INFERENCE WITH A SINGLE TREATED CLUSTER

ANDREAS HAGEMANN

ABSTRACT. I introduce a generic method for inference about a scalar parameter in research designs with a finite number of heterogeneous clusters where only a single cluster received treatment. This situation is commonplace in difference-in-differences estimation but the test developed here applies more broadly. I show that the test controls size and has power under asymptotics where the number of observations within each cluster is large but the number of clusters is fixed. The test combines weighted, approximately Gaussian parameter estimates with a rearrangement procedure to obtain its critical values. The weights needed for most empirically relevant situations are tabulated in the paper. Calculation of the critical values is computationally simple and does not require simulation or resampling. The rearrangement test is highly robust to situations where some clusters are much more variable than others. Examples and an empirical application are provided.

JEL classification: C01, C22, C32
Keywords: cluster-robust inference, difference in differences, two-way fixed effects, clustered data, dependence, heterogeneity

1. INTRODUCTION

Studies with difference-in-differences estimation that arguably compare a single treated group to multiple control groups are routinely published in prominent journals. Between 2017 and 2021, this study design came up repeatedly in the *American Economic Review*, *Journal of Political Economy*, *Quarterly Journal of Economics*, and *Review of Economic Studies*: Dustmann, Schönberg, and Stuhler (2017) compare the German-Czech border region to distant German regions; Cunningham and Shah (2018)

Date: October 18, 2024. I would like to thank Connor Dowd, Bruno Ferman, Sarah Miller, Pepe Montiel Olea, Jonathan Roth, Bernard Salanié, Cyrus Samii, Jeffrey Wooldridge, co-editor Francesca Molinari, and three anonymous reviewers for helpful comments. Meng-Hsuan Hsieh and Candice Wang provided excellent research assistance. All errors are my own. `R` and `Stata` commands are available at `https://hgmn.github.io/rea`. I thank Dr. Michael Rice, Dr. Scott Regenbogen, and the staff at Michigan Medicine for their excellent care and continual support.

compare Rhode Island to other US states; Johnston and Mas (2018) compare Missouri to other US states; Cengiz, Dube, Lindner, and Zipperer (2019) compare Washington state to other US states; Deryugina and Molitor (2020) compare New Orleans to similar cities; Cameron, Seager, and Shah (2020) compare East Java to similar districts; Giorcelli and Moser (2020) compare Lombardy-Venetia to other early 19th century regions in present-day Italy; Cooper, Scott Morton, and Shekita (2020) compare New York state to other US states; Mastrobuoni (2020) compares Milan to other Italian cities; and Rubin and Rubin (2021) compare articles published in the discontinued *Journal of Business* to articles in other top finance journals. Statistical inference in this context is challenging and the results of some studies have been questioned specifically because they only have a single treated group. For instance, Ham and Ueda (2021) argue that the influential work of Garthwaite, Gross, and Notowidigdo (2014) does not properly account for having only Tennessee as the treated unit. Kaestner (2016, 2021) criticizes several studies of the Massachusetts health care reform and Deryugina and Molitor (2020) for the same reason.

The primary challenge for inference with one treated and multiple control groups is that the groups are typically large economic units such as states, villages, or other geographic regions. Observations within each of these groups likely depend on one another in unobservable ways and therefore require the researcher to cluster at the group level. With one treated cluster, currently available inferential procedures assume identically distributed clusters or other undesirable homogeneity conditions that are unlikely to hold in empirical practice. In an attempt to avoid statistical issues stemming from having a single treated cluster, researchers therefore routinely resort to splitting large groups into smaller clusters that are presumed to be independent. However, numerical evidence by Bertrand, Duflo, and Mullainathan (2004), MacKinnon and Webb (2017), and others suggests that ignoring dependence or heterogeneity may lead to heavily distorted inference. In both cases, the actual size of the test can exceed its nominal level by several orders of magnitude, i.e., nonexistent effects are far too likely to show up as highly significant. Part of the underlying problem is that most available inference procedures achieve consistency by requiring the number of clusters to go to infinity, which is difficult to justify when the clusters are states or regions.

In this paper, I introduce an asymptotically valid method for inference with a single treated cluster that allows for heterogeneity of unknown form. The number of observations within each cluster is presumed to be large but the total number of clusters is fixed. The method, which I refer to as a *rearrangement test*, applies to standard difference-in-differences estimation and other settings where treatment occurs in a single cluster and the treatment effect is identified by between-cluster comparisons. The key theoretical insight for the rearrangement test is that a mild restriction on some but not all of the heterogeneity in two samples of independent normal variables allows testing the equality of their means even if one sample consists of only a single observation. I prove that this is possible for empirically relevant levels of significance if the other sample consists of at least twenty observations. The test is feasible with even fewer observations if other restrictions are strengthened. The rearrangement test compares the data to a reordered version of itself after attaching a special weight to the sample with a single observation. The weights needed for most standard situations are tabulated in the paper and calculating additional weights is computationally simple. I also show that the weights remain approximately valid if the two samples of independent heterogeneous normal variables arise as a distributional limit. I exploit this result in the context of cluster-robust inference by constructing asymptotically normal cluster-level statistics to which the rearrangement test can be applied. The resulting test is consistent against all fixed alternatives to the null, powerful against $1/\sqrt{n}$ local alternatives, and does not require simulation or resampling.

Inference based on cluster-level estimates goes back at least to Fama and MacBeth (1973). Their approach is generalized and formally justified by Ibragimov and Müller (2010, 2016), who construct t statistics from cluster-level estimates and show that these statistics can be compared to Student t critical values. Canay, Romano, and Shaikh (2017) obtain null distributions by permuting the signs of cluster-level statistics under symmetry assumptions. Hagemann (2022) permutes cluster-level statistics directly but adjusts inference to control for the potential lack of exchangeability. All of these methods allow for a fixed number of large and heterogeneous clusters but require several treated clusters. At conventional significance levels, Canay et al. (2017) and Hagemann (2022) require at least four treated clusters. Ibragimov and Müller's (2016) approach remains valid with as few as two treated clusters. The rearrangement test

complements these methods because it relies on the same type of high-level condition on the cluster-level statistics but is explicitly designed for a single treated cluster. It does not readily extend to multiple treated clusters. Other methods that are valid with a fixed number of clusters are the tests of Bester, Conley, and Hansen (2011) and a cluster-robust version of the wild bootstrap (see, e.g, Cameron, Gelbach, and Miller, 2008; Djogbenou, MacKinnon, and Nielsen, 2019) analyzed by Canay, Santos, and Shaikh (2020). However, these papers rely on strong homogeneity conditions across clusters that are not needed here.

Several approaches for inference have been developed specifically for difference-in-differences estimation. Conley and Taber (2011) provide a method that is valid with a single treated cluster and infinitely many control clusters under strong independence and homogeneity conditions that justify an exchangeability argument. Ferman and Pinto (2019) extend this approach to estimators based on comparisons of means where the form of heteroskedasticity is known exactly. Another extension by Ferman (2020) allows for spatial correlation while maintaining Conley and Taber's exchangeability condition. The rearrangement test differs from these methods because it does not require infinitely many control clusters, does not rely on exchangeability conditions, and allows for completely unknown forms of heterogeneity. Other approaches due to MacKinnon and Webb (2019, 2020) use randomization (permutation) inference for difference-in-differences estimation and other models with few treated clusters. They test "sharp" (Fisher, 1935) nulls under randomization hypotheses and asymptotics where the number of clusters is eventually infinite. In contrast, the present paper is able to test conventional nulls in a setting with finitely many clusters.

The remainder of the paper is organized as follows: Section 2 introduces the rearrangement test in a stylized model. Section 3 discusses the practical implementation in several examples. Section 4 establishes the validity or approximate validity of the rearrangement test under explicit regularity conditions. Section 5 illustrates the finite sample behavior of the new test in simulations and in data used by Garthwaite et al. (2014), who analyze the effects of a large-scale disruption of public health insurance in Tennessee. Section 6 concludes. The appendix contains auxiliary results and proofs.

I will use the following notation. $1\{A\}$ is an indicator function that equals one if A is true and equals zero otherwise. Cardinality of a set A is denoted by $|A|$. Limits are as $n \to \infty$ unless noted otherwise and $\rightsquigarrow$ denotes convergence in distribution.

2. Stylized example

In this section, I construct a test for the equality of means of two samples of independent heterogeneous normal variables where one sample consists of only a single observation. The other sample has finitely many observations. I use this framework in the next section to analyze the situation where the "observations" are cluster-level summary statistics and only one cluster received treatment.

Consider q independent variables $X_{0,1}, \dots, X_{0,q}$ with $X_{0,k} \sim N(\mu_0, \sigma_k^2)$ for $1 \leqslant k \leqslant q$. Independently, there is an additional variable $X_1 \sim N(\mu_1, \sigma^2)$. I interpret this as a two-sample problem with "control" sample $X_{0,1}, \dots, X_{0,q}$ and "treatment" sample X_1, although all of the following still applies if these roles are reversed. The objective is to test the null hypothesis of equality of means,

$$H_0 \colon \mu_1 = \mu_0,$$

without knowledge of $\mu_0, \sigma, \sigma_1, \dots, \sigma_q$ and without assuming that these quantities can be consistently estimated. I account for the uncertainty about μ_0 by recentering the data $X = (X_1, X_{0,1}, \dots, X_{0,q})$ with $\bar{X}_0 = q^{-1} \sum_{k=1}^{q} X_{0,k}$ to define

$$S(X, w) = \Big((1+w)(X_1 - \bar{X}_0), (1-w)(X_1 - \bar{X}_0), X_{0,1} - \bar{X}_0, \dots, X_{0,q} - \bar{X}_0 \Big) \quad (2.1)$$

for some known weight $w \in (0,1)$ that will be chosen shortly. If $X_1 - \bar{X}_0 > 0$, then w increases $(1+w)(X_1 - \bar{X}_0)$ and decreases $(1-w)(X_1 - \bar{X}_0)$. If $X_1 - \bar{X}_0 < 0$, these effects are reversed. The idea underlying the test is that if the decreased version of $X_1 - \bar{X}_0$ is still large in comparison to $X_{0,1} - \bar{X}_0, \dots, X_{0,q} - \bar{X}_0$, then this size difference is unlikely to be only due to heterogeneity in $\sigma^2, \sigma_1^2, \dots, \sigma_q^2$ but provides evidence for the alternative $H_1 \colon \mu_1 > \mu_0$. The test performs this comparison by effectively using $S(X, w)$ as if it were the data. I show below that choosing an appropriate w in $S(X, w)$ allows me to bound the size of this test at a predetermined significance level.

The test is constructed as follows. For a given vector $s \in \mathbb{R}^d$, let $s_{(1)} \leqslant \cdots \leqslant s_{(d)}$ be the ordered entries of s. Denote by $s \mapsto s^{\triangledown} = (s_{(d)}, \dots, s_{(1)})$ the operation of

rearranging the components of s from largest to smallest. The test uses $S(X,w)$ and its rearranged version $S(X,w)^{\triangledown}$ in the difference-of-means statistic

$$s=(s_1,\dots,s_{q+2}) \mapsto T(s)=\frac{s_1+s_2}{2}-\frac{1}{q}\sum_{k=1}^{q}s_{k+2} \tag{2.2}$$

to define the test function

$$\varphi(X,w)=1\Big\{T\Big(S(X,w)\Big)=T\Big(S(X,w)^{\triangledown}\Big)\Big\}. \tag{2.3}$$

The test, which I refer to as *rearrangement test*, rejects if $\varphi(X,w)=1$ and does not reject otherwise. As stated, the test is against the alternative $H_1\colon \mu_1>\mu_0$. For a test against $H_1\colon \mu_1<\mu_0$, simply use $\varphi(-X,w)$. These alternatives can be combined to provide a two-sided test. I describe the exact implementation below equation (2.6) ahead. Also note that the first difference of means in (2.3) simplifies to $T(S(X,w))=X_1-\bar{X}_0$ but $T(S(X,w)^{\triangledown})$ is in general a complicated function of w.

The rearrangement test can be interpreted a permutation test that treats $S=S(X,w)$ as if it were the sample and uses the largest feasible critical value from the permutation distribution of $T(S)$ for the test decision. Here, a permutation of S is a reordering of the elements of S according to some rule g. Denote such a permutation by gS, denote the set of every possible g (including the identity, i.e., leaving S unchanged) by G, and denote by $T(S)_{(1)}\leqslant T(S)_{(2)}\leqslant\cdots\leqslant T(S)_{(j)}\leqslant\cdots\leqslant T(S)_{(|G|)}$ the ordered values of the permutation statistics $T(gS)$ as g varies over G. A permutation test rejects if $T(S)>T(S)_{(j)}$ for a pre-specified j, with a larger j corresponding to a more conservative test. Choosing $j=|G|$ is never appropriate here because $T(S)$ is among the permutation statistics, which forces $T(S)\leqslant T(S)_{(|G|)}$ for every S. The largest feasible $T(S)_{(j)}$ for a permutation test is therefore determined by the largest j such that $T(S)_{(j)}<T(S)_{(|G|)}$. This j is not necessarily $|G|-1$ because of potential ties but is characterized as a j for which the equivalence "$T(S)>T(S)_{(j)}$ if and only if $T(S)=T(S)_{(|G|)}$" holds. Because $g\mapsto T(gS)$ with T as in (2.2) is maximized by ordering S from largest to smallest, $T(S)_{(|G|)}$ equals $T(S^{\triangledown})$ and therefore rejecting if $T(S)=T(S^{\triangledown})$ as in (2.3) is indeed equivalent to rejecting if $T(S)>T(S)_{(j)}$ with the largest feasible $T(S)_{(j)}$. The rearrangement test is different from a classical permutation test, which would use exchangeability conditions on S to determine an appropriate critical value $T(S)_{(j)}$. This is not possible in the present context because

the S constructed here is far from exchangeable. Instead, the rearrangement test uses a large critical value to guarantee size control in the presence of heterogeneity but then fine-tunes the test through w so that it is not unnecessarily conservative.

