A simple way to assess inference methods*

Bruno Ferman†

Sao Paulo School of Economics - FGV

First Draft: December 15th, 2019

This Draft: October 19th, 2020

Please click here for the most recent version

Abstract

We propose a simple way to assess the quality of asymptotic approximations required for inference methods. Our assessment can detect problems when the asymptotic theory that justifies the inference method is invalid and/or provides a poor approximation given the design of the empirical application. It can be easily applied to a wide range of applications. We analyze in detail the cases of differences-in-differences with few treated clusters, stratified experiments, shift-share designs, and matching estimators. If widely used by applied researchers, this assessment has the potential of substantially reducing the number of papers that are published based on misleading inference.

Keywords: inference, asymptotic theory, cluster robust variance estimator, differences-in-differences, field experiment, stratification, shift-share design, matching estimator

JEL Codes: C12; C21

*I would like to thank Arun Advani, Kirill Borusyak, Andres Santos, Marcelo Fernandes, Lucas Finamor, Raymond Fisman, Peter Hull, Toru Kitagawa, Michael Leung, Marcelo Medeiros, Duda Mendes, Marcelo Moreira, Vitor Possebom, Marcelo Sant'Anna, Yuri Saporito, Rodrigo Soares, Rodrigo Targino, and participants at the PUC-Chile seminar and at the EEA virtual conference for excellent comments and suggestions. Luis Alvarez and Lucas Barros provided exceptional research assistance. I also thank Pedro Ogeda for discussing with me an application that led me to think about this assessment for the first time.

†email: bruno.ferman@fgv.br; address: Sao Paulo School of Economics, FGV, Rua Itapeva no. 474, Sao Paulo - Brazil, 01332-000; telephone number: +55 11 3799-3350

1 Introduction

The decision among different alternatives for inference generally presents important trade-offs in terms of the assumptions in which different inference methods are valid, and the asymptotic approximations different methods rely on. As a concrete example, consider the decision about using a cluster robust variance estimator (CRVE). When we use CRVE, we allow errors to be correlated within cluster. However, inference based on CRVE is generally justified by an asymptotic theory in which the number of clusters — not the total number of observations — goes to infinity.[1] A crucial question for applied researchers then is: how many clusters are enough for reliable inference using CRVE? While there are some "rules of thumb" for deciding whether or not we have "enough" clusters, this question becomes even more subtle when we take into account that design details, such as variation in cluster sizes and the leverage of covariates, directly impact the quality of such approximations.[2] Therefore, inference based on CRVE may be unreliable even in settings where the number of clusters is usually considered as large enough, so that most researchers would not suspect there is a problem. Overall, whether we consider CRVE or any other inference method that relies on asymptotic theory, it is not always obvious whether the asymptotic approximations are reasonable in specific empirical applications.

We propose a practical and very simple way to assess the quality of asymptotic approximations required for different inference methods in a wide variety of empirical applications. The idea is to estimate the model under the null hypothesis, and generate simulations considering draws of a random variable in place of the residuals.[3] For each simulation, we estimate the parameter of interest and conduct inference in the same way as we would do with the

[1] See, for example, Arellano (1987), Carter et al. (2017), Cameron and Miller (2015), Hansen and Lee (2019), Liang and Zeger (1986), MacKinnon and Webb (2019b), and Wooldridge (2003). See also Barrios et al. (2012) and Abadie et al. (2017) for a design-based approach to clustering. In this case, the assumptions would be on the assignment mechanisms instead of on the errors, but we would still have asymptotic validity when the number of clusters goes to infinity.

[2] See, for example, MacKinnon and Webb (2017), Carter et al. (2017), Conley and Taber (2011), Chesher and Jewitt (1987), and Young (2018).

[3] Such random draws may simply be iid normal random variables. Another alternative is to sample with replacement from the distribution of the residuals.

original data. Then we calculate the proportion of times in which the null would be rejected in a large number of simulations. By construction, the null hypothesis is valid given this sampling framework. Moreover, when we increase the number of simulations, this assessment converges in probability to the size of the test, conditional on the design of the empirical application, but for the distribution of the errors assumed in the simulations. Therefore, for an α-level test, we should expect a rejection rate of approximately α in these simulations if the asymptotic theory that justifies the inference method provides a good approximation given the design of the empirical application. In contrast, we should expect significant distortions if the asymptotic theory is invalid and/or provides a poor approximation given the empirical design. This assessment is very easy to implement, and can be easily modified to accommodate different estimation strategies and alternative sampling schemes.[4]

Importantly, this assessment is uninformative about the plausibility of assumptions on the structure of the errors that different inference methods rely on. If we consider again the CRVE case, the main assumption usually considered in the literature for such inference method is that errors can be correlated within clusters, but uncorrelated across clusters.[5] Our idea is to simulate a sampling framework such that the underlying assumptions for asymptotic validity of the inference method hold. Therefore, by construction, this assessment would not inform about whether such assumptions are reasonable or not. Moreover, the assessment should not provide the exact level of the test, unless we consider the true distribution for the errors. Overall, we do not see these limitations as fundamental problems, as we see this assessment as a first screening. If this assessment uncovers a rejection rate significantly larger than the level of the test using a simple distribution for the errors, then this would be a strong indication that the inference method is not reliable for the specific empirical application, and the researcher should consider using alternative inference

[4]A simple code that implements the inference assessment in Stata can be found at `https://sites.google.com/site/brunoferman/home`.

[5]CRVE may also be asymptotically valid under alternative sets of assumptions. For example, Barrios et al. (2012) show that such procedure remains valid when there is between cluster correlations if the independent variable of interest is randomly assigned at the cluster level.

methods. However, if the assessment is close to α, then this would not provide a definite indication that the inference method is reliable. In this case, the researcher would still have to justify that other assumptions/conditions that would not be captured by this assessment are reasonable for the particular empirical application.

We present this assessment in details for the case of OLS regressions. With minor adjustments, however, our assessment is applicable to a wider range of applications. Then we present a series of examples in which our assessment can be used. We first consider the case of "differences-in-differences" (DID) with few treated clusters. As an empirical illustration, we consider the analysis of the effects of the Massachusetts 2006 health care reform on mortality rates. Sommers et al. (2014) showed that this reform let to a reduction in mortality and, as described by Kaestner (2016), received widespread media attention given the importance of this result. However, Kaestner (2016) revisited this analysis, and concluded that there was no evidence that the reform caused a reduction on mortality. Importantly, our assessment would have indicated an over-rejection on the order of 60% in the main analysis considered by Sommers et al. (2014), immediately raising a red flag that the inference method they were using was problematic. We analyze in detail how our assessment could have been used in this applications to help decide among a variety of different inference methods, and avoid some problems related to unreliable inference. We highlight in this example that there are some problems that our inference assessment would easily detect, and some other problems that it would not.

We then consider the case of stratified experiments where, for example, treatment assignment is at the school level, but data is at the student level. Chaisemartin and Ramirez-Cuellar (2019) showed that variance estimators usually considered by applied researchers may underestimate the true variance by a factor of 2 in this setting. As they report based on a survey of published papers that used paired randomized experiments, around 33% of the regressions such that the authors from these papers found a 5%-level significant effect are not significant at 5% when using a valid estimator for the standard errors. This is an extremely important finding, as they show that a relevant fraction of academic evidence based on such design is not properly controlling for the probability of false positives. We show that our assessment, if widely used among applied researchers, would have indicated that inference in these cases is misleading even before an econometrics paper was written uncovering this important problem. More generally, our assessment has the potential of preventing the accumulation of published papers based on misleading inference in a wide variety of settings, which would only later be (potentially) uncovered. This example also shows that our assessment can detect problems even in settings in which there is a very large number of clusters, so researchers would likely not suspect inference could be off because asymptotic approximations would be unreliable. Finally, our assessment can be informative if alternative inference procedures that are asymptotically valid in this case are unreliable when the asymptotic theory provides a poor approximation given the data in hand. This may happen when the number of strata is small, and is aggravated when school-level covariates are included.[6]

We also consider inference in shift-share designs. We show that our assessment can be informative about whether inference methods as the ones proposed by Adão et al. (2019) and Borusyak et al. (2018) are reliable in specific shift-share design applications. While these inference methods should always be preferred relative to alternatives such as CRVE or wild bootstrap when they are reliable, we describe an application in which our assessment would indicate that inference based on CRVE or wild bootstrap would be preferable. We also present evidence that focusing on the assessment for a specific significance level α may be misleading when assessing inference methods that impose the null hypothesis to estimate the standard errors. We show that, in this case, an assessment for a 5% level test may (apparently) control well for size, while an assessment for the same inference method, but

[6]For this setting, Carter et al. (2017) derive an effective number of clusters statistic that would also be informative about how good the asymptotic approximation is when we consider an asymptotically valid inference method. However, their assessment would not detect problems when we consider an asymptotically invalid inference method, which would be detected by our assessment. Moreover, our assessment is valid for a wider range of applications.

for a 10% level test, may detect large distortions. This happens because the apparent size control for the smaller significance level comes from counterbalancing positive and negative biases in the test size. We recommend, therefore, considering the whole distribution of p-values generated by the assessment to provide a more careful evaluation of the inference method. This result has also important implications for the presentation of Monte Carlo simulations more generally. Finally, we consider the case of matching estimators, showing that our assessment can be easily adjusted to applications that do not rely on OLS, and that it can also be informative about the power of different inference methods.

While seemingly related to a bootstrap, the idea we propose is conceptually different. Instead of trying to recover the distribution of the estimator using the bootstrap simulations, we use these simulations to assess whether an alternative inference procedure is reliable. To understand this difference, consider the case of inference based on heteroskedasticity-robust standard errors. A residual bootstrap in this setting, for example, would provide valid asymptotic inference — when the number of observations goes to infinity — under strong assumptions, including homoskedasticity. Instead of using residual-bootstrap simulations to recover the distribution of the estimator, our idea in this case would be to use these simulations to assess whether inference based on heteroskedasticity-robust standard errors, which is asymptotically valid under weaker conditions, would be reliable in a given empirical application. Our assessment is also conceptually different from the idea of using bootstrap to calculate critical values, which would generally only be valid asymptotically. Likewise, our assessment is conceptually different from the idea of Monte Carlo tests, where the goal is to provide valid finite sample inference under some assumptions on the distribution of the errors (see Dufour and Khalaf (2007) for a survey on MC tests). Following the same rationale, our assessment differs from permutation tests. In contrast to those methods, our idea is to inform about whether asymptotic approximations are reliable. If our assessment detects that a given inference method is unreliable, then the researcher should consider different alternatives for inference, even if these alternatives rely on stronger assumptions, as we exemplify in the applications presented in Section 3.

Evaluations of whether asymptotic approximations are reliable have been considered for specific inference methods. For example, Chesher and Jewitt (1987) study the bias of heteroskedasticity-robust standard errors, and recommend that users should examine measures of leverage to avoid taking an over-optimistic view of the accuracy attained in estimation. For the CRVE, Carter et al. (2017) derive a measure of effective number of clusters, which takes into account not only the number of clusters, but also other features of the design of the empirical application. In contrast to these other efforts, our assessment can be used to evaluate asymptotic approximations in a wide variety of applications, instead of being specific to particular examples. Moreover, it provides a natural metric to evaluate whether inference methods are reliable. It reflects the over-rejection one would face by using the inference method, if the errors had the distribution considered in the assessment.

Similar strategies have also been used to perform Monte Carlo simulations based on real datasets.[7] Most related to our proposal, Young (2020) uses Monte Carlo simulations to analyze the distribution of instrumental variables estimators from papers published in the journals of the American Economic Association. However, to the extent of our knowledge, the procedure we describe has never been proposed and analyzed as a general way for applied researchers to assess inference methods. We describe in details what we can and what we cannot learn from such assessment, and analyze how it can be used in specific applications to reduce the number of papers published based on misleading inference. Moreover, we show that presenting rejection rates for a specific significance level can be misleading, particularly when we consider inference methods that impose the null.

The remainder of this paper proceeds as follows. We describe in details the proposed

[7]For example, Young (2016) considers a similar strategy to evaluate the improvements of his proposed inference method on regressions from published experimental papers, Chaisemartin and Ramirez-Cuellar (2019) consider random draws of the treatment allocation in paired randomized control trials, while Adão et al. (2019) simulate random shocks in shift-share designs. There are also a number of recent papers proposing alternative ways to construct Monte Carlo simulations based on empirical applications, such as Huber et al. (2016), Busso et al. (2014), and Athey et al. (2019). See also Advani et al. (2019) for a critical analysis of empirical Monte Carlo studies for estimator selection.

assessment for the case of OLS regressions in Section 2. In Section 3, we present different applications in which our assessment can be used. We consider the cases of DID with few treated cluster (Section 3.1), field experiments (Section 3.2), shift-share designs (Section 3.3), and matching estimators (Section 3.4). Section 4 concludes.

