A SIMPLE WAY TO ASSESS INFERENCE METHODS*

Bruno Ferman†

Sao Paulo School of Economics - FGV

First Draft: December 15th, 2019

This Draft: April 23rd, 2021

Abstract

We propose a simple way to assess whether inference methods are reliable. The 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 show that, despite being a simple idea and despite its limitations, this assessment has the potential of making scientific evidence more reliable, if it becomes widely used by applied researchers. We analyze in detail the cases of differences-in-differences with few treated clusters, shift-share designs, weighted OLS, stratified experiments, and matching estimators.

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 Alberto Abadie, Arun Advani, Kirill Borusyak, Xavier D'Haultfoeuille, Marcelo Fernandes, Avi Feller, Lucas Finamor, Raymond Fisman, Peter Hull, Toru Kitagawa, Michael Leung, Marcelo Medeiros, Duda Mendes, Marcelo Moreira, Vitor Possebom, Marcelo Sant'Anna, Andres Santos, Rodrigo Soares, and participants at the PUC-Chile, Chicago, and PIMES seminars, 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 credibility of scientific research depends crucially on the control of false-positive results. However, we may have an excess of false-positive results if inference is based on incorrect or unreliable methods.[1] Such problems may arise when (i) inference is based on methods that rely on unrealistic assumptions, (ii) inference is based on asymptotic theory when the asymptotic approximation is poor, and/or (iii) the inference method is invalid even asymptotically. In such cases, we can have an accumulation of scientific evidence based on misleading inference.

We propose that applied researchers use a practical and very simple way to assess whether an inference method is reliable in specific applications. The main idea is to consider a simple distribution for the errors, and assess whether inference methods are reliable using simulations with the same structure of the empirical application. While the idea of using simulations to evaluate statistical methods — sometimes with data generating processes based on real applications — is not new, we show that, in common settings, an assessment based on such simulations is invariant to the scale of the error and to the values of parameters not being tested, so there is no gain in correctly specifying these features of the data.[2] We also show that there is generally little gain in correctly specifying more complex structures on the errors, such as within-cluster correlation when we consider inference based on cluster robust variance estimator (CRVE). Moreover, the assessment does not require specifying the distribution of covariates. In such cases, the assessment can be computed by simply replacing

[1] Other reasons include the possibility of p-hacking and publication bias (e.g., Christensen and Miguel (2018), Brodeur et al. (2016), and Brodeur et al. (2018)). The approach we propose in this paper is focused on cases of excess of false-positive results that arise from incorrect or unreliable inference, and would not be informative about these other potential problems.

[2] A non-exhaustive list of papers that consider simulations based on real applications include Huber et al. (2016), Busso et al. (2014), Young (2016), and Chaisemartin and Ramirez-Cuellar (2019). These papers, however, consider such simulations in the context of methodological papers, and not as an assessment for applied researchers. The idea of relying on simulation studies tailored to the features of the data at hand has been proposed by Athey et al. (2020) and Blair et al. (2019) to select among alternative estimators and to diagnose research designs, and has also been considered as a part of a "workflow" for Bayesian analysis (e.g., Gelman et al. (2020)). See also Advani et al. (2019) for a critical analysis of the idea of using empirical Monte Carlo studies for estimator selection. Also, Young (2020) uses MC simulations to analyze the distribution of instrumental variables estimators in published papers.

the outcome variable with an iid standard normal random variable, which we recommend as a default.

We show that the simplicity of such procedure presents a series of advantages. First, conditioning on a good assessment does not imply distortions for subsequent testing. In contrast, we show that, if we consider more complex simulations in which we attempt to learn about the distribution of the errors based on the residuals, then conditioning on a good assessment can exacerbate size distortions.[3] Second, having such a simple assessment as a default implies less room for applied researchers to cherry pick simulations in which their inference methods would look good. This would not be the case if we consider more complex alternatives that depend on the specification of a series of tuning parameters. Moreover, the proposed assessment is very easy to implement, and can be applied in different empirical applications with minimal adaptation.

Another contribution is to provide strong evidence from a series of applications showing that, despite being a simple idea and despite its limitations, the widespread use of this procedure has the potential of making scientific evidence more reliable. We analyze in detail the cases of differences-in-differences with few treated clusters, shift-share designs, weighted OLS, stratified experiments, and matching estimators. We show a series of settings in which the assessment would detect problems for inference even when applied researchers would likely not suspect that inference is problematic. We also show that scientific evidence on important topics may severely underestimate uncertainty, even after going thorough peer-review processes. As an illustration, we show a series of published papers on the effects of the Massachusetts 2006 health reform in which the assessment suggests over-rejections on the order of 60%. In such cases, the widespread use of the assessment we propose would have led researchers to consider alternative inference methods that are more suitable to their applications. Moreover, we provide evidence that applied papers may recurrently base

[3]There are a series of papers that analyze the implications of pre-testing on subsequent testing, including, for example, Andrews (2018), Guggenberger (2010), and Roth (2019). In contrast to the settings analyzed in these papers, when we consider such simple assessment, conditioning on a satisfactory assessment does affect the true size of the test, because the assessment does not depend on the realization of the errors.

inference on unreliable inference methods, even years after the publication of econometrics papers raising concerns about such inference methods.

We also present novel evidence that analyzing rejection rates 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) suggest that an inference methods is reliable, while an assessment for the same inference method, but for a 10% level test, detects large distortions. This happens because imposing the null implies a downward bias in the rejection rates that is stronger when we consider smaller significant levels. We recommend, therefore, considering the assessment for different significance levels 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.

We also provide important contributions that are specific to the burgeoning literature on shift-share regression designs. Adão et al. (2019) provide important theoretical results showing that commonly used standard errors for shift-share regressions can be underestimated if errors are spatially correlated. They provide a series of simulation studies based on the empirical setting considered by Autor et al. (2013), suggesting rejection rates on the order of 55% for 5% nominal level tests in this application. We revisit the simulations considered by Adão et al. (2019), and show that their simulations can severely overstate the relevance of such problem. Moreover, while the alternative inference method they suggest should always be preferred relative to alternatives such as CRVE when they are reliable, we describe an application in which we conclude that inference based on CRVE should be preferable. This example also illustrates that the analysis of simulation studies may not be so straightforward. Therefore, there is value in considering a simpler assessment as the one we propose, with a clear understanding of its limitations, if such assessment is going to be widely used by applied researchers.

Finally, we also show that asymptotic approximations may be substantially less reli-

able when we consider weighted OLS specifications, relative to standard OLS specifications. While MacKinnon and Webb (2017) provides evidence that the use of rules of thumb to determine the appropriate number of clusters may be misleading for specifications with disaggregated data or when there few treated clusters, we provide a novel setting in which such rules of thumb may be misleading, even when we consider specifications with aggregate data.

Given the simplicity of the assessment we propose, it is natural that it presents some limitations. Importantly, this assessment is uninformative about the plausibility of assumptions on the structure of the errors that the inference methods rely on. For example, if we consider the case of CRVE, the main assumption usually considered in the literature for such inference method is that errors can be correlated within clusters, but uncorrelated across clusters.[4] Our idea in this case 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, which would likely not be the case. These limitations imply that the assessment may suggest that the inference method is reliable when it actually is not. Likewise, it may also be that the assessment suggests distortions even when the true size of the test is good.

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 indicate that the researcher should proceed with caution. In such cases, researchers should consider alternative inference methods, or should argue and provide credible evidence that alternative simulations in which the original inference method looks reliable provide a better approximation for their empirical applications. If instead the assessment is close to α, then this

[4]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.

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. Importantly, this assessment should not preclude the use of alternative assessments that could detect problems it would not be able to detect.

While seemingly related to other methods, such as bootstrap, permutation test, and Monte Carlo (MC) test, the idea we propose is conceptually different.[5] The main goal in these approaches is to derive an inference method that is valid. In contrast, our goal is to assess whether an inference method is reliable. In common settings, the inference method we are assessing would have the advantage of being valid under weaker assumptions relative to methods that are valid in finite samples, but the disadvantage of only being valid asymptotically. In such cases, the assessment would be informative about whether such asymptotic approximations are reasonable, or whether we should consider alternative methods that depend on stronger assumptions. Note also that the assessment can actually be used to evaluate whether specific bootstrap methods are reliable. For example, considering the case of DID, the assessment can reveal that block bootstrap is unreliable when we have few treated clusters.

Our proposal is also related to a series of papers that evaluate whether asymptotic approximations are reliable 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, the assessment we propose can be used to evaluate asymptotic approximations in a wide variety of applications, instead of being specific to particular examples. Moreover, it provides a nat-

[5]Note that MC test is a different concept than MC simulations. See Dufour and Khalaf (2007) for a survey on MC tests.

ural 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. For the case of CRVE, we also show an application in Section 3.4 that the assessment we propose would detect problems, while the effective number of clusters proposed by Carter et al. (2017) would not.

