbf{X},\boldsymbol{\beta}_0\neq\boldsymbol{\beta}) < g_{i^*}\} + [1-P(I_1)]\Pr_{\mathbf{D}}\{g(\mathbf{X},\boldsymbol{\beta}_0\neq\boldsymbol{\beta}) < g_{i^*-1}\} \underset{N\to\infty}{>}$$

$$P(I_1)G_{k,\lambda}\left(\frac{i^*}{D+1}\right) + [1-P(I_1)]G_{k,\lambda}\left(\frac{i^*-1}{D+1}\right) > G_{k,\lambda}\left(\frac{P(I_1)i^*}{D+1} + \frac{[1-P(I_1)](i^*-1)}{D+1}\right) = G_{k,\lambda}(1-\alpha),$$

where in the second line we again use the fact that $G_{k,\lambda}(a)$ is a convex function of a, while 1-α is the expectation of the "random variable" that takes on the values $i^*/(D+1)$ and $(i^*\text{-}1)/(D+1)$ with probabilities $P(I_1)$ and 1- $P(I_1)$. In the special cases where $i^* = 1$ or $D+1$, using the same techniques we have

$$(\text{D.19})\ i^*=1:\quad P(I_1)\Pr_{\mathbf{D}}\{g(\mathbf{X},\boldsymbol{\beta}_0\neq\boldsymbol{\beta})<g_1\}+[1-P(I_1)]\underbrace{\Pr_{\mathbf{D}}\{g(\mathbf{X},\boldsymbol{\beta}_0\neq\boldsymbol{\beta})<g_0\}}_{=0}$$

$$\underset{N\to\infty}{>}P(I_1)G_{k,\lambda}\left(\frac{1}{D+1}\right)+[1-P(I_1)]\underbrace{G_{k,\lambda}(0)}_{=0}>G_{k,\lambda}\left(\frac{P(I_1)}{D+1}+\frac{[1-P(I_1)]*0}{D+1}\right)=G_{k,\lambda}(1-\alpha)$$

$$i^*=D+1:\quad P(I_1)\underbrace{\Pr_{\mathbf{D}}\{g(\mathbf{X},\boldsymbol{\beta}_0\neq\boldsymbol{\beta})<g_{D+1}\}}_{=1}+[1-P(I_1)]\Pr_{\mathbf{D}}\{g(\mathbf{X},\boldsymbol{\beta}_0\neq\boldsymbol{\beta})<g_D\}$$

$$\underset{N\to\infty}{>}P(I_1)\underbrace{G_{k,\lambda}(1)}_{=1}+[1-P(I_1)]G_{k,\lambda}\left(\frac{D}{D+1}\right)>G_{k,\lambda}\left(P(I_1)*1+\frac{[1-P(I_1)]D}{D+1}\right)=G_{k,\lambda}(1-\alpha).$$

So we see that the coverage probability of an incorrect null for tests based upon g(), equivalently $\tau^{\subseteq}$, is asymptotically greater than that of the conventional test, resulting in lower power, as stated in (R4).

Regarding the double limit $D,N\to\infty$, as noted above the integral in (D.16) can be thought of as the expectation of a function across the random variable a. Its variance is given by

$$(\text{D.20})\ \sigma^2(a)=\int_0^1 a^2\frac{D!a^{i-1}(1-a)^{D-i}}{D-i!i-1!}\mathrm{da}-\frac{i^2}{(D+1)^2}=\frac{i(i+1)}{(D+1)(D+2)}-\frac{i^2}{(D+1)^2}=\frac{i(D+1)-i^2}{(D+1)^2(D+2)}\le\frac{1}{4(D+2)}.$$

which converges to 0 as $D\to\infty$, while its mean in (D.17) converges to 1 - α for $i=i^*$ or i^*-1. As $\Pr_{\mathbf{D}}\{g(\mathbf{X},\boldsymbol{\beta}_0\neq\boldsymbol{\beta})<a\}$ is non-decreasing in a, with $\Pr_{\mathbf{D}}\{g(\mathbf{X},\boldsymbol{\beta}_0\neq\boldsymbol{\beta})<0\}=0$ & $\Pr_{\mathbf{D}}\{g(\mathbf{X},\boldsymbol{\beta}_0\neq\boldsymbol{\beta})<1\}=1$, we can construct bounds on the left-hand side integral in (D.16) for $i=i^*$ or i^* - 1 using