When I study the statistical properties of the test $\varphi(X, w)$ in Section 4, I assume that the variances σ_k^2 of the $X_{0,k}$, $1 \leqslant k \leqslant q$, are bounded away from zero by some $\underline{\sigma}^2 > 0$ for all but one $X_{0,k}$ with possibly zero variance. The reason for this restriction is that if two (or more) $X_{0,k}$ had zero variance, this could be seen in the data because the $X_{0,k}$ have the same mean and two (or more) $X_{0,k}$ would therefore be identical. In contrast, a single zero variance cannot be detected. I also restrict the variance σ^2 of X_1 to be bounded above by some $\bar{\sigma}^2 < \infty$ because letting $\sigma \to \infty$ in $\varphi(X, w)$ would have the same effect as setting all σ_k^2 equal to zero. Under the null hypothesis, the distribution of $\varphi(X, w)$ is then determined by the unknown value of

$$\lambda \in \Lambda := \{(\mu_0, \sigma, \sigma_1, \dots, \sigma_q) \in \mathbb{R} \times (0, \infty)^{q+1} : \sigma \leqslant \bar{\sigma} \text{ and } \sigma_k \geqslant \underline{\sigma} \text{ for all } k \text{ but one}\}.$$

Under the alternative, the distribution of $\varphi(X, w)$ also depends on the treatment effect $\delta = \mu_1 - \mu_0$. I write $\mathrm{E}_{\lambda,\delta}$ and $P_{\lambda,\delta}$ to emphasize this dependence but occasionally drop subscripts to prevent clutter.

Theorem 4.1 in Section 4 shows that the rejection probability $\mathrm{E}_{\lambda,0}\varphi(X, w)$ under the null hypothesis satisfies

$$\mathrm{E}_{\lambda,0}\varphi(X, w) \leqslant \xi_q(w, \varrho),$$

where $\xi_q(w, \varrho)$ (defined in (4.1) ahead) is a function of the weight w, the number of control observations q, and the maximal relative heterogeneity $\varrho := \bar{\sigma}/\underline{\sigma}$ of treated and untreated observations. The parameter ϱ is user chosen and has a simple interpretation: it restricts how much more variable X_1 can be relative to the $X_{0,k}$ in the extreme case when one of the σ_k equals zero and the remaining σ_k are all equal to the lower limit $\underline{\sigma}$. This is the worst-case scenario for the test because X_1 is then likely to be very large on accident in comparison to the $X_{0,k}$. In that scenario, a ϱ of 5 simply means that the variance of X_1 can be up to $5^2 = 25$ times larger than the variances of all but one of the $X_{0,k}$ and "infinitely more variable" than the remaining $X_{0,k}$. However, even at $\varrho = 1$ or below the rearrangement test is robust against some heterogeneity because there are no restrictions on how much *less* variable X_1 can be than $X_{0,1}, \dots, X_{0,q}$.

The existence of the bound $\xi_q(w, \varrho)$ in Theorem 4.1 implies that the rearrangement test controls size, i.e.,

$$\sup_{\lambda \in \Lambda} \mathrm{E}_{\lambda,0} \varphi(X, w) \leqslant \alpha,$$

whenever q, w, and ϱ are such that $\xi_q(w, \varrho) \leqslant \alpha$ for the desired significance level α. Because a w that satisfies $\xi_q(w, \varrho) = \alpha$ is not necessarily unique and because Theorem 4.1 suggests that power against the alternative $H_1 : \mu_1 > \mu_0$ for w near one can be low, it is sensible to choose the smallest feasible w, denoted by

$$w_q(\alpha, \varrho) = \inf\Big\{w \in (0,1) : \xi_q(w, \varrho) = \alpha\Big\}, \tag{2.4}$$

in the definition of the rearrangement test function for a test of size α,

$$x \mapsto \varphi_\alpha(x) \coloneqq \varphi\Big(x, w_q(\alpha, \varrho)\Big). \tag{2.5}$$

The test φ_α also depends on ϱ but this is suppressed here to prevent clutter. Table 1 lists values of $w_q(\alpha, \varrho)$ for common choices of α as a function of ϱ and q. They guarantee

$$\sup_{\lambda \in \Lambda} \mathrm{E}_{\lambda,0} \varphi_\alpha(X) \leqslant \alpha. \tag{2.6}$$

The list is not exhaustive and additional values can be easily calculated by numerical integration. No simulation or optimization over Λ is needed. Software that performs the calculations can be found at `https://hgmn.github.io/rea`.

Table 1 shows that the rearrangement test is available in a wide variety of situations depending on the desired significance level and tolerance for heterogeneity. For instance, a test with a 10% significance level is already available with $q = 10$ control observations. A 5% level test becomes available at $q = 15$, a 1% level test at $q = 20$, and for $q \geqslant 25$ there are essentially no restrictions to the level and underlying heterogeneity. This provides two avenues for implementation:

(1) Choose a desired maximal degree of heterogeneity ϱ and make test decisions based on this choice.
(2) Determine at which degree of maximal heterogeneity the null hypothesis can no longer be rejected.

The first option is in line with the broader literature on testing in situations where the researcher has to take a stand on the value of a parameter that cannot be estimated

TABLE 1. Weights $w_q(\alpha, \varrho)$ as defined in (2.4) that guarantee size control at α for a given maximal degree of heterogeneity $\varrho = \bar{\sigma}/\underline{\sigma}$ for different values of q.

		q								
α	$\bar{\sigma}/\underline{\sigma}$	10	15	20	25	30	35	40	45	49
.10	2	.6333	.4010	.3294	.2829	.2475	.2188	.1948	.1742	.1562
	3		.6098	.5543	.5221	.4983	.4792	.4632	.4495	.4375
	4		.7127	.6669	.6418	.6238	.6094	.5974	.5871	.5781
	5		.7732	.7344	.7137	.6991	.6876	.6779	.6697	.6625
	6		.8129	.7792	.7615	.7493	.7396	.7316	.7248	.7188
	7		.8409	.8111	.7957	.7851	.7768	.7700	.7641	.7590
	8		.8616	.8350	.8213	.8120	.8048	.7987	.7936	.7891
	9		.8776	.8536	.8413	.8329	.8265	.8211	.8165	.8125
.05	2		.5752	.5020	.4615	.4318	.4081	.3884	.3715	.3568
	3		.7287	.6703	.6414	.6213	.6054	.5923	.5810	.5712
	4		.8024	.7541	.7314	.7161	.7041	.6942	.6858	.6784
	5		.8450	.8042	.7854	.7729	.7633	.7554	.7486	.7428
	6		.8727	.8374	.8213	.8108	.8028	.7962	.7905	.7856
	7		.8921	.8610	.8469	.8379	.8310	.8253	.8205	.8163
	8		.9064	.8786	.8661	.8582	.8521	.8471	.8429	.8392
	9		.9173	.8923	.8811	.8739	.8685	.8641	.8604	.8571
.025	2		.6981	.6049	.5656	.5387	.5175	.5001	.4852	.4723
	3			.7400	.7111	.6926	.6784	.6667	.6568	.6482
	4			.8069	.7838	.7696	.7588	.7501	.7426	.7362
	5			.8466	.8273	.8157	.8071	.8001	.7941	.7889
	6			.8728	.8563	.8465	.8393	.8334	.8284	.8241
	7			.8914	.8770	.8685	.8622	.8572	.8529	.8493
	8			.9053	.8924	.8849	.8795	.8751	.8713	.8681
	9			.9160	.9045	.8978	.8929	.8890	.8856	.8828
.01	2			.6986	.6543	.6286	.6092	.5935	.5801	.5686
	3			.8058	.7709	.7527	.7396	.7290	.7201	.7124
	4			.8578	.8290	.8147	.8047	.7968	.7901	.7843
	5			.8882	.8636	.8519	.8438	.8374	.8321	.8275
	6			.9080	.8866	.8767	.8699	.8645	.8601	.8562
	7			.9219	.9030	.8943	.8885	.8839	.8801	.8768
	8			.9322	.9153	.9076	.9024	.8984	.8951	.8922
	9			.9401	.9248	.9179	.9133	.9097	.9067	.9042
.005	2			.7642	.7029	.6764	.6576	.6426	.6300	.6191
	3				.8042	.7847	.7719	.7618	.7534	.7461
	4				.8544	.8389	.8290	.8214	.8150	.8096
	5				.8842	.8713	.8632	.8571	.8520	.8477
	6				.9040	.8929	.8861	.8809	.8767	.8731
	7				.9180	.9082	.9024	.8980	.8943	.8912
	8				.9284	.9198	.9146	.9107	.9075	.9048
	9				.9365	.9287	.9241	.9207	.9178	.9154

Note: Missing cells mean that the test is not recommended or not feasible. The vertical lines are discussed above Proposition 4.3.

without additional restrictions on the data; see, e.g., Kitagawa (2015), Kolesár and Rothe (2018), and Armstrong and Kolesár (2021). The second option can be viewed as sensitivity analysis. Implementing the test in this way has a meaningful interpretation because a result that is robust to a tenfold larger standard deviation in the treated observation relative to the control sample is more credible than a result that only survives a twofold difference in standard deviation. This second option leaves it up to the reader to decide whether the results are convincing.

The test decision itself is simple. Determine $w = w_q(\alpha, \varrho)$ for a given number of control observations q, desired significance level α, and tolerance for heterogeneity ϱ. For this w, compute $S = S(X, w)$ as in (2.1) and reorder the entries of S from largest to smallest to obtain $S^\triangledown$. For an α-level test of $\mu_1 = \mu_0$, reject in favor of $\mu_1 > \mu_0$ if $T(S) = T(S^\triangledown)$ as defined in (2.2). For a one-sided test with level α against $\mu_1 < \mu_0$, reject if $T(-S) = T((-S)^\triangledown)$. For a two-sided test with level 2α, reject in favor of $\mu_1 \neq \mu_0$ if either

$$T(S) = T(S^\triangledown) \text{ or } T(-S) = T((-S)^\triangledown). \tag{2.7}$$

If desired, increase ϱ until the null hypothesis can no longer be rejected against the alternative of interest. The test decision is monotonic in ϱ, i.e., if $\varrho' > \varrho$ lead to the same test decision, then the decision does not change for any value between ϱ and ϱ'. `R` and `Stata` commands that implement the test for any choice of ϱ and find the largest feasible ϱ are available at `https://hgmn.github.io/rea`. For a given ϱ, it is also possible to compute p-values as $\hat{p}_X = \inf\{\alpha : \varphi_\alpha(X) = 1\}$. These p-values provide the smallest significance level under which the null hypothesis would be rejected and satisfy $P_{\lambda,0}(\hat{p}_X \leqslant u) \leqslant u$ for all $\lambda \in \Lambda$.

3. Practical implementation

In this section, I use three examples to describe how to apply the rearrangement test introduced in the previous section to research designs with a finite number of large, heterogeneous clusters when only a single cluster received treatment.

The examples discussed below have several generic features: Data from $q + 1$ large clusters (e.g., states, industries, or villages, possibly observed over more than one time period) are available. Data are dependent within clusters but independent across

clusters. The exact form of dependence is unknown and not presumed to be estimable. An intervention took place during which one cluster received treatment and q clusters did not. The quantity of interest is a treatment effect or an object related to it that can be represented by a scalar parameter δ. Because the entire cluster was treated, this parameter is only identified up to a location shift θ_0 within the treated cluster and therefore only the left-hand side of

$$\theta_1 = \theta_0 + \delta$$

can be identified from this cluster. If the treated cluster would have behaved similarly to the untreated clusters in the absence of an intervention, then θ_0 can be identified from each untreated cluster. Pairwise comparison then identifies δ. The objective is to test $H_0\colon \theta_1 = \theta_0$ or, equivalently,

$$H_0\colon \delta = 0.$$

The rearrangement test in this context relies on the availability of an estimate $\hat{\theta}_1$ of θ_1 and estimates $\hat{\theta}_{0,1}, \ldots, \hat{\theta}_{0,q}$ of θ_0 so that

$$\hat{\theta}_n = (\hat{\theta}_1, \hat{\theta}_{0,1} \ldots, \hat{\theta}_{0,q})$$

is approximately a vector of independent but potentially heterogeneous normal variables (defined precisely in equation (4.3) ahead) that can be used as if it were the data vector X from Section 2. Independence of the estimates $\hat{\theta}_1, \hat{\theta}_{0,1} \ldots, \hat{\theta}_{0,q}$ can be ensured by using only data from one cluster per estimate because the clusters are assumed to be independent. Approximate normality is a reasonable assumption if the estimates consist of (possibly weighted) averages, differences of averages, or other quantities to which a central limit theorem can be applied. For the central limit theory, the number of observations per cluster is presumed to be large but the number of clusters is fixed. With $\hat{\theta}_n$ now playing the role of X from the previous section, we can rewrite the vector S, defined in (2.1), as

$$S(\hat{\theta}_n, w) = \Big((1+w)(\hat{\theta}_1 - \bar{\theta}_0), (1-w)(\hat{\theta}_1 - \bar{\theta}_0), \hat{\theta}_{0,1} - \bar{\theta}_0, \ldots, \hat{\theta}_{0,q} - \bar{\theta}_0\Big),$$

where $\bar{\theta}_0 = q^{-1} \sum_{k=1}^{q} \hat{\theta}_{0,k}$.

The following examples focus on the construction of θ_1, θ_0, and $\hat{\theta}_n$ in several standard applications. The $\hat{\theta}_n$ is the main input for the rearrangement test described in Algorithm 3.4 below, which contains a general step-by-step procedure for inference about a scalar parameter in research designs with a finite number of large, heterogeneous clusters and a single treated cluster.