2 A simple way to assess inference methods

We present the main ideas of our proposed assessment for the OLS estimator. However, the assessment is applicable to a wider range of applications with minor adjustments, as we show in the example in Section 3.4. Consider a simple model

$$y_i = \mathbf{x}_i\boldsymbol{\beta} + \epsilon_i, \tag{1}$$

where y_i is an outcome, $\mathbf{x}_i$ is an $1 \times K$ vector of covariates, and $\boldsymbol{\beta}$ is the parameter of interest. We observe $\{y_i, \mathbf{x}_i\}$ for a sample of $i = 1, ..., N$ observations. Let $\mathbf{y} = [y_1 \ ... \ y_N]'$, $\mathbf{X} = [x_1 \ ... \ x_N]'$, and $\boldsymbol{\epsilon} = [\epsilon_1 \ ... \ \epsilon_N]'$.

It is well known that the OLS estimator for $\boldsymbol{\beta}$ is unbiased if we assume that $\mathbb{E}[\boldsymbol{\epsilon}|\boldsymbol{X}] = 0$. Moreover, it is possible to draw finite sample inference if we impose strong assumptions on the errors, such as normality, homoskedasticity, and non-autocorrelation (e.g., Greene (2003)).[8] Relaxing those assumptions, however, generally entails difficulties for inference in finite samples. See, for example, discussions about the Behrens-Fisher problem (Behrens (1929), Fisher (1939), Scheffe (1970), Wang (1971), and Lehmann and Romano (2008)).

An often-used alternative to assuming such strong conditions on the errors is to rely on asymptotic theory. For example, heteroskedasticity-robust variance estimator (EHW from hereon), under some assumptions, is asymptotically valid when the number of observations goes to infinity, even when we relax the normality and homoskedasticity assumptions (Eicker

[8] We consider the properties of the estimator $\widehat{\boldsymbol{\beta}}$ in a repeated sampling framework over the distribution of $\boldsymbol{\epsilon}$. See Remark 5 for a discussion of our assessment if we consider a design-based approach for inference.

(1967), Huber (1967), and White (1980)). Cluster-robust standard errors allow for correlation between observations in the same cluster, and can be asymptotically valid when the number of clusters goes to infinity (Liang and Zeger (1986)). Other alternatives to allow for temporal or spatial dependence include, for example, Newey and West (1987) and Conley (1999). However, it is not always trivial to determine whether the asymptotic approximations these inference methods are based on are reliable in specific empirical applications.

We propose a simple way to assess whether the asymptotic theory that an inference method is based on is correct and/or the asymptotic approximation is reliable. The basic idea is to estimate the model under the null, and then replace the residuals with another random variable. For example, we can simply consider random draws from iid standard normal random variables. We discuss alternative options in Remark 1. Then we calculate the proportion of times such inference method would reject the null in a large number of simulations. We show in Section 3 examples in which our assessment can be easily modified for cases in which the estimator is not based on OLS, and when alternative sampling schemes are considered.

Let the null hypothesis be given by $\mathbf{R}\boldsymbol{\beta} = \mathbf{q}$, for a $J \times K$ matrix $\mathbf{R}$ and a $J \times 1$ vector $\mathbf{q}$. A step-by-step procedure to calculate our assessment is given by:

- Step 1: estimate model (1) imposing the null hypothesis. Let $\widehat{\boldsymbol{\beta}}_0 = \underset{\mathbf{b} \in \mathbb{R}^K:\ \mathbf{Rb}=\mathbf{q}}{\text{argmin}} \frac{1}{N}\sum_{i=1}^{N} (y_i - \mathbf{x}_i\mathbf{b})^2$.
- Step 2: store the predicted values from the restricted regression in Step 1, $\mathbf{X}\widehat{\boldsymbol{\beta}}_0$.
- Step 3: do $\mathcal{B}$ iterations of this step. In each step:
    - Step 3.1: draw a random vector $\boldsymbol{\epsilon}^b$ from a chosen distribution, and put it in place of the residuals from Step 1.[9] We generate $\mathbf{y}^b = \mathbf{X}\widehat{\boldsymbol{\beta}}_0 + \boldsymbol{\epsilon}^b$.
    - Step 3.2: estimate the unrestricted model with $\mathbf{y}^b$ instead of $\mathbf{y}$.

[9] For example, we can simply consider random draws from iid $N(0,1)$ random variables. Alternatively, we could resample with replacement from the estimated residuals.

- Step 3.3: test the null hypothesis using the inference method that is being assessed for a significance level of α. Store whether the null is rejected in this draw.

- Step 4: our assessment for this inference method is given by the proportion of the $\mathcal{B}$ simulations in which the null is rejected.

A simple code that implements the inference assessment can be found at `https://sites.google.com/site/brunoferman/home`. This code can be easily modified to accommodate different estimation strategies and alternative sampling schemes.

The data from the simulations in Step 3 is generated by a DGP such that the null hypothesis is valid, and that has the same empirical design (for example, number of observations, $\mathbf{X}$, sampling weights, and so on) as the real empirical application, except for the distribution of the errors. By construction, when the number of simulations $\mathcal{B}$ goes to infinity, our assessment converges in probability to the size of a test based on such inference method, conditional on the empirical design, but given the distribution of the errors considered in the simulations. Note that, for this assessment, we can consider a number of simulations as large as we want, so we can control the sampling error coming from the simulations. If in Step 3.1 we draw the errors from a distribution that satisfies the assumptions for asymptotic validity of the inference method, then we should expect a rejection rate close to α for an α-level test if the test is asymptotically valid and such asymptotic theory provides a good approximation given the empirical design. In contrast, we should expect large distortions in the assessment if the asymptotic theory is invalid and/or the asymptotic theory provides a poor approximation given the empirical design.[10]

By construction, the estimator considered in Step 3.2, say $\widehat{\boldsymbol{\beta}}^b$, is such that $\mathbf{R}\widehat{\boldsymbol{\beta}}^b - \mathbf{q} = \mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\epsilon}^b$, while the residuals of this regression are given by $\widehat{\boldsymbol{\epsilon}}^b = \left(\mathbb{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right)\boldsymbol{\epsilon}^b$. Therefore, changing the scale of the distribution of $\boldsymbol{\epsilon}^b$ will not affect the relative magnitude

[10]In Appendix A.1 we show that, under some regularity conditions, inference based on EHW standard errors is asymptotically valid conditional on $\mathbf{X}$, given such distribution considered in the assessment. Therefore, we should expect assessments close to α when assessing inference based on EHW standard errors if the asymptotic theory provides a good approximation for the empirical design.

between $\mathbf{R}\widehat{\boldsymbol{\beta}}^b - \mathbf{q}$ and $\widehat{\boldsymbol{\epsilon}}^b$, implying that, for most inference methods, the assessment will be numerically invariant to the scale of the distribution of the errors. This is true even if we do not consider a normal distribution for the errors. A non-exhaustive list in which this will be the case include, for example, inference methods based on heteroskedasticity-robust standard errors, cluster-robust standard errors, and the standard errors proposed by Adão et al. (2019) and Borusyak et al. (2018). This will also be the case for bootstrap methods.

Other variations in the distribution for the errors, however, might potentially lead to different assessments. Consider a simple example of a regression of y_i on a dummy variable x_i with iid sampling. In this case, it is well known that the OLS estimator would be given by the difference in means $\widehat{\beta} = \frac{1}{N_1}\sum_{i \in \mathcal{I}_1} y_i - \frac{1}{N_0}\sum_{i \in \mathcal{I}_0} y_i$, where N_w ($\mathcal{I}_w$) is the number (set) of observations with x_i equal to $w \in \{0,1\}$. Moreover, the variance of this estimator is given by $var(\widehat{\beta}|\mathbf{X}) = \frac{1}{N_1}\sigma_1^2 + \frac{1}{N_0}\sigma_0^2$, where $\sigma_w^2 = var(\epsilon_i | x_i = w)$, for $w \in \{0,1\}$. The EHW estimator for this variance is given by $\widehat{var(\widehat{\beta}|\mathbf{X})} = \frac{1}{N_1}\widehat{\sigma}_1^2 + \frac{1}{N_0}\widehat{\sigma}_0^2$, where $\widehat{\sigma}_w^2 = \frac{1}{N_w}\sum_{i \in \mathcal{I}_w} \hat{\epsilon}_i^2$. Therefore, a t-test based on such standard errors converges in distribution to a standard normal, providing asymptotically valid inference when both N_1 and N_0 goes to infinity.

Consider the example above in a setting with $N_1 = 5$ and $N_0 = 100$. If we consider an iid normal homoskedastic distribution for the errors, then our assessment would indicate a rejection rate of around 13% for a 5%-level test using EHW standard errors. If, however, $\sigma_0^2 = 100 \times \sigma_1^2$, then the assessment would be very close to 5%. This happens because, in this case, most of the variability of the estimator would come from observations with $x_i = 0$, and we would have a relatively large sample with $x_i = 0$ observations to estimate its distribution. Alternatively, if σ_1^2 is 100 times larger, then the assessment would indicate a rejection rate greater than 13%. This simple example shows that considering a simpler case in which errors are normally distributed and homoskedastic would *not* generally provide a lower bound to the true size of an inference method.

Such dispersion in the assessment depending on the degree of heteroskedasticity occurs because N_1 is small (even though $N_1 + N_0$ is reasonably large), so the asymptotic theory that justifies EHW standard errors does not provide a good approximation in this setting. We show in Appendix A.2 that, assuming a normal distribution, the rejection rate for an α level test converges uniformly to α when $N_1, N_0 \rightarrow \infty$, irrespectively of σ_1^2 and σ_0^2. Therefore, in this example, one could also consider alternative distributions for the error, by changing the ratio σ_1^2/σ_0^2, and report the maximum of the different assessments. Given the uniform convergence, we should still expect that this maximum over different assessments is close to α if the asymptotic theory provides a good approximation.[11] In this case, the assessment would provide a worst-case scenario for the asymptotic approximation assuming that errors are normally distributed. It would also be possible to consider the assessment relaxing the normal distribution for the errors. However, if we do not impose *any* restriction on such distributions, then we would always be able to find a distribution with heavy enough tails such that the rejection rate is much greater than α for any given (N_1, N_0), as Bahadur and Savage (1956) show for a simpler case of inference concerning a population mean. In this case, the assessment considering all possible distributions would be uninformative.

We do not see the possibility that different distributions for the errors lead to different assessments as a fundamental problem. We see this assessment as a first screening to evaluate whether an inference method is reliable. If we find large distortions when we consider simulations with, for example, simple iid standard normal errors, then this should be a strong indicative that the asymptotic theory that justifies the inference method is unreliable, and that the inference method should not be used. The fact that there could be alternative distributions for the errors in which the level of the test would be close to α should not provide a good excuse to continue relying on the inference method in this case. Likewise, an assessment close to α considering, for example, normal errors should not be taken as definite evidence that the inference is reliable. For example, in this case the inference method may not be reliable if we are considering an application in which we expect the distribution of the errors to have heavier tails.

[11] In this case, a larger number of simulations would be required to keep the simulation sampling error constant.

In simple examples, it may be possible to derive uniform convergence of rejection rates under some moment restrictions, and derive worst-case scenarios conditional on a set of possible distributions for the errors. In the comparison of means example discussed above, for example, it would be possible to derive a worst-case scenario by varying the heteroskedasticity if we restrain to normally distributed errors. However, in more complex applications, the set of distributions we should consider may not be that clear. The advantage of considering the assessment using a simple distribution for the error, such as iid standard normal, is that it is simpler for applied researchers to use the assessment. While the assessment in this case would not generally provide a worst-case scenario for a set of distributions, it would still detect cases in which one should definitely be concerned about the quality of the asymptotic approximation, as we present in Section 3.

Moreover, as we consider by construction a distribution for the errors that satisfies the assumptions of the inference method, this assessment would obviously not detect violations of the inference method related to such specific assumptions. For example, if we consider the case of clustered standard errors, the assessment would be completely uninformative about the possibility of correlation across clusters. We discuss this issue in detail for the case of inference in DID with few treated cluster in Section 3.1.

Overall, we see this assessment as a first screening to prevent evidence based on obviously misleading inference. On the one hand, an assessment much larger than α should provide strong indication that another inference method should be used. On the other hand, if the assessment is close to α, researchers should still justify and provide evidence that that other assumptions/conditions that would not be captured by the assessment are reasonable for the particular empirical application.

Remark 1 Two alternatives for the distribution of ϵ_i would be to resample with replacement from the estimated residuals, as in a residual bootstrap, or to multiply the residuals by 1 with probability 0.5 and by -1 with probability 0.5, as in a wild bootstrap. Under some conditions (including homoskedasticity), resampling from the estimated residuals would asymptotically recover the distribution of the errors. However, in finite samples, the assessment based on these two approaches may be unreliable and/or depend on the realization of the errors in the original regression. We show in Appendix A.3 two simple examples. In one example, the assessment would be close to 5% even when the inference method is highly unreliable. In the other example, the assessment would tend to be closer to 5% exactly when we would (incorrectly) reject the null. In this case, if we restrict to cases in which the assessment is close to 5%, then the assessment would prevent us from using an unreliable inference method in some cases, but we would have a *larger* over-rejection in case the assessment turns out to be close to 5%. This is something that would not happen if we consider a distribution for the errors that does not depend on the realization of the errors.