Finally, while recent work on statistical decision theory suggests that there might be too much focus on test size (Manski, 2019), it remains important to understand whether standard errors presented in applied work reliably capture uncertainty about our estimates, and the assessment we propose can be useful in this direction. Relatedly, in a recent paper, Broderick et al. (2020) propose an interesting approach to measure the extent to which the conclusions from a study are robust to the removal of a small fraction of the sample. We see our approaches as complementary, in that there are settings in which the assessment we propose would detect problems while their approach would not, and vice versa.[6]

The remainder of this paper proceeds as follows. We describe in details the proposed assessment for the case of OLS regressions in Section 2. In Section 3, we present different applications in which the assessment can be used. We consider the cases of DID with few treated cluster (Section 3.1), shift-share designs (Section 3.2), weighted OLS (Section 3.3), field experiments (Section 3.4), and matching estimators (Section 3.5). 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.

[6]For example, in the empirical application we consider in Section 3.4, the inference problem comes from the fact that the standard errors are asymptotically invalid, so we may have applications with size distortions in which the conclusions remain robust to the removal of small fractions of the data. Likewise, consider a DID setting with one treated cluster, as in Section 3.1. If we have many observations per cluster, then the results may remain robust to the removal of small fractions of the data, even though we should expect very large size distortions from inference based on CRVE.

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)).[7] 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 (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. For example, CRVE is generally asymptotically valid when the number of clusters goes to infinity. A crucial question for applied researchers then is: how many clusters are enough for reliable

[7] We consider the properties of the estimator $\widehat{\boldsymbol{\beta}}$ in a repeated sampling framework over the distribution of $\boldsymbol{\epsilon}$. See Remark 3 for a discussion of the assessment if we consider a design-based approach for inference.

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.[8] 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 (see, for example, Section 3.3). Moreover, in some cases it may be that the inference method is invalid even asymptotically (see, for example, Section 3.4).

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. 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 $\mathbf{q}$. The basic idea is to choose a $\widetilde{\boldsymbol{\beta}}$ that satisfies $\mathbf{R}\widetilde{\boldsymbol{\beta}} = \mathbf{q}$, and a distribution for the error $\widetilde{F}(\boldsymbol{\epsilon})$ that satisfies the assumptions on the errors for the inference method that is being assessed. In most settings, this distribution can simply be iid standard normal. Then we simulate new datasets $\mathbf{y}^b = \mathbf{X}\widetilde{\boldsymbol{\beta}} + \boldsymbol{\epsilon}^b$, where $\boldsymbol{\epsilon}^b$ is drawn from $\widetilde{F}(\boldsymbol{\epsilon})$, as in a MC simulation or a bootstrap, and for each draw we test the null using the inference method that is being assessed. The assessment is the proportion of times we reject the null using such inference method in a large number of simulations. We discuss alternative options for the distribution of the error in Remarks 1 and 2, and alternative sampling schemes (for example, by resampling covariates instead of errors) in Remarks 3 and 4. We also show in Section 3 examples in which the assessment can be easily modified for cases in which the estimator is not based on OLS.

A step-by-step procedure to calculate the assessment is given by:

- Step 1: choose $\widetilde{\boldsymbol{\beta}}$ such that $\mathbf{R}\widetilde{\boldsymbol{\beta}} = \mathbf{q}$, and $\widetilde{F}(\boldsymbol{\epsilon})$ such that the assumptions for the inference method being assessed are satisfied.

- Step 2: do $\mathcal{B}$ iterations of this step. In each step:

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

- Step 2.1: draw a random vector $\boldsymbol{\epsilon}^b$ from the distribution $\widetilde{F}(\boldsymbol{\epsilon})$, and generate $\mathbf{y}^b = \mathbf{X}\widetilde{\boldsymbol{\beta}} + \boldsymbol{\epsilon}^b$.
- Step 2.2: estimate the model with $\mathbf{y}^b$ in place of $\mathbf{y}$.
- Step 2.3: 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 3: the 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 same idea can also be used to assess the reliability of confidence intervals. For example, consider the case in which $\widetilde{\boldsymbol{\beta}}$ is a scalar. In this case, in Step 2.3 we would construct confidence intervals with the method that is being assessed, and store whether these confidence intervals include the parameter $\widetilde{\boldsymbol{\beta}}$ chosen in Step 1. This would provide an assessment of the coverage of the confidence intervals.

The data from the simulations in Step 2 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, the 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. Since $\widetilde{F}(\boldsymbol{\epsilon})$ is chosen to satisfy the assumptions for asymptotic validity of the inference method, 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.[9] 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.

When we consider a linear model and we are testing a null hypothesis regarding a linear combination of the parameter $\boldsymbol{\beta}$, the estimator considered in Step 2.2, 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 $\widehat{\boldsymbol{\epsilon}}^b = (\mathbb{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\,\boldsymbol{\epsilon}^b$. Therefore, $\mathbf{R}\widehat{\boldsymbol{\beta}}^b - \mathbf{q}$ and the residuals in the simulations will be invariant with respect to the choice of $\widetilde{\boldsymbol{\beta}}$ (provided $\mathbf{R}\widetilde{\boldsymbol{\beta}} = \mathbf{q}$), and the relative magnitude between $\widehat{\boldsymbol{\beta}}^b$ and $\widehat{\boldsymbol{\epsilon}}^b$ will be invariant to the scale of the distribution of $\boldsymbol{\epsilon}^b$. This is true even if we do not consider a normal distribution for the errors. Therefore, for most inference methods, the assessment will be numerically invariant to the choice of $\widetilde{\boldsymbol{\beta}}$ and to the scale of the distribution of 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. See Remark 5 for cases in which this may not be the case.

Therefore, in common settings in which we want to test the null that a specific coefficient is equal to zero in a linear model, we can consider a simple case in which $\widetilde{\boldsymbol{\beta}} = 0$ and errors are iid normal with variance one. There is no gain in correctly specifying the *scale* of the errors and how other covariates correlate with the outcome, because the assessment is invariant to these features. In this case, the assessment becomes very easy to implement. We would just have to replace the vector of outcomes in the original data with a vector of iid standard normal random variable. Importantly, the use of normal errors does not necessarily mean

[9] This is true if the inference method is asymptotically valid conditional on a sequence of $\mathbf{X}$. As an example, 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. See Remark 3 for more details on that.

that we believe errors are normal in specific applications. Rather, we see that as a simple way to implement an assessment that provides a "low bar" to check if inference is problematic (see discussions in Remarks 1 and 2 regarding the use of alternative distributions).

Other variations in the distribution for the errors, however, might potentially lead to different assessments. For example, if the true distribution of the errors has heavier tails than a normal distribution, then the assessment using normal errors may understate inference problems. 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 also show in Appendix A.2 an example in which the assessment based on homoskedastic errors may actually overstate inference problems if the true errors are heteroskedastic.

Therefore, we may have that the assessment suggests that the inference method controls well for size when the true size is larger than α. This may happen either because the true distribution of the errors has different characteristics relative to the distribution considered in the assessment, or because other assumptions required by the inference method are invalid. Alternatively, the assessment may suggest relevant over-rejection even when the true size of the test is good.

Overall, we do not see those as fundamental problems, because we see this assessment as only 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 seen as a strong indicative that the asymptotic theory that justifies the inference method is unreliable, and that the researcher should, at least, proceed with caution. In such cases, the researcher might consider using alternative inference methods, being careful in case such alternative inference methods rely on stronger assumptions. In many cases, the assessment we propose sheds some light on important trade-offs related

to asymptotic theory and assumptions when considering alternative inference methods, as we present in the applications in Section 3. Alternatively, researchers may try to justify that alternative simulations that lead to assessments closer to α are more reasonable in his/her particular application.[10] Note that the possibility that researchers continue to rely on an inference method even when the default assessment is larger than α does not mean that the use of the assessment is innocuous. In such cases, applied researchers would have to provide convincing evidence that the inference method they consider is reliable, despite being unreliable in a simple setting in which errors are iid standard normal (see discussion in the application considered in Section 3.3).

If instead 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. Importantly, the use of the assessment we propose should not preclude the use of alternative diagnosis methods or alternative simulations that may detect problems it is unable to detect.

In simple examples, as in the one considered in Appendix A.2, it may be possible to derive worst-case scenarios conditional on a set of possible distributions for the 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. As we present in the applications from Section 3, despite its simplicity and all potential limitations, a widespread use of the assessment among applied researchers based on such simple distribution can already go a long way in detecting inference problems in a wide range of settings.

