Inference with Difference-in-Differences Revisited *

Mike Brewer
University of Essex
Institute for Fiscal Studies

Thomas F. Crossley
Koc University
Institute for Fiscal Studies
University of Cambridge

Robert Joyce
Institute for Fiscal Studies

May 1, 2013

Abstract

A growing literature has focused on inference in difference-in-differences (DiD) designs with grouped errors. This literature has been pessimistic about obtaining hypothesis tests of the correct size, particularly with few groups. We provide Monte Carlo evidence for three points. First, it is possible to obtain tests of the correct size, even with few groups, using methods that are very straightforward to implement. Second, the main problem in DiD designs with grouped errors is instead low power to detect real effects. Third, feasible GLS estimation combined with robust inference methods can increase power considerably whilst maintaining correct test size - again, even with few groups. These findings are robust to a range of data generating processes. We therefore recommend that applied researchers think seriously about efficiency, rather than just consistency and test size, when using these designs.

JEL Classification: C12, C21.

Mike Brewer, Institute for Social and Economic Research, University of Essex, Colchester, Essex, CO4 3SQ.

Thomas F. Crossley, Istanbul Institute for Fiscal Studies, 7 Ridgmount Street, London, WC1E 7AE.

Robert Joyce, Institute for Fiscal Studies, 7 Ridgmount Street, London, WC1E 7AE.

*We would like to thank Matthias Parey, Jonathan Shaw, seminar participants at the Institute for Fiscal Studies and the Work Pensions and Labour Economics Group at the University of Sheffield for helpful comments. All remaining errors are our own. The research was supported by Programme Evaluation for Policy Analysis, a Node of the National Centre for Research Methods supported by the UK Economic and Social Research Council. See http://www.ifs.org.uk/centres/PEPA.

1 Introduction

Difference-in-differences (DiD) designs are extremely common as a way of estimating the effects of policies or programmes (henceforth 'treatment effects'). A recent literature has highlighted that failure to appropriately quantify the uncertainty surrounding DiD estimates can lead to dramatically misleading inference (e.g. Bertrand et al, 2004). In particular, researchers will tend to reject true null hypotheses with a probability that is far higher than the nominal size of the hypothesis test. The literature has suggested that obtaining tests that are close to the correct size requires non-standard techniques, and that it may not be possible with a small number of groups (Angrist and Pischke, 2009; Bertrand et al, 2004; Cameron et al, 2008).

In this paper we report evidence from Monte Carlo simulations that suggests a different conclusion. We make three main points. First, tests of the correct size in typical DiD settings can in fact be obtained with very straightforward methods that are trivial to implement with standard statistical software (in fact, STATA's cluster-robust inference implements these methods by default). This is true even with few groups. Second, these techniques have very low power to detect real treatment effects. Thus the real challenge for inference with DiD designs is power rather than size. Third, we show that substantial gains in power can be achieved using feasible GLS. Moreover, if feasible GLS is combined with robust inference methods, test size can still be controlled in a way that is robust to having small numbers of groups, and to violations of the parametric assumptions about the error process implicit in FGLS estimation. We therefore recommend that applied researchers using DiD designs pay careful attention not just to consistency and test size, but also to the efficiency of their estimators.

DiD designs often use micro-data but estimate the effects of a treatment which varies only at a group level at any point in time (e.g. variation in policy across US states). A consequence is that within-group correlation of errors can substantially increase the true level of uncertainty surrounding the treatment effect (e.g. Angrist and Pischke, 2009; Donald and Lang, 2007; Moulton, 1990; Wooldridge, 2003). Furthermore, errors are likely to exhibit serial correlation within groups - an issue for inference highlighted in Bertrand et al (2004). For example, in an evaluation of a labour market policy implemented at a regional level, a researcher should worry both that people in the same region at the same time are affected by common labour market shocks (unrelated to the policy), and that these regional shocks are serially correlated. In a well-cited Monte Carlo study using US earnings data, Bertrand et al (2004) show that accounting only for grouped errors at the state-time level whilst ignoring serial correlation led to a 44% probability of rejecting a true null hypothesis using a nominal 5% level test.