Remark 2 We also recommend that the researcher presents the assessment for different significance levels. As we show in Section 3.3, it is possible that the assessment looks good when $\alpha = 0.05$, but uncovers large over-rejection when $\alpha = 0.1$. Therefore, checking different significance levels can provide a more accurate assessment of the inference method. An alternative way to report our assessment is to calculate the p-value of the test in Step 3.3, and report the maximum over-rejection over the significance levels $\alpha \in [0, 1]$. A disadvantage of this measure is that it does not have the straightforward interpretation as opposed to, for example, the rejection rate we would expect for a test with a given significance α (if the errors followed the distribution used in the assessment).

Remark 3 In addition to using our assessment to check whether different inference methods provide correct test size, our assessment can also be used to check the power of different tests against specific alternative hypotheses. In this case, one would first estimate the model imposing the *alternative* hypothesis, and then run simulations testing the null hypothesis. An important caveat is that, unless we consider in the simulations the true distribution for the errors, our assessment will generally *not* approximate the true power of the test against this alternative. In particular, the power in this case will not be invariant to scale changes in the distribution of the errors. However, in a setting in which a researcher has more than one reliable method for inference, such assessment may be informative about which test should be used taking the power of the tests in our assessments into account.

Remark 4 In case our assessment detects a relevant over-rejection for a given inference method, it might be tempting to use the simulations to adjust the test so that it controls for size. For example, one could use the $\mathcal{B}$ simulations to determine a new critical value so that the assessment would give a 5% rejection rate. We consider this strategy with caution. By construction, this strategy would generate a test with correct size *if the distribution for the errors used in the simulations were correct.* However, we will generally not be able to say anything about the true size of the test, as we have no guarantee that the distribution of the errors chosen for the simulations is the correct one.[12] If the assessment detects significant problems for a given inference method, then we recommend considering alternative methods, even if they rely on stronger assumptions, as discussed in the examples in Section 3.

Remark 5 We rationalize the assessment in this section considering uncertainty based on a repeated sampling framework over the distribution of ϵ_i in equation (1). This differs from a design-based approach considered by Abadie et al. (2020) and Abadie et al. (2017). In their setting, they consider a finite population setting in which potential outcomes are fixed, and uncertainty comes from the assignment of $\mathbf{x}_i$ and from a random sampling from the finite population. In this case, the variance of the estimator depends crucially on the estimand of interest. Consider for simplicity the case in which $\mathbf{x}_i \in \{0, 1\}$, and let N_1 (N_0) be the number of treated (control) observations in the sample. As Abadie et al. (2020) show, whether the estimand of interested is the descriptive difference between treated and controls, the average treatment effect in the population, or the average treatment effect in the sample, the variance of the estimator is given by $S_1^2/N_1 + S_0^2/N_0$ minus a correction term, where S_w^2 is the population variance of potential outcomes when $\mathbf{x}_i = w$. As they also show, when $N_1, N_0 \to \infty$, the EHW variance estimator converges in probability to $S_1^2/N_1 + S_0^2/N_0$,

[12]Since the idea of the assessment is to check whether the inference method is reliable for a given sample size, it would generally not be possible to consistently estimate the distribution of the errors.

which would be conservative in case the correction terms are relevant. In this setting, our assessment would be informative about whether the EHW variance estimator provides a good approximation to $S_1^2/N_1 + S_0^2/N_0$. If this variance is conservative, then this could partially compensate over-rejections due to poor approximations of the EHW variance estimator. However, we do not see that as a good reason to rely on the EHW variance estimator if the asymptotic approximation is poor. Overall, we see our assessment as informative in this setting even if we consider a design-based approach for inference. Moreover, we can also consider our assessment using different sampling schemes, as we discuss in Section 3.3.

3 Applications

We consider the use of our assessment in a series of applications. First, we consider in Section 3.1 the case of DID with few treated clusters. We then consider in Section 3.2 the case of stratified randomized control trials. In Section 3.3, we consider the case of shift-share designs. Finally, we consider in Section 3.4 the case of matching estimators.

3.1 Differences-in-Differences with Few Treated Units

As a first empirical illustration of potential problems that our inference assessment would be able to detect, and also of potential problems that our assessment would fail to detect, we consider an analysis of the Massachusetts 2006 health care reform. This reform was analyzed by Sommers et al. (2014) using a DID design comparing 14 Massachusetts counties with 513 control counties that were selected based on a propensity score to be more similar with the treated counties.[13] Sommers et al. (2014) found a reduction of 2.9%-4.2% in mortality in Massachusetts relative to the controls after the reform (depending on whether covariates are included). They reported standard errors clustered at the state level, and also consid-

[13]The propensity score used age distribution, sex, race/ethnicity, poverty rate, median income, unemployment, uninsured rate, and baseline annual mortality as predictors. We take this first selection step as given in our analysis. We find similar results if we consider a DID regression using all counties, so that there is no pre-selection of control counties.

ered standard errors clustered at the county level in their online appendix. Their inference procedures were then revisited by Kaestner (2016), and then by Ferman (2020). Kaestner (2016) considered randomization inference tests at both the state and county levels. He found substantially larger p-values, ranging from 0.22 to 0.78, concluding that there is no evidence that the reform caused significant reductions in mortality. Ferman (2020) showed that the p-values from Kaestner (2016) were over-estimated due to variation in population sizes, but under-estimated due to spatial correlation (in the case of randomization inference at the county level), also concluding that the evidence is not statistically significant.

We first apply our assessment to the inference methods considered by Sommers et al. (2014). When we consider clustering at the state level, our assessment using simple iid normal errors indicates a rejection rate of 63%. This would provide an immediate conclusion that such inference procedure is not reliable, and that alternative inference methods should be considered. While nowadays this conclusion might be unsurprising for many applied researchers given that there is only one treated cluster, this was possibly not that obvious before recent papers by, for example, Conley and Taber (2011) and MacKinnon and Webb (2017). Moreover, the timing of these publications reveals a potential lag from the time in which inference problems are uncovered in econometrics papers, and the widespread knowledge of these conclusions for applied researchers, editors, and referees. This simple example highlights that our assessment may be used to easily detect problems in inference methods even before econometrics papers are written uncovering such problems, and may remain important for preventing problems even after such econometrics papers have been published.

Given the conclusion that CRVE at the state level is unreliable, researchers should consider alternative methods that do not rely on an asymptotic theory in which the number of treated states goes to infinity. Such alternatives, however, would inherently rely on stronger assumptions on the errors, as it would not be possible to allow for unrestricted within-state correlation and unrestricted heteroskedasticity with a single treated state. Importantly, our inference assessment will not generally be informative about whether such stronger assumptions on the errors are valid, because the errors used in the assessment must satisfy the assumptions in which the inference method rely on. Therefore, researchers should provide other arguments or evidence specific to their application to justify the validity of such assumptions, as we discuss below.

For example, considering cluster at a finer level (in this case, at the county level) would rely on an asymptotic theory in which the number of treated counties goes to infinity, but would not allow for state-level shocks. Our assessment would be informative about whether 14 treated counties is enough for such asymptotic approximation to be reliable. In this case, the assessment for a 5% test is around 10%, still suggesting some over-rejection, but at a much lower degree relative to CRVE at the state level. If we had more treated counties, then our assessment would get closer to 5%. However, the assessment would be mute about the possibility of state-level shocks.[14] In this case, Ferman (2020) proposed another test, which is specific for this kind of settings to detect spatial correlation, that detected that clustering at the county level would not be reliable due to spatial correlation in this application.

Another alternative could be relying on randomization inference type of procedures. A standard permutation test, as considered by Kaestner (2016), is exact when treatment is randomly assigned (e.g., Fisher (1971)). Conley and Taber (2011) propose an inference method that is similar in spirit to the idea of permutation tests, and is valid in DID settings if errors are iid across units. Therefore, if we consider an assessment based on iid errors for the method proposed by Conley and Taber (2011) (or a permutation test), then we would trivially have an assessment close to 5%. However, as discussed above, there would still be potential problems that the assessment would not capture. If we consider a permutation test at the county level, then state-level shocks would invalidate an important assumption of these inference methods. As described above, we would have to rely on alternative ways to assess whether this is a problem. In this particular application, Ferman (2020) finds evidence

[14] Note that by allowing the distribution of the errors in the simulations to have state-level shocks, we could find an assessment as close to one as we want. We would just have to increase the variance of the state-level shocks. We do not see that as informative, unless we have some information on how large state-level shocks may be relative to the idiosyncratic shocks.

that state-level shocks are relevant.

Moreover, whether we consider a permutation test at the state or county level, variation in population sizes would likely lead to heteroskedasticity in the state-time aggregate model, which would also invalidate simple permutation tests (see Ferman and Pinto (2019)). Therefore, checking whether population sizes are heterogeneous would indicate whether this is a problem. The alternative inference method proposed by Ferman and Pinto (2019) corrects for heteroskedasticity generated by variation in population sizes, but does not allow for unrestricted heteroskedasticity. This is an important restriction on the errors that would not be detected by the assessment. In a recent paper, Hagemann (2020) proposes an interesting alternative that allows for unrestricted heteroskedasticity, even when there is only a single treated cluster. However, relaxing this assumption on the errors generally comes at a cost of lower power, particularly when we expect that the treated state has a relatively lower variance. Therefore, there are important trade-offs that should be evaluated when considering these two alternatives. In this particular application, we find no evidence of statistically significant effects of the Massachusetts 2006 health reform at usual significance levels, whether we consider the methods proposed by Ferman and Pinto (2019) or Hagemann (2020).

More generally, if we have N_1 treated and N_0 control states, then there would be important trade-offs between relying on CRVE at the state level (which imposes weaker assumptions on the errors, but relies on large N_1 and N_0) and the other approaches we considered above (which impose stronger assumptions or may have lower power, but do not require large N_1). Whether N_1 is "large enough" to rely on CRVE is not something well defined, and our assessment can be used to evaluate this trade off and help applied researchers decide on which inference method to use. In this particular case, if the assessment for CRVE is close to 5%, then we would have some support to use this method, as it relies on weaker assumptions. Importantly, N_1 and N_0 will generally not be the only characteristics of the empirical application that matter for determining whether the asymptotic approximations for CRVE are reliable. Other characteristics, such as covariates (see Section 3.2), sampling weights (see Section 3.3), and others may also be relevant. Our assessment takes all of those characteristics into account.

Overall, our assessment provides a simple and widely applicable way of checking *some* problems related to inference methods. By being simple and applicable to a wide range of applications, it can be widely used by applied researchers, providing a first check on whether an inference method is reliable. However, we emphasize that there may be other potential problems that the assessment would not detect. In these cases, detecting such problems would require a deeper introspection on the assumptions the inference method relies on, and/or other assessments that would be specific to the particular application, as we described above.

3.2 Stratified randomized control trials

Consider a setting in which we have a total of N schools, and those schools are divided into S strata of G schools each, so $N = G \times S$. For each strata, exactly half of the schools receive treatment, while the other half are assigned as controls. For simplicity, consider that each school has n students. A sensible approach in this setting is to estimate the treatment effect using OLS regression of the outcome on a treatment dummy and strata fixed effects. It is well-known that one should take into account that the error term is likely correlated among students within the same schools. In this case, one could consider relying on CRVE at the school level. However, Chaisemartin and Ramirez-Cuellar (2019) show that inference based on CRVE at the school level in this case leads to significant over-rejection when G is small. They recommend clustering at the strata level to solve this problem.

We present a simple Monte Carlo study to show that our assessment can be informative in this setting. First, we show that our assessment would easily detect the problem raised by Chaisemartin and Ramirez-Cuellar (2019) for the case of small G. Moreover, we show that clustering at the strata level comes at a cost. While clustering at the strata level corrects for this finite G problem, this means a fewer number of clusters to estimate the variance.

We show that our assessment can be informative about which of the inference methods would be more reliable, if any, given the design of the empirical application. Also, in more complex designs the number of clusters would not be the only relevant variable to determine whether such asymptotic approximation should be reliable. As explored by MacKinnon and Webb (2017) and Carter et al. (2017), for example, such approximations become poorer when there are large variations in cluster sizes. See also the discussion from Conley and Taber (2011), Ferman and Pinto (2019), and MacKinnon and Webb (2019a) for cases in which there is a large number of clusters, but there are only few treated clusters. Moreover, inclusion of covariates — in particular those that vary at the school level — effectively reduces the number of degrees of freedom for the estimation of the standard errors, implying that a larger number of clusters should be necessary so that such asymptotic approximations become reliable. This is related to the discussion on leverage, considered by Chesher and Jewitt (1987). Our assessment takes all of these features into account.

We consider simulations where we vary the total number of schools $N \in \{12, 20, 40, 100, 400\}$. In all cases, we set $n = 10$. In panel A of Table 1, we consider the case in which schools are stratified in pairs. In column 1, we present our assessment if we consider for inference CRVE at the school level. We generate simulations with iid standard normal random variables to construct our assessment.[15] When there are 12 schools, the assessment would detect a rejection rate of 23% for an 5%-level test. This could reflect that the inference method is not asymptotically valid and/or the asymptotic approximation is poor given a research design with 12 schools divided in 6 strata. When we consider a setting with 400 schools, we still find a significant over-rejection, which is consistent with the theoretical result from Chaise-

[15] Since we consider a setting with school-level covariates and homogeneous cluster sizes, the OLS estimator with student-level data is numerically the same as the OLS estimator using school-level averages. Moreover, standard errors clustered at school level in the student-level regression are, up to a difference in the degrees-of-freedom correction, equivalent to heteroskedasticity-robust standard errors in the school-level regression. Therefore, given the discussion in Section 2, our measure will be invariant to changes in the distribution for the error that leads to a scale change in the school-level average of the errors, even if we consider a distribution with within-school correlation. Moreover, given the discussion in Appendix A.2, if we assume that errors are multivariate normal and independent across schools, then the assessment will converge to α uniformly with respect to the parameters of the distribution, even if we consider distributions with within-school correlation and heteroskedasticity.

martin and Ramirez-Cuellar (2019), showing that CRVE calculated in this Stata command is not asymptotically valid. Note that calculating the effective number of clusters proposed by Carter et al. (2017) would not detect a problem, since the problem in this case is related to the way the CRVE is calculated.