Another advantage of considering the assessment based on iid standard normal errors as a

[10] Alternative simulations may consider, for example, alternative distributions for the errors (see Section 3.3), consider unconditional inference, or treat the covariates as stochastic and the outcomes as fixed, as discussed in Remarks 3 and 4.

default is that it reduces the possibility that researchers look for specific simulations in which the assessment is close to α. It is important, though, to have some flexibility to consider the assessment with different simulations, because it may be that alternative simulations better approximate a specific empirical application. However, if we have a setting in which the assessment suggests relevant over-rejection using the default simulation, but does not detect substantial problems in alternative simulations, then the applied researcher would have to justify why his/her empirical application is better approximated by these alternative simulations.

Remark 1 Considering an iid distribution in more complex settings, such as when we are assessing CRVE, is not as limiting as it may appear at first glance. Consider a setting in which cluster sizes are homogeneous, and we do not have individual-level covariates. If we restrict to multivariate normal distributions for the errors that are iid across clusters, then the assessment would be invariant to changes in the within-cluster correlation.[11] The main idea here is that, while CRVE relaxes the assumption of independence within cluster, it remains asymptotically valid when the number of *clusters* goes to infinity, even if we consider a distribution for the errors that is iid within clusters. Therefore, considering iid errors for the assessment still provides a first screening on the inference method, while avoiding more complex specifications of within-cluster correlations.

Remark 2 An advantage of considering the default assessment we propose is that the assessment becomes independent from the realization of the errors in the real data. Therefore, the true size of the test would be the same whether we condition on a good assessment or not.[12] In contrast, if we consider more complex simulations in which we attempt to learn about the distribution of the errors based on the estimated residuals, then conditioning on a good assessment may affect the size of the test, and may even exacerbate over-rejection problems. We present in Appendix A.3 examples in which this may happen when we consider

[11] If we have variation in cluster sizes, then the assessment may vary when we change the intra-cluster correlation for the same reason why heteroskedasticity may change the assessment when we consider EHW.

[12] In this case, we consider a test conditional on the covariates $\mathbf{X}$.

two alternatives for the distribution of ϵ_i: resampling with replacement from the estimated residuals, as in a residual bootstrap, or multiplying the residuals by 1 with probability 0.5 and by -1 with probability 0.5, as in a wild bootstrap.

Remark 3 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), where potential outcomes are fixed, and uncertainty comes from the assignment of $\mathbf{x}_i$ and from a random sampling from the finite population. In their setting, if we consider the case in which $\mathbf{x}_i \in \{0, 1\}$, then EHW standard errors would be asymptotically valid when N_1 and N_0 increase, although they may be conservative depending on the estimand of interest. Note, however, that EHW standard errors are also asymptotically valid under a repeated sampling of $\boldsymbol{\epsilon}$ conditional on $\mathbf{X}$ exactly when N_1 and N_0 increase. Likewise, Adão et al. (2019) consider a repeated sampling framework for shift-share designs in which the shocks are stochastic, while potential outcomes are fixed. As we show in Section 3.2, however, their inference method would also be asymptotically valid in a framework where we condition on $\mathbf{X}$ and resample errors. These examples show that the assessment resampling errors may be informative about whether we should proceed with caution, even when we have a design-based approach for uncertainty in mind.

Remark 4 Related to Remark 3, alternative simulations based on finite population settings in which $\mathbf{X}$ is considered as stochastic and $\mathbf{Y}$ is fixed has been used, for example, by Chaisemartin and Ramirez-Cuellar (2019) in the context of stratified field experiments. They consider permutations of the treatment assignment, and then evaluate an inference procedure in each permutation. However, they consider such simulations in the context of a methodological paper, and not as a recommendation for applied researchers to evaluate inference methods in their specific applications. The fact that they find many published papers with inference problems — together with the evidence from Young (2018) — provides clear evidence that the use of such simulations to assess inference methods is not widespread

among applied researchers, even when we restrict to papers based on field experiment. Likewise, Adão et al. (2019) consider simulations to check the reliability of different inference methods in shift-share design applications considering $\mathbf{X}$ as stochastic. Again, such simulations are considered in the context of a methodological paper, and not as a recommendation for applied researchers. Moreover, as we discuss in Section 3.2, there are a couple of subtleties in the use of such simulations in a shift-share design setting. Overall, the assessment we propose is more straightforward to use in applications in which the researcher does not know the distribution of $\mathbf{X}$. We stress that the use of the assessment based on resampling errors does not preclude the use of alternative assessments, for example, based on resampling covariates. As we show in Section 3.2, the set of problems an assessment is able to detect depends on how it is constructed. Therefore, we indicate the assessment resampling errors as only a first screening, and we emphasize that it should not preclude the use of alternative assessments.

Remark 5 As shown above, in common applications the assessment is invariant to scale changes in the distribution of the errors, and to the choice of $\widetilde{\boldsymbol{\beta}}$ (provided the null is valid). This property also holds, for example, for the nearest neighbor matching estimator, considered in Section 3.5. However, this property may not hold in non-linear models or in settings in which the null is a non-linear function of the parameters. In such settings, we recommend that $\widetilde{\boldsymbol{\beta}}$ is chosen as a constrained estimator where the null is imposed, and that the scale of the errors is estimated from the residuals. Overall, this means that, in such settings, there would be additional reasons why the assessment may differ from the true size of the test being assessed.

Remark 6 We also recommend that the researcher presents the assessment for different significance levels. As we show in Section 3.2, 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. One

possibility is to consider the distribution of p-values in the simulations, which should be close to uniform $[0, 1]$ if the inference method is reliable.

Remark 7 In case the assessment detects a relevant over-rejection for a given inference method, one possibility is to use the simulations to adjust the critical value of the test, so that it controls for size. By construction, this strategy would generate a test with correct size if the distribution for the errors used in the simulations were correct. However, this approach should be considered with caution, because we will generally have no guarantee that the distribution of the errors chosen for the simulations is the correct one.[13] More generally, the main goal of the assessment is to warn about the possibility that an inference method is unreliable, and not to be a general solution to inference problems. The idea is that, if this first screening suggests a problem, then applied researchers should consider the use of alternative inference methods that are more suitable for their specific application, carefully analyzing the assumptions that such alternative inference methods rely on. In case no suitable existing inference method is available in a given setting, then the assessment would indicate that new inference methods should be developed for such settings.

3 Applications

We consider the use of the assessment in a series of applications. First, we consider in Section 3.1 the case of DID with few treated clusters. In Section 3.2, we consider the case of shift-share designs. In Section 3.3, we consider the case of weighted OLS. We then consider in Section 3.4 the case of stratified randomized control trials. Finally, we consider in Section 3.5 the case of matching estimators. Overall, these empirical applications provide clear evidence that, despite being a simple idea and despite its limitations, the widespread use of this procedure has the potential of making scientific evidence more reliable.

[13] 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.

3.1 Differences-in-Differences with Few Treated Clusters

As a first empirical illustration of potential problems that the inference assessment would be able to detect, and also of potential problems that the assessment would fail to detect, we consider an analysis of the Massachusetts 2006 health care reform. This reform was analyzed in a series of papers using a DID design in which MA is the treated state (Sommers et al. (2014), Miller (2012), Niu (2014), Courtemanche and Zapata (2014), and Kolstad and Kowalski (2012)). For example, Sommers et al. (2014) consider 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.[14] They found a reduction of 2.9%-4.2% in mortality in Massachusetts relative to the controls after the reform, and they reported standard errors clustered at the state level (they also considered 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, 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 the assessment to the inference methods considered by Sommers et al. (2014). When we consider clustering at the state level, the assessment using simple iid normal errors indicates a rejection rate of 63%. Therefore, this simple assessment would have provided an immediate conclusion that such inference procedure is not reliable, and that alternative inference methods should be considered. Importantly, a series of other papers

[14]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.

analyzing the Massachusetts 2006 health care reform rely on similar research designs, and are subject to the same problem of presenting standard errors that are severely underestimated.[15] In addition to presenting results based on CRVE, Kolstad and Kowalski (2012) also consider confidence intervals based on block bootstrap. Block bootstrap is one of the recommendations from Bertrand et al. (2004) for taking serial correlation into account in DID settings. However, in settings with only one treated cluster, this method leads to substantial over-rejection. The assessment we propose would also be able detect that this bootstrap method is unreliable in this setting.

Interestingly, 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. In this particular example, the problem in considering clustered standard errors at the state level with few treated clusters was discussed at least since Conley and Taber (2011). This simple example highlights that the 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. In the first case, the assessment should prompt new developments in econometrics to deal with such problems, while in the second case it should lead applied researchers to consider alternatives that are more suitable to their specific applications. This example also illustrates that scientific evidence on important topics can be based on misleading inference, even after going through peer-review processes.