A simple approach to deal with both cross-sectional and serial correlation in within-group errors would be to use the formula for a cluster-robust variance matrix due to Liang and Zeger (1986). This is consistent and Wald statistics which use it are asymptotically normal, but the asymptotics apply as the number of clusters tends to infinity. By clustering at the group level rather than the group-time level to account for serial correlation, one is often left with few clusters. The finite sample (i.e. few-clusters) performance of this approach - an empirical question - then becomes crucial, and the literature to date has come to pessimistic conclusions about it. Bertrand et al (2004) and Cameron et al (2008) use US earnings data and generate placebo state-level treatments before estimating their 'effects'. Forming t-statistics using cluster-robust standard errors (CRSEs), they obtain 9% and 11% rejection rates using nominal 5% level tests with samples from 10 and 6 US states respectively.[1] This is a considerable improvement over using OLS standard errors, when rejection rates are more than 40%. But it is still approximately double the nominal test size.

We first show that simple modifications to the above procedure can dramatically improve test size with few clusters. One can apply a simple scaling factor to the OLS residuals that are plugged into the CRSE formula, and use critical values from a t distribution with degrees of freedom equal to the number of groups minus one, rather than a standard normal. When this is done, our Monte Carlo simulations show that true test size is within about one percentage point of nominal test size with 50, 20, 10 or 6 groups. We further show that this holds under a wide range of error processes.

Various alternative techniques for achieving correct test size have been proposed and/or tested (Bertrand et al, 2004; Cameron et al, 2008; Donald and Lang, 2007; Bester et al, 2011). Of these, only a wild cluster bootstrap-t procedure has been shown to produce tests of approximately the right size in the typical DiD setup considered in this paper (see Section 2) when the number of groups is as small as six (Cameron et al, 2008). Like using CRSEs, this is theoretically robust to heteroscedasticity and arbitrary patterns of error correlation within clusters, and to variation in error processes across clusters. But it is less trivial to implement, computationally more intensive, and still slightly over-rejects when the number of groups is small.

Our second point is that, while it is straightforward to obtain the correct size, power to detect real effects is a serious concern. When we plug appropriately scaled residuals into the CRSE formula and compare the resulting quasi-t statistics to critical values from a t distribution with degrees of freedom equal to the number of groups minus one, we find that DiD designs have very low power. For example, with a large 30-year panel of US earnings data from 6 states, a policy implemented by half of the states that raised earnings by 2% would be detected with only 7.3% probability (using a test of size 0.05). The policy would have to increase earnings by 16% if the null of no policy effect is to be rejected with 80% probability.

[1] Both papers first account for cross-sectional within-group error correlation by aggregating to the group-time level, taking mean residuals within each group-time cell from a regression of earnings on individual-level characteristics. This is a straightforward way to deal with this problem and is appropriate in typical DiD settings where the number of observations per group-time cell is large. (It will also be the approach taken in this paper.) The remaining issue for inference is dealing with a finite number of groups and any serial correlation in group-time shocks.

Finally, we show that substantial gains in power can be achieved using feasible GLS. In particular, one can increase power by modelling the serial correlation of unobservables inherent in typical DiD designs. Test size can still be controlled in a way that is robust to having small numbers of groups, and to violations of the parametric assumptions about the error process implicit in FGLS estimation, using the straightforward cluster-robust inference technique described above.

The paper proceeds as follows. Section 2 describes the standard econometric setup in DiD designs that we consider. Section 3 details the Monte Carlo design we use to test different inference methods, and outlines the methods that we test. Section 4 presents ands discusses the results of our Monte Carlo simulations. Section 5 summarises and concludes.