In column 2 of Table 1, we present our assessment when we consider inference based on CRVE at the strata level. In this case, we find over-rejection (10%) when there are 12 schools. However, when the number of strata increases, our assessment becomes close to 5%. For example, it is 6% when there are 100 schools, and 5.11% when there are 400 schools. This is consistent with the fact that such inference procedure is asymptotically valid, but that 12 schools do not provide a large enough sample so that this asymptotic approximation becomes reliable.

In settings with very few strata, Chaisemartin and Ramirez-Cuellar (2019) recommend using randomization inference. This is indeed an interesting alternative when the number of strata is very small, but we recall that randomization inference tests are generally valid for a more narrowly defined null hypothesis. Moreover, they do not directly provide standard errors. Therefore, there are some gains in considering clustered standard errors at the strata level if they are reliable, even when exact inference methods are available.

In panel B, we consider a case in which the N schools are divided in S strata of $G = 4$ schools each. As expected, the assessment presents a lower over-rejection relative to the case of paired experiments when we consider CRVE at the school level. However, we still detect over-rejection even when N is very large. When we consider inference based on CRVE at the strata level, the assessment shows that such inference method is reliable when N is very large. However, it detects a larger over-rejection for $N \leq 40$ relative to the case with paired experiments. This is consistent with the intuition that, for a given N, the number of clusters is larger in paired experiments. Therefore, a larger N is necessary so that the asymptotic approximation becomes reliable when we consider $G = 4$. Finally, in panel C we present the case in which N schools are divided into $S = 2$ strata. In this case, our assessment detects that CRVE at the strata level becomes unreliable even when N is large, which is consistent with the fact that we have only two clusters to estimate the CRVE in this case. In contrast, our assessment suggests that inference based on CRVE at the school level is reliable in this case when we have $N \geq 40$.

We also consider the case in which there are five school-level covariates in the model. For each (N, S, G) cell, we generate one single draw for such school-level covariates, and then proceed with the simulations to calculate our assessment conditional on this draw for the covariates. We present our assessments for the case with covariates in columns 3 and 4. In this case, the assessment detects that the inference methods that are asymptotically valid when $N \to \infty$ (CRVE at the strata level in Panels A and B, and at the school level in Panel C) remain reliable when N is very large. However, it also detects that a larger N is necessary so that the inference methods remain reliable relative to the case without covariates. For example, when $N = 20$ in paired experiments, our assessment indicates an over-rejection of 7.4% for the case without covariates, but 27% for the case with covariates.

The results presented in columns 3 and 4 from Table 1 are based on one single draw of the school-level covariates for each (N, S, G) cell. We consider now whether different draws of the covariates could lead to different assessments on the quality of the asymptotic approximation. For the setting $(N, S, G) = (40, 20, 2)$, we consider the assessment for 100 different draws of the covariates. We present in Figure 1 the pdf of our assessment in this case. Our assessment indicates an over-rejection ranging from 10% to 16%, depending on the specific draw of the covariates. This variation in assessments is not simply generated by the fact that we are considering a finite number (10,000) of simulations in this case. We can strongly reject the null hypothesis that the assessment is the same for all draws of covariates (p-value < 0.01). This shows that the number of schools and the number of school-level covariates are not sufficient to determine the finite-sample distortion we would have if we consider inference based on CRVE at the strata level. The particular draw of the school-level covariates will matter, as it would determine the amount of variation we still have for the treatment variable after we partial out the school-level covariates and the fixed effects. Our assessment will be informative about the specific empirical setting at hand, which includes the particular draw of the covariates. For the case of clustered standard errors, Carter et al. (2017) developed an effective number of clusters statistics. We present in Figure 2.A the scatterplot of our assessment measure and the effective number of clusters. The two measures are highly correlated (correlation coefficient of -0.75), showing that our assessment detects a more serious problem for inference exactly when the effective number of cluster is smaller. However, the effective number of clusters proposed by Carter et al. (2017) would not detect a problem with standard errors clustered at the school level, which is detected by our assessment.

When we consider 100 draws of the covariates for the $(N, S, G) = (400, 200, 2)$ scenario, then the assessment would be closer to 5%, and would be much less disperse (see Figure 1). In this case, it would range from 5% to 6%, and we cannot reject the null that the assessment is the same for all draws of the covariates (p-value $= 0.71$). Therefore, most of the variation in the assessments in this setting comes from the fact that we consider only a finite number of simulations. While there is still variation across covariates draws, the number of effective clusters is always large, which implies that the assessment is close to 5% for all draws (see Figure 2.B). This is consistent with the fact that a test based on CRVE at the strata level is asymptotically valid.

If treatment effects are heterogeneous, then Abadie et al. (2017) and Bai et al. (2019) show that t-tests may be conservative. Importantly, if we consider a distribution for the errors as we did in our simulations, then our assessment would not be able to detect this problem with the inference method. This is because we are implicitly assuming homogeneous treatment effects in our simulations. Assuming other distributions for the errors would allow us to detect that the test may be conservative. However, we would not necessarily recommend that one should try our assessment with a wide variety of distributions for the errors. We stress that our assessment should be seen as a first screening for inference methods, and that it will generally not be able to detect all potential problems that inference methods may have.

3.3 Shift-share designs

Shift-share designs are regression specifications in which one studies the impact of a set of shocks (shifters) on units differentially exposed to them, with the exposure measured by a set of weights (shares). Prominent examples include Bartik (1991), Blanchard and Katz (1992), Card (2001), and Autor et al. (2013).

Adão et al. (2019) show that inference based on heteroskedasticity-robust or cluster-robust standard errors, which were commonly used in such applications, can lead to over-rejection if units with similar shares have correlated errors, or if the treatment effects are heterogeneous. Adão et al. (2019) and Borusyak et al. (2018) propose interesting alternatives to estimate the standard errors in this settings, which allows for heterogeneous treatment effects and for units with similar shares to have correlated errors.

They show that their standard errors are asymptotically valid when the number of shocks goes to infinity, if the size of each shifter becomes asymptotically negligible. Another assumption their method relies on is that shocks are independent. This assumption can be relaxed to allow for correlated shocks within specific clusters of sectors. In this case, however, the asymptotic theory would be based on the number of clusters of sectors — not the number of sectors — going to infinity. Therefore, similar to the case of CRVE, there is a trade-off between relaxing the assumption on the correlation of shocks, and having fewer "clusters of shocks" to estimate the standard errors. Overall, it may not be trivial to determine whether such asymptotic theory — which depends not only on the number of shocks, but also on the relevance of each shock — provides a good approximation in specific empirical applications. Moreover, as of now, there is no statistic that can be used to evaluate whether such approximation is reliable. We show that our assessment can be informative in this setting.

The theory behind the inference method proposed by Adão et al. (2019) is based on resampling shocks, while holding potential outcomes as fixed. We can easily adapt our assessment to consider simulations with random draws of the shocks. In this setting, this is essentially what Adão et al. (2019) do in their simulations. Our assessment, in this case, can be interpreted as the rejection rate of a given inference method when the distribution of shocks is the one considered in the assessment. Alternatively, we can construct our assessment based on resampling errors, as presented in Section 2. We consider both alternatives in the applications below. Importantly, one should be aware about which potential problems for the inference methods the assessment would detect, and which problems it would not detect, depending on how it is constructed.

We consider three different applications of shift-share designs. The first one, from Autor et al. (2013), studies the effects of changes in sector-level Chinese import competition on labor market outcomes across U.S. Commuting Zones. This is one of the empirical applications considered by Adão et al. (2019). The second one exploits the 1990 trade liberalization in Brazil as a natural experiment, which has been used in a series of papers (e.g., Kovak (2013), Dix-Carneiro and Kovak (2017), and Dix-Carneiro et al. (2018)). Finally, we also consider an application from Acemoglu and Restrepo (2020), who estimate the effects of exposure to robots on local labor market outcomes.

We first present in Table 2 our inference assessment for CRVE, which is the inference method originally considered in these applications. When we consider our assessment based on resampling shocks for a 5%-level test, we find large over-rejection for the specifications considered in columns 1 to 6, ranging from 27% to 70%. This is the same kind of exercise considered by Adão et al. (2019). Not surprisingly, we find similar results. However, we do not take that as direct evidence that CRVE leads to such substantial over-rejection in these applications. Consider a population model $y_i = \beta x_i + \epsilon_i$, where x_i is the variable constructed based on the shift-share design. The problem highlighted by Adão et al. (2019) comes from the fact that ϵ_i may include other shocks that might be correlated among units with similar shares, and possible heterogeneous effects of the shocks. Intuitively, when we consider randomly drawn shocks to construct $\tilde{x}_i$, and hold y_i fixed, the population model would be $y_i = \gamma\tilde{x}_i + \tilde{\epsilon}_i$. Since, by construction, y_i and $\tilde{x}_i$ are independent, we have that $\gamma = 0$, which implies $\tilde{\epsilon}_i = \beta x_i + \epsilon_i$. Therefore, we have that the over-rejection in these simulations captures not only the potential spatial correlation in ϵ_i, but also the fact that, whenever $\beta \neq 0$, such simulations induce, by construction, an additional spatial correlation in the population error $\tilde{\epsilon}_i$, which is given by βx_i. We present this idea in more details in Appendix A.4.

An interesting way to assess whether spatially correlated errors pose significant distortions for CRVE is to resample shocks in simulations with placebo outcomes that could share the same correlation structure of the real outcome for the error, but that are not correlated with the shift-share variable. For example, one could consider pre-shock measures of the outcome variable y_i. This is similar in spirit to the idea of pre-testing in differences-in-differences applications (see, for example, Roth (2019) and Ferman (2019)). In this case, we would have $\beta = 0$, and the simulations with random shocks would not have an additional term βx_i in the error. Such assessment, however, would not be informative about the possibility of over-rejection due to heterogeneous treatment effects.

We present in columns 7 and 8 of Table 2 our inference assessment for CRVE for the placebo exercise considered by Acemoglu and Restrepo (2020), where they estimate the relationship between exposure to robots and labor market outcomes *before* 1990. In this case, our inference assessments become closer to 5%. This is consistent with the argument that the assessment when we consider the effects on labor market outcomes after 1990 over-estimates the relevance of spatially correlated shocks. Even for these placebo outcomes, however, the assessment indicates relevant over-rejection, particularly for the specification that uses sampling weights (column 8). While this could indicate presence of spatially correlated shocks, note that we also find similar over-rejection when our assessment is constructed based on resampling errors (see Appendix Table A.1). This suggests that the over-rejection we detect in this case comes mainly from the number of clusters not being large enough, which implies that the asymptotic theory for CRVE does not provide a good approximation.

Interestingly, even though 48 clusters would usually be considered as sufficient for CRVE to be reliable, given this design — which includes variation in cluster sizes, covariates and sampling weights for some specifications —, our assessment based on resampling errors indicates that we should still expect over-rejection (see Appendix Table A.1). It is also interesting that our assessment detects a much larger over-rejection for the specifications with sampling weights. While weighting by population size may have the advantage of improving the variance of the estimator, our results indicate that this comes at a cost of making the asymptotic approximation less reliable. We find exactly the same pattern for the application based on trade liberalization, showing that this cost may be relevant even in applications with 91 clusters. Our assessment can be used to inform about whether this potential cost is relevant in specific applications.

We consider then the use of wild bootstrap, which has been widely used in empirical settings with few clusters following the work by Cameron et al. (2008). As with CRVE, wild bootstrap allows for correlation within clusters, but not across clusters. The use of wild bootstrap has been justified both based on frameworks where the number of clusters goes to infinity (Djogbenou et al. (2019)), and where the number of clusters is fixed and the number of observations within clusters goes to infinity (Canay et al. (2018)). We use our assessment to check whether either one of these theories provides a good approximation to justify the use of wild bootstrap. Our assessment based on resampling shocks for wild bootstrap is relatively close to 5% (around 8.5%) for the specification without sampling weights (column 7 of Table 2), but still detects large over-rejection (around 19.5%) for the specification with sampling weights (column 8 of Table 2). These results indicate that the use of wild bootstrap in the application from Acemoglu and Restrepo (2020) is reliable when we consider specifications without sampling weights, unless there are substantial heterogeneous treatment effects. In this case, they would still reject the null with a p-value of 0.001.