Example 3.1 (Regression with cluster-level treatment). Consider a linear regression model

$$Y_{i,k} = \theta_0 + \delta D_k + \beta_k' X_{i,k} + U_{i,k},$$

where i indexes individuals within cluster k. There are $q+1$ clusters and individuals in cluster $k = q+1$ received treatment ($D_{q+1} = 1$) but those in $1 \leqslant k \leqslant q$ did not ($D_k = 0$). The parameter of interest δ on the treatment indicator D_k can be interpreted as an average treatment effect under suitable conditions. See, e.g., Słoczyński (2018, 2020) and references therein for a precise discussion. The regression may also include covariates $X_{i,k}$ that vary within each cluster and have coefficients β_k that may vary across clusters. The condition $\mathrm{E}(U_{i,k} \mid D_k, X_{i,k}) = 0$ identifies $\theta_1 = \theta_0 + \delta$ within the treated cluster and θ_0 within the untreated clusters. The preceding display can then be written as

$$Y_{i,k} = \begin{cases} \theta_0 + \beta_k' X_{i,k} + U_{i,k}, & 1 \leqslant k \leqslant q, \\ \theta_1 + \beta_k' X_{i,k} + U_{i,k}, & k = q+1. \end{cases}$$

View these as $q+1$ separate regressions and use the least squares estimates of the constants θ_1 and θ_0 as the vector $\hat{\theta}_n = (\hat{\theta}_1, \hat{\theta}_{0,1} \ldots, \hat{\theta}_{0,q})$ described above. This $\hat{\theta}_n$ can be viewed as approximately normal if the dependence within each cluster is weak enough for a suitable central limit theorem to apply. □

Example 3.2 (Difference in differences). Consider the panel model

$$Y_{i,t,k} = \theta_0 I_t + \delta I_t D_k + \beta_k' X_{i,t,k} + \zeta_{i,k} + U_{i,t,k}, \tag{3.1}$$

where i indexes individuals in unit $k \in \{1, \ldots, q+1\}$ at time $t \in \{0, 1\}$. Treatment occurred between periods 0 and 1. Right-hand side variables are a post-intervention indicator $I_t = 1\{t = 1\}$, a treatment indicator D_k that equals 1 if unit k ever received treatment, individual fixed effects $\zeta_{i,k}$, and other covariates $X_{i,t,k}$ that for every k vary

at least before or after the intervention. The collection of pre and post intervention data from unit k forms the k-th cluster. View each cluster as a separate regression and rewrite (3.1) in first differences as

$$\Delta Y_{i,k} = \begin{cases} \theta_0 + \beta_k' \Delta X_{i,k} + \Delta U_{i,k}, & 1 \leqslant k \leqslant q, \\ \theta_1 + \beta_k' \Delta X_{i,k} + \Delta U_{i,k}, & k = q+1, \end{cases}$$

where $\Delta Y_{i,k} = Y_{i,1,k} - Y_{i,0,k}$ and so on. Provided $\mathrm{E}(\Delta U_{i,k} \mid \Delta X_{i,k}) = 0$, the data identify $\theta_1 = \theta_0 + \delta$ in a treated cluster and θ_0 in an untreated cluster. The least squares estimates $\hat{\theta}_1$ and $\hat{\theta}_{0,k}$ of the parameters θ_1 and θ_0 are suitable cluster-level estimates if $\hat{\theta}_n = (\hat{\theta}_1, \hat{\theta}_{0,1}, \ldots, \hat{\theta}_{n,q})$ satisfies a central limit theorem. □

Example 3.3 (Two-way fixed effects). Construction of θ_0, θ_1, and $\hat{\theta}_n$ is not always as straightforward as in the preceding examples. For instance, consider the two-way fixed effects model

$$Y_{t,k} = \delta I_t D_k + \eta_t + \zeta_k + U_{t,k}, \tag{3.2}$$

where I_t is a post-intervention indicator, $D_k = 1\{k = q+1\}$ is a treatment indicator, and η_t and ζ_k are time and cluster fixed effects, respectively. Neither θ_0 nor θ_1 are present in (3.2) but construction of $\hat{\theta}_n$ is still possible. Let $\bar{Y}_{-,k}$ and $\bar{Y}_{+,k}$ be time averages of $Y_{t,k}$ pre and post intervention, respectively, and define $\Delta\bar{Y}_k = \bar{Y}_{+,k} - \bar{Y}_{-,k}$. The fixed effects estimator $\hat{\delta}$ can be written as $\hat{\delta} = \Delta\bar{Y}_{q+1} - \sum_{k=1}^{q} \Delta\bar{Y}_k / q$, which suggests using $\Delta\bar{Y}_{q+1}$ as estimate of $\theta_1 = \mathrm{E}\Delta\bar{Y}_{q+1}$ and $\Delta\bar{Y}_k$ for $1 \leqslant k \leqslant q$ as estimates of $\theta_0 = \mathrm{E}\Delta\bar{Y}_k$. Equivalently, the $\Delta\bar{Y}_k$ can be computed in $q+1$ separate artificial regressions of $Y_{t,k}$ on a constant and the post-intervention indicator I_t,

$$Y_{t,k} = \begin{cases} \zeta + \theta_0 I_t + \mathrm{error}_{t,k}, & 1 \leqslant k \leqslant q, \\ \zeta + \theta_1 I_t + \mathrm{error}_{t,k}, & k = q+1, \end{cases} \tag{3.3}$$

where ζ is the intercept in each regression. The least squares estimates of θ_0 and θ_1 satisfy $\hat{\theta}_{0,k} = \Delta\bar{Y}_k$ for $1 \leqslant k \leqslant q$ and $\hat{\theta}_1 = \Delta\bar{Y}_{q+1}$. If (3.2) includes covariates, then these covariates can also be included in (3.3). The estimates $\hat{\theta}_1, \hat{\theta}_{0,1}, \ldots, \hat{\theta}_{0,q}$ can be viewed as approximately normal if (3.2) comes from individual-level data aggregated to the cluster level even if the number of time periods is fixed. □

Remark (Nonlinear models). The methodology presented here also includes nonlinear models because the parameter δ does not need to be interpretable by itself. For example, suppose the model in Example 3.1 is the latent model in a binary choice framework with symmetric link function F and $\beta_k \equiv \beta$. Then $F(\theta_0 + \delta + \beta' x) - F(\theta_0 + \beta' x)$ for some x may be the treatment effect of interest but $H_0\colon \delta = 0$ still determines whether the treatment effect is zero or not. Estimates of θ_0 and $\theta_1 = \theta_0 + \delta$ from these models typically do not have closed form in the presence of covariates but generally have asymptotic representations to which a central limit theorem can be applied. □

The following procedure shows how to use $\hat{\theta}_n$ such as those computed in Examples 3.1-3.3 in the rearrangement test. By Theorem 4.4 ahead, this procedure provides an asymptotically α-level test in the presence of a finite number of large clusters when only a single cluster received treatment. The test is able to detect all fixed alternatives and has power against $1/\sqrt{n}$-local alternatives. Recall that ϱ here measures how much more variable the estimate from the treated cluster $\hat{\theta}_1$ can be relative to the second-least variable control cluster estimate $\hat{\theta}_{0,k}$. A ϱ of 5 means that the (asymptotic) variance of $\hat{\theta}_1$ can be up to $5^2 = 25$ times larger. There is no restriction on how much *less* variable $\hat{\theta}_1$ can be than any of the other estimates and $\hat{\theta}_1$ can be infinitely more variable than the least variable control cluster.

Algorithm 3.4 (Rearrangement test). (1) Use Table 1 or the provided software to obtain w for the number of available control clusters q, desired significance level α, and maximal tolerance for heterogeneity, e.g., $\varrho = 2$.

(2) Compute for each untreated cluster $k = 1, \ldots, q$ an estimate $\hat{\theta}_{0,k}$ of θ_0 and compute an estimate $\hat{\theta}_1$ of θ_1 from the treated cluster so that the difference $\theta_1 - \theta_0$ is the treatment effect of interest. (See Examples 3.1-3.3 above.) Use $\hat{\theta}_n = (\hat{\theta}_1, \hat{\theta}_{0,1} \ldots, \hat{\theta}_{0,q})$ as if it were X in (2.1) to compute $S = S(\hat{\theta}_n, w)$ with w as in Step (1). Note that $\bar{X}_0$ is replaced here by $q^{-1} \sum_{k=1}^{q} \hat{\theta}_{0,k}$.

(3) Reorder the entries of S from largest to smallest. Denote this by $S^{\triangledown}$ as defined above (2.2). Compute $T(S)$ and $T(S^{\triangledown})$ as in (2.2).

(4) Reject $H_0\colon \theta_1 = \theta_0$ in favor of
   (a) $H_1\colon \theta_1 > \theta_0$ if $T(S) = T(S^{\triangledown})$.
   (b) $H_1\colon \theta_1 < \theta_0$ if $T(-S) = T((-S)^{\triangledown})$.

(c) $H_1\colon \theta_1 \neq \theta_0$ if either $T(S) = T(S^{\triangledown})$ or $T(-S) = T((-S)^{\triangledown})$ but use $\alpha/2$ in Step (1). □

This procedure can also be used as sensitivity analysis if inference was originally performed with a method designed for a finer level of clustering, e.g., at the county level instead of the state level. In that case Algorithm 3.4 can illustrate how well the results of the original test hold up if there is dependence across counties. For this type of analysis, one could start at $\varrho = 0$ or $\varrho = 1$ and increase ϱ until the null hypothesis can no longer be rejected. This is informative because a result that holds up to a potentially $\varrho^2 = 25$ times larger variance is more credible than a result that only holds if $\varrho^2 = 1$, i.e., if $\hat{\theta}_1$ cannot be more variable than all but one $\hat{\theta}_{0,k}$. `R` and `Stata` commands that implement Algorithm 3.4 and the sensitivity analysis for any choice of ϱ are available at `https://hgmn.github.io/rea`.

Finally, it is important to note that the assumptions maintained by the rearrangement test do not collapse to the assumption of homogeneity for any value of ϱ^2. If homogeneity of the components of $\hat{\theta}_n$ is assumed and robustness against heterogeneity is not required, there is no advantage to using the rearrangement test. Instead, one can simply perform a standard permutation test using $\hat{\theta}_n$ as the data (or alternatively use the Conley and Taber (2011) test, see Example 5.1 ahead). For the permutation test, define the difference-of-means statistic $\bar{T}(\hat{\theta}_n) = \hat{\theta}_1 - \sum_{k=1}^{q} \hat{\theta}_{0,k}/q$ and let $g_k\hat{\theta}_n$ be the action of switching the location of $\hat{\theta}_1$ and $\hat{\theta}_{0,k}$ in $\hat{\theta}_n$. Let $\bar{T}(\hat{\theta}_n)_{(1)} \leqslant \bar{T}(\hat{\theta}_n)_{(2)} \leqslant \ldots \leqslant \bar{T}(\hat{\theta}_n)_{(q+1)}$ be the ordered values of $(\bar{T}(\hat{\theta}_n), \bar{T}(g_1\hat{\theta}_n), \ldots, \bar{T}(g_q\hat{\theta}_n))$ and let $\lceil a \rceil$ be the smallest integer greater than or equal a. Using arguments as in Canay et al. (2017) or Hagemann (2023), it is then straightforward to show that

$$\bar{T}(\hat{\theta}_n) > \bar{T}(\hat{\theta}_n)_{(\lceil (1-\alpha)(q+1) \rceil)} \tag{3.4}$$

is an asymptotically α-level test under assumptions maintained in Theorem 4.4 if the (asymptotic) variances of $\hat{\theta}_1, \hat{\theta}_{0,1}, \ldots, \hat{\theta}_{0,q}$ are all the same. The rearrangement test can be viewed as a version of this test that allows for these variances to differ.

The next section provides the theoretical justification for the rearrangement test. Readers mostly interested in applying the test can move ahead to Section 5, where I illustrate how the rearrangement test performs in Monte Carlo experiments.

4. Theory

Section 4.1 analyzes the size and power of the rearrangement test in the normal model introduced in Section 2. Section 4.2 connects these results with the situation where normality is an asymptotic approximation.

4.1. Inference with heterogenous normal variables. I start by establishing a bound on the probability $\mathrm{E}_{\lambda,0}\varphi(X,w)$ that the rearrangement test function $\varphi(X,w)$, defined in (2.3), rejects under the null hypothesis $H_0\colon \mu_1=\mu_0$ for the normal model. This bound is valid uniformly in the parameters $\lambda\in\Lambda$ and holds for a fixed number of control observations q. I also show that the test has power against the alternative $H_1\colon \mu_1>\mu_0$. Results in the other direction follow by considering $\mathrm{E}_{\lambda,-\delta}\varphi(-X,w)$ instead of $\mathrm{E}_{\lambda,\delta}\varphi(X,w)$. The bound relies on the joint normality of X combined with the location and scale invariance property $\varphi(X,w)=\varphi((X-\mu_0 1_{q+1})/\sigma,w)$, where 1_{q+1} is a $(q+1)$-vector of ones. The location invariance is forced by the recentering of X with $\bar{X}_0$, which effectively removes μ_0 from the list of nuisance quantities. The scale invariance is ensured by how the data transformation S and test statistic T (introduced at the beginning of Section 2) enter the test function φ. Scale invariance reduces the unit-dependent unknowns $\sigma,\sigma_1,\dots,\sigma_q$ to the more tractable ratios $\sigma_1/\sigma,\dots,\sigma_q/\sigma$. While I only discuss results for S and T because of these convenient properties, it should be noted that other statistics and transformations may also lead to valid tests. Let Φ and ϕ denote the normal distribution and density functions, respectively.

Theorem 4.1 (Size and power). *Let $X_1,X_{0,1},\dots,X_{0,q}$ be independent with $X_1\sim N(\mu_0+\delta,\sigma^2)$ and $X_{0,k}\sim N(\mu_0,\sigma_k^2)$ for $1\leqslant k\leqslant q$. If $\delta=0$, then for all $w\in(0,1)$,*

$$\sup_{\lambda\in\Lambda}\mathrm{E}_{\lambda,0}\varphi(X,w)\leqslant\xi_q(w,\varrho):=\frac{1}{2^{q+1}}+\int_0^\infty\Phi\big((1-w)\varrho y\big)^{q-1}\phi(y)dy \tag{4.1}$$
$$+\min_{t>0}\left(\Phi\left(\sqrt{q-1}wt\right)^{q-1}+2\Phi(-qt)\right).$$

Furthermore, for every $\lambda\in\Lambda$ and $w\in(0,1)$, we have $\lim_{\delta\to\infty}\mathrm{E}_{\lambda,\delta}\varphi(X,w)=1$ and $\lim_{\delta\to\infty}\mathrm{E}_{\lambda,\delta}\varphi(X,1)=0$.