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, generally rely on stronger assumptions on the errors. Importantly, the inference assessment will not generally be informative

[15]This is the case for Miller (2012), Niu (2014), Courtemanche and Zapata (2014), and Kolstad and Kowalski (2012). In addition to presenting results based on CRVE, Courtemanche and Zapata (2014) also present inference based on a permutation test, similar to what was proposed by Kaestner (2016).

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. The 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. However, the assessment would be mute about the possibility of state-level shocks.[16] In this case, Ferman (2020) proposed another assessment, 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. 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 Conley and Taber (2011) at the county level, then state-level shocks would invalidate an important assumption of this method, as in the case in which we consider CRVE at the county level. Moreover, whether we consider Conley and Taber (2011) 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 this method

[16]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.

(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) at the state level corrects for heteroskedasticity generated by variation in population sizes, but does not allow for unrestricted heteroskedasticity. This is again 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. 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). However, whether N_1 is "large enough" to rely on CRVE is not something well defined. The assessment can shed some light on this trade-off, and help applied researchers decide on which inference method to use. If the assessment for CRVE is close to 5%, then we would have some support to use this method. Since inference based on CRVE relies on weaker assumptions and/or has more power relative to alternatives that are valid with fixed N_1, it should be preferred in case it is reliable. 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 3.4), sampling weights (see Section 3.3), and others may also be relevant. The assessment takes all of those characteristics into account.

Overall, the assessment provides a simple and widely applicable way of detecting *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 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 are 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. We show that the 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 the assessment to consider simulations with random draws of the shocks. In this setting, this is what Adão et al. (2019) do in their simulations. The 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 continue to construct the assessment based on resampling errors, by simply replacing the outcome variable with an iid standard normal random variable. In Appendix A.4, we show that the inference method proposed by Adão et al. (2019) is also asymptotically valid in this sampling framework exactly when the number of shocks goes to infinity and the size of each shifter becomes asymptotically negligible. Therefore, a default assessment based on resampling errors and conditional on covariates would still be informative about whether this inference method is reliable in specific applications, even though the original theory that justifies this method is based on resampling shocks.[17] We consider both alternatives to construct the assessment. Importantly, as we show below, 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),

[17] Similar to the discussion in Remark 1, the idea here is that, while these standard errors are robust to spatial correlation, they remain valid when errors are iid. In this case, the idea of the assessment resampling errors is to inform whether the asymptotic theory — which depends on the number of sectors going to infinity whether errors are spatially correlated or not — provides a good approximation to specific applications.

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 1 the inference assessment for CRVE, which is the inference method originally considered in these applications. When we consider the 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%.[18] This is the same kind of exercise considered by Adão et al. (2019). Not surprisingly, we find similar results. However, differently from Adão et al. (2019), we do not take that as direct evidence that CRVE leads to such substantial distortions in test size in these applications. As we formally show in Appendix A.5, such simulations may confound the true treatment effect of the shift-share variable with spatially correlated shocks. Therefore, we may find size distortions in such simulations even when errors are not spatially correlated.

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)).[19] In this case, the true treatment effect would be zero, and the simulations with random shocks would not confound treatment effects with spatial correlation. Such assessment, however, would not be informative about the possibility of over-rejection due to heterogeneous treatment effects.

[18] We consider iid standard normal shocks. As discussed in Section 2, we could potentially consider alternative distributions for the shocks. For example, Borusyak and Hull (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 the inference assessment is not to recover the true distribution of the test, but to assess whether the inference method is reliable. See also Remark 2 for potential problems in running the assessment with a distribution for the errors based on the idea of a wild bootstrap.

[19] Roth (2019) shows that, in the DID setting, pre-testing may exacerbate the problem of violations of parallel trends in case it fails to detect such violations due to sampling noise. In contrast, if parallel trends hold, but CRVE is invalid due to spatial correlation, then Ferman (2019) shows that failing to detect spatial correlation problems does not exacerbate the over-rejection problem.

We present in columns 7 and 8 of Table 1 the 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, the 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. For these placebo outcomes, we still find some over-rejection for the specification without population weights (around 11% for a 5% test), and a larger over-rejection for the specification with population weights (around 26% for a 5% test). While this could indicate presence of spatially correlated shocks, note that we also find similar over-rejection when the assessment is constructed based on resampling errors. This suggests that the over-rejection we detect in this case comes mainly from the number of clusters not being large enough. In this case, Acemoglu and Restrepo (2020) would still reject the null with a p-value smaller than 0.01 when considering the specification from column 5, which is relatively more reliable (Appendix Table A.1). We analyze the differences between considering standard and weights OLS in more detail in Section 3.3.

Differently from CRVE, 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, these methods 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 the assessment we propose can be informative in this case.

Adão et al. (2019) propose two alternatives for the estimation of their proposed standard errors. One in which the residuals used to estimate the standard errors are based on the original regression without imposing the null (we call that AKM), and another one in which they impose the null imposed to estimate the residuals (AKM0).[20] For the specifications

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

based on Autor et al. (2013), the assessments based on resampling shocks are close to 5%, particularly when we impose the null. These results replicate the findings from Adão et al. (2019). The assessment resampling errors, however, indicates some over-rejection when we consider the standard errors without imposing the null. As we show in Appendix A.4, this assessment should be close to α if we have many sectors that are asymptotically negligible, and if the sequence of realized shocks is consistent with an underlying distribution of shocks that are independent across sectors (Assumption A.3.(iii) in Appendix A.4). The assessment resampling shocks considers, by construction, shocks independent across sectors, while the assessment resampling errors does not impose this condition. Therefore, contrasting the two types of assessments, this suggests that the assumption that shocks are independent should be considered with caution. More generally, this example illustrates that different assessments may detect different problems. Therefore, it is crucial to understand the set of problems an assessment is able to detect, depending on how it is constructed. Moreover, since different assessments may detect different problems, this example highlights that the assessment we propose should be seen as only a first screening, and should not preclude the use of alternative assessments.

When we consider the use of these inference methods for the other two applications, the assessments suggest more severe problems. When the null is not imposed, we find over-rejections ranging from 21% to 79% for a 5%-level test, whether we consider the assessment resampling shocks or errors. Note that this inference method is similar to the one proposed by Borusyak et al. (2018), which was considered as a robustness by Acemoglu and Restrepo (2020). In this application, we find assessments of 35% and 43%, depending on the specification. Therefore, in this application, inference based on CRVE (as Acemoglu and Restrepo (2020) consider in their main tables), seems to be more reliable than the new inference methods, especially when we consider the OLS regressions with no population weights.

When the null is imposed, the assessment is generally greater than 11% for the specifications based on Dix-Carneiro et al. (2018) and Adão et al. (2019). An exception is the

specification considered in column 4, which indicates an assessment of 3.4% (2.9% for the assessment based on resampling errors). 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 (49% for the assessment based on resampling errors).

To analyze that further, we present in Figure 1 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 the null leads to under-rejection when we consider a 5%-level test, as presented in Table 1, but 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 a strong evidence that this inference method is not reliable in this application.

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 may be more reliable, as we show for the application from Acemoglu and Restrepo (2020). Other alternative in this case would be the randomiza-

tion inference type of test proposed by Borusyak and Hull (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 would be reasonable. Overall, here again we have to deal with non-trivial trade-offs in terms of asymptotic theory and assumptions when selecting among different inference methods, and the assessment we propose can be used to shed some light on which inference method should be used in such applications.

3.3 Weighted OLS

As we show in Section 3.2, the assessments for CRVE resampling errors suggest substantially more distortions in the weighted OLS specifications for the applications from Dix-Carneiro et al. (2018) and Acemoglu and Restrepo (2020). This suggests the possibility of very large distortions when using CRVE with weighted OLS, even when we have a reasonably large number of clusters, and when the assumptions on the errors are valid.

To understand how weights may affect the quality of asymptotic approximations, consider a sample $\{Y_i\}_{i=1}^N$, where $Y_i \overset{iid}{\sim} N(0,1)$. We estimate the mean of Y_i with a weighted average, where the first half of the observations receives weight of one, and the other half receives weight of $W > 1$. In this case, when $W \to \infty$, this essentially means that this weighted average would only be based on $N/2$ observations, implying that asymptotic approximations would be poorer, particularly if N is not very large. For example, simulating a t-test using the asymptotic critical value in this setting with $N = 10$ and $W = 10$, we find rejection rates of 8% when we do not use weights, and 13% when we consider a weighted average. This is consistent with our findings in Section 3.2.