$$(\text{D.21})\ \Pr_{\mathbf{D}}\left\{\begin{matrix}g(\mathbf{X},\boldsymbol{\beta}_0\neq\boldsymbol{\beta})<\\ i/(D+1)-\sqrt[4]{D}\sigma(a)\end{matrix}\right\}\left[1-D^{-½}\right]\le\Pr_{\mathbf{D}}\left\{\begin{matrix}g(\mathbf{X},\boldsymbol{\beta}_0\neq\boldsymbol{\beta})<\\ 0\end{matrix}\right\}\Pr\{A\}+\Pr_{\mathbf{D}}\left\{\begin{matrix}g(\mathbf{X},\boldsymbol{\beta}_0\neq\boldsymbol{\beta})<\\ i/(D+1)-\sqrt[4]{D}\sigma(a)\end{matrix}\right\}[1-\Pr\{A\}]$$

$$\le\ \Pr_{\mathbf{D}}\{g(\mathbf{X},\boldsymbol{\beta}_0\neq\boldsymbol{\beta})<g_i\}\ \le$$

$$\Pr_{\mathbf{D}}\left\{\begin{matrix}g(\mathbf{X},\boldsymbol{\beta}_0\neq\boldsymbol{\beta})<\\ i/(D+1)+\sqrt[4]{D}\sigma(a)\end{matrix}\right\}[1-\Pr\{A\}]+\Pr_{\mathbf{D}}\left\{\begin{matrix}g(\mathbf{X},\boldsymbol{\beta}_0\neq\boldsymbol{\beta})<\\ 1\end{matrix}\right\}\Pr\{A\}\le\Pr_{\mathbf{D}}\left\{\begin{matrix}g(\mathbf{X},\boldsymbol{\beta}_0\neq\boldsymbol{\beta})<\\ i/(D+1)+\sqrt[4]{D}\sigma(a)\end{matrix}\right\}\left[1-D^{-½}\right]+D^{-½},$$

$$\text{where }\Pr\{A\}=\Pr\left\{\left|a-i/(D+1)\right|\ge\sqrt[4]{D}\sigma(a)\right\}\le D^{-½}\text{ by Chebyshev's Inequality}$$

Furthermore, the probability the incorrect null lies in the *RCI*, given by the first line of (D.18), is bounded from above and below (respectively) by $\Pr_{\mathbf{D}}\{g(\mathbf{X},\boldsymbol{\beta}_0\neq\boldsymbol{\beta})<g_{i^*}\}$ & $\Pr_{\mathbf{D}}\{g(\mathbf{X},\boldsymbol{\beta}_0\neq\boldsymbol{\beta})<g_{i^*-1}\}$. The probability the incorrect null lies in the *CCI* is given by the probability $\tau^{\subseteq}(\mathbf{X},\boldsymbol{\beta}_0^{\subseteq})$ is less than $x(1\text{-}\alpha)$, i.e. the critical value for the 1-α percentile of the chi-squared with k degrees of freedom. So, we can say that

$$(\text{D.22})\ \Pr_{\mathbf{D}}\left\{g(\mathbf{X},\boldsymbol{\beta}_0)<\frac{i^*-1}{D+1}-\sqrt[4]{D}\sigma(a)\right\}\left[1-D^{-½}\right]-\Pr_{\mathbf{D}}\{\tau^{\subseteq}(\mathbf{X},\boldsymbol{\beta}_0^{\subseteq})<x(1-\alpha)\}$$

$$\le\Pr\{\boldsymbol{\beta}_0(\neq\boldsymbol{\beta})\in RCI(1-\alpha,D)\}-\Pr\{\boldsymbol{\beta}_0(\neq\boldsymbol{\beta})\in CCI(1-\alpha)\}\ \le$$

$$\Pr_{\mathbf{D}}\left\{g(\mathbf{X},\boldsymbol{\beta}_0)<\frac{i^*}{D+1}+\sqrt[4]{D}\sigma(a)\right\}\left[1-D^{-½}\right]+D^{-½}-\Pr_{\mathbf{D}}\{\tau^{\subseteq}(\mathbf{X},\boldsymbol{\beta}_0^{\subseteq})<x(1-\alpha)\}.$$