The theorem implies that the rearrangement test controls size, i.e.,

$$\sup_{\lambda\in\Lambda}\mathrm{E}_{\lambda,0}\varphi(X,w)\leqslant\alpha,$$

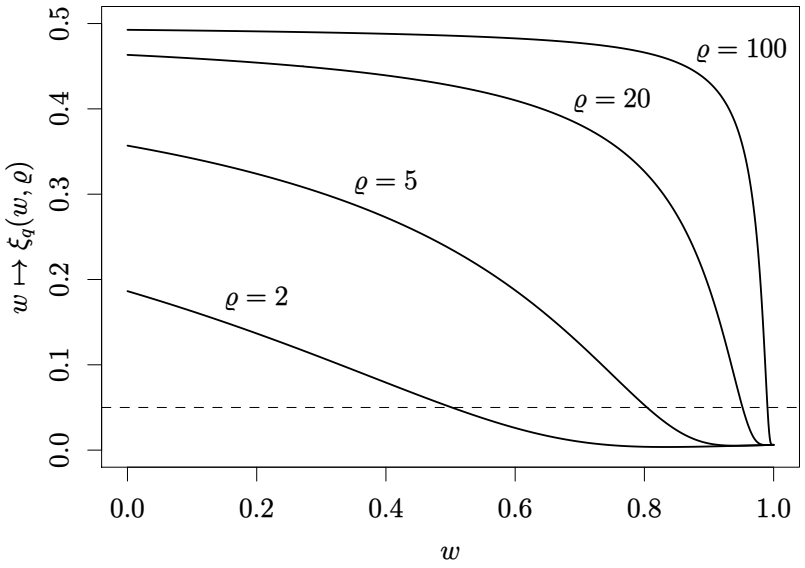

FIGURE 1. Solid lines show the size bound $\xi_q(w, \varrho)$ at $q = 20$ control observations as a function of the weight w for different values of the maximal heterogeneity ϱ. The dashed line equals .05.

whenever q, w, and ϱ are such that $\xi_q(w, \varrho) \leqslant \alpha$ for the desired significance level α. The bound $\xi_q(w, \varrho)$ has several properties that make this possible. In particular, it is monotonically increasing in ϱ and decreasing in q. The reason for the monotonicity is that if X_1 can be more variable than $X_{0,1}, \ldots, X_{0,q}$, then the burden of proof to show "$\mu_1 > \mu_0$" as opposed to "$\mu_1 = \mu_0$ with a large realization of X_1" becomes necessarily higher. A large q can ameliorate this effect somewhat because it removes uncertainty about μ_0. The bound also tends to be decreasing in $w \in [0, 1]$ because the integral generally dominates the other components, but can increase slightly in some situations. This is illustrated in Figure 1, where $w \mapsto \xi_q(w, \varrho)$ (solid lines) is essentially decreasing over the entire domain except for $\varrho = 2$ and $w \geqslant .85$. Most importantly, it can be seen that $w \mapsto \xi_q(w, \varrho)$ decreases enough to dip below the desired significance level $\alpha = .05$ (dashed line) for all values of ϱ. As q increases (not shown), $w \mapsto \xi_q(w, \varrho)$ is pushed towards zero but the shape of the function does not change meaningfully with q. The w at which $\xi_q(w, \varrho) = \alpha$ is generally unique for most empirically relevant α and does not exist in some situations. This can be seen in Figure 1, where $w \mapsto \xi_q(w, \varrho)$ crosses $\alpha = .05$ only once for each ϱ but, for example, $\xi_q(w, 2)$ remains below .2 so that $\xi_q(w, 2) = .2$ is never attained. In the latter case, w can simply be set to zero

because size is controlled as long as $\xi_q(w, 2)$ is below the desired significance level. The software package implements the test in this way if ϱ is small but q or α are relatively large. However, the size of the test may be below α in that case. If needed, this can be remedied by increasing ϱ until $\xi_q(w, \varrho) = \alpha$. If the w at which $\xi_q(w, \varrho) = \alpha$ is not unique, the test function φ_α, defined in (2.5), chooses the smallest possible w that satisfies $\xi_q(w, \varrho) = \alpha$.

Theorem 4.1 also provides information about the interplay between w and the test under the alternative. In particular, it shows that the rearrangement test has power against $H_1 : \mu_1 > \mu_0$ for every $w \in (0, 1)$ but the power declines sharply at $w = 1$. I therefore explore the behavior of the test with w near 1 further in the following result. It provides a lower bound on the power of the test for fixed δ.

Proposition 4.2 (Lower bound on power). *Let $X_1, X_{0,1}, \ldots, X_{0,q}$ be independent with $X_1 \sim N(\mu_0 + \delta, \sigma^2)$ and $X_{0,k} \sim N(\mu_0, \sigma_k^2)$ for $1 \leqslant k \leqslant q$. For every $w \in (0, 1)$, $\sigma, \sigma_1, \ldots, \sigma_q > 0$, and $\delta > 0$,*

$$\inf_{\mu_0 \in \mathbb{R}} \mathrm{E}_{\lambda,\delta} \varphi(X, w) \geqslant 2^q \sup_{t \geqslant 0} \Phi\left(\frac{\delta}{\sigma} - \frac{1+w}{1-w} t\right) \prod_{k=1}^{q} \left(\Phi\left(\frac{\sigma}{\sigma_k} t\right) - 0.5\right).$$

The supremum is attained on $t \in (0, \infty)$. The right-hand side is strictly positive and converges to 1 as $\delta \to \infty$.

The lower bound shows that the test exhibits a standard relationship between the signal δ and the noise components $\sigma_1, \ldots, \sigma_q$. Power is low if the signal relative to σ is weak or the noise in the control group relative to σ is strong. The latter relationship is in contrast to Theorem 4.1, where small σ_k relative to σ were problematic. In addition, the bound also clarifies that w dampens δ through the function $w \mapsto (1 + w)/(1 - w)$, which is arbitrarily large for w sufficiently close to 1. A w very close to 1 can therefore drown out a large treatment effect even if the noise coming from the control observations is mild. (The role of the supremum is simply to find the best possible balance for a given set of parameters.) This provides the motivation for choosing the smallest possible w in (2.4) when defining φ_α in (2.5). It is also worth noting that the bound is tight enough to converge to 1 as $\delta \to \infty$ and to 0 as $w \to 1$.

I now discuss some technical aspects of the bound $\sup_{\lambda\in\Lambda} \mathrm{E}_{\lambda,0}\varphi(X,w) \leqslant \xi_q(w,\varrho)$, introduced in Theorem 4.1, that forms the theoretical underpinning for the rearrangement test. The components of $\xi_q(w,\varrho)$,

$$\underbrace{\min_{t>0}\left(\Phi\left(\sqrt{q-1}wt\right)^{q-1}+2\Phi(-qt)\right)}_{\text{(i)}}+\underbrace{\frac{1}{2^{q+1}}}_{\text{(ii)}}+\underbrace{\int_0^\infty \Phi\big((1-w)\varrho y\big)^{q-1}\phi(y)dy}_{\text{(iii)}}, \qquad (4.2)$$

all have simple interpretations: Component (i) is a uniform bound on the distance between $\mathrm{E}_{\lambda,0}\varphi_\alpha(X)$ and the rejection probability of an oracle version of the test where μ_0 is known and used instead of $\bar{X}_0$.[1] The rejection probability of the oracle test can be represented as the probability of two disjoint events, one of which cannot exceed (ii). Component (iii) bounds the probability of the other event uniformly in Λ and, in particular, there is a $\lambda\in\Lambda$ such that this probability attains (iii). Taken together, $\xi_q(w,\varrho)$ can therefore be roughly viewed as a tight bound up to the two adjustments (i) and (ii). These adjustments are generally small relative to (iii) for moderately large q. I use Table 1 to illustrate their relative size. In the table, empty cells correspond to situations where there is either $\xi_q(w,\varrho)>\alpha$ or more than $\alpha/2$ of $\xi_q(w_q(\alpha,\varrho),\varrho)$ is taken up by (i)+(ii). Cells to the left of vertical lines are settings where between $\alpha/2$ and $\alpha/10$ of the bound are taken up by (i)+(ii). The lack of tightness in the remaining cells, as measured by (i)+(ii), is less than $\alpha/10$. For these cells $\sup_{\lambda\in\Lambda}\mathrm{E}_{\lambda,0}\varphi_\alpha(X)$ approximately equals α. As the table shows, $\xi_q(w_q(\alpha,\varrho),\varrho)$ is an essentially tight bound for $\sup_{\lambda\in\Lambda}\mathrm{E}_{\lambda,0}\varphi_\alpha(X)$ for $q\geqslant 30$. The bound is also nearly tight for values of q as small as 15 as long as ϱ is not too large. As a referee points out, component (iii) also cannot exceed $1/2$, which effectively rules out significance levels above $1/2$. The $1/2$ is reached as $\varrho\to\infty$ and is equivalent to a situation where the σ_k are all equal to zero while σ is positive. In that case, (iii) is the probability that the mean-zero normal variable $X_1-\mu_0$ exceeds $X_{0,k}-\mu_0$, which now has point mass at 0. That probability is equal to $1/2$.

Inspection of the proof of Theorem 4.1 also reveals that if the parameter space is shrunk to $\Lambda\cap\{\sigma_k\geqslant\underline{\sigma}\text{ for all }k\}$ to remove the potential zero variance for one of the

[1]The minimizer does not have closed form but is easily found numerically. In particular, at $t=1/q$, $\Phi(\sqrt{q-1}wt)^{q-1}+2\Phi(-qt)<\Phi(1/\sqrt{q})^{q-1}+2\Phi(-1)<1$ for $q>2$. Because $\Phi(\sqrt{q-1}wt)^{q-1}+2\Phi(-qt)\geqslant 1$ at $t\in\{0,\infty\}$, the minimization problem always has an interior solution. This also implies that the bound as a whole is a smooth function of w and ϱ.

variables, the bound (4.1) in the theorem can be improved slightly to

$$\frac{1}{2^{q+1}} + \int_0^\infty \Phi\Big((1-w)\varrho y\Big)^q \phi(y)dy + \min_{t>0}\bigg(\Phi\Big(\sqrt{q}wt\Big)^q - \Phi\Big(-\sqrt{q}wt\Big)^q + 2\Phi(-qt)\bigg).$$

For the majority of values in Table 1, this decreases the weight by less than .001. However, when $q \leqslant 20$, removing the possibility of a zero variance can meaningfully lower the bound for larger values of ϱ. The software packages therefore also give the option to use this bound instead of (4.1).

Finally, before concluding this section, I show that the rearrangement test remains approximately valid for random vectors X_n converging in distribution to the random vector $X = (X_1, X_{0,1}, \dots, X_{0,q})$ described in Theorem 4.1. The reason is that $\mathrm{E}\varphi(X_n, w)$ and $\mathrm{E}\varphi(X, w)$ eventually coincide whenever X has independent entries and a smoothly distributed first entry. The X in Theorem 4.1 easily satisfies these conditions, which makes $\varphi_\alpha(X_n)$ asymptotically an α-level test.

Proposition 4.3 (Large sample approximation). *Let $X_1, X_{0,1}, \dots, X_{0,q}$ be independent and let X_1 have a continuous distribution. If $X_n \rightsquigarrow X$, then $\mathrm{E}\varphi(X_n, w) \to \mathrm{E}\varphi(X, w)$ for every $w \in (0,1)$.*

4.2. Inference with a single treated cluster. I will now show that the cluster-level statistics $\hat{\theta}_n$ introduced in Section 3 can be used together with the results in the previous section to perform a consistent test as the sample size n grows large. The test is not limited to parameters estimated by least squares as in Examples 3.1-3.3. Instead, consistency relies on the condition that a centered and scaled version of some estimate $\hat{\theta}_n$ converges to a $(q+1)$-dimensional normal distribution,

$$\sqrt{n}\left(\frac{\hat{\theta}_1 - \theta_1}{\sigma(\theta_1)}, \frac{\hat{\theta}_{0,1} - \theta_0}{\sigma_1(\theta_0)}, \dots, \frac{\hat{\theta}_{0,q} - \theta_0}{\sigma_q(\theta_0)}\right) \overset{\theta}{\rightsquigarrow} N(0, I_{q+1}), \tag{4.3}$$

where $\overset{\theta}{\rightsquigarrow}$ denotes weak convergence under $\theta = (\theta_1, \theta_0)$. For fixed θ, the display can be interpreted as $\sqrt{n}(\hat{\theta}_1 - \theta_1, \dots, \hat{\theta}_{0,1} - \theta_0, \dots, \hat{\theta}_{0,q} - \theta_0) \rightsquigarrow N(0, \mathrm{diag}(\sigma^2, \sigma_1^2, \dots, \sigma_q^2))$ to include the case that one of the $\sigma_1, \dots, \sigma_q$ may be zero as in Theorem 4.1.

Condition (4.3) is similar to the high-level conditions imposed by Ibragimov and Müller (2010, 2016) and Canay et al. (2017). The key difference is that these papers do not allow for only a single treated cluster. A common feature is that the σ and

$\sigma_1, \dots, \sigma_q$ are not assumed to be known or estimable by the researcher. This is important for applications because consistent variance estimation generally requires knowledge of an explicit ordering of the dependence structure within each cluster. While time-dependent data are automatically ordered, it may be difficult or impossible to infer or credibly assume an ordering of the data within states or villages. A condition like (4.3) can be established under weak (short-range) dependence conditions that only require *existence* of a potentially unknown ordering for which the dependence of more distant units decays sufficiently fast. El Machkouri, Volný, and Wu (2013) present convenient moment bounds and limit theorems for this situation. For more results in this direction, see also Bester et al. (2011) and references therein. In general, the convergence in (4.3) also implicitly requires the number of observations in all clusters to grow with the sample size n. However, the clusters are not required to have similar or even identical sizes. Another noteworthy feature of condition (4.3) is the diagonal covariance matrix of the limiting distribution. It is the only independence condition that is imposed on the clusters.

I now show that under the joint convergence (4.3), a rearrangement test that uses $\hat{\theta}_n$ is asymptotically of level α with a single treated cluster and a fixed number of control clusters. The test $\varphi_\alpha(\hat{\theta}_n)$, as defined in (2.5), has power against all fixed alternatives $\theta_1 = \theta_0 + \delta$ with $\delta > 0$ and local alternatives $\theta_1 = \theta_0 + \delta/\sqrt{n}$ converging to the null. In the latter situation, θ_0 is fixed and $\theta = (\theta_0 + \delta/\sqrt{n}, \theta_0)$ implicitly depends on n. The convergence in (4.3) is then a statement about an entire sequence $(\theta_0 + \delta/\sqrt{n}, \theta_0)$ instead of a single point. Results for alternatives with $\delta < 0$ follow from the same result by considering $\varphi_\alpha(-\hat{\theta}_n)$. These tests can be combined as in Algorithm 3.4 for a two-sided test that has power against fixed and local alternatives from either direction.