However, one of the reasons for using sampling weights may be that observations with lower variance should receive larger weights, in order to improve precision. This may be the case, for example, when weights are given by population sizes. In this case, this may com-

pensate part of the rationale above on why sampling weights may lead to poorer asymptotic approximations. In the example above with $N = 10$ and $W = 10$, if $Y_i \overset{iid}{\sim} N(0, 0.1)$ for the observations that received weight $W = 10$, then the rejection rate using the weighted average would be 11%, slightly lower than the rejection rate we had when observations were homoskedastic. However, the test using simple average would still have a lower over-rejection in this case (7.6%). When N increases, all of those test, whether we use homoskedastic or heteroskedastic errors, and whether we use simple or weighted averages, converge to have a rejection rate of 5%.

In light of this simple example, we revisit the applications from Dix-Carneiro et al. (2018) and Acemoglu and Restrepo (2020), analyzed in Section 3.2. We abstract from the possibility of spatial correlation in shift-share designs, so we can focus on the consequences of using weighted OLS when inference is based on CRVE, even when the assumptions for CRVE are valid. We consider the assessment resampling errors.

In both applications, Y_i are averages for a given region level, and weights are given by population sizes. The first line of Table 2 presents the assessment using homoskedastic errors. Again, we find evidence of relevant over-rejection, particularly when we consider weighted OLS. If we had that the individual-level errors were independent within region, then we should expect that the variance of Y_i would be proportional to $1/M_i$, where M_i is the population of region i. We present in the second line of Table 2 the assessment in which errors are normally distributed with variance $1/M_i$. For the specification from Acemoglu and Restrepo (2020), the assessment continues to suggest large distortions for the weighted regressions, but much lower than when we consider homoskedastic errors. For the weighted regression from Dix-Carneiro et al. (2018), the assessment based on heteroskedastic errors would suggest that the inference method is reasonably reliable ($\approx$ 7%), providing more support to rely on such inference method than when we consider the assessment based on homoskedastic errors ($\approx$ 15%).

This means that the assessment based on homoskedastic errors for the weighted OLS

specification from Dix-Carneiro et al. (2018) may incorrectly suggest that CRVE is more unreliable than it actually is. As discussed in Section 2, if the applied researcher can convincingly argue that the heteroskedastic errors better approximate his/her empirical application, then he/she would have some support to rely on CRVE in this setting. In this case, however, this would depend on whether the researcher can convincingly provide such evidence.

If we observed the errors ϵ_i, and $\epsilon_i \sim N(0, b/M_i)$ for some constant b, then a regression of ϵ_i^2 on a constant and $1/M_i$ would provide unbiased estimators for zero and b. Since we do not observe ϵ_i, we can consider instead the residuals $\hat{\epsilon}_i$, and regress $\hat{\epsilon}_i^2$ on a constant and $1/M_i$. A problem here is that, with a finite number of observations, these estimators may not be unbiased. Still, this provides some evidence on whether the heteroskedastic distribution for the errors considered in the assessment is reasonable. When we consider this regression, we find estimators for b that are positive (confirming the intuition that regions with larger populations have lower variance), but we can strongly reject the null hypotheses that the constants are equal to zero. If we consider the assessment with errors $\epsilon_i \sim N(0, \hat{a} + \hat{b}/M_i)$, where $(\hat{a}, \hat{b})$ are the OLS estimator of $\hat{\epsilon}_i^2$ on a constant and $1/M_i$, then the assessment would be closer to the case in which errors are homoskedastic (third line of Table 2). Therefore, in this case, we would not have strong evidence that inference based on CRVE is reliable, even though there are some distributions for the errors in which the assessment would be relatively close to 5%.

Importantly, in settings in which the researcher is able to provide some support that the distribution for the errors used in the assessment that is closer to 5% is more reasonable, whether we would like to conclude that the inference method is reliable depends on how conservative we want to be regarding the possibility of concluding that the inference method is reliable when it may actually be unreliable. Note that such differences depending on whether errors are homoskedastic or heteroskedastic only arise because the number of clusters is not very large. If we had a larger number of clusters, then the assessments would be close to 5% irrespectively of whether we use weighted OLS, and of whether we use homoskedastic or

heteroskedastic errors. We present that in Panels B and C of Table 2, where we replicate the structure of the empirical application, and consider the clusters between different replications as independent (so we have two or four times the number of clusters relative to the original application).

Overall, this example highlights that researchers should have some room for considering alternative distributions for the errors instead of fixing a default of iid standard normal when constructing the assessment. However, departures from such default should be well justified by the applied researchers. As we present in this example, whether it is reasonable to focus on the assessment using heteroskedatic errors depends crucially on specific details of the application.

3.4 Stratified randomized control trials

As another empirical application, 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 the assessment can be informative in this setting. First, we show that the assessment would easily detect the problem raised by Chaisemartin and Ramirez-Cuellar (2019) for the case of small G. Given the evidence from Section 3.1, we expect that a number of papers will continue to circulate without correcting

for this problem, even after it has been presented in an econometrics paper. Therefore, if widely used by applied researchers, we expect that the assessment will remain relevant to prevent papers from circulating based on misleading inference due to this problem. In this case, applied researchers would find an assessment larger than α, and this should lead them to consider econometrics papers that discuss this issue, accelerating the diffusion of this knowledge.

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 the 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 (2019) 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). The 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 3, we consider the case in which schools are stratified in pairs. In column 1, we present the assessment if we consider for inference CRVE at the school level. We rely on the default assessment, by resampling the outcome variable from iid standard normal random variables. When there are 12 schools, the assess-

ment would suggest 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 Chaisemartin 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 3, we present the 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, the 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 in finite samples for a more narrowly defined null hypothesis. Moreover, they do not directly provide standard errors. Depending on the choice of the test statistic, permutation tests may also be asymptotically valid for weaker null hypotheses (e.g., Wu and Ding (2020)). However, considering iid normal variables for the assessment, it would not be informative about whether such permutation tests are reliable for weaker nulls. Finally, depending on the estimand of interest, the researchers should cluster at different levels to take uncertainty into account (e.g., Abadie et al. (2017) and Deeb and de Chaisemartin (2020)). Therefore, there are some gains in considering cluster-robust standard errors, if they are reliable, even when exact randomization 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 extreme case in which N schools are divided into $S = 2$ strata. In this case, the 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, the 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 the assessment conditional on this draw for the covariates.[21] We present the 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, the 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 3 are based on one single draw of the school-level covariates for each (N, S, G) cell. We consider now whether different

[21] This draw was generated from five independent standard normal variables at the school level.

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 2 the pdf of the assessment in this case. The 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. The 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 3.A the scatterplot of the assessment measure and the effective number of clusters. The two measures are highly correlated (correlation coefficient of -0.75), showing that the 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 the assessment.

When we consider 100 draws of the covariates for the $(N, S, G) = (400, 200, 2)$ scenario, the assessment would be closer to 5%, and would be much less disperse (see Figure 2). 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 3.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 the assessment would not be able to detect this problem. This is because we are implicitly assuming homogeneous treatment effects in our simulations when we consider a default with iid standard normal outcomes. As we emphasize, the assessment should be seen as a first screening for inference methods, and it will generally not be able to detect all potential problems that inference methods may have. An important advantage of considering an assessment by simply replacing the outcome variable with an iid standard normal variable is that it becomes easier to implement, and becomes widely applicable with minimal adaptation. Considering more complex simulations could potentially uncover other problems. We stress that, being a first screening, the simple assessment we propose does not preclude the use of alternative assessments that may detect problems it would not be able to detect.

3.5 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, the assessment can be used to provide some evidence on 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, the 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.[22] In the absence of finite sample bias, these tests would have a size smaller or equal to α% 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 the assessment for those tests would never be greater than α%, it would be informative about the extent to which these tests are conservative. Overall, the assessment can inform about the potential trade-offs between different inference procedures in a setting of matching estimators with

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

few treated observations.

4 Concluding remarks

We propose an assessment for inference methods that is very easy to implement, and that can be used in a wide range of applications. This assessment can detect problems when the asymptotic theory that justifies an inference method is invalid and/or provides a poor approximation given the design of the empirical application. However, this assessment will not be able to detect all potential problems an inference method may face. Moreover, in finite samples, the applied researcher will generally not have all necessary information to derive the true size of a test. Therefore, we have to consider the possibility of two kinds of errors: (i) that the assessment suggests that the inference method controls well for size when there are relevant size distortions, and (ii) that it suggests significant distortions even when the true size of the test is good.