For the right-hand side of (D.22), with $v=i^*/(D+1)+D^{¼}\sigma(a)$ we can say:

$$(\text{D.23})\ |\Pr_{\mathbf{D}}\{g(\mathbf{X},\boldsymbol{\beta}_0)<v\}[1-D^{-½}]+D^{-½}-\Pr_{\mathbf{D}}\{\tau^{\subseteq}(\mathbf{X},\boldsymbol{\beta}_0^{\subseteq})<x(1-\alpha)\}|\ \le$$

$$|D^{-½}|+|\Pr_{\mathbf{D}}\{g(\mathbf{X},\boldsymbol{\beta}_0)<v\}-G_{k,\lambda}(v)|+|G_{k,\lambda}(v)-G_{k,\lambda}(1-\alpha)|+|G_{k,\lambda}(1-\alpha)-\Pr_{\mathbf{D}}\{\tau^{\subseteq}(\mathbf{X},\boldsymbol{\beta}_0^{\subseteq})<x(1-\alpha)\}|,$$

$$\text{where }\lim_{D\to\infty}D^{-½}=0,\ \lim_{D\to\infty}v=1-\alpha,\ \lim_{N\to\infty}\Pr_{\mathbf{D}}\{g(\mathbf{X},\boldsymbol{\beta}_0)<v\}=G_{k,\lambda}(v),\ \&\ \lim_{N\to\infty}\Pr_{\mathbf{D}}\{\tau^{\subseteq}(\mathbf{X},\boldsymbol{\beta}_0^{\subseteq})<x(1-\alpha)\}=G_{k,\lambda}(1-\alpha).$$

The convergence of $\Pr_{\mathbf{D}}\{g(\mathbf{X},\boldsymbol{\beta}_0 \neq \boldsymbol{\beta}) < a\}$ is pointwise, but since the x^* used in (D.15) is increasing in a and for large enough D v can be made to lie arbitrarily close to 1-α, by choosing x^* to be the maximum for that neighbourhood $|\Pr_{\mathbf{D}}\{g(\mathbf{X},\boldsymbol{\beta}_0) < v\} - G_{k,\lambda}(v)|$ can be made arbitrarily small for sufficiently large N. So, for each α & Δ there exists an $M(\alpha, \Delta)$ such that for all N & $D > M(\alpha, \Delta)$ each of the four absolute values on the second line is less than $\Delta/4$ and the sum is less than Δ. A similar decomposition allows us to show that the absolute value of the left-hand side of (D.22) can be made less than Δ, establishing that for any α in (0,1) and $\Delta > 0$ it is possible to find an $M(\alpha, \Delta)$ such that for N & $D > M(\alpha, \Delta)$ $|\Pr\{\boldsymbol{\beta}_0 (\neq \boldsymbol{\beta}) \in RCI(1-\alpha, D)\} - \Pr\{\boldsymbol{\beta}_0 (\neq \boldsymbol{\beta}) \in CCI(1-\alpha)\}| < \Delta$, as stated in (R4).

The proof above mirrors the intuition given in the text, although we finesse complications by using an unobserved auxiliary variable that smoothes the finite sample permutation distribution. Asymptotically the order statistics are a random variable with expectation equal to the 1-α percentile of the chi-squared. For a test of a true null, the conventional coverage probability is an accurate linear function in the percentiles of the chi-squared, and hence using the order statistics to evaluate significance produces a 1-α rejection probability. For tests of false nulls, the conventional coverage probability is a convex function of the percentiles of the chi-squared, and hence by Jensen's inequality using the order statistics produces higher coverage probabilities and lower power, unless the number of draws D is taken to ∞, so that the distribution of the relevant order statistic converges to a point.