2 Setup

We follow Bertrand et al (2004), Cameron et al (2008) and Hansen (2007) in using data on women aged 25 to 50 in their fourth interview month in the Merged Outgoing Rotation Group of the Current Population Survey. Our data include all 50 US states and the period 1979 to 2008 inclusive (i.e. G = 50 and T = 30). We focus primarily on log(earnings) as the dependent variable. We also consider the case where a binary employment indicator is the dependent variable in a linear probability model.[2] This covers the two most common outcomes of interest in DiD studies, according to a survey of the applied literature in Bertrand at al (2004).

We consider the standard linear DiD model

$$E_{igt} = \alpha_g + \delta_t + \beta T_{gt} + \gamma W_{igt} + \upsilon_{igt}, \tag{1}$$

where α_g and δ_t capture group (state) and time (year) fixed effects, β is the treatment effect of interest for a treatment which varies at the group-time level only, W_{igt} are individual-level control variables[3], and ν_{igt} is the unobserved individual-level earnings shock.

We focus on inference. Hence we assume that the OLS DiD estimator based on equation 1 is unbiased, i.e. $E(\upsilon_{igt}|\alpha_g, \delta_t, T_{gt}, W_{igt}) = 0$ so that $E(\hat{\beta}^{OLS}) = \beta$. This is ensured in our Monte Carlo simulations because we generate placebo treatments randomly.

The problem we seek to address is that the ν_{igt} may not be iid within groups. Some of the variation in ν_{igt} may occur at the group-time level. The DiD estimator is therefore effectively attempting to distinguish

[2] This gives us samples based upon the 750,127 women with strictly positive earnings and the 1,170,522 women with non-missing employment status respectively.

[3] We just use a quartic in age.

between the effects of a group-time level treatment and between-group differences in the evolution of group-time shocks. In addition, group-time shocks may be serially correlated. Both cross-sectional and serial correlation in within-group shocks are highly likely in many DiD applications, including the primary example used in this paper where groups are US states and the outcome of interest is earnings. We want to accurately quantify the additional uncertainty about β that this causes.

As in the aforementioned papers, we account for the grouping of errors at the state-time level by collapsing the data to this level. We run a regression using the micro-data of E_{igt} on W_{igt}, and take the mean residual within each state-time cell.[4] We denote this Y_{gt}. The variation across state-time cells in Y_{gt} is then essentially due to variation in state-time shocks. Hence we proceed with the model

$$Y_{gt} = \alpha_g + \delta_t + \beta T_{gt} + \varepsilon_{gt}, \tag{2}$$

where ε_{gt} is the state-time shock.[5],[6] Rewriting the model in this way highlights that, with both a treatment and a shock that vary at the group-time level, the OLS estimator of the treatment effect is consistent only as $G \rightarrow \infty$, regardless of the number of observations in the micro-data.

3 Methods

3.1 Experimental design

In our first set of Monte Carlo simulations, we repeatedly resample states with replacement from the CPS data and randomly choose half of the states to be 'treated'.[7] For all treated states in each Monte Carlo replication, the placebo treatment is applied in the same randomly chosen year[8] and in all subsequent years. We estimate the 'effect' of this placebo treatment by estimating equation 2. We initially use OLS, and

[4] Given large state-time cells (average cell sizes are 500 and 780 when the dependent variables are log(earnings) and employment status respectively), this averages out the iid shock component precisely. We ignore any estimation error from this procedure.

[5] We recommend this first step not only to make the estimation simpler computationally. We find that, even with moderate numbers of groups, test size can not be reliably controlled if one attempts to conduct cluster-robust inference straight from the micro-data (i.e. if one tries to account for all cross-sectional and serial correlation in within-group errors in a single step). This issue was also evident in the results of Cameron et al (2008) and is noted in Hansen (2007).

[6] If one is unsure whether this grouped error problem exists, Wooldridge (2006) points out that one could test for it. If the error term is dropped from equation 2, this imposes a set of (GT-1) restrictions on the data which can be used to estimate β by minimum distance. One can then test the over-identifying restrictions.