The possibility that the assessment fails to detect problems should be acknowledge by applied researchers, and means that an assessment close to α does not immediately guarantees that the inference method is reliable. The researcher should justify and provide evidence that potential problems that are not captured by the assessment are not relevant in their setting. Moreover, an assessment larger than α using a default distribution for the errors, such as iid standard normal, should be seen as an important warning that such inference method may be unreliable, but does not provide a definite conclusion that it is unreliable. In this case, in order to continue relying on such inference method, the applied researcher should provide a good justification why the inference method remains reliable, despite such evidence from the assessment using a default distribution for the errors. Our example with weighted OLS regressions goes in this direction. An interesting avenue for future research is to identify specific settings in which inference methods are reliable even when the assessment with a default distribution for the errors suggests they are not, and to propose alternative

ways to assess the reliability of such inference methods in these settings.

Overall, as illustrated in a series of applications, despite all potential limitations, the widespread use of this assessment has the potential of making scientific evidence substantially more reliable.

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.

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

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

Andrews, I. (2018). Valid Two-Step Identification-Robust Confidence Sets for GMM. *The Review of Economics and Statistics*, 100(2):337–348.

Athey, S., Imbens, G., Metzger, J., and Munro, E. (2020). 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.

Bertrand, M., Duflo, E., and Mullainathan, S. (2004). How much should we trust differences-in-differences estimates? *Quarterly Journal of Economics*, page 24975.

Blair, G., Cooper, J., Coppock, A., and Humphreys, M. (2019). Declaring and diagnosing research designs. *American Political Science Review*, 113(3):838?859.

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

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

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

Broderick, T., Giordano, R., and Meager, R. (2020). An automatic finite-sample robustness metric: Can dropping a little data change conclusions?

Brodeur, A., Cook, N., and Heyes, A. (2018). Methods Matter: P-Hacking and Causal Inference in Economics. IZA Discussion Papers 11796, Institute of Labor Economics (IZA).

Brodeur, A., LÃ©, M., Sangnier, M., and Zylberberg, Y. (2016). Star wars: The empirics strike back. *American Economic Journal: Applied Economics*, 8(1):1–32.

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.

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

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.

Christensen, G. and Miguel, E. (2018). Transparency, reproducibility, and the credibility of economics research. *Journal of Economic Literature*, 56(3):920–80.

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.

Courtemanche, C. J. and Zapata, D. (2014). Does universal coverage improve health? the massachusetts experience. *Journal of Policy Analysis and Management*, 33(1):36–69.

Deeb, A. and de Chaisemartin, C. (2020). Clustering and external validity in randomized controlled trials.

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.

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.

Gelman, A., Vehtari, A., Simpson, D., Margossian, C. C., Carpenter, B., Yao, Y., Kennedy, L., Gabry, J., Burkner, P.-C., and Modrak, M. (2020). Bayesian workflow.

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

Guggenberger, P. (2010). The impact of a hausman pretest on the asymptotic size of a hypothesis test. *Econometric Theory*, 26(2):369–382.

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

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.

Kolstad, J. T. and Kowalski, A. E. (2012). The impact of health care reform on hospital and preventive care: Evidence from massachusetts. *Journal of Public Economics*, 96(11):909 – 929. Fiscal Federalism.

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. (2019). Randomization Inference for Difference-in-Differences with Few Treated Clusters. *Journal of Econometrics, Forthcoming*.

Manski, C. F. (2019). Treatment choice with trial data: Statistical decision theory should supplant hypothesis testing. *The American Statistician*, 73(sup1):296–304.

Miller, S. (2012). The impact of the massachusetts health care reform on health care use among children. *American Economic Review*, 102(3):502–07.

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

Niu, X. (2014). Health insurance and self-employment: Evidence from massachusetts. *ILR Review*, 67(4):1235–1273.

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.

Wu, J. and Ding, P. (2020). Randomization tests for weak null hypotheses in randomized experiments.

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: **Shift-share designs - resampling errors**

	China shock		Trade liberalization		Exposure to robots			
					Main effects		Placebos	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A: Assessment resampling shocks								
CRVE								
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
AKM								
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								
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
Panel B: Assessment resampling errors								
CRVE								
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
AKM								
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								
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 presents the assessment when we consider inference based on CRVE, AKM, and AKM0, for different applications. In Panel A, the assessment is based on random draws of iid standard normal shocks, while in Panel B it is based on random iid standard normal errors. 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).

Table 2: **Assessment for unweighted vs weighted OLS**

	Trade liberalization		Exposure to robots			
			Main effects		Placebos	
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: original data						
$\epsilon \sim N(0,1)$	0.060	0.148	0.100	0.319	0.105	0.388
$\epsilon \sim N(0,1/M_i)$	0.053	0.069	0.059	0.135	0.060	0.157
$\epsilon \sim N(0,\hat{a}+\hat{b}/M_i)$	0.056	0.140	0.095	0.313	0.098	0.388
Panel B: duplicated data						
$\epsilon \sim N(0,1)$	0.052	0.104	0.074	0.162	0.075	0.197
$\epsilon \sim N(0,1/M_i)$	0.053	0.062	0.051	0.093	0.055	0.105
$\epsilon \sim N(0,\hat{a}+\hat{b}/M_i)$	0.057	0.105	0.066	0.169	0.070	0.195
Panel C: quadruplicated data						
$\epsilon \sim N(0,1)$	0.053	0.076	0.061	0.102	0.070	0.115
$\epsilon \sim N(0,1/M_i)$	0.053	0.061	0.049	0.074	0.055	0.081
$\epsilon \sim N(0,\hat{a}+\hat{b}/M_i)$	0.050	0.077	0.065	0.099	0.068	0.118
Weighted	No	Yes	No	Yes	No	Yes
# of clusters (orig. data)	91	91	48	48	48	48
# of observations (orig. data)	411	411	722	722	722	722

Notes: this table presents the assessment for CRVE, using different distributions for the errors, for the specifications considered in columns 3 to 8 of Table 1. The first distribution is simple iid standard normal. The second one is a normal distribution with variance $1/M_i$, where M_i is the weight of observation i. The third one is a normal with variance $\hat{a}+\hat{b}/M_i$, where $(\hat{a},\hat{b})$ are the OLS estimator of $\hat{\epsilon}_i^2$ on a constant and $1/M_i$. In panel B, we duplicate the data, and consider clusters for the original data and for the replication as independent. For example, in columns 1 and 2, this leads to a model with 182 clusters and 822 observations. In panel C, we quadruplicate the data.

Table 3: **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.

Figure 1: **Assessment of AKM0 inference method - shift-share design**

Figure 1.A: assessment resampling shocks

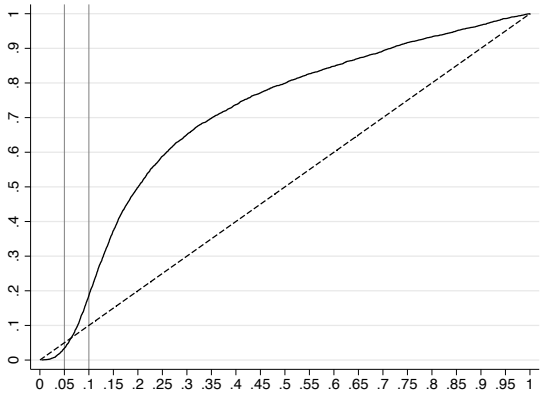
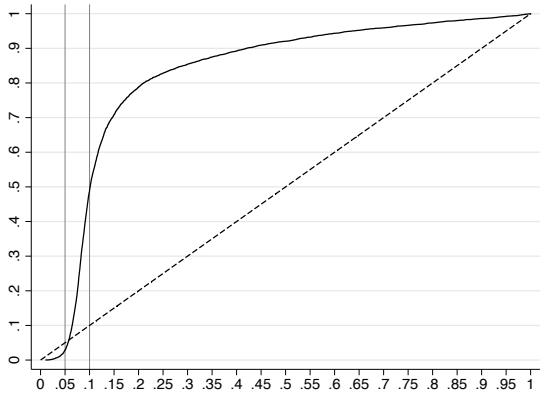
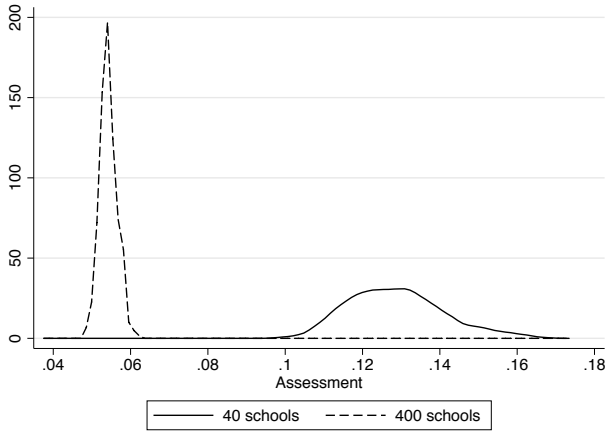