Theorem 4.4 (Consistency and local power). *Suppose*

$$\sqrt{n}(\hat{\theta}_1 - \theta_1, \dots, \hat{\theta}_{0,1} - \theta_0, \dots, \hat{\theta}_{0,q} - \theta_0) \rightsquigarrow N\Big(0, \operatorname{diag}(\sigma^2, \sigma_1^2, \dots, \sigma_q^2)\Big)$$

with $\bar{\sigma} \geqslant \sigma$, at most one $\sigma_k = 0$ for $1 \leqslant k \leqslant q$, and $\sigma_k \geqslant \underline{\sigma} > 0$ for all remaining k. If $\theta_1 = \theta_0$ and $\varrho = \bar{\sigma}/\underline{\sigma}$, then

$$\lim_{n\to\infty} \mathrm{E}\varphi_\alpha(\hat{\theta}_n) \leqslant \alpha, \qquad \text{every } \alpha, \varrho \text{ with } 0 < w_q(\alpha, \varrho) < 1,$$

where $w_q(\alpha, \varrho)$ is defined in (2.4). If $\theta_1 > \theta_0$, then $\mathrm{E}\varphi_\alpha(\hat\theta_n) \to 1$. If (4.3) holds with $\theta = (\theta_0 + \delta/\sqrt{n}, \theta_0)$ and the $\sigma, \sigma_1, \dots, \sigma_q$ are continuous and positive at θ_0, then

$$\lim_{n\to\infty} \mathrm{E}\varphi_\alpha(\hat\theta_n) \geqslant 2^q \sup_{t\geqslant 0} \Phi\left(\left(\frac{\delta}{\sigma(\theta_0)} - \frac{1 + w_q(\alpha,\varrho)}{1 - w_q(\alpha,\varrho)} t\right)\right) \prod_{k=1}^{q} \left(\Phi\left(\frac{\sigma(\theta_0)}{\sigma_k(\theta_0)} t\right) - 0.5\right) > 0.$$

Remarks. (i) Because $\varphi_\alpha(\hat\theta_n) = 1$ if and only if $\varphi_\alpha(a(\hat\theta_n - \theta_0 1_{q+1})) = 1$, where $a > 0$ and 1_{q+1} is a $(q+1)$-vector of ones, the $\sqrt{n}$-rate in (4.3) and in the theorem can be replaced by any other rate as long as the asymptotic normal distribution in (4.3) is still attained. Several semiparametric or nonstandard estimators are therefore covered by the theorem.

(ii) It is sometimes of interest in applications to test the null hypothesis $H_0\colon \theta_1 = \theta_0 + \gamma$ for a given γ. In that case, define $\Gamma = (\gamma 1\{k = 1\})_{1\leqslant k\leqslant q+1}$ and reject if $\varphi_\alpha(\hat\theta_n - \Gamma) = 1$. Replace θ_0 by $\theta_0 + \gamma$ in Theorem 4.4 and use part (i) of this remark to see that this leads to a consistent test. Confidence intervals for $\delta = \theta_1 - \theta_0$ can be obtained by inverting these tests for a given ϱ. By construction, all values of γ that cannot be rejected form an asymptotic $1 - \alpha$ confidence interval.

(iii) The tests in Ibragimov and Müller (2010, 2016) and Canay et al. (2017) are (or can be made) location and scale invariant and as such have the same features as those described in parts (i) and (ii) of this remark. The key difference is that their tests do not allow for a single treated cluster. □

I now discuss how the high-level condition (4.3) can be verified in an application. The specific example I use is difference-in-differences estimation but the arguments presented here apply more broadly. See also Canay et al. (2017) and Hagemann (2022) for similar types of arguments in other models.

Example 4.5 (Difference in differences, cont.). This example discusses the $\hat\theta_n$ constructed in Example 3.2 in more detail. Let n_k be the number of individuals in cluster k so that $n = 2\sum_{k=1}^{q+1} n_k$ is the total sample size. In the absence of covariates (i.e., $\beta_k \equiv 0$ in (3.1)), the centered and scaled least squares estimate in a control cluster under H_0 can be expressed as

$$\sqrt{n}(\hat\theta_{0,k} - \theta_0) = \left(\frac{n}{n_k}\right)^{1/2} n_k^{-1/2} \sum_{i=1}^{n_k} \Delta U_{i,k}.$$

The same is true for $\sqrt{n}(\hat{\theta}_1 - \theta_0)$ with $k = q+1$ on the right-hand side of the display. If the number of individuals per cluster is large in the sense that $n/n_k \to c_k \in (0, \infty)$ for $1 \leqslant k \leqslant q+1$, then condition (4.3) already holds if $n^{-1/2}(\sum_{i=1}^{n_k} U_{i,0,k}, \sum_{i=1}^{n_k} U_{i,1,k})$ is independent across $1 \leqslant k \leqslant q+1$ and has a non-degenerate normal limiting distribution for each k. The latter condition can be ensured with a central limit theorem for spatially dependent data. See, e.g., Jenish and Prucha (2009) and El Machkouri et al. (2013) for appropriate results. If the number of individuals per cluster is small, then Theorem 4.1 implies that the rearrangement test can still be applied under the assumption that $((U_{i,0,k})_{1 \leqslant i \leqslant n_k}^T, (U_{i,1,k})_{1 \leqslant i \leqslant n_k}^T)$ is multivariate normal for $1 \leqslant k \leqslant q+1$. This last condition may be strong but serves to illustrate that $\hat{\theta}_1$ and $\hat{\theta}_{0,k}$ need not even be consistent for the test to be valid.

Now consider pooled cross sections with n_k individuals in period 0, m_k individuals in period 1, and $\zeta_{i,k} \equiv \zeta_k$. The calculations in the preceding paragraph still apply with minor modifications after replacing n_k in period 1 by m_k. The analysis is no longer in first differences but the underlying conditions are essentially identical as long as $n/n_k \to c_k \in (0, \infty)$ and $n/m_k \to c_k' \in (0, \infty)$ for $1 \leqslant k \leqslant q+1$, where n is the total sample size. If the number of individuals available post intervention $m = \sum_{k=1}^{q+1} m_k$ is relatively small in the sense that $m/n_k \to 0$ and $m/m_k \to c_k' \in (0, \infty)$, the scale invariance discussed in the remarks below Theorem 4.4 allows replacement of the $\sqrt{n}$ in (4.3) by $\sqrt{m}$. Then (4.3) holds if $n_k^{-1/2} \sum_{i=1}^{n_k} U_{i,0,k} = O_P(1)$ and $m_k^{-1/2} \sum_{i=1}^{m_k} U_{i,1,k}$ obeys a central limit theorem for $1 \leqslant k \leqslant q+1$. The same argument applies with the roles of n_k and m_k reversed if relatively few individuals are available pre intervention.

The calculations in the preceding two paragraphs can be generalized to include covariates and additional time periods at the expense of more involved notation and non-singularity conditions. The same types of arguments also apply if each cluster consists of one or few units over many time periods, although the conditions for time dependence are generally less involved. See Dedecker et al. (2007) for a comprehensive overview. These remarks and the calculations in this example also apply to the models in Examples 3.1 and 3.3. □

5. Numerical results

This section explores the finite-sample behavior of the rearrangement test in two experiments. Example 5.1 compares the rearrangement test to the widely used Conley and Taber (2011) test, which is designed specifically for two-way fixed effects and applies to models with a single treated cluster. Example 5.2 applies the rearrangement test to the results of Garthwaite et al. (2014). The discussion focuses on one-sided tests to the right but the results apply more generally.

Example 5.1 (Two-way fixed effects, cont.; Conley and Taber, 2011). Following Conley and Taber (2011, sec. V), the data are generated from the two-way fixed effects model

$$Y_{t,k} = \delta I_t D_k + \beta X_{t,k} + \eta_t + \zeta_k + U_{t,k}, \tag{5.1}$$

where I_t is a post-intervention indicator, D_k is a treatment indicator, $X_{t,k}$ is a covariate, and η_t and ζ_k are time and cluster fixed effects, respectively. The covariate is constructed as $X_{t,k} = D_k/2 + Z_{t,k}$, where the $Z_{t,k}$ are iid copies of a standard normal variable. The error term satisfies

$$U_{t,k} = \gamma U_{t-1,k} + \sigma^{1\{k=q+1\}} V_{t,k}, \tag{5.2}$$

where the $V_{t,k}$ are iid standard normal and $k = q+1$ is the one cluster that received treatment. The baseline model uses $\eta_t \equiv 0 \equiv \zeta_k$, ten time periods with four post-intervention periods, and, unless stated otherwise, $\gamma = .5$, $\beta = 1$, and $\delta = 0$. I do not consider all of Conley and Taber's variations of their model but expand upon their analysis by investigating smaller numbers of control clusters q and values of σ other than one. In the latter situation, the Conley-Taber test can be expected to fail because it relies heavily on homogeneity of all clusters in absence of an intervention.

To make this last statement more precise, I now briefly review the Conley and Taber (2011) test. Let $\bar{U}_{-,k}$ and $\bar{U}_{+,k}$ be time averages of $U_{t,k}$ pre and post intervention and define $\Delta\bar{U}_k = \bar{U}_{+,k} - \bar{U}_{-,k}$. In the absence of covariates ($\beta = 0$), the fixed effects estimator $\hat{\delta}$ in (5.1) can be written as $\hat{\delta} = \delta + \Delta\bar{U}_{q+1} - \sum_{k=1}^{q} \Delta\bar{U}_k/q$, where $\sum_{k=1}^{q} \Delta\bar{U}_k/q$ is small in probability as $q \to \infty$ under regularity conditions imposed by Conley and Taber. Their main identifying assumption is that the distribution of the $U_{t,k}$ is such that $\Delta\bar{U}_{q+1}$ and $\Delta\bar{U}_k$ have identical distributions for every k. This allows them to

approximate the distribution of $\delta + \Delta\bar{U}_{q+1}$ by $\delta + \Delta\bar{U}_k$ as $q \to \infty$. Conley and Taber's conditions fail, e.g., if a $U_{t,k}$ from any control cluster $k = 1, \dots, q$ in one time period t is more or less variable than $U_{t,q+1}$, i.e., $\sigma \neq 1$ in (5.2). The problem can be remedied (as $q \to \infty$) if the exact form of the heterogeneity is known (Ferman and Pinto, 2019; Ferman, 2020) but this is not assumed here. If covariates are present, all of the tests mentioned in this paragraph still apply but the expressions for $\hat{\delta}$ become more involved. In particular, a Conley-Taber test with one treated cluster can be computed as follows:

(1) Regress the outcome on $I_t D_k$, time and cluster fixed effects, and other covariates. Denote the coefficient on $I_t D_k$ by $\hat{\delta}$.
(2) Split the residuals by cluster and run, for each of the q control clusters separately, regressions of the residuals on a constant and I_t.
(3) Compute the $1 - \alpha$ empirical quantile of the q coefficients on I_t. Reject $H_0 \colon \delta = 0$ if $\hat{\delta}$ is larger than that quantile.

In contrast to the test described above equation (3.4), this is not a proper permutation test because it does not include data from the treated cluster in the null distribution. As a result, the Conley-Taber test will tend to over-reject when q is small even if the residuals in Step (2) are iid. I illustrate this property below.

As outlined in Example 3.3, the rearrangement test computes $q+1$ separate artificial regressions of $Y_{t,k}$ on a constant, the post-intervention indicator I_t, and covariates,

$$Y_{t,k} = \begin{cases} \zeta + \theta_0 I_t + \beta X_{t,k} + \text{error}_{t,k}, & 1 \leqslant k \leqslant q, \\ \zeta + \theta_1 I_t + \beta X_{t,k} + \text{error}_{t,k}, & k = q + 1. \end{cases} \tag{5.3}$$

Because $\delta = \theta_1 - \theta_0$, I apply the rearrangement test to the least squares estimates $\hat{\theta}_{0,1}, \dots, \hat{\theta}_{0,q}$ and $\hat{\theta}_1$ of θ_0 and θ_1, respectively. I view (5.1) as coming from individual-level data aggregated to the cluster level with a fixed number of time periods. The estimates $\hat{\theta}_1, \hat{\theta}_{0,1}, \dots, \hat{\theta}_{0,q}$ should therefore be approximately normal for the rearrangement test to apply. To test deviations from this assumption in finite samples, I also consider a situation where the innovations $V_{t,k}$ in (5.2) are $\chi^2_2/2$ variables centered at zero. These innovations are asymmetric but still have unit variance.

Figure 2 shows the rejection frequencies of a true null hypothesis $H_0 \colon \delta = 0$ as a function of $\sigma \in \{1, 1.05, 1.1, \dots, 2.5\}$ for the two tests at the 5% level (short-dashed

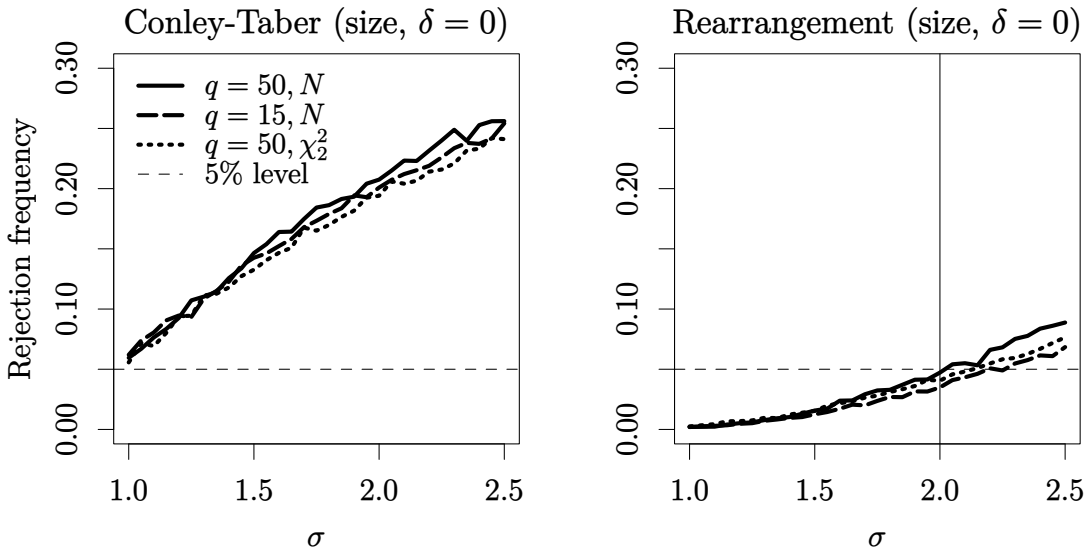

FIGURE 2. Rejection frequencies of a true null as a function of the heterogeneity σ for the Conley-Taber test (left) and the rearrangement test (right) with (i) $q = 50$ control clusters and normal errors (solid lines), (ii) $q = 15$ and normal errors (long-dashed), and (iii) $q = 50$ and chi-squared errors (dotted). The short-dashed line equals .05. The rearrangement test uses $\varrho = 2$ (vertical line).

lines). The assumptions of the Conley-Taber test (left) hold as $q \to \infty$ when $\sigma = 1$ but are violated at any sample size as soon as $\sigma > 1$. The rearrangement test (right) here uses $\varrho = 2$ (vertical line). The assumptions of the rearrangement test are violated as soon as $\sigma > 2$. The figure shows rejection rates in 10,000 Monte Carlo experiments for each horizontal coordinate with (i) $q = 50$ control clusters (solid lines), (ii) $q = 15$ (long-dashed), and (iii) $q = 50$ but the $V_{t,k}$ are iid copies of a $(\chi^2_2 - 2)/2$ variable (dotted). Both methods were faced with the same data. As can be seen, the Conley-Taber test over-rejected slightly at $\sigma = 1$ but quickly became unusable as σ increased. It exceeded a 10% rejection rate at about $\sigma = 1.25$. At $\sigma = 2.5$, the Conley-Taber test falsely discovered a nonzero effect in about 25% of all cases. In contrast, the rearrangement test was able to reject at or below the nominal level of the test as long as $\sigma \leqslant \varrho$. For $\sigma > \varrho$, the rearrangement test eventually started to over-reject. It performed worst at $\sigma = 2.5$, where it rejected in 6.8-8.8% of all cases.