R5: The conventional test yields a p-value of 1 - $F\chi^2_k(\tau^{\subseteq}(\mathbf{X},\boldsymbol{\beta}_0^{\subseteq}))$. The randomization p-value is given by $(G+u(E+1))/(D+1)$. Let g_i and e_i denote the iid binary (0,1) random variables that indicate whether on the i^{th} random permutation $\mathbf{T}$ of $\mathbf{X}$ $\tau^{\subseteq}(\mathbf{T},\boldsymbol{\beta}_0)$ is greater than or equal, respectively, to $\tau^{\subseteq}(\mathbf{X},\boldsymbol{\beta}_0^{\subseteq})$, and $E_{\mathbf{T}}$ and $\Pr_{\mathbf{T}}$ the expectation and probability across permutations $\mathbf{T}$ of $\mathbf{X}$, with $E_{\mathbf{T}}(g_i) = 1 - \Pr_{\mathbf{T}}\{\tau^{\subseteq}(\mathbf{T},\boldsymbol{\beta}_0) \leq \tau^{\subseteq}(\mathbf{X},\boldsymbol{\beta}_0^{\subseteq})\}$ & $E_{\mathbf{T}}(e_i) = \Pr_{\mathbf{T}}\{\tau^{\subseteq}(\mathbf{T},\boldsymbol{\beta}_0) = \tau^{\subseteq}(\mathbf{X},\boldsymbol{\beta}_0^{\subseteq})\}$. We have:

$$\text{(D.24)}\quad \left|\frac{G+u(E+1)}{D+1} - [1 - F\chi^2_k(\tau^{\subseteq}(\mathbf{X},\boldsymbol{\beta}_0^{\subseteq}))]\right| \leq \left|\frac{Du-G}{D(D+1)}\right| + \left|\frac{uE}{D+1}\right| + \left|\frac{G}{D} - 1 + F\chi^2_k(\tau^{\subseteq}(\mathbf{X},\boldsymbol{\beta}_0^{\subseteq}))\right|$$
$$\leq \underbrace{\left|\frac{1}{D+1}\right|}_{\text{"}a\text{"}} + \underbrace{\left|\frac{E}{D} - E_{\mathbf{T}}(e_i)\right|}_{\text{"}b\text{"}} + \underbrace{\left|E_{\mathbf{T}}(e_i)\right|}_{\text{"}c\text{"}} + \underbrace{\left|\frac{G}{D} - E_{\mathbf{T}}(g_i)\right|}_{\text{"}d\text{"}} + \underbrace{\left|E_{\mathbf{T}}(g_i) - 1 + F\chi^2_k(\tau^{\subseteq}(\mathbf{X},\boldsymbol{\beta}_0^{\subseteq}))\right|}_{\text{"}e\text{"}}.$$

As each draw of $\mathbf{T}$ is iid, and G and E equal the sum of the realizations of g_i and e_i, by the Strong Law of Large Numbers "b" and "d" almost surely converge to 0 as $D \rightarrow \infty$. For tests of nulls within a finite $\sqrt{N}$ neighbourhood of $\boldsymbol{\beta}$, $\tau^{\subseteq}(\mathbf{T},\boldsymbol{\beta}_0)$ almost surely converges in distribution to the chi-squared with k degrees of freedom as $N \rightarrow \infty$. So almost surely 1 - $\Pr_{\mathbf{T}}\{\tau^{\subseteq}(\mathbf{T},\boldsymbol{\beta}_0) \leq \tau^{\subseteq}(\mathbf{X},\boldsymbol{\beta}_0^{\subseteq})\}$ converges to 1 - $F\chi^2_k(\tau^{\subseteq}(\mathbf{X},\boldsymbol{\beta}_0^{\subseteq}))$ and $\Pr_{\mathbf{T}}\{\tau^{\subseteq}(\mathbf{T},\boldsymbol{\beta}_0) = \tau^{\subseteq}(\mathbf{X},\boldsymbol{\beta}_0^{\subseteq})\}$ to 0.[56] Consequently, for each Δ and $\tau^{\subseteq}(\mathbf{X},\boldsymbol{\beta}_0^{\subseteq})$ there exists an $M(\Delta, \tau^{\subseteq}(\mathbf{X},\boldsymbol{\beta}_0^{\subseteq}))$ such that almost surely for all N & $D > M(\Delta, \tau^{\subseteq}(\mathbf{X},\boldsymbol{\beta}_0^{\subseteq}))$ each of the absolute values in the second row of (D.24) is less than $\Delta/5$, establishing (R5).

[56] If the latter does not hold, the asymptotic cumulative distribution is discontinuous and cannot converge to the chi-squared, establishing a contradiction.