Figure 1.B: assessment resampling errors



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

Figure 2: **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 3: **Assessment vs effective number of clusters**

Figure 3.A: 40 schools

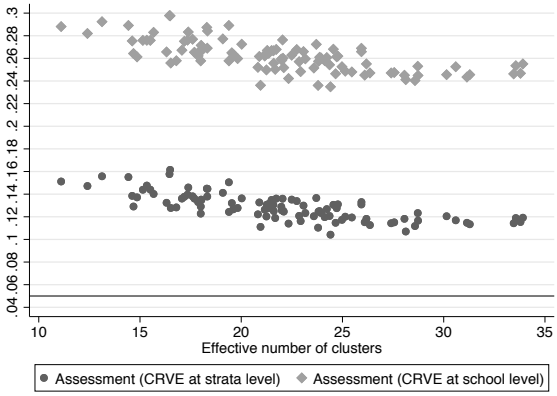
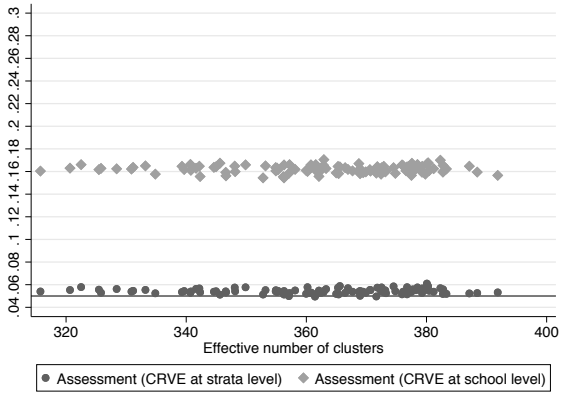


Figure 3.B: 400 schools



Notes: This figure presents scatterplots of the 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 $(N, S, G) = (400, 200, 2)$.

A Online 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 $\frac{1}{N}\sum_{i=1}^N \mathbf{x}_i'\mathbf{x}_i \to \mathbf{Q}$, $\frac{1}{N}\sum_{i=1}^N \sigma^2(\mathbf{x}_i)\mathbf{x}_i'\mathbf{x}_i \to \mathbf{A}$, $\limsup \frac{1}{N}\sum_{i=1}^N \|\mathbf{x}_i\|_2^4 < \infty$ *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,* $Pr\left(t < a|\{\mathbf{x}_i\}_{i\in\mathbb{N}}\right) \to \Phi(a)$ *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 Simple example of comparison of means

A.2.1 Example

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 the 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.2.2 that, assuming a normal distribution, the rejection rate for an α level test converges uniformly to α when $N_1, N_0 \to \infty$, 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. 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.

A.2.2 Uniform convergence

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 converges 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

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

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,

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

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 $e > 0$,

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

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, the 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 $\hat{\beta}^b = -\frac{1}{N-1}\sum_{i\neq 1} g_i^b \hat{e}_i$. 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.3.3 Assessment based on sign changes - real application

We consider the use of the assessment based on sign changes using the specification from Acemoglu and Restrepo (2020) considered in column 4 of Table 2. If the errors were iid standard normal, we know from Table 2 that the true size of the test is 0.319. In this section, we assume that the true model is such that errors are iid standard normal. We then draw 1000 realizations of such errors, and for each draw, we calculate the assessment based

on sign changes. When we do not impose the null to estimate the residuals, the median assessment is 15% which is substantially lower than the true size of the test, and in 10% of the realizations the assessment would be smaller than 8%, failing to indicate substantial problems for inference. Moreover, there is a negative and statistically significant correlation between the assessment and the absolute value of the t-statistic of the original regression (corr=−0.418). When we impose the null to estimate the residuals, the assessment becomes larger (median=0.540), and there is a positive and statistically significant correlation between the assessment the absolute value of the t-statistic (corr=0.569). The results reinforce the idea that, in finite samples, an assessment using a distribution that is based on the estimated residuals may depend on the realization of the errors.

A.4 Shift-share designs conditional on shares and shocks

We consider the simpler case with no covariates to simplify the proofs, as Adão et al. (2019) consider in their Section 4.1. In this case, the researcher runs an OLS regression $Y_i = \beta X_i + \epsilon_i$, where $X_i = \sum_{s=1}^{S} w_{is}\mathcal{X}_s$. We assume throughout that $\sum_{s=1}^{S} w_{is} \leq 1$ for all i. The main difference relative to the setting considered by Adão et al. (2019) is that we consider the properties of the estimator conditional on a sequence $\{w_{is}\}_{i=1,s=1}^{N,S}$ and $\{\mathcal{X}_s\}_{s=1}^{S}$.

We assume that the error term ϵ_i is iid standard normal and that $\beta = 0$. The idea is to show that, under some technical conditions, the standard error proposed by Adão et al. (2019) is valid under the sample scheme considered in the assessment based on resampling iid standard normal errors. We consider the following assumptions on the sequences of shares and shocks.

Assumption A.3 *(i)* $\max_s n_s / \sum_{t=1}^{S} n_t \to 0$, *where* $n_s = \sum_{i=1}^{N} w_{is}$, *(ii)* $\frac{1}{N}\sum_{i=1}^{N} X_i^2 \to lim\left(\frac{1}{N}\sum_i \sum_s w_{is}^2 \mathcal{X}_s^2\right) = Q > 0$, *(iii)* $\mathcal{X}_s$ *is uniformly bounded.*

Assumption A.3(i) is one of the main assumptions considered by Adão et al. (2019) for their asymptotic theory. It implies that the size of each sector, n_s, becomes asymptotically

negligible. Assumption A.3(ii) resembles Assumption A.1(ii) from the online appendix of Adão et al. (2019). It is a regularity condition so that the shocks generate enough variation in X_i, and the denominator of $\hat{\beta}$, divided by N, does not converge to zero. Considering the distribution of shocks (that is, before we condition on the sequence of shares and shocks), the condition that $\frac{1}{N}\sum_{i=1}^{N} X_i^2 \to \lim\left(\frac{1}{N}\sum_i \sum_s w_{is}^2 \mathcal{X}_s^2\right)$ would be satisfied with probability one if $\mathcal{X}_s$ are independent with mean zero, under some technical conditions. Note that independence of shocks is another of the crucial assumptions for the inference methods proposed by Adão et al. (2019). The condition that these variable have mean zero is made for simplification, as Adão et al. (2019) do in their Section 4.1. Finally, the condition that $\mathcal{X}_s$ is uniformly bounded is made for ease of exposition.

Overall, Assumption A.3 impose conditions that we should expect to be satisfied in the framework considered by Adão et al. (2019), if we condition on X_i instead of conditioning on potential outcomes. The advantage of conditioning on potential outcomes, as Adão et al. (2019) do, is that they do not impose any restriction on the spatial dependence of the errors. However, our goal here is different. We want to show that the inference method they propose remains asymptotically valid conditional on the sequences of shares and shocks, if we consider a simple iid normal distribution for the errors. The goal is to show that we should expect an assessment resampling errors close to be α if the conditions in Assumption A.3 are satisfied and provide a good approximation for the empirical application. If the assessment resampling errors is significantly larger than α, then this would suggest that an asymptotic theory that relies on the number of sectors diverging and the size of each sector being asymptotically negligible does not provide a good approximation to the empirical application. It may also suggest that the assumption that shocks are independent is not valid. Therefore, even though the theory from Adão et al. (2019) is based on resampling shocks, the assessment resampling errors would still be informative in this case. The following proposition establish this result.

Proposition A.2 *Let* $\hat{\beta} = \sum_{i=1}^{N} X_i Y_i / \sum_{i=1}^{N} X_i^2$, *and* $\hat{\sigma}^2_{AKM} = \sum_{s=1}^{S} \mathcal{X}_s^2 \left(\sum_{i=1}^{N} w_{is}\hat{\epsilon}_i\right)^2 / \left(\sum_{i=1}^{N} X_i^2\right)^2$, *and let* $\mathcal{F}$ *be the set of information from shares and shocks. Suppose Assumption A.3 holds,*

and that Y_i is iid $N(0,1)$. Then, conditional on $\mathcal{F}$, $t = \hat{\beta}/\hat{\sigma}_{AKM}$ converges in distribution to a standard normal random variable.

Proof.