The rearrangement test is designed to be robust against heterogeneity of unknown form. If σ were known, then the tests of Ferman and Pinto (2019) and Ferman (2020) could be used. Ferman and Pinto (2019) combine the idea behind the Conley-Taber

TABLE 2. Rejection frequencies of a true null for specifications (1)-(5) in Example 5.1 for (i) the rearrangement test with $\varrho = 2$ (R), (ii) a permutation test that presumes homogeneity (Perm), (iii) the Conley-Taber test (CT), (iv) the Ferman-Pinto test with correctly specified variance (FP-C), and (v) the Ferman-Pinto test with incorrectly specified variance (FP-I).

		$\sigma = 2$					$\sigma = 1$				
		R	Perm	CT	FP-C	FP-I	R	Perm	CT	FP-C	FP-I
$q = 25$	(1)	.050	.169	.232	.077	.355	.002	.043	.083	.083	.240
	(2)	.047	.166	.231	.077	.353	.002	.041	.080	.080	.239
	(3)	.047	.171	.227	.077	.360	.001	.042	.084	.084	.240
	(4)	.048	.158	.223	.081	.349	.004	.039	.080	.080	.224
	(5)	.045	.167	.228	.072	.351	.002	.041	.084	.084	.242
$q = 50$	(1)	.044	.176	.211	.054	.340	.002	.042	.062	.062	.218
	(2)	.042	.177	.210	.057	.341	.003	.040	.065	.065	.220
	(3)	.048	.175	.210	.060	.339	.002	.038	.059	.059	.217
	(4)	.042	.160	.193	.057	.331	.003	.042	.064	.064	.208
	(5)	.043	.177	.207	.057	.343	.002	.038	.059	.059	.214

test with a bootstrap but focus on the situation where heterogeneity only comes from differences in cluster sizes. Ferman (2020) considers more general situations where the heterogeneity is known up to an estimable parameter. Neither of these cases is assumed here and neither paper suggests using their test when the variance is not known or not estimable. I follow Ferman (2020, Section 3) and rescale the q coefficients from step (3) of the Conley-Taber test (as described above equation (5.3)) to have the same variance as the coefficient from the treated cluster. To compare this test to the rearrangement test, I conducted experiments in five variations of the model used for Figure 2 when $q \in \{25, 50\}$ and $\sigma \in \{1, 2\}$:

(1) Baseline model (5.1) and (5.2), $\gamma = .5$, $V_{t,k}$ standard normal.
(2) Everything as in (1) but $\gamma = .1$.
(3) Everything as in (1) but $\gamma = .9$.
(4) Everything as in (1) but $V_{t,k}$ iid $\chi^2_2/2$ centered at zero.
(5) Everything as in (1) but $X_{t,k} = D_k W_{t,k} + Z_{t,k}$, $W_{t,k}$ iid standard normal.

Table 2 shows rejection frequencies of a true null hypothesis in 10,000 Monte Carlo experiments per entry for specifications (1)-(5) with the following tests:

R: Rearrangement test with $\varrho = 2$.

Perm: Permutation test described in equation (3.4). The test assumes homogeneity under the null.

CT: Conley-Taber test.

FP-C: Ferman-Pinto test with correctly specified heterogeneity where the researcher knows σ.

FP-I: Ferman-Pinto test with incorrectly specified heterogeneity. The test incorrectly specifies .5 instead of σ.

As can seen, the Conley-Taber test again over-rejected slightly even when the clusters were homogeneous but this issue disappeared when q was large. When the clusters were heterogeneous, the Conley-Taber test over-rejected severely. The Ferman-Pinto test used here is a rescaled Conley-Taber test. It performed well when the variance was known but rejected far too many true null hypotheses when the variance was misspecified. In contrast, the rearrangement test was able to control size in all situations but was conservative when the clusters were homogeneous. The permutation test is valid under homogeneity for fixed q; it had size close to nominal level when the clusters were homogeneous. It over-rejected under heterogeneity but substantially less than then Conley-Taber test.

I now turn to the performance of the rearrangement test under the alternative. (I discuss the behavior of the Conley-Taber and Ferman-Pinto tests under the alternative towards the end of this example.) I consider the same models as before but use nonzero δ. Figure 3 shows the results with $\delta = 2$ (left) and $\delta = 3$ (right). The baseline model is again model (i) with $q = 50$ control clusters, standard normal $V_{t,k}$, and time dependence set to $\gamma = .5$ (solid lines). The other models deviate from (i) in the following ways: (ii) uses $q = 15$ (long-dashed), (iii) lowers the time dependence to $\gamma = .1$ (short-dashed grey), (iv) increases the time dependence to $\gamma = .9$ (solid grey), and (v) changes the innovations to $(\chi^2_2 - 2)/2$ (dotted). As can be seen, having to guard against near arbitrary heterogeneity of unknown form made it difficult to detect a relatively small treatment effect (left) when the number of control clusters was low, the distribution of the innovations was non-normal, or the treatment effect was obfuscated by strong time dependence. However, the rearrangement test reliably detected smaller treatment effects when the time dependence was relatively weak. Increasing the treatment effect (right) improved detection rates substantially and uniformly across

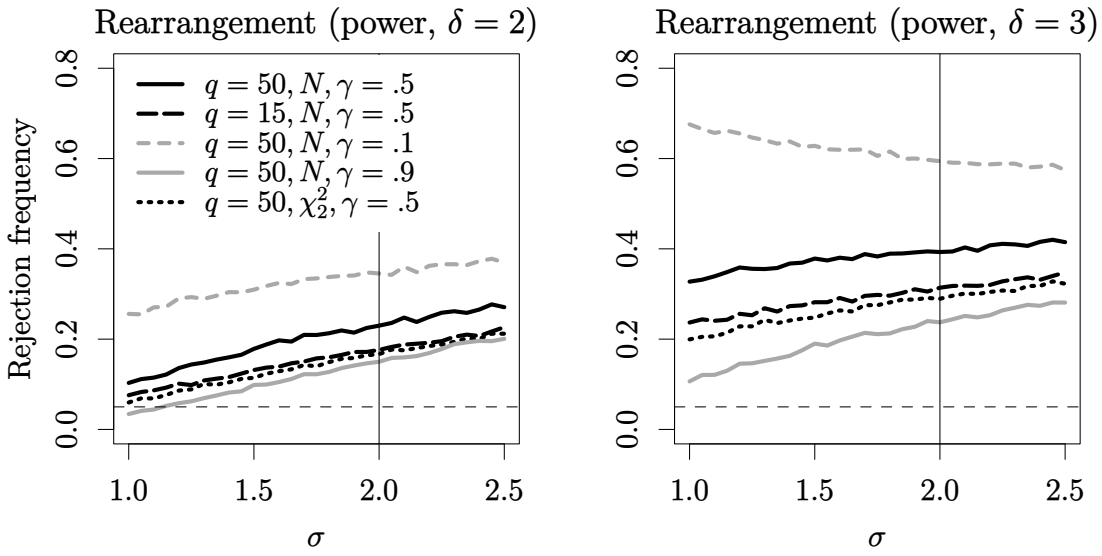

FIGURE 3. Rejection frequencies of the rearrangement test ($\varrho = 2$) under the alternative as a function of the heterogeneity σ at $\delta = 2$ (left) and $\delta = 3$ (right) with (i) and (ii) as in Figure 2, (iii) is (i) with weak time dependence $\gamma = .1$ (short-dashed grey), (iv) is (i) with strong time dependence $\gamma = .9$ (solid grey) (v) is (i) with chi-squared errors (dotted). The short-dashed line equals .05.

models, with strong time dependence again being the most challenging situation. The rearrangement test now had considerable power even when only 15 control clusters were available, the innovations were asymmetric, or the time dependence was not extreme. Power was very high when there was little time dependence.

Figures 2 and 3 also illustrate two noteworthy aspects of the rearrangement test: (1) The inequality the rearrangement is based on is nearly tight (as discussed in the paragraph below equation (4.2)) in the sense that it cannot be meaningfully improved upon unless q is very small. This can be seen in the right panel of Figure 2, where the rejection rate of the test was essentially at or slightly below nominal level when $\sigma = \varrho$. (2) Rejection rates under the null hypothesis increase with σ but this does not necessarily translate into increased rejection rates under the alternative for large σ. This is seen in the right panel of Figure 3, where the power decreases with σ in the presence of weak time dependence ($\gamma = .1$).

Finally, I investigate the trade-off between size, power, and robustness of the rearrangement test for different degrees of heterogeneity imposed on the test when the

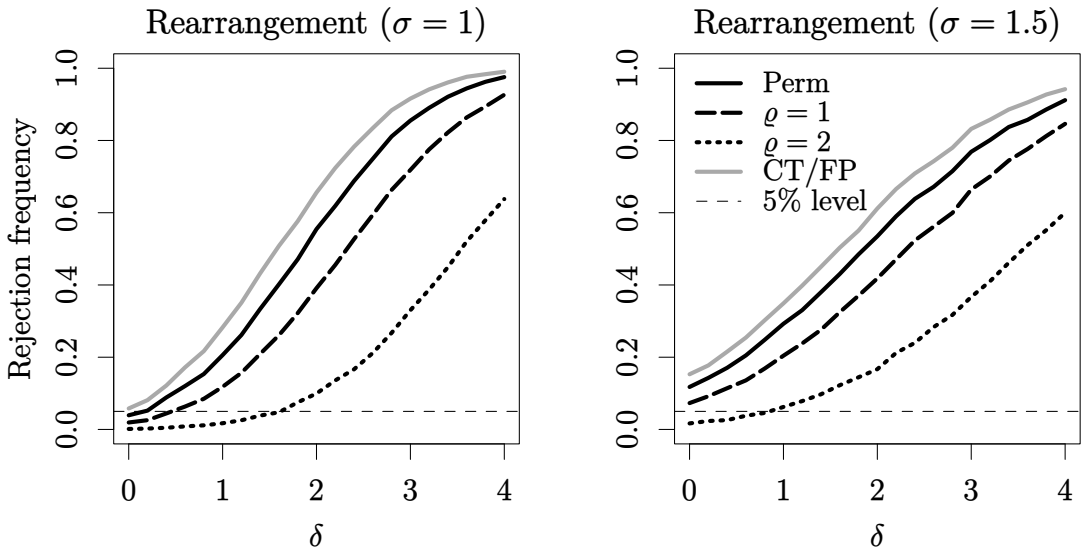

FIGURE 4. Rejection frequencies of (i) the permutation test (solid lines), the rearrangement test with (ii) $\varrho = 1$ (dashed), and (iii) $\varrho = 2$ (dotted) as a function of the treatment effect δ at $\sigma = 1$ (left) and $\sigma = 1.5$ (right). Grey lines are the Conley-Taber/Ferman-Pinto test with correctly (left) and incorrectly (right) specified heterogeneity. The null hypothesis is true at $\delta = 0$.

underlying data are homogeneous. To this end, I used the baseline model (5.1) and (5.2) with $q = 50$, $\gamma = .5$, and standard normal $V_{t,k}$. Figure 4 shows the rejection rates of the rearrangement test with $\varrho = 1$ (dashed lines) and $\varrho = 2$ (dotted) compared to the permutation test (solid black) for treatment effects $\delta \in \{0, .2, .4, \ldots, 4\}$ at $\sigma = 1$ (left) and $\sigma = 1.5$ (right). The permutation test assumes homogeneity. The null hypothesis is true at $\delta = 0$. Each δ coordinate uses 10,000 Monte Carlo repetitions. As can be seen, assuming heterogeneity when there is none is costly in terms of power but assuming homogeneity when there is heterogeneity is costly in terms of size. In particular, when there is homogeneity (left), then imposing that the treated cluster cannot be more variable than the control clusters ($\varrho = 1$) led to a mild power loss compared to correctly assuming homogeneity. Allowing the treated cluster to be much more heterogeneous ($\varrho = 2$) was more costly. When some heterogeneity was present (right), the rearrangement test with $\varrho = 1$ over-rejected slightly but the rearrangement test with $\varrho = 2$ was able to control size while remaining powerful against deviations from the null.

The grey lines in Figure 4 are the Conley-Taber/Ferman-Pinto test. I specify $\sigma = 1$ for the Ferman-Pinto test in both panels. The Conley-Taber and Ferman-Pinto tests are then identical but slightly misspecified in the right panel where the true σ equals 1.5. Because q was large, these tests did not over-reject under the null when the test was correctly specified. However, they over-rejected substantially and more than the other tests when they were misspecified. This lack of size control translated into higher rejection rates under the alternative.