[7] In treating exactly half of the states, we follow the main approach in Bertrand et al (2004) and Cameron et al (2008). Note that this is the most favourable possible choice in terms of the resulting precision of treatment effect estimates, as it maximises between-group variation in treatment status. In other words, unequal numbers of treated and untreated groups would generally imply even *lower* power to detect real effects in DiD designs than we show is the case in this paper. Note also that randomly choosing states for placebo treatment means that there is no cross-cluster spatial correlation in treatment status. As shown by Barrios et al (2012), this means that inference is robust to errors in correctly specifying the clusters, i.e. robust to some correlation in the error terms across states. Hence, we will not be confusing the impacts of inadequately accounting for grouped errors with the impacts of incorrectly specifying the groups, even though we are using real data in which there is likely to be correlation between the earnings shocks of people in geographical proximity but in different states.

[8] The treatment year is chosen from a uniform distribution between 1988 and 2002.

later feasible GLS, for estimation. Our interest lies in the performance of different methods for performing inference about β, both in terms of type 1 and type 2 error (i.e. test size and power to detect real effects). To examine the effects of having differing numbers of groups, we run variants where we resample 50, 20, 10 and 6 states.

We first report how often the null hypothesis of no treatment effect is rejected using tests of nominal size 0.05 when using different inference methods. We show that tests of correct size can be achieved, even with as few as six groups, by forming a t-statistic using a simple variant of Liang and Zeger's (1986) cluster-robust standard error estimator and comparing it to critical values from a t_{G-1} distribution.

We check the robustness of this result to alternative data generating processes. We repeat the same Monte Carlo design, but use simulated state-time shocks rather than those from the CPS, allowing them to evolve according to an AR(1) process where we vary both the amount of serial correlation and the degree of non-normality in the white noise.

Following Bertrand et al (2004) and Cameron et al (2008), we then look at power by reporting how often the null of no effect is rejected when there is approximately a 2% effect on earnings or a 2 percentage point effect on employment rates, i.e. when we add $0.02T_{gt}$ to Y_{gt}.

For the method that we show produces tests of correct size even with as few as six groups, we look at power in more detail. We compute minimum detectable effects (MDEs) as first defined in Bloom (1995): the smallest effects that would lead to a rejection of the null hypothesis (of no effect) with a given probability. To do this, we use the same Monte Carlo procedures as described above to simulate the distribution of the t-statistic under the null hypothesis.[9] For power of x%, the MDE depends only on the (100-x)th centile of this distribution, the critical values from the t_{G-1} distribution, and the standard error estimator (see later). We therefore recover the entire relationship between power and MDEs. We do this for DiD designs with varying numbers of groups.

Finally, we show how power can be improved by using FGLS rather than OLS estimation. We compare MDEs for the cases of FGLS and OLS and show that they are substantially smaller when using FGLS. Using the same 'placebo' treatment Monte Carlos as described above, we also show that the power improvement can be gained without distorting test size, even if the parametric assumptions implicit in FGLS estimation are incorrect.

3.2 Approaches to inference that we compare

As a baseline we show what would happen if one used the standard OLS estimator of the standard error of $\hat{\beta}$ and compared the resulting t-statistic to standard normal critical values. This effectively assumes that

[9] This is necessary because, with few clusters, the t-statistic generally has an unknown distribution.

the ε_{gt} in equation 2 are iid, i.e. it ignores serial correlation.

We then focus primarily on inference based on variants of Liang and Zeger's (1986) cluster-robust standard error (CRSE) estimator. Their formula for a cluster-robust variance matrix is

$$\hat{V}_{CR} = (X'X)^{-1}(\sum_{g=1}^{G} X_g u_g u_g^{'} X_g^{'})(X'X)^{-1}, \tag{3}$$

where X is the regressor matrix, X_g is the regressor matrix for group g, and u_g is the vector of regression residuals for group g. This is consistent, and Wald statistics based on it are asymptotically normal, as $G \to \infty$. But it is biased, and the bias can be substantial when G is small. Intuitively, residuals will tend to be both smaller in magnitude and less correlated within clusters than the true errors because of model overfitting. Therefore CRSEs calculated using equation 3 will tend to be biased downwards.