We consider throughout the proof that $\mathbb{E}^*[]$ is the expectation conditional on $\mathcal{F}$. First, note that $\mathbb{E}^*[X_i\epsilon_i] = 0$ and $var^*[X_i\epsilon_i] = X_i^2$. Since $\epsilon_i|\mathcal{F} \sim N(0,1)$, we have that

$$\left(\hat{\beta}|\mathcal{F}\right) = \left(\frac{\sum_i X_i\epsilon_i}{\sum_i X_i^2}|\mathcal{F}\right) \sim N\left(0, \frac{1}{\sum_i X_i^2}\right), \tag{14}$$

while the AKM variance estimator is given by

$$\hat{\sigma}^2_{\text{AKM}} = \frac{1}{N}\frac{\frac{1}{N}\sum_{s=1}^S \mathcal{X}_s^2\left(\sum_{i=1}^N w_{is}\hat{\epsilon}_i\right)^2}{\left(\frac{1}{N}\sum_{i=1}^N X_i^2\right)^2}. \tag{15}$$

Now note that

$$\begin{aligned}\frac{1}{N}\sum_{s=1}^S \mathcal{X}_s^2\left(\sum_{i=1}^N w_{is}\hat{\epsilon}_i\right)^2 &= \frac{1}{N}\sum_{s=1}^S \mathcal{X}_s^2\left(\sum_{i=1}^N w_{is}\epsilon_i\right)^2 + \frac{1}{N}\sum_{s=1}^S \mathcal{X}_s^2\left(\sum_{j=1}^N w_{js}X_j(\beta-\hat{\beta})\right)\left(\sum_{i=1}^N w_{is}\epsilon_i\right) \\ &\quad + \frac{1}{N}\sum_{s=1}^S \mathcal{X}_s^2\left(\sum_{j=1}^N w_{js}X_j(\beta-\hat{\beta})\right)^2. \end{aligned} \tag{16}$$

We show that the first term in the RHS of equation 16 converges in probability to Q, while the other two terms are $o_p(1)$. Note that

$$\mathbb{E}^*\left[\frac{1}{N}\sum_{s=1}^S \mathcal{X}_s^2\left(\sum_{i=1}^N w_{is}\epsilon_i\right)^2\right] = \mathbb{E}^*\left[\frac{1}{N}\sum_{s=1}^S\sum_{i=1}^N\sum_{j=1}^N \mathcal{X}_s^2 w_{is}w_{js}\epsilon_i\epsilon_j\right] \tag{17}$$

$$= \frac{1}{N}\sum_{s=1}^S\sum_{i=1}^N \mathcal{X}_s^2 w_{is}^2. \tag{18}$$

Moreover, we have that the

$$
\begin{aligned}
var^*\left[\frac{1}{N}\sum_{s=1}^{S}\mathcal{X}_s^2\left(\sum_{i=1}^{N}w_{is}\epsilon_i\right)^2\right] &= \frac{1}{N^2}\sum_{i=1}^{N}var^*(\epsilon_i^2)\left[\sum_{s=1}^{S}\mathcal{X}_s^2 w_{is}^2\right]^2 + && (19)\\
&\quad +\frac{2}{N^2}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}var^*(\epsilon_i)var^*(\epsilon_j)\left[\sum_{s=1}^{S}\mathcal{X}_s^2 w_{is}w_{js}\right]^2 && (20)\\
&\leq \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left[\sum_{s=1}^{S}w_{is}w_{js}\right]^2 \leq \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{s=1}^{S}w_{is}w_{js} && (21)\\
&= \frac{\sum_{s=1}^{S}n_s}{N^2} \leq \frac{\max_s n_s}{N} \to 0. && (22)
\end{aligned}
$$

Therefore, we have that the first term converges in probability to Q. Now note that the second term in the RHS of equation 16 is given by $(\beta-\hat{\beta})\frac{1}{N}\sum_{s=1}^{S}\sum_{i=1}^{N}\sum_{j=1}^{N}\mathcal{X}_s^2 X_j\epsilon_i w_{is}w_{js}$, where $(\beta-\hat{\beta})=o_p(1)$. Given that $\mathcal{X}_s$ is bounded, for some constant K,

$$
\begin{aligned}
&var^*\left(\frac{1}{N}\sum_{s=1}^{S}\sum_{i=1}^{N}\sum_{j=1}^{N}\mathcal{X}_s^2 X_j\epsilon_i w_{is}w_{js}\right) = \frac{1}{N^2}\sum_{i=1}^{N}\left(\sum_{s=1}^{S}\sum_{j=1}^{N}\mathcal{X}_s^2 X_j w_{is}w_{js}\right)^2 && (23)\\
&\leq K\frac{1}{N^2}\sum_{i=1}^{N}\left(\sum_{s=1}^{S}\sum_{j=1}^{N}w_{is}w_{js}\right)^2 = K\frac{1}{N^2}\sum_{i=1}^{N}\left(\sum_{s=1}^{S}w_{is}\left(\sum_{j=1}^{N}w_{js}\right)\right)^2 && (24)\\
&= K\frac{1}{N^2}\sum_{i=1}^{N}\left(\sum_{s=1}^{S}w_{is}n_s\right)^2 \leq K\frac{\max_s n_s}{N^2}\sum_{i=1}^{N}\left(\sum_{s=1}^{S}w_{is}\right)^2 && (25)\\
&\leq K\frac{\max_s n_s}{N} \to 0. && (26)
\end{aligned}
$$

Therefore, $(\beta-\hat{\beta})\frac{1}{N}\sum_{s=1}^{S}\sum_{i=1}^{N}\sum_{j=1}^{N}\mathcal{X}_s^2 X_j\epsilon_i w_{is}w_{js} = o_p(1)O_p(1)$. Finally, for some other

constant K, the third term in the RHS of equation 16 is given by

$$
\begin{aligned}
(\beta-\hat{\beta})^2\frac{1}{N}\sum_{s=1}^{S}\mathcal{X}_s^2\left(\sum_{j=1}^{N}w_{js}X_j\right)^2 &= (\beta-\hat{\beta})^2\frac{1}{N}\sum_{s=1}^{S}\sum_{j=1}^{N}\sum_{i=1}^{N}\mathcal{X}_s^2 w_{js}w_{is}X_jX_i && (27)\\
&\leq K\left(\sqrt{N}(\beta-\hat{\beta})\right)^2\frac{1}{N^2}\sum_{s=1}^{S}\sum_{j=1}^{N}\sum_{i=1}^{N}w_{js}w_{is} && (28)\\
&= K\left(\sqrt{N}(\beta-\hat{\beta})\right)^2\frac{\sum_{s=1}^{S}n_s}{N^2} && (29)\\
&\leq K\left(\sqrt{N}(\beta-\hat{\beta})\right)^2\frac{\max_s n_s}{N} = O_p(1)o(1). && (30)
\end{aligned}
$$

Combining all these results, we have that $\frac{1}{N}\sum_{s=1}^{S}\mathcal{X}_s^2\left(\sum_{i=1}^{N}w_{is}\hat{\epsilon}_i\right)^2 \to_p Q$.

Now consider

$$
t=\frac{\hat{\beta}}{\hat{\sigma}_{\text{AKM}}}=\frac{\sqrt{N}\frac{1}{N}\sum_{i=1}^{N}X_i\epsilon_i}{\sqrt{\frac{1}{N}\sum_{s=1}^{S}\mathcal{X}_s^2\left(\sum_{i=1}^{N}w_{is}\hat{\epsilon}_i\right)^2}}. \tag{31}
$$

Conditional on $\mathcal{F}$, the numerator converges in distribution to a normal with mean zero and variance Q, while the denominator converges in probability to $\sqrt{Q}$. Therefore, conditional on $\mathcal{F}$, t converges in distribution to a standard normal variable. ■

A.5 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 originally 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

$$\mathbb{V}_{\text{true}} \equiv var\left(\hat{\delta}|\{Y_i\}_{i=1}^N\right) = \frac{4}{F(F-2)} \sum_{f=1}^{F} \left(\beta \mathbb{1}\{f \in \mathcal{T}\} - \beta/2 + \bar{\epsilon}_f - \bar{\epsilon}\right)^2, \tag{32}$$

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

$$\mathbb{V}_{\text{CRVE}} = \frac{4}{N(N-2)} \sum_{i=1}^{N} \left(\beta \mathbb{1}\{i \in \Lambda_f \text{ such that } f \in \mathcal{T}\} - \beta/2 + \epsilon_j - \bar{\epsilon}\right)^2. \tag{33}$$

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,

$$F\left(\mathbb{V}_{\text{true}} - \mathbb{V}_{\text{CRVE}}\right) = \beta^2 \left[\frac{F}{F-2} - \frac{F}{N-2}\right] + o(1). \tag{34}$$

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.

A.6 Appendix Tables

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

					Exposure to robots			
	China shock		Trade liberalization		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
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
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
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, AKM, and AKM0 for the applications considered in Table 1.