I also conducted a large number of additional experiments under the null and the alternative. I considered (not shown) other distributions for $V_{t,k}$ and other values of the AR(1) coefficient γ, the number of time periods, the number of post-intervention periods, and the number of control clusters. However, I found that these changes had little impact on the results. The Conley-Taber test performed well when there was no heterogeneity but over-rejected wildly otherwise. More results on the Conley-Taber test can be found in Canay et al. (2017), who come to the same conclusion in their experiments. The Ferman-Pinto test performed well when the variance was specified correctly. The rearrangement test continued to be highly robust to heterogeneity as long as ϱ was not chosen to be much too small. Among the specifications I considered, the number of control clusters had the highest impact on the size and power of the rearrangement test, with $q \geqslant 30$ leading to the best results. □

Example 5.2 (Health insurance and labor supply; Garthwaite et al., 2014). In this example, I use the rearrangement test to reanalyze the results of Garthwaite et al. (2014). They use a difference-in-differences design to study the effects of a large-scale disruption of public heath insurance on labor supply. Their design exploits that in 2005 approximately 170,000 adults in Tennessee (roughly 4% of the state's non-elderly, adult population) abruptly lost access to TennCare, the state's public health insurance system. Garthwaite et al. use data from the 2001-2008 March Current Population Survey to determine health insurance and work status for the years 2000-2007. The comparison groups for Tennessee are the 16 other Southern states[2] defined by the U.S. Census Bureau.

[2]The Southern states are Alabama, Arkansas, Delaware, the District of Columbia, Florida, Georgia, Kentucky, Louisiana, Maryland, Mississippi, North Carolina, Oklahoma, Tennessee, Texas, Virginia, South Carolina, and West Virginia.

TABLE 3. Effects of TennCare disenrollment in Garthwaite et al. (2014, Table II.A) with their auto-correlation robust bootstrap standard errors (top) and the largest ϱ at which a rearrangement test robust to arbitrary correlation within states and over time still detects an effect (bottom).

	(1)	(2)	(3)	(4)	(5)	(6)
	Has public health insurance	Employed	Employed working <20 hours per week	Employed working ⩾20 hours per week	Employed working 20-35 hours per week	Employed working ⩾35 hours per week
$\hat{\delta}$	−0.046	0.025	−0.001	0.026	0.001	0.025
s.e.	(0.010)	(0.011)	(0.004)	(0.010)	(0.007)	(0.011)
p-val.	[0.000]	[0.019]	[0.621]	[0.011]	[0.453]	[0.020]
	Rearrangement test: largest ϱ at which $H_0\colon \delta = 0$ is rejected					
α	("×" indicates that $H_0\colon \delta = 0$ cannot be rejected for any $\varrho \geqslant 0$)					
.10	2.331	1.339	×	1.486	×	×
.05	1.707	0.986	×	1.093	×	×

The main treatment effect in Garthwaite et al. (2014, their β in their equation (1)) can be estimated as δ in

$$Y_{t,k} = \theta_0 I_t + \delta I_t D_k + \zeta_k + U_{t,k},$$

where $Y_{t,k}$ is a state-by-year mean of an outcome of interest for state k in year t, $I_t = 1\{t \geqslant 2006\}$ is a post-intervention indicator, and D_k equals one for an observation from Tennessee and equals zero otherwise. There are $17 \times 8 = 136$ state-by-year means in total. Garthwaite et al. estimate the model in the preceding display by least squares and conduct inference about δ with bootstrap standard errors that are compared to Student t critical values with 16 degrees of freedom. Their preferred bootstrap first draws states with replacement and then draws individuals within those states with replacement. This type of inference accounts for autocorrelation within individuals over time but generally requires the number of clusters to be infinite for the asymptotics. This bootstrap also does not account for potential dependence within states.

I replicate the findings of Garthwaite et al. (2014) in the top panel of Table 3. They estimate the causal effect of the TennCare disenrollment on the probability of (1) having public health insurance, (2) being employed, and (3)-(6) being employed for a certain number of hours per week. I show their bootstrap standard errors

in parentheses but report one-sided p-values in brackets instead of their two-sided p-values. In (1) the alternative is a negative effect, for (2)-(6) the alternative is positive. Garthwaite et al. find a highly significant 4.6 percentage point decrease for (1) and mostly significant positive effects for (2)-(6). They document an approximately 2.5 percentage point increase in employment and find the same effect if the outcome is restricted to individuals working more than 20 hours or more than 35 hours a week. All three effects are significant at the 5% level. The inference in Garthwaite et al. shows no significant effect for individuals working less than 20 hours or 20-35 hours.

I now apply the rearrangement test. I view each state over time as a single cluster and run 17 separate least squares regressions of the form

$$\begin{aligned} Y_{t,k} &= \theta_0 I_t + \zeta_k + U_{t,k}, \qquad 1 \leqslant k \leqslant 16, \\ Y_{t,k} &= \theta_1 I_t + \zeta_k + U_{t,k}, \qquad k = 17, \end{aligned}$$

to obtain $\hat{\theta}_{0,k}$ $(1 \leqslant k \leqslant 16)$ from each of the Southern states except Tennessee and $\hat{\theta}_1$ from Tennessee $(k = 17)$. Note that the ζ_k are now the constant terms in each regression. To perform the test, I start with $\varrho = 0$ and increase ϱ by .001 in Algorithm 3.4 as long as the null hypothesis $H_0 \colon \delta = 0$ is still rejected. The bottom panel of Table 3 shows the largest feasible value of ϱ for outcomes (1)-(6). At the 10% level, the result in (1) survives an up to $2.331^2 \approx 5.4$ times larger variance in the estimate from Tennessee relative to the second-least variable control cluster estimate. The result in (2) holds if Tennessee has a $1.339^2 \approx 1.8$ times larger variance and (4) holds even with an up to $1.486^2 \approx 2.2$ times larger variance. At the 5% level, these three results remain valid with smaller ϱ but the result in (2) only survives if the estimate from Tennessee is at most slightly less variable than the second-least variable control cluster estimate. The results in (3) and (5) confirm findings in Garthwaite et al. (2014) in that they are not significant at any level and for any value of ϱ.

A noteworthy situation occurs in (6), where the rearrangement test disagrees sharply with the significant effect found by Garthwaite et al. (2014). The rearrangement test finds no effect at any significance level and for any ϱ. In contrast, the effects in (2) and (6) are not only essentially identical but also have identical standard errors. (The p-values differ slightly because of rounding.) This also illustrates that the rearrangement test differs fundamentally from inference based on t statistics and resampling.

In sum, the rearrangement test robustly confirms—with one exception—the results of Garthwaite et al. (2014). There is statistical evidence of increased employment concentrated among individuals working at least 20 hours per week even if one accounts for arbitrary dependence within states and over time. The results hold up to substantial heterogeneity across clusters even if the number of clusters is treated as fixed for the analysis. It is also worth noting that ϱ only restricts heterogeneity in one direction. All of the results presented here are robust to arbitrary heterogeneity in any other direction and to Tennessee being infinitely more variable than the least variable control cluster. □

6. Conclusion

I introduce a generic method for inference about a scalar parameter in research designs with a finite number of large, heterogeneous clusters where only a single cluster received treatment. This situation is commonplace in difference-in-differences estimation but the test developed here applies more generally. I show that the test asymptotically controls size and has power in a setting where the number of observations within each cluster is large but the number of clusters is fixed. The test combines independent, approximately Gaussian parameter estimates from each cluster with a weighting scheme and a rearrangement procedure to obtain its critical values. The weights needed for most empirically relevant situations are tabulated in the paper. The critical values are computationally simple and do not require simulation or resampling. The test is highly robust to situations where some clusters are much more variable than others. Examples and an empirical application are provided.

Appendix A. Proofs

Proof of Theorem 4.1. Choose any $\lambda \in \Lambda$ and $w \in (0,1)$. Let $S(X,w) = S = (S_1, \ldots, S_{q+2})$. By continuity, we have $T(S) = T(S^{\triangledown})$ if and only if $S_1 + S_2 = S_{(q+2)} + S_{(q+1)}$ and $\sum_{k=1}^{q} S_{k+2} = \sum_{k=1}^{q} S_{(k)}$ almost surely. Conclude that

$$\mathrm{E}_{\lambda,0}\varphi(X,w) = P_{\lambda,0}\Big(\min\{(1+w)(X_1 - \bar{X}_0), (1-w)(X_1 - \bar{X}_0)\} > \max_k (X_{0,k} - \bar{X}_0)\Big).$$

Because of the centering, we can without loss of generality assume $\mu_0 = 0$. Define $X_{1,1} = (1+w)X_1$ and $X_{1,2} = (1-w)X_1$. Use monotonicity of maximum and minimum

to express the right-hand side of the preceding display as $P_{\lambda,0}(\min\{X_{1,1} - w\bar{X}_0, X_{1,2} + w\bar{X}_0\} > X_{0,(q)})$. Let $s^2 = \sum_{k=1}^q \sigma_k^2$ and denote by $\tilde{\varphi}(X, w)$ an infeasible version of the test function $\varphi(X, w)$ that replaces $\bar{X}_0$ by μ_0. The inequality $|1\{a > b\} - 1\{c > b\}| \leqslant 1\{|a - b| \leqslant |a - c|\}$ for $a, b, c \in \mathbb{R}$ and the triangle inequality then imply that for every $t > 0$

$$\sup_{\lambda \in \Lambda} \Big| \mathrm{E}_{\lambda,0} \varphi(X, w) 1\{|\bar{X}_0| \leqslant st\} - \mathrm{E}_{\lambda,0} \tilde{\varphi}(X, w) 1\{|\bar{X}_0| \leqslant st\} \Big|$$

cannot exceed

$$\sup_{\lambda \in \Lambda} P_{\lambda,0}\Big(|X_{1,(1)} - X_{0,(q)}| \leqslant |\min\{X_{1,1} - w\bar{X}_0, X_{1,2} + w\bar{X}_0\} - X_{1,(1)}|, |\bar{X}_0| \leqslant st \Big).$$

By monotonicity, this is at most $\sup_{\lambda \in \Lambda} P_{\lambda,0}(|X_{1,(1)} - X_{0,(q)}| \leqslant wst)$. Note that $X_{1,(1)}$ is negatively skewed and $X_{0,(q)}$ positively skewed. Because $X_{1,(1)}$ and $X_{0,(q)}$ are independent, $P_{\lambda,0}(|X_{1,(1)} - X_{0,(q)}| \leqslant wst)$ is largest when $X_{1,(1)}$ has the least skew. This happens at $\sigma = 0$ and implies

$$\sup_{\lambda \in \Lambda} P_{\lambda,0}(|X_{1,(1)} - X_{0,(q)}| \leqslant wst) = \sup_{\lambda \in \Lambda} P_{\lambda,0}(|X_{0,(q)}| \leqslant wst).$$

The probability on the right is the supremum of $\prod_{k=1}^q \Phi(wst/\sigma_k) - \prod_{k=1}^q \Phi(-wst/\sigma_k)$ over $\lambda \in \Lambda$. Because s/σ_k is decreasing in σ_k, the entire expression must be decreasing in σ_k and the supremum in the preceding display is therefore attained at $\sigma_1 = \cdots = \sigma_{q-1} = \underline{\sigma}$ and $\sigma_q = 0$. Conclude that $\sup_{\lambda \in \Lambda} P_{\lambda,0}(|X_{1,(1)} - X_{0,(q)}| \leqslant wst) \leqslant \Phi(\sqrt{q-1}wt)^{q-1}$. Because

$$\Big| \mathrm{E}_{\lambda,0} \varphi(X, w) 1\{|\bar{X}_0| > st\} - \mathrm{E}_{\lambda,0} \tilde{\varphi}(X, w) 1\{|\bar{X}_0| > st\} \Big| \leqslant P(|\bar{X}_0| > st) = 2\Phi(-qt)$$

and because all bounds so far are valid for every t, it follows that

$$\sup_{\lambda \in \Lambda} \Big| \mathrm{E}_{\lambda,0} \varphi(X, w) - \mathrm{E}_{\lambda,0} \tilde{\varphi}(X, w) \Big| \leqslant \min_{t > 0} \Big(\Phi\Big(\sqrt{q-1}wt\Big)^{q-1} + 2\Phi(-qt) \Big).$$

Now consider $\mathrm{E}_{\lambda,0} \tilde{\varphi}(X, w) = P_{\lambda,0}(X_{1,(1)} > X_{0,(q)})$, which can be expressed as

$$P\Big((1 - w)X_1 > X_{0,(q)}, X_1 > 0\Big) + P\Big((1 + w)X_1 > X_{0,(q)}, X_1 < 0\Big).$$

The second term on the right is at most $P(X_{0,(q)} < 0, X_1 < 0) = \Phi(0)^{q+1} = 2^{-q-1}$. Use independence to write the first term of the preceding display as

$$\int_0^\infty \prod_{k=1}^{q} \Phi\left(\frac{(1-w)\sigma y}{\sigma_k}\right)\phi(y)dy \leqslant \int_0^\infty \Phi\left(\frac{(1-w)\bar{\sigma} y}{\underline{\sigma}}\right)^{q-1}\phi(y)dy,$$

where the inequality follows because the integrand is increasing in σ, decreasing in σ_k, and at most one σ_k can be arbitrarily close to zero. Combine the bounds on $\mathrm{E}_{\lambda,0}\tilde{\varphi}(X,w)$ and $\mathrm{E}_{\lambda,0}\varphi(X,w) - \mathrm{E}_{\lambda,0}\tilde{\varphi}(X,w)$ to obtain the bound ξ_q.