Differently from CRVE and wild bootstrap, an important advantage of the method proposed by Adão et al. (2019) and Borusyak et al. (2018) in this setting is that it allows for presence of not only spatially correlated shocks, but also heterogeneous treatment effects. Therefore, if reliable in a given application, this method should always be preferred relative to other alternatives. However, it is not trivial to determine whether the asymptotic theory these inference methods rely on provides a good approximation. We show that our assessment can be informative in this case. We consider the method proposed by Adão et al. (2019) without the null imposed (AKM) and with the null imposed (AKM0).[16] Since the theory behind this inference method is based on resampling shocks, we focus on our assessment based on such resampling scheme. For the specifications based on Autor et al. (2013), our assessment is close to 5%, particularly when we impose the null. These results replicate the findings from Adão et al. (2019). When we consider the use of this inference method for the other two applications, however, our assessment indicates substantial problems. When the null is not imposed, we find over-rejection ranging from 21% to 63% for an 5%-level test.[17] This inference method is similar to the one proposed by Borusyak et al. (2018), and considered as a robustness by Acemoglu and Restrepo (2020). When the null is imposed, our assessment is generally greater than 11%, with the exception of the specification considered in column 4, which indicates an assessment of 3.4%. While at first glance this would suggest that AKM0 can be reliably used in the specification considered in columns 4, note that we would find a rejection rate of almost 19% if we considered a 10%-level test.

To analyze that further, we present in Figure 3.A the cdf of p-values in the simulations from the specification considered in column 4, when we use AKM0. If the asymptotic theory is valid, and the asymptotic approximation is good, then we should expect that the distribution of p-values follow an uniform $[0, 1]$ random variable. In this case, imposing

[16] Borusyak et al. (2018) also consider a version of their inference method with the null imposed.

[17] As discussed in Section 2, we could potentially consider alternative distributions for the shocks. For example, Borusyak and Hulll (2020) consider a wild bootstrap to approximate better the true DGP of the shocks in their simulations. We stress, however, that the main goal of our inference assessment is not to recover the true distribution of the test, but to assess whether the inference method is reliable. In the applications considered in columns 3 to 6, it becomes clear that such methods would not be reliable in these settings based on a simpler distribution for the shocks.

the null leads to under-rejection when we consider a 5%-level test, as presented in Table 2, but we should expect large over-rejection if we consider tests with a larger significance level. The intuition for this result is that, by imposing the null, the further away $\hat{\beta}$ (the unrestricted estimator) is from the null, the larger the sum of squared residuals when the null is imposed. Therefore, the variance of $\hat{\beta}$ will be over estimated exactly for the cases in which $\hat{\beta}$ is large, generating a downward bias on the rejection rates under the null that counterbalances other potential upward biases in the test. This effect will be particularly relevant when $\hat{\beta}$ is further away from the null, which is exactly the cases in which the test would reject at a low significance level. This is why we find under-rejection when α is lower and over-rejection when α is higher.

Since we cannot guarantee that the threshold in which this test is conservative would be the same if we considered the true distribution for the shocks, we take that as evidence that the inference method in this case is not reliable. Note that the number of clusters of sectors in this application is the same as in the application considered in column 2. As shown in Figure 3.B, the cdf of p-values is much closer to follow a uniform distribution when we consider the application from column 2. Again, this suggests that the number of clusters of sectors alone is not sufficient to determine whether the asymptotic approximations for the method proposed by Adão et al. (2019) are reliable.

Overall, these results suggest that it is not trivial to determine whether different inference methods are reliable in shift-share designs. If the methods proposed by Adão et al. (2019) and Borusyak et al. (2018) prove reliable, then they should be preferred relative to other alternatives, as they impose less restrictive assumptions on the errors and on the treatment effects. In some cases, however, CRVE or wild bootstrap may be more reliable than the methods proposed by Adão et al. (2019) and Borusyak et al. (2018), as we show for the application from Acemoglu and Restrepo (2020). Other alternative in this case would be the randomization inference type of test proposed by Borusyak and Hulll (2020). The test they propose has the advantage of being valid with few or concentrated shocks, but requires specification of the shock assignment mechanism. This may be the only alternative if the inference methods proposed by Adão et al. (2019) and Borusyak et al. (2018) are unreliable, and we do not have evidence to support that CRVE or wild bootstrap would be reasonable. Overall, here again we have to deal with trade offs in terms of asymptotic theory and assumptions on the errors when selecting among different inference methods, and our assessment can be used to inform which inference method should be used in such applications.

3.4 Matching estimators

As a final example, we consider the case of matching estimators. Abadie and Imbens (2006) derive the asymptotic distribution of the nearest-neighbor matching estimator when the number of treated and control observations goes to infinity. While they allow for settings in which the number of control observations grows at a faster rate than the number of treated observations, their asymptotic approximations may be unreliable if the number of treated observations is very small, as analyzed by Ferman (2019). In this setting, our assessment can be used to inform whether the number of treated observations is sufficiently large so that inference based on such asymptotic approximations is reliable. Since this is not an OLS estimator, it is not possible to follow the exact procedure outlined in Section 2. However, it is straightforward to adapt this procedure to this setting. For example, in this case one could simply considering iid standard normal draws for the outcome variables. Such assessment would then provide the size of the test based on the asymptotic distribution derived by Abadie and Imbens (2006), given the set of covariates used by the matching estimator, if outcomes followed the distribution considered in the simulations. Importantly, our assessment would not be informative about the finite sample bias of the matching estimator as, by construction, the estimator would be unbiased given this distribution of outcomes.

If the assessment reveals a relevant over-rejection due to a small number of treated observations, then we could consider two alternative inference methods proposed by Ferman (2019), that are asymptotically valid when the number of treated observations is fixed, and the number of control observations goes to infinity. These tests are based on the theory of randomization tests under an approximate symmetry assumption, developed by Canay et al. (2017). One test relies on permutations, while the other relies on group transformations given by sign changes. Importantly, if we consider a setting in which the number of treated observations is fixed and the number of control observations goes to infinity, these tests rely on stronger assumptions on the errors, exposing again relevant trade offs in the choice among different inference methods.[18] In the absence of finite sample bias, these tests would have a size smaller or equal to $\alpha\%$ even in finite samples. However, as Ferman (2019) shows, these tests may be too conservative if there are few group transformations, which would translate into poor power. The number of group transformations will depend on the number of treated observations, the number of nearest neighbors used in the estimation, and the number of shared nearest neighbors across treated observations. In this case, while our assessment for those tests would never be greater than $\alpha\%$, it would be informative about the extent to which these tests are conservative. Overall, our assessment can inform about the potential trade-offs between different inference procedures in a setting of matching estimators with few treated observations.

4 Concluding remarks

We propose a simple way to assess whether inference methods are reliable in specific empirical applications. Our assessment may detect whether the design of empirical applications is well approximated by the asymptotic theory that justifies specific inference methods. If widely used by applied researchers, our assessment has the potential of substantially reducing the number of papers that are published based on misleading inference. As an example, the widespread use of our assessment could have prevented the large number of significant results that failed to prove significant in field experiments if we considered randomization

[18] As Ferman (2019) shows, these tests are valid under weaker assumptions if the number of treated observations also increases.

tests, as uncovered by Young (2018). Moreover, even though permutation tests are exact in randomized experiments, they are valid to test a more narrowly defined null hypothesis, and they do not directly provide a measure for the standard error of the estimator. Therefore, even though randomization tests are available in randomized experiments, there is still a gain of assessing whether inference based on asymptotic methods is reliable in specific applications.

Our assessment can also detect cases in which the asymptotic theory is invalid. Therefore, we recommend that the use of our assessment should not be restricted to settings where the researcher believes the sample may not be large enough to justify the asymptotic approximation (e.g., when there are few clusters when we consider the use of CRVE). Rather, it should be used even when one has an arguably large sample, given that it can potentially detect problems in inference methods that remain even asymptotically. As an example, a widespread use of our procedure in paired experiments with a very large number of pairs could have detected the problem uncovered by Chaisemartin and Ramirez-Cuellar (2019) decades earlier, potentially preventing a large numbers of published papers based on misleading inference.

References

Abadie, A., Athey, S., Imbens, G. W., and Wooldridge, J. (2017). When should you adjust standard errors for clustering? Working Paper 24003, National Bureau of Economic Research.

Abadie, A., Athey, S., Imbens, G. W., and Wooldridge, J. M. (2020). Sampling-based versus design-based uncertainty in regression analysis. *Econometrica*, 88(1):265–296.

Abadie, A. and Imbens, G. W. (2006). Large sample properties of matching estimators for average treatment effects. *Econometrica*, 74(1):235–267.

Abadie, A., Imbens, G. W., and Zheng, F. (2014). Inference for misspecified models with fixed regressors. *Journal of the American Statistical Association*, 109(508):1601–1614.

Acemoglu, D. and Restrepo, P. (2020). Robots and jobs: Evidence from us labor markets. *Journal of Political Economy*, 0(ja):null.

Advani, A., Kitagawa, T., and Słoczyński, T. (2019). Mostly harmless simulations? using monte carlo studies for estimator selection. *Journal of Applied Econometrics*, 34(6):893–910.

Adão, R., Kolesár, M., and Morales, E. (2019). Shift-Share Designs: Theory and Inference*. *The Quarterly Journal of Economics*, 134(4):1949–2010.

Arellano, M. (1987). Computing robust standard errors for within-groups estimators. *Oxford Bulletin of Economics and Statistics*, 49(4):431–434.

Athey, S., Imbens, G., Metzger, J., and Munro, E. (2019). Using wasserstein generative adversarial networks for the design of monte carlo simulations.

Autor, D. H., Dorn, D., and Hanson, G. H. (2013). The china syndrome: Local labor market effects of import competition in the united states. *American Economic Review*, 103(6):2121–68.

Bahadur, R. R. and Savage, L. J. (1956). The nonexistence of certain statistical procedures in nonparametric problems. *Ann. Math. Statist.*, 27(4):1115–1122.

Bai, Y., Romano, J. P., and Shaikh, A. M. (2019). Inference in experiments with matched pairs.

Barrios, T., Diamond, R., Imbens, G. W., and Kolesar, M. (2012). Clustering, spatial correlations, and randomization inference. *Journal of the American Statistical Association*, 107(498):578–591.

Bartik, T. (1991). *Who Benefits from State and Local Economic Development Policies?* W.E. Upjohn Institute for Employment Research.

Behrens, W. U. (1929). Ein beitrag zur fehlerberechnung bei wenigen beobachtungen. *Landwirtschaftliche Jahrbucher.*, 68:807–837.

Blanchard, O. and Katz, L. (1992). Regional evolutions. *Brookings Papers on Economic Activity*, 23(1):1–76.

Borusyak, K., Hull, P., and Jaravel, X. (2018). Quasi-Experimental Shift-Share Research Designs. Papers 1806.01221, arXiv.org.

Borusyak, K. and Hulll, P. (2020). Non-Random Exposure to Natural Experiments:Theory and Applications. Technical report.

Busso, M., DiNardo, J., and McCrary, J. (2014). New Evidence on the Finite Sample Properties of Propensity Score Reweighting and Matching Estimators. *The Review of Economics and Statistics*, 96(5):885–897.

Cameron, A. C., Gelbach, J. B., and Miller, D. L. (2008). Bootstrap-based Improvements for Inference with Clustered Errors. *The Review of Economics and Statistics*, 90(3):414–427.

Cameron, A. C. and Miller, D. L. (2015). A practitioner's guide to cluster-robust inference. *Journal of Human Resources*, 50(2):317–372.

Canay, I. A., Romano, J. P., and Shaikh, A. M. (2017). Randomization tests under an approximate symmetry assumption. *Econometrica*, 85(3):1013–1030.

Canay, I. A., Santos, A., and Shaikh, A. M. (2018). The wild bootstrap with a "small" number of "large" clusters. CeMMAP working papers CWP27/18, Centre for Microdata Methods and Practice, Institute for Fiscal Studies.

Card, D. (2001). Immigrant inflows, native outflows, and the local labor market impacts of higher immigration. *Journal of Labor Economics*, 19(1):22–64.

Carter, A. V., Schnepel, K. T., and Steigerwald, D. G. (2017). Asymptotic behavior of a t-test robust to cluster heterogeneity. *The Review of Economics and Statistics*, 99(4):698–709.

Chaisemartin, C. and Ramirez-Cuellar, J. (2019). At What Level Should One Cluster Standard Errors in Paired Experiments? *arXiv e-prints*, page arXiv:1906.00288.

Chesher, A. and Jewitt, I. (1987). The bias of a heteroskedasticity consistent covariance matrix estimator. *Econometrica*, 55(5):1217–1222.

Conley, T. (1999). Gmm estimation with cross sectional dependence. *Journal of Econometrics*, 92(1):1 – 45.

Conley, T. G. and Taber, C. R. (2011). Inference with Difference in Differences with a Small Number of Policy Changes. *The Review of Economics and Statistics*, 93(1):113–125.

Dix-Carneiro, R. and Kovak, B. K. (2017). Trade liberalization and regional dynamics. *American Economic Review*, 107(10):2908–46.

Dix-Carneiro, R., Soares, R. R., and Ulyssea, G. (2018). Economic shocks and crime: Evidence from the brazilian trade liberalization. *American Economic Journal: Applied Economics*, 10(4):158–95.

Djogbenou, A. A., MacKinnon, J. G., and Nielsen, M. (2019). Asymptotic theory and wild bootstrap inference with clustered errors. *Journal of Econometrics*, 212(2):393 – 412.

Dufour, J.-M. and Khalaf, L. (2007). *Monte Carlo Test Methods in Econometrics*, chapter 23, pages 494–519. John Wiley & Sons, Ltd.

Eicker, F. (1967). Limit theorems for regressions with unequal and dependent errors. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, pages 59–82, Berkeley, Calif. University of California Press.

Ferman, B. (2019). Inference in differences-in-differences: How much should we trust in independent clusters?

Ferman, B. (2019). Matching Estimators with Few Treated and Many Control Observations. *arXiv e-prints*, page arXiv:1909.05093.

Ferman, B. (2020). Inference in differences-in-differences with few treated units and spatial correlation.

Ferman, B. and Pinto, C. (2019). Inference in differences-in-differences with few treated groups and heteroskedasticity. *The Review of Economics and Statistics*, 101(3):452–467.

Fisher, R. A. (1939). The comparison of sample with possibly unequal variances. *Annals of Eugenics*, 9(2):380–385.

Fisher, R. A. (1971). *The Design of Experiments*. Hafner Publishing Company, United States, 8^{th} edition edition.

Greene, W. H. (2003). *Econometric Analysis*. Pearson Education, fifth edition.

Hagemann, A. (2020). Inference with a single treated cluster.

Hansen, B. E. and Lee, S. (2019). Asymptotic theory for clustered samples. *Journal of Econometrics*, 210(2):268–290.

Huber, M., Lechner, M., and Mellace, G. (2016). The finite sample performance of estimators for mediation analysis under sequential conditional independence. *Journal of Business & Economic Statistics*, 34(1):139–160.

Huber, P. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, pages 221–233, Berkeley, Calif. University of California Press.

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

Kovak, B. K. (2013). Regional effects of trade reform: What is the correct measure of liberalization? *American Economic Review*, 103(5):1960–76.

Lehmann, E. and Romano, J. (2008). *Testing Statistical Hypotheses*. Springer Texts in Statistics. Springer New York.

Liang, K.-Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1):13–22.

MacKinnon, J. G. and Webb, M. D. (2017). Wild bootstrap inference for wildly different cluster sizes. *Journal of Applied Econometrics*, 32(2):233–254.

MacKinnon, J. G. and Webb, M. D. (2019a). Randomization Inference for Difference-in-Differences with Few Treated Clusters. *Journal of Econometrics, Forthcoming.*

MacKinnon, J. G. and Webb, M. D. (2019b). When and How to Deal with Clustered Errors in Regression Models. Working Paper 1421, Economics Department, Queen's University.

Newey, W. and West, K. (1987). A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, 55(3):703–08.

Roth, J. (2019). Pre-test with caution: Event-study estimates after testing for parallel trends.

Scheffe, H. (1970). Practical solutions of the behrens-fisher problem. *Journal of the American Statistical Association.*, 65:1501–1508.

Sommers, B. D., Long, S. K., and Baicker, K. (2014). Changes in mortality after massachusetts health care reform. *Annals of Internal Medicine*, 160(9):585–593. PMID: 24798521.

Wang, Y. Y. (1971). Probabilities or the type l errors of the welch tests for the behrens-fisher problem. *Journal of the American Statistical Association.*, 66:605–608.

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

Wooldridge, J. M. (2003). Cluster-sample methods in applied econometrics. *American Economic Review*, 93(2):133–138.

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

Young, A. (2018). Channeling Fisher: Randomization Tests and the Statistical Insignificance of Seemingly Significant Experimental Results*. *The Quarterly Journal of Economics*, 134(2):557–598.

Young, A. (2020). Consistency without Inference: Instrumental Variables in Practical Application.

Table 1: **Stratified field experiment - CRVE**

	Without covariates		With covariates	
	School cluster	Strata cluster	School cluster	Strata cluster
# of schools	(1)	(2)	(3)	(4)
Panel A: $G = 2$, $S = N/2$				
$N = 12$	0.231	0.102	1.000	1.000
$N = 20$	0.196	0.074	0.427	0.273
$N = 40$	0.179	0.067	0.248	0.115
$N = 100$	0.165	0.060	0.188	0.072
$N = 400$	0.154	0.051	0.164	0.054
Panel B: $G = 4$, $S = N/4$				
$N = 12$	0.129	0.192	0.359	0.483
$N = 20$	0.118	0.126	0.218	0.196
$N = 40$	0.102	0.091	0.136	0.111
$N = 100$	0.090	0.063	0.098	0.065
$N = 400$	0.083	0.050	0.084	0.052
Panel C: $G = N/2$, $S = 2$				
$N = 12$	0.109	0.305	0.368	0.469
$N = 20$	0.079	0.304	0.149	0.326
$N = 40$	0.057	0.302	0.084	0.299
$N = 100$	0.053	0.298	0.058	0.300
$N = 400$	0.050	0.298	0.052	0.298

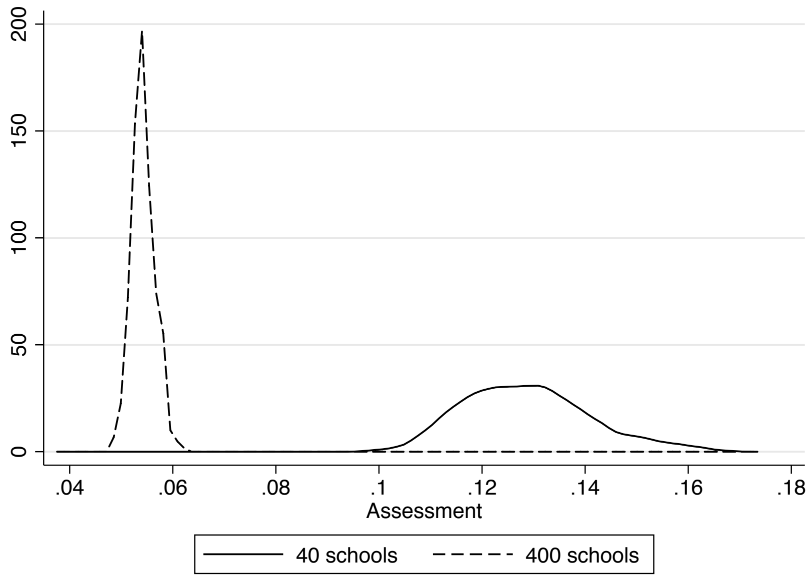
Notes: this table presents the assessment of different inference methods in a stratified field experiment. We consider a 5% test. Treatment effect is estimated by OLS regression of the outcome variable on the treatment dummy and strata fixed effects for columns 1 and 2, and on the treatment dummy, strata fixed effects, and five school-level covariates in columns 3 and 4. Each line presents the assessment of the inference method for a given set (N, S, G), where each school has ten observations. Columns 1 and 3 consider the CRVE at the school level (Stata command `areg` command with the `cluster(school)` option), while columns 2 and 4 consider the CRVE at the strata level (Stata command `xtreg` with the `fe` option). For each cell, we fixed the covariates, and generate 10,000 simulations for the outcome variable from an iid normal distribution. We present in the table the proportion of simulations such that the null would be rejected for a given inference method. Columns 3 and 4 are derived based on a single realization of the five school-level covariates.

Table 2: **Shift-share designs**

	China shock		Trade liberalization		Exposure to robots			
					Main effects		Placebos	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Estimate	-0.489	-0.489	-1.976	-2.443	-0.516	-0.448	-0.217	0.006
CRVE								
Standard error	0.076	0.076	0.822	0.723	0.118	0.059	0.151	0.070
p-value	0.000	0.000	0.016	0.001	0.000	0.000	0.152	0.930
Inference Assessment								
5% test	0.273	0.273	0.332	0.702	0.430	0.471	0.116	0.263
10% test	0.369	0.369	0.500	0.759	0.515	0.546	0.189	0.349
Wild bootstrap								
p-value	0.000	0.000	0.027	0.002	0.001	0.001	0.198	0.928
Inference Assessment								
5% test	0.238	0.238	0.279	0.665	0.376	0.336	0.085	0.194
10% test	0.327	0.327	0.457	0.732	0.455	0.424	0.144	0.283
AKM								
Standard error	0.164	0.148	0.311	0.112	0.053	0.030	0.070	0.054
p-value	0.003	0.001	0.000	0.000	0.000	0.000	0.002	0.908
Inference Assessment								
5% test	0.076	0.103	0.540	0.631	0.353	0.429	0.227	0.214
10% test	0.130	0.162	0.600	0.673	0.420	0.509	0.301	0.295
AKM0								
Standard error	0.139	0.166	0.873	1.366	0.226	0.221	0.106	0.056
p-value	0.000	0.003	0.024	0.074	0.022	0.043	0.041	0.912
Inference Assessment								
5% test	0.041	0.034	0.208	0.034	0.291	0.364	0.112	0.127
10% test	0.086	0.085	0.391	0.186	0.374	0.463	0.200	0.221
Weighted	Yes	Yes	No	Yes	No	Yes	No	Yes
# of clusters	48	48	91	91	48	48	48	48
# of observations	1444	1444	411	411	722	722	722	722
# of sectors	770	770	20	20	19	19	19	19
# of clusters of sectors	136	20	20	20	19	19	19	19

Notes: this table presents the estimates, standard errors, and p-values when we consider inference based on CRVE, wild bootstrap, AKM, and AKM0. The assessment is based on random draws of iid standard normal shocks. Then we calculate either the rejection rate for a 5%- or 10%-level test. In column 1, we present the specification presented in column 1 of Table 5 from Adão et al. (2019), which is based on the application from Autor et al. (2013). In column 2, we present the same specification as in column 1, but with clusters for 2-digit industries. In columns 3 and 4 we present specifications 1 and 2 of Table 2 from Dix-Carneiro et al. (2018). In columns 5 and 6 we present specifications 4 and 6 of Table 2 from Acemoglu and Restrepo (2020). In columns 7 and 8 we present specifications 2 and 4 of Table 4 from Acemoglu and Restrepo (2020).

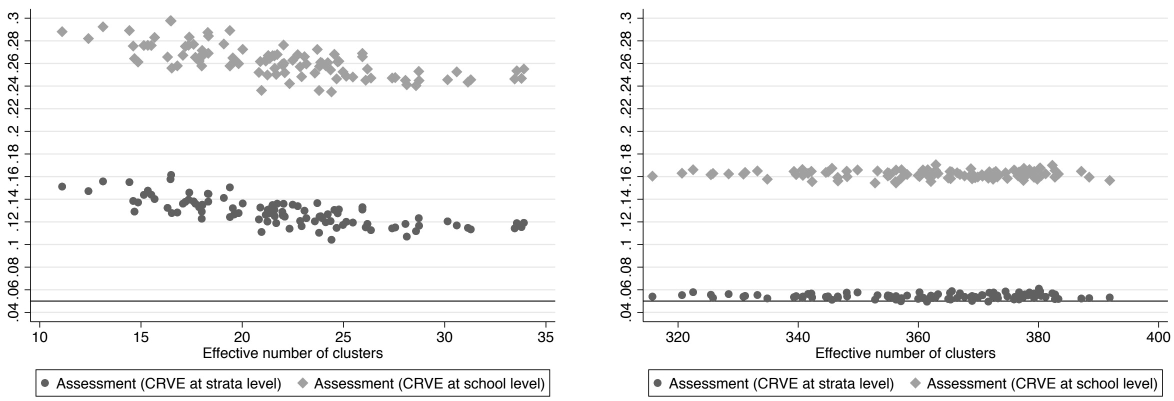
Figure 1: **Distribution of assessment**



Notes: This figure presents the pdf of the assessment for 100 different draws for the covariates. We calculate the assessment for the regression including fixed effects and covariates, with standard errors clustered at the strata level. For each of draw of the covariates, the assessment is calculated based on 10,000 simulations. We consider the scenarios $(N, S, G) = (40, 20, 2)$ and $(N, S, G) = (400, 200, 2)$.

Figure 2: **Assessment vs effective number of clusters**

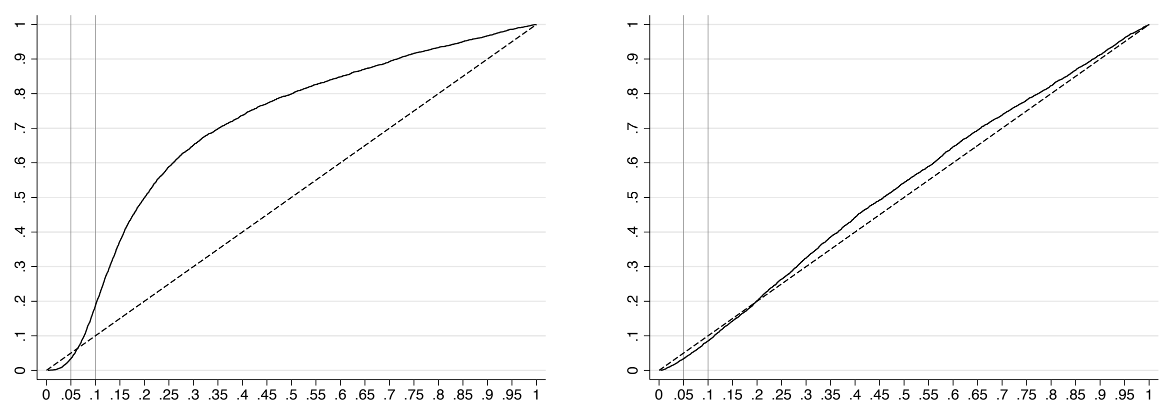
Figure 2.A: 40 schools   Figure 2.B: 400 schools



Notes: This figure presents scatterplots of our assessment and the effective number of clusters proposed by Carter et al. (2017) for 100 different draws for the covariates. We present information for standard errors clustered at the strata level and at the school level. We consider the scenarios $(N, S, G) = (40, 20, 2)$ and $(N, S, G) = (400, 200, 2)$.