Now consider the alternative. We still have

$$\mathrm{E}_{\lambda,\delta}\varphi(X,w) = P_{\lambda,\delta}\Big(\min\{(1+w)(X_1 - \bar{X}_0), (1-w)(X_1-\bar{X}_0)\} > \max_k (X_{0,k} - \bar{X}_0)\Big).$$

Because $1\{\min\{(1+w)(X_1-\bar{X}_0),(1-w)(X_1-\bar{X}_0)\} > \max_k(X_{0,k}-\bar{X}_0)\} \to 1$ almost surely as $\delta \to \infty$ for $w \in (0,1)$, dominated convergence implies $\mathrm{E}_{\lambda,\delta}\varphi(X,w) \to 1$. At $w = 1$, $\min\{2(X_1-\bar{X}_0),0\} - \max_k(X_{0,k}-\bar{X}_0) \to -\max_k(X_{0,k}-\bar{X}_0)$ almost surely as $\delta \to \infty$. This limit has a continuous distribution function at 0. At $w = 1$, the Slutsky lemma implies that the preceding display converges to $P(0 > \max_k(X_{0,k} - \bar{X}_0)) = P(\bar{X}_0 > \max_k X_{0,k}) = 0$, as required. □

Proof of Proposition 4.2. Let $A_t = \bigcap_{k=1}^q \{-t < X_{0,k} \leqslant t\}$ for some $t > 0$. As above, assume without loss of generality that $\mu_0 = 0$ and recall that $\mathrm{E}_{\lambda,\delta}\varphi(X,w) = P_{\lambda,\delta}(\min\{X_{1,1} - w\bar{X}_0, X_{1,2} + w\bar{X}_0\} > X_{0,(q)})$. For every fixed t, this is strictly larger than

$$P\Big(\min\{X_{1,1} - w\bar{X}_0, X_{1,2} + w\bar{X}_0\} > X_{0,(q)}, A_t\Big) \geqslant P\Big(\min\{X_{1,1}, X_{1,2}\} - wt > t, A_t\Big)$$

because $X_{0,(q)} \leqslant t$ and $|\bar{X}_0| \leqslant t$. By independence and because $t > 0$, the display can be expressed as

$$P_{\lambda,\delta}\left(X_1 > \frac{1+w}{1-w}t\right)P_\lambda(A_t) = P_{\lambda,\delta}\left(X_1 > \frac{1+w}{1-w}t\right)\prod_{k=1}^{q}\Big(\Phi(t/\sigma_k) - \Phi(-t/\sigma_k)\Big).$$

By symmetry, this simplifies to

$$\Phi\left(\left(\frac{1+w}{1-w}t - \delta\right)/\sigma\right)2^q\prod_{k=1}^{q}\Big(\Phi(t/\sigma_k) - 0.5\Big)$$

and, because t was arbitrary, it must be true that

$$\mathrm{E}_{\lambda,\delta}\varphi(X,w) \geqslant 2^q \sup_{t\geqslant 0} \Phi\left(\left(\delta - \frac{1+w}{1-w}t\right)/\sigma\right) \prod_{k=1}^{q}\big(\Phi(t/\sigma_k) - 0.5\big).$$

Replace t by $t\sigma$ to obtain the bound in the proposition.

The quantity inside the supremum is continuous on $[0,\infty]$, equals zero at $t=0$ and $t=\infty$, and is strictly positive on $t\in(0,1)$. The space $[0,\infty]$ with the order topology is compact and the supremum must therefore be attained on $t\in(0,\infty)$ to not contradict the extreme value theorem. The supremum in the preceding display is therefore a maximum over $t\in(0,\infty)$ for every fixed $\delta\in[0,\infty)$ and the maximized function is a continuous function of δ on $[0,\infty]$ by the Berge maximum theorem. As $\delta\to\infty$, the supremum is attained at $t=\infty$ and the right-hand side of the display equals one. □

Proof of Proposition 4.3. Let $S(X_n,w) = S_n = (S_{1,n},\dots,S_{q+2,n})$. We cannot have

$$\min\{S_{1,n},S_{2,n}\} < \max\{S_{3,n},\dots,S_{q+2,n}\}$$

and $T(S_n) = T(S_n^\triangledown)$ at the same time. Moreover, the reverse inequality implies $T(S_n) = T(S_n^\triangledown)$. Conclude that

$$\begin{aligned}\mathrm{E}\varphi(X_n,w) = {}& P\Big(\min\{S_{1,n},S_{2,n}\} > \max\{S_{3,n},\dots,S_{q+2,n}\}\Big)\\ &+ P\Big(T(S_n) = T(S_n^\triangledown), \min\{S_{1,n},S_{2,n}\} = \max\{S_{3,n},\dots,S_{q+2,n}\}\Big).\end{aligned}$$

By the assumed weak convergence and the continuous mapping theorem, we have $S(X_n,w)\rightsquigarrow S(X,w) = (S_1,\dots,S_{q+2})$. Use the continuous mapping theorem again to deduce

$$\min\{S_{1,n},S_{2,n}\} - \max\{S_{3,n},\dots,S_{q+2,n}\} \rightsquigarrow \min\{S_1,S_2\} - \max\{S_3,\dots,S_{q+2}\}.$$

The right-hand side can be expressed as

$$h_{X_{0,1},\dots,X_{0,q}}(X_1) \coloneqq \min\{(1+w)(X_1-\bar X_0),(1-w)(X_1-\bar X_0)\} - \max_k (X_{0,k}-\bar X_0),$$

where $x\mapsto h_{X_{0,1},\dots,X_{0,q}}(x)$ is strictly increasing and continuous for almost every realization of $X_{0,1},\dots,X_{0,q}$ and therefore has a strictly increasing and continuous inverse $h^{-1}_{X_{0,1},\dots,X_{0,q}}$ almost everywhere. Independence implies that the distribution function of

the preceding display equals $x \mapsto \mathrm{E}\Phi(h^{-1}_{X_{0,1},\ldots,X_{0,q}}(x)/\sigma)$, which is continuous by dominated convergence. Conclude that $h_{X_{0,1},\ldots,X_{0,q}}(X_1)$ must have a continuous distribution function at 0 so that

$$P(\min\{S_{1,n}, S_{2,n}\} - \max\{S_{3,n}, \ldots, S_{q+2,n}\} > 0) \to \mathrm{E}\varphi(X, w)$$

and $P(\min\{S_{1,n}, S_{2,n}\} - \max\{S_{3,n}, \ldots, S_{q+2,n}\} = 0) \to 0$. Combine these two results to obtain $\mathrm{E}\varphi(X_n, w) \to \mathrm{E}\varphi(X, w) + 0$, as desired. □

Proof of Theorem 4.4. Let $X_{1,n} = \sqrt{n}(\hat{\theta}_1 - \theta_1)$ and $X_{0,k,n} = \sqrt{n}(\hat{\theta}_{0,k} - \theta_0)$ for $1 \leqslant k \leqslant q$. By assumption, $X_n = (X_{1,n}, X_{0,1,n}, \ldots, X_{0,q,n}) \rightsquigarrow X$. Because $x \mapsto \varphi_\alpha(x)$ is invariant to multiplication of x with positive constants, we have $\varphi_\alpha(\hat{\theta}_n) = \varphi_\alpha(X_n)$ if $\theta_1 = \theta_0$. By Proposition 4.3 and Theorem 4.1, this implies $\mathrm{E}\varphi_\alpha(\hat{\theta}_n) \to \mathrm{E}\varphi_\alpha(X) \leqslant \alpha$ under the null hypothesis.

Suppose $\theta_1 = \theta_0 + \delta/\sqrt{n}$. Let $x \mapsto S_\alpha(x) = S(x, w_q(\alpha, \varrho))$ and $\Delta = (\delta 1\{k = 1\})_{1 \leqslant k \leqslant q+1}$. By the assumed continuity and the Slutsky lemma, we have $X_n + \Delta \overset{\theta}{\rightsquigarrow} X + \Delta$. Because $\sqrt{n} S_\alpha(\hat{\theta}_n) = S_\alpha(X_n + \Delta)$ and φ_α is invariant to scaling of S by positive constants, it follows from Proposition 4.3 that $\mathrm{E}\varphi_\alpha(\hat{\theta}_n) = \mathrm{E}\varphi_\alpha(X_n + \Delta) \to \mathrm{E}\varphi_\alpha(X + \Delta)$, to which the lower bound developed in Proposition 4.2 can be applied.

Now suppose $\delta = \theta_1 - \theta_0 > 0$. Let $\bar{X}_{0,n} = q^{-1} \sum_{k=1}^{q} X_{0,k,n}$. Because $X_n/\sqrt{n} \rightsquigarrow 0$, the continuous mapping theorem implies that

$$\min\{(1 + w)(X_{1,n} + \delta - \bar{X}_{0,n}), (1 - w)(X_{1,n} + \delta - \bar{X}_{0,n})\} - \max_k (X_{0,k,n} - \bar{X}_{0,n})$$

divided by $\sqrt{n}$ converges weakly to $\min\{(1 + w)\delta, (1 - w)\delta\}$. Because zero is a continuity point of the distribution of this degenerate variable unless $\delta = 0$, conclude that $\mathrm{E}\varphi_\alpha(\hat{\theta}_n) \to 1$ by the same arguments as in Propsition 4.3. □

References

Armstrong, T. and M. Kolesár (2021). Finite-sample optimal estimation and inference on average treatment effects under unconfoundedness. *Econometrica 89*, 1141–1177.

Bertrand, M., E. Duflo, and S. Mullainathan (2004). How much should we trust differences-in-differences estimates? *Quarterly Journal of Economics 119*, 249–275.

Bester, C. A., T. G. Conley, and C. B. Hansen (2011). Inference with dependent data using cluster covariance estimators. *Journal of Econometrics 165*, 137–151.

Cameron, A. C., J. B. Gelbach, and D. L. Miller (2008). Bootstrap-based improvements for inference with clustered errors. *Review of Economics and Statistics 90*, 414–427.

Cameron, L., J. Seager, and M. Shah (2020). Crimes against morality: Unintended consequences of criminalizing sex work. *The Quarterly Journal of Economics 136*, 427–469.

Canay, I., J. P. Romano, and A. M. Shaikh (2017). Randomization tests under an approximate symmetry assumption. *Econometrica 85*, 1013–1030.

Canay, I. A., A. Santos, and A. M. Shaikh (2020). The wild bootstrap with a "small" number of "large" clusters. *Review of Economics and Statistics*, forthcoming.

Cengiz, D., A. Dube, A. Lindner, and B. Zipperer (2019). The effect of minimum wages on low-wage jobs. *The Quarterly Journal of Economics 134*, 1405–1454.

Conley, T. G. and C. R. Taber (2011). Inference with "difference in differences" with a small number of policy changes. *Review of Economics and Statistics 93*, 113–125.

Cooper, Z., F. Scott Morton, and N. Shekita (2020). Surprise! out-of-network billing for emergency care in the united states. *Journal of Political Economy 128*, 3626–3677.

Cunningham, S. and M. Shah (2018). Decriminalizing indoor prostitution: Implications for sexual violence and public health. *The Review of Economic Studies 85*, 1683–1715.

Dedecker, J., P. Doukhan, G. Lang, J. R. León, S. Louhichi, and C. Prieur (2007). *Weak Dependence: With Examples and Applications*. Springer.

Deryugina, T. and D. Molitor (2020). Does when you die depend on where you live? evidence from hurricane katrina. *American Economic Review 110*, 3602–33.

Djogbenou, A. A., J. G. MacKinnon, and M. Ø. Nielsen (2019). Asymptotic theory and wild bootstrap inference with clustered errors. *Journal of Econometrics 212*, 393–412.

Dustmann, C., U. Schönberg, and J. Stuhler (2017). Labor supply shocks, native wages, and the adjustment of local employment. *The Quarterly Journal of Economics 132*, 435–483.

El Machkouri, M., D. Volný, and W. B. Wu (2013). A central limit theorem for stationary random fields. *Stochastic Processes and their Applications 123*, 1–14.

Fama, E. F. and J. D. MacBeth (1973). Risk, return, and equilibrium: Empirical tests. *Journal of Political Economy 81*, 607–636.

Ferman, B. (2020). Inference in differences-in-differences with few treated units and spatial correlation. Sao Paulo School of Economics FGV working paper, `arXiv:2006.16997`.

Ferman, B. and C. Pinto (2019). Inference in differences-in-differences with few treated groups and heteroskedasticity. *Review of Economics and Statistics 101*, 452–467.

Fisher, R. A. (1935). "The coefficient of racial likeness" and the future of craniometry. *Journal of the Royal Anthropological Institute of Great Britain and Ireland 66*, 57–63.

Garthwaite, C., T. Gross, and M. J. Notowidigdo (2014). Public health insurance, labor supply, and employment lock. *Quarterly Journal of Economics 129*, 653–696.

Giorcelli, M. and P. Moser (2020). Copyrights and creativity: Evidence from italian opera in the napoleonic age. *Journal of Political Economy 128*, 4163–4210.

Hagemann, A. (2022). Permutation inference with a finite number of heterogeneous clusters. *Review of Economics and Statistics*, forthcoming.

Hagemann, A. (2023). Inference on quantile processes with a finite number of clusters. *Journal of Econometrics*, forthcoming.

Ham, J. C. and K. Ueda (2021). The employment impact of the provision of public health insurance: A further examination of the effect of the 2005 TennCare contraction. *Journal of Labor Economics 39*, S199–S238.

Ibragimov, R. and U. Müller (2010). t-statistic based correlation and heterogeneity robust inference. *Journal of Business & Economic Statistics 28*, 453–468.

Ibragimov, R. and U. Müller (2016). Inference with few heterogenous clusters. *Review of Economics and Statistics 98*, 83–06.

Jenish, N. and I. R. Prucha (2009). Central limit theorems and uniform laws of large numbers for arrays of random fields. *Journal of Econometrics 150*, 86–98.

Johnston, A. C. and A. Mas (2018). Potential unemployment insurance duration and labor supply: The individual and market-level response to a benefit cut. *Journal of Political Economy 126*, 2480–2522.

Kaestner, R. (2016). Did Massachusetts health care reform lower mortality? No according to randomization inference. *Statistics and Public Policy 3*, 1–6.

Kaestner, R. (2021). Alive and kicking: Mortality of New Orleans Medicare enrollees after hurricane Katrina. *Econ Journal Watch 18*, 35–51.

Kitagawa, T. (2015). A test for instrument validity. *Econometrica 83*, 2043–2063.

Kolesár, M. and C. Rothe (2018). Inference in regression discontinuity designs with a discrete running variable. *American Economic Review 108*, 2277–2304.

MacKinnon, J. G. and M. D. Webb (2017). Wild bootstrap inference for wildly different cluster sizes. *Journal of Applied Econometrics 32*, 233–254.

MacKinnon, J. G. and M. D. Webb (2019). Wild bootstrap randomization inference for few treated clusters. *Advances in Econometrics 39*, 61–85.

MacKinnon, J. G. and M. D. Webb (2020). Randomization inference for difference-in-differences with few treated clusters. *Journal of Econometrics 218*, 435–450.

Mastrobuoni, G. (2020). Crime is terribly revealing: Information technology and police productivity. *The Review of Economic Studies 87*, 2727–2753.

Rubin, A. and E. Rubin (2021). Systematic bias in the progress of research. *Journal of Political Economy 129*, 2666–2719.

Słoczyński, T. (2018). A general weighted average representation of the ordinary and two-stage least squares estimands. Working paper, Department of Economics, Brandeis University.

Słoczyński, T. (2020). Interpreting OLS estimands when treatment effects are heterogeneous: Smaller groups get larger weights. *Review of Economics and Statistics*, forthcoming.

University of Michigan Ross School of Business, 701 Tappan Ave, Ann Arbor, MI 48109, USA. Tel.: +1 (734) 615-6663. Fax: +1 (734) 764-2769

Email address: hagem@umich.edu

URL: umich.edu/~hagem