Figure 3: **Assessment of AKM0 inference method - shift-share design**

Figure 3.A: Dix-Carneiro et al. (2018) (Column 4 of Table 2)   Figure 3.B: Autor et al. (2013) (Column 2 of Table 2)



Notes: This figure presents the CDFs of the p-values in the simulations using AKM0 for inference. The dashed line is the CDF of an uniform $[0, 1]$ random variable. Figures A presents the CDF for the application from Dix-Carneiro et al. (2018), presented in column 4 of Table 2. Figure B presents the CDF for the application from Autor et al. (2013), presented in column 2 of Table 2. Assessments are calculated based on resampling shocks.

A Appendix

A.1 Asymptotic validity of EHW se's conditional on X

Consider the model

$$y_i = \mathbf{x}_i \boldsymbol{\beta} + \epsilon_i, \tag{2}$$

where y_i is an outcome, $\mathbf{x}_i$ is an $1 \times K$ vector of covariates, and $\boldsymbol{\beta}$ is the parameter of interest. We consider the asymptotic validity of EHW standard errors conditional on a fixed sequence $\{\mathbf{x}_i\}_{i \in \mathbb{N}}$. We consider the following assumptions.

Assumption A.1 $\mathbb{E}[\epsilon_i | \{\mathbf{x}_i\}_{i \in \mathbb{N}}] = 0$, *var*$(\epsilon_i | \{\mathbf{x}_i\}_{i \in \mathbb{N}}) = \sigma^2(\mathbf{x}_i)$, *and cov*$(\epsilon_i, \epsilon_j | \{\mathbf{x}_i\}_{i \in \mathbb{N}}) = 0$ *for* $i \neq j$.

Assumption A.2 $\frac{1}{N} \sum_{i=1}^N \mathbf{x}_i' \mathbf{x}_i \rightarrow \mathbf{Q}$, $\frac{1}{N} \sum_{i=1}^N \sigma^2(\mathbf{x}_i) \mathbf{x}_i' \mathbf{x}_i \rightarrow \mathbf{A}$, $\limsup \frac{1}{N} \sum_{i=1}^N \|\mathbf{x}_i\|_2^4 < \infty$ *and, for some* $\delta > 0$, $\limsup \frac{1}{N} \sum_{i=1}^N \mathbb{E}\left[\|\mathbf{x}_i' \epsilon_i\|_2^{2+\delta} | \{\mathbf{x}_i\}_{i=1}^N \right] < \infty$, *where* $\mathbf{Q}$ *and* $\mathbf{A}$ *are positive definite matrices.*

Proposition A.1 *Let* $\{y_i, \mathbf{x}_i\}_{i=1}^N$ *be defined by equation (2), and consider the distribution of* $t = (\mathbf{c}'(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})) / \left(\mathbf{c}' \widehat{var}(\widehat{\boldsymbol{\beta}}) \mathbf{c} \right)^{1/2}$ *conditional on the sequence* $\{\mathbf{x}_i\}_{i \in \mathbb{N}}$, *where* $\mathbf{c} \in \mathbb{R}^K$, $\widehat{\boldsymbol{\beta}}$ *is the OLS estimator of* y_i *on* $\mathbf{x}_i$, *and* $\widehat{var}(\widehat{\boldsymbol{\beta}})$ *is the EHW variance estimator of* $\widehat{\boldsymbol{\beta}}$. *Then, under Assumptions A.1 and A.2,* $Pr\left(t < a | \{\mathbf{x}_i\}_{i \in \mathbb{N}}\right) \rightarrow \Phi(a)$ *for all* $a \in \mathbb{R}$, *where* $\Phi(a)$ *is the CDF of a standard normal.*

Proof.

Note that

$$\mathbf{c}'\widehat{\boldsymbol{\beta}} = \mathbf{c}'\boldsymbol{\beta} + \mathbf{c}'\mathbf{Q}^{-1} \left(\frac{1}{N} \sum_{i=1}^N \mathbf{x}_i' \epsilon_i \right) + \mathbf{c}' \left(\left(\frac{1}{N} \sum_{i=1}^N \mathbf{x}_i' \mathbf{x}_i \right)^{-1} - \mathbf{Q}^{-1} \right) \left(\frac{1}{N} \sum_{i=1}^N \mathbf{x}_i' \epsilon_i \right), \tag{3}$$

From Assumptions A.1 and A.2, it follows that, conditional on $\{\mathbf{x}_i\}_{i=1}^N$, $\mathbf{c}'\widehat{\boldsymbol{\beta}} \to_p \mathbf{c}'\boldsymbol{\beta}$. Now we show that, conditional on $\{\mathbf{x}_i\}_{i=1}^N$,

$$\frac{\sum_{i=1}^N \tilde{\mathbf{c}}'\mathbf{x}_i'\epsilon_i}{\left(\sum_{i=1}^N \sigma^2(\mathbf{x}_i)\tilde{\mathbf{c}}'\mathbf{x}_i'\mathbf{x}_i\tilde{\mathbf{c}}\right)^{1/2}} \to_d N(0,1), \tag{4}$$

where $\tilde{\mathbf{c}}' = \mathbf{c}'\mathbf{Q}$. Define $s_i = \tilde{\mathbf{c}}'\mathbf{x}_i'\epsilon_i / \left(\sum_{i=1}^N \sigma^2(\mathbf{x}_i)\tilde{\mathbf{c}}'\mathbf{x}_i'\mathbf{x}_i\tilde{\mathbf{c}}\right)^{1/2}$. Then we have that $\mathbb{E}[s_i|\{\mathbf{x}_i\}_{i=1}^N] = 0$, and $\sum_{i=1}^N \mathbb{E}[s_i^2|\{\mathbf{x}_i\}_{i=1}^N] = 1$. Therefore, we only need that $\sum_{i=1}^N \mathbb{E}\left[|s_i|^{2+\delta}|\{\mathbf{x}_i\}_{i=1}^N\right] \to 0$ to apply the Lyapunov CLT. Note that, from Assumptions A.1 and A.2,

$$\sum_{i=1}^N \mathbb{E}\left[|s_i|^{2+\delta}|\{\mathbf{x}_i\}_{i=1}^N\right] = \frac{1}{N^{\delta/2}} \frac{\frac{1}{N}\sum_{i=1}^N \mathbb{E}\left[|\tilde{\mathbf{c}}'\mathbf{x}_i'\epsilon_i|^{2+\delta}|\{\mathbf{x}_i\}_{i=1}^N\right]}{\left(\frac{1}{N}\sum_{i=1}^N \sigma^2(\mathbf{x}_i)\tilde{\mathbf{c}}'\mathbf{x}_i'\mathbf{x}_i\tilde{\mathbf{c}}\right)^{1+\delta/2}} \to 0, \tag{5}$$

implying that (4) holds.

Finally, we only need to show that, conditional on the sequence $\{\mathbf{x}_i\}_{i\in\mathbb{N}}$,

$$\left(\frac{1}{N}\sum_{i=1}^N \mathbf{x}_i'\mathbf{x}_i\right)^{-1} \left(\frac{1}{N}\sum_{i=1}^N (y_i - \mathbf{x}_i\widehat{\boldsymbol{\beta}})^2\mathbf{x}_i'\mathbf{x}_i\right) \left(\frac{1}{N}\sum_{i=1}^N \mathbf{x}_i'\mathbf{x}_i\right)^{-1} \to_p \mathbf{Q}^{-1}\mathbf{A}\mathbf{Q}^{-1}. \tag{6}$$

This follows from $\widehat{\boldsymbol{\beta}} \to_p \boldsymbol{\beta}$ and from Assumption A.2. Combining all results, we have that $Pr\left(t < a|\{\mathbf{x}_i\}_{i\in\mathbb{N}}\right) \to \Phi(a)$ for all $a \in \mathbb{R}$. ■

If the CEF of y_i conditional on $\mathbf{x}_i$ is not linear, then we may not have $\mathbb{E}[\epsilon_i|\{\mathbf{x}_i\}_{i\in\mathbb{N}}] = 0$ for all i. In this case, we can consider inference either relative to $\boldsymbol{\beta}$ defined as the population OLS coefficient, or relative to $\boldsymbol{\beta}(\{\mathbf{x}_i\}_{i=1}^N)$, defined based on the sample $\{\mathbf{x}_i\}_{i=1}^N$. The first parameter provides the best linear approximation to the CEF using the population distribution of $\mathbf{x}_i$ as weights, while the second one provides the best linear approximation to the CEF using the sample distribution of $\mathbf{x}_i$ as weights. See Abadie et al. (2014) for details. If we focus on the conditional parameter $\boldsymbol{\beta}(\{\mathbf{x}_i\}_{i=1}^N)$, then a test based on the test statistic t, conditional on $\{\mathbf{x}_i\}_{i=1}^N$, would be asymptotically conservative. If the focus is on $\boldsymbol{\beta}$, then conditional on

$\{\mathbf{x}_i\}_{i\in\mathbb{N}}$, the asymptotic distribution of t may not be $N(0,1)$. In this case, the asymptotic distribution would be given by normal with variance smaller than one, and with a mean that depends on $\{\mathbf{x}_i\}_{i\in\mathbb{N}}$. If we integrate over the distribution of $\{\mathbf{x}_i\}_{i\in\mathbb{N}}$, then we would recover an asymptotic distribution that is standard normal. Overall, this does not invalidate the assessment in this setting. Note that Assumption A.1 is satisfied given the distribution on the errors assumed in the assessment. Therefore, considering the distribution of the error considered in the assessment, we should expect reliable inference conditional on $\{\mathbf{x}_i\}_{i\in\mathbb{N}}$ if N is large enough. If the assessment detects large distortions in this case, then this would be an important indication that inference based on EHW is unreliable, whether Assumption A.1 is valid or not. As an alternative, it is also possible to consider an assessment in which we consider simulations of $(y_i, \mathbf{x}_i)$. In this case, the assessment would provide the size of an unconditional test with the chosen distribution considered for $(y_i, \mathbf{x}_i)$.

A.2 Uniform convergence with normal distribution

Let $y_i = x_i\beta + \epsilon_i$, where $x_i \in \{0,1\}$, $\epsilon_i | x_i = w \sim N(0,\sigma_w^2)$ for $w \in \{0,1\}$, and the sample $\{x_i,\epsilon_i\}_{i=1}^N$ is iid. Since x_i in this case is a dummy variable, we have that the CEF is linear, so Assumption A.1 in Appendix Section A.1 holds. Therefore, we know that inference conditional on $\{x_i\}_{i=1}^N$ is asymptotically valid, provided that both the number of treated and control observations diverge. We show that, in this case, under normality, the size of a test based on EHW standard errors converge to α uniformly in σ_0^2 and σ_1^2.

Let N_w ($\mathcal{I}_w$) be the number (set) of observations with x_i equal to $w \in \{0,1\}$. Under the null $\beta = 0$, the t-statistic using EHW standard errors is given by

$$t = \frac{\frac{1}{N_1}\sum_{i\in\mathcal{I}_1}\epsilon_i - \frac{1}{N_0}\sum_{i\in\mathcal{I}_0}\epsilon_i}{\sqrt{\frac{1}{N_1}\hat{\sigma}_1^2 + \frac{1}{N_0}\hat{\sigma}_0^2}} = \left(\frac{\frac{1}{N_1}\sum_{i\in\mathcal{I}_1}\epsilon_i - \frac{1}{N_0}\sum_{i\in\mathcal{I}_0}\epsilon_i}{\sqrt{\frac{1}{N_1}\sigma_1^2 + \frac{1}{N_0}\sigma_0^2}}\right)\left(\frac{\sqrt{\frac{1}{N_1}\sigma_1^2 + \frac{1}{N_0}\sigma_0^2}}{\sqrt{\frac{1}{N_1}\hat{\sigma}_1^2 + \frac{1}{N_0}\hat{\sigma}_0^2}}\right), \tag{7}$$

where $\hat{\sigma}_w^2 = \frac{1}{N_w}\sum_{i\in\mathcal{I}_w}\hat{\epsilon}_i^2$. Note that, conditional on $\mathbf{X}$, $\frac{\frac{1}{N_1}\sum_{i\in\mathcal{I}_1}\epsilon_i - \frac{1}{N_0}\sum_{i\in\mathcal{I}_0}\epsilon_i}{\sqrt{\frac{1}{N_1}\sigma_1^2+\frac{1}{N_0}\sigma_0^2}} \sim N(0,1)$.

Therefore, since $\frac{1}{N_w}\sum_{i\in\mathcal{I}_w}\epsilon_i$ and $\hat{\sigma}_w^2$ are independent,

$$P(t\leq a|\mathbf{X})=P\left(\frac{\frac{1}{N_1}\sum_{i\in\mathcal{I}_1}\epsilon_i-\frac{1}{N_0}\sum_{i\in\mathcal{I}_0}\epsilon_i}{\sqrt{\frac{1}{N_1}\sigma_1^2+\frac{1}{N_0}\sigma_0^2}}\leq a\frac{\sqrt{\frac{1}{N_1}\hat{\sigma}_1^2+\frac{1}{N_0}\hat{\sigma}_0^2}}{\sqrt{\frac{1}{N_1}\sigma_1^2+\frac{1}{N_0}\sigma_0^2}}\middle|\mathbf{X}\right)=\Phi\left(a\frac{\sqrt{\frac{1}{N_1}\hat{\sigma}_1^2+\frac{1}{N_0}\hat{\sigma}_0^2}}{\sqrt{\frac{1}{N_1}\sigma_1^2+\frac{1}{N_0}\sigma_0^2}}\right). \quad (8)$$

We show that $\frac{\sqrt{\frac{1}{N_1}\hat{\sigma}_1^2+\frac{1}{N_0}\hat{\sigma}_0^2}}{\sqrt{\frac{1}{N_1}\sigma_1^2+\frac{1}{N_0}\sigma_0^2}}\to_p 1$ uniformly in σ_1^2 and σ_0^2. Note that

$$\frac{\frac{1}{N_1}\hat{\sigma}_1^2+\frac{1}{N_0}\hat{\sigma}_0^2}{\frac{1}{N_1}\sigma_1^2+\frac{1}{N_0}\sigma_0^2}=\gamma\frac{\hat{\sigma}_1^2}{\sigma_1^2}+(1-\gamma)\frac{\hat{\sigma}_0^2}{\sigma_0^2},\text{ where }\gamma=\frac{1}{1+\frac{N_1}{N_0}\frac{\sigma_0^2}{\sigma_1^2}}. \quad (9)$$

We know that $\sum_{i\in\mathcal{I}_w}\frac{\hat{\epsilon}_i^2}{\sigma_w^2}|\mathbf{X}\sim\chi^2_{N_1-1}$. Let $\eta_1\sim\chi^2_{N_1-1}$ and $\eta_0\sim\chi^2_{N_0-1}$, where η_1 and η_0 are independent. Then, for any $e>0$,

$$\begin{aligned}
P\left(\left|\gamma\frac{\hat{\sigma}_1^2}{\sigma_1^2}+(1-\gamma)\frac{\hat{\sigma}_0^2}{\sigma_0^2}-1\right|>e\middle|\mathbf{X}\right) &= P\left(\left|\gamma\left(\frac{\eta_1}{N_1}-1\right)+(1-\gamma)\left(\frac{\eta_0}{N_0}-1\right)\right|>e\right) && (10)\\
&\leq \frac{1}{e^2}\mathbb{E}\left[\left(\gamma\left(\frac{\eta_1}{N_1}-1\right)+(1-\gamma)\left(\frac{\eta_0}{N_0}-1\right)\right)^2\right] && (11)\\
&\leq \frac{1}{e^2}\mathbb{E}\left[\left(\frac{\eta_1}{N_1}-1\right)^2\right]+\frac{1}{e^2}\mathbb{E}\left[\left(\frac{\eta_0}{N_0}-1\right)^2\right]+ && (12)\\
&\quad +2\left|\mathbb{E}\left[\left(\frac{\eta_1}{N_1}-1\right)\left(\frac{\eta_0}{N_0}-1\right)\right]\right|=o(1), && (13)
\end{aligned}$$

where the last inequality comes from $\gamma\in[0,1]$. Therefore, for any sequence $\mathbf{X}$ such that N_1 and $N_0\to\infty$, we have that $\frac{\frac{1}{N_1}\hat{\sigma}_1^2+\frac{1}{N_0}\hat{\sigma}_0^2}{\frac{1}{N_1}\sigma_1^2+\frac{1}{N_0}\sigma_0^2}\to_p 1$ uniformly in σ_1^2 and σ_0^2. Since $\Phi(.)$ and $\sqrt{.}$ are continuous functions, it follows that $P(t\leq a|\mathbf{X})\to\Phi(a)$ for any sequence $\mathbf{X}$ such that $N_1,N_0\to\infty$. Therefore, our assessment using EHW standard errors converge to α uniformly in σ_1^2 and σ_0^2.

A.3 Simple examples with alternative distributions for the errors

A.3.1 Assessment based on resampling residuals

We present a very simple example in which we construct the distribution for the errors by resampling with replacement the residuals. Suppose Y_i is iid log-normal, with the mean normalized to zero. We consider testing the null hypothesis the the $\mathbb{E}[Y_i] = 0$ with a t-test when $N = 20$. Based on simulations using this distribution, we find a rejection rate of 15% for a 5% test.

We consider the assessment resampling with replacement the residuals. We consider 5000 draws of $\{Y_i\}_{i=1}^{20}$, and for each draw we calculate the assessment based on 1000 draws from the estimated residuals. We find large variation in the assessment depending on the realization of the original sample, with the first percentile being at 5.9% and the 99 percentile at 35.9%. In 78% of the simulations, the assessment would be greater than 8%, suggesting that the inference method may be unreliable. Interestingly, however, the assessment would be less likely to indicate a rejection rate greater than 8% when the null would be rejected in the original data. When the null would not be rejected in the original data, we would have a 19% chance of having an assessment smaller than 8%, while this probability increases to 41% when then null was (incorrectly) rejected.

Therefore, suppose a researcher only considers the test if the assessment is close to 5% (say, if it is smaller than 8%). In this case, in 78% of the time the assessment would prevent the researcher from using an inference method that is unreliable. However, conditional on having an assessment close to 5%, the researcher would face a probability of rejecting the null would be 29%, which is *higher* than the unconditional size of the test. Such distortion would not happen if the distribution of the errors used in the assessment were independent from the realization of the errors. However, the assessment using iid standard normal errors in this case would not detect a large problem in this setting.

A.3.2 Assessment based on sign changes

We consider now a very simple example in which the assessment would be misleading if we construct the distribution for the errors by multiplying the residuals by +1 and -1, as in a wild bootstrap. Let $Y_i = \gamma + \beta D_i + \epsilon_i$, where $D_i = 1$ for $i = 1$, and $D_i = 0$ for $i > 1$. We want to assess whether EHW standard errors are reliable for inference. The assessment considering iid standard normal errors or resampling with replacement from the residuals would clearly indicate that EHW standard errors are unreliable in this case.

Now consider a distribution for the errors by multiplying the estimated residuals by -1 or +1 with equal probabilities. Note that $\hat{e}_1 = 0$ in this case, which implies that $\hat{\beta}^b = -\frac{1}{N-1}\sum_{i\neq 1} g_i \hat{e}_i$. In this case, if the number of controls is large enough, then the assessment would be close to 5%. We would find similar results when the number of treated observations is greater than one, if the errors of the treated observations happen to be very similar (in this case, the residuals for the treated observations would be close to zero). If we consider the same strategy, but with residuals from the restricted regression, then we could have an assessment close to 5% if $\hat{\beta} \approx 0$.

A.4 Placebo evidence in shift-share designs

We consider the placebo exercise to evaluate the performance of CRVE in shift-share design applications. Following the idea from Adão et al. (2019), we consider placebo samples in which the outcome and the shares remain fixed, and we randomly draw placebo shifters. We show that this exercise can falsely detect spatial correlation problems in the errors when the shift-share variable has an effect different from zero.

Let $Y_i = \beta X_i + \epsilon_i$, where $X_i = \sum_{f=1}^{F} w_{if}\mathcal{X}_f$, $w_{if} \geq 0$ for all f, and $\sum_{f=1}^{F} w_{if} = 1$. We consider a simplified version of the shift-share regression in order to point out that this exercise induces an over-rejection if there is a significant effect of the explanatory variable X_i in the original model. Suppose observations $i = 1, ..., N$ are partitioned into equally-sized groups $\Lambda_1, ..., \Lambda_F$, with $w_{if} = 1$ if $i \in \Lambda_f$, and $w_{if} = 0$ otherwise. Assume also that

$\mathcal{X}_f \in \{0,1\}$. This way, the model is similar to the one considered by Ferman (2019) in his Appendix A.4. We show that such placebo exercise would lead to over-rejection if $\beta \neq 0$, even if $\{\epsilon_i\}_{i=1}^N$ were drawn from a distribution in which the errors are independent. We assume for simplicity that $\sum_{f=1}^F \mathcal{X}_f = F/2$, and consider random draws of $\widetilde{\mathcal{X}}_f$ such that $\sum_{f=1}^F \widetilde{\mathcal{X}}_f = F/2$,

Let $\hat{\delta}$ be the estimator of the placebo regression. Therefore, we have from Lemma 5 from Barrios et al. (2012) that $\mathbb{E}\left[\hat{\delta}|\{Y_i\}_{i=1}^N\right] = 0$, and

$$\mathbb{V}_{\text{true}} \equiv var\left(\hat{\delta}|\{Y_i\}_{i=1}^N\right) = \frac{4}{F(F-2)}\sum_{f=1}^{F}\left(\beta\mathbb{1}\{f\in\mathcal{T}\} - \beta/2 + \bar{\epsilon}_f - \bar{\epsilon}\right)^2, \tag{14}$$

where $\mathcal{T}$ is the set of sectors such that $\mathcal{X}_f = 1$ (in the original data), $\bar{\epsilon}_f$ is the average of ϵ_i for $i \in \Lambda_f$, and $\bar{\epsilon}$ is the average across all i.

Likewise, if we consider CRVE at the observation level (not at the sector level), it would asymptotically recover

$$\mathbb{V}_{\text{CRVE}} = \frac{4}{N(N-2)}\sum_{i=1}^{N}\left(\beta\mathbb{1}\{i\in\Lambda_f \text{ such that } f\in\mathcal{T}\} - \beta/2 + \epsilon_j - \bar{\epsilon}\right)^2. \tag{15}$$

Consider now a sequence in which $F \to \infty$, where we maintain the number of observations in each Λ_f fixed, and that $\sum_{f=1}^F \mathcal{X}_f = F/2$. Given the assumption that ϵ_i was drawn from a distribution in which errors are independent, and assuming that such distribution has finite fourth moments, we have that the sequence $\{\epsilon_i\}_{i\in\mathbb{N}}$ is such that, with probability one,

$$F\left(\mathbb{V}_{\text{true}} - \mathbb{V}_{\text{CRVE}}\right) = \beta^2\left[\frac{F}{F-2} - \frac{F}{N-2}\right] + o(1). \tag{16}$$

Therefore, except for the case in which $F/(N-2) \to 1$, which is a setting in which a correction for the spatial correlation such as the one considered by Adão et al. (2019) would not be necessary, CRVE would asymptotically underestimate the variance of the true distribution of $\hat{\delta}$ whenever $\beta \neq 0$. In this case, the placebo exercise would reveal over-rejection even if the underlying distribution of ϵ_i did not exhibit spatial correlation.

Table A.1: **Shift-share designs - resampling errors**

	China shock		Trade liberalization		Exposure to robots: Main effects		Exposure to robots: Placebos	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Estimate	-0.489	-0.489	-1.976	-2.443	-0.516	-0.448	-0.217	0.006
CRVE								
Standard error	0.076	0.076	0.822	0.723	0.118	0.059	0.151	0.070
p-value	0.000	0.000	0.016	0.001	0.000	0.000	0.152	0.930
Inference Assessment								
5% test	0.102	0.102	0.061	0.147	0.092	0.320	0.102	0.385
10% test	0.165	0.165	0.116	0.221	0.152	0.398	0.168	0.471
Wild bootstrap								
p-value	0.000	0.000	0.027	0.002	0.001	0.001	0.198	0.928
Inference Assessment								
5% test	0.052	0.052	0.049	0.067	0.050	0.066	0.052	0.054
10% test	0.099	0.099	0.106	0.125	0.102	0.131	0.104	0.122
AKM								
Standard error	0.164	0.148	0.311	0.112	0.053	0.030	0.070	0.054
p-value	0.003	0.001	0.000	0.000	0.000	0.000	0.002	0.908
Inference Assessment								
5% test	0.163	0.211	0.570	0.791	0.386	0.556	0.316	0.605
10% test	0.235	0.283	0.625	0.821	0.454	0.615	0.391	0.660
AKM0								
Standard error	0.139	0.166	0.873	1.366	0.226	0.221	0.106	0.056
p-value	0.000	0.003	0.024	0.074	0.022	0.043	0.041	0.912
Inference Assessment								
5% test	0.069	0.047	0.336	0.029	0.160	0.198	0.052	0.084
10% test	0.153	0.129	0.516	0.489	0.300	0.377	0.199	0.266
Weighted	Yes	Yes	No	Yes	No	Yes	No	Yes
# of clusters	48	48	91	91	48	48	48	48
# of observations	1444	1444	411	411	722	722	722	722
# of sectors	770	770	20	20	19	19	19	19
# of clusters of sectors	136	20	20	20	19	19	19	19

Notes: this table replicates Table 2 but calculating the assessment by resampling errors instead of resampling shocks.