A typical way of attempting to reduce small-G bias (or to eliminate it under special circumstances) is effectively to scale up the residuals before plugging them into equation 3. The default in STATA is to scale by $\sqrt{\frac{G(N-1)}{(G-1)(N-K)}}$, where N is the total number of observations and k is the number of parameters.[10] With large N, this is approximately equivalent to $\sqrt{G/(G-1)}$: the additional $\sqrt{(N-1)/(N-k)}$ is a degrees of freedom correction which makes a negligible difference in large samples. For brevity we refer to residuals scaled in this way simply as $\sqrt{G/(G-1)}$ residuals. This would lead to an unbiased CRSE estimator only under very special circumstances (see Bell and McCaffrey, 2002) so should be viewed as a bias-*reducing* correction. The same applies to a second, data-dependent scaling of u_g proposed in Bell and McCaffrey (2002),[11] and extended in Imbens and Kolesár (2012). This data-dependent scaling requires a full rank condition for X_g which is violated in the typical DiD setup considered here, both due to the inclusion of fixed group effects and because control groups are never treated. We therefore do not consider it in this paper.

For CRSEs formed using unscaled and $\sqrt{G/(G-1)}$ residuals, we show rejection rates when comparing the resulting t-statistics against critical values from both a standard normal and a t distribution with $G-1$ degrees of freedom. The former reference distribution is correct asymptotically as $G \to \infty$ (conditional on having a consistent estimator of the standard error), so the implicit assumption when using it is that G is large enough for the asymptotics to be a reliable guide. The latter is a common small-G correction, again used by STATA for Wald tests and confidence intervals. As one expects with finite sample methods, it does not have an exact theoretical justification under very general conditions.

However, recent work by Bester et al (2011) did provide theoretical justification in certain small-G

[10] When one uses the "vce(cluster *clustvar*)" option in a regression.

[11] This minimises the expected sum of squared differences between the scaled residuals and the true errors in the baseline case where errors are iid.

settings for the *combination* of CRSEs using $\sqrt{G/(G-1)}$-scaled residuals and t_{G-1} critical values. Their asymptotics apply as group size tends to infinity, holding the number of groups fixed. Despite the familiar result that a CRSE estimator is not consistent with fixed G, they show that plugging $\sqrt{G/(G-1)}$-scaled residuals into the CRSE formula nevertheless produces a covariance matrix which converges to a limiting random variable under certain conditions. Crucially, the resulting t-statistic turns out to have an asymptotic t_{G-1} distribution.[12] This result relies on a homogeneity requirement which, in the context of the canonical DiD setup with a binary treatment, would entail that each group has to be treated the same proportion of the time.[13] Clearly this is violated where some control groups are never treated. But the results we present in the following section suggest that, in practice, the Bester et al approach extends well to the standard DiD case (in terms of getting test size right).

For comparison, following Cameron et al (2008) we also consider the wild cluster bootstrap-t procedure.[14] Those authors found this to be the best of a large number of inference techniques in settings with few groups, in terms of test size. It outperformed various other bootstrap-based approaches, as well as inference based upon t-statistics formed with CRSEs. But that paper did not consider the $\sqrt{G/(G-1)}$ residual correction, and only standard normal critical values were used in t-tests.

4 Results

4.1 Rejection rates when the null is true

Table 1 contains results from our first Monte Carlo simulations, using the CPS log(earnings) data. It shows the rate with which the null of no effect is rejected when generating placebo treatments, estimating equation 2 by OLS, and using methods to perform inference about β. All hypothesis tests are of nominal size 0.05. Hence, rejection rates that deviate significantly from 0.05 indicate incorrect test size. We use 5000 replications. Simulation standard errors are shown in parentheses. The standard error for an estimated rejection rate $\hat{r}$ is $se(\hat{r}) = \sqrt{\hat{r}(1-\hat{r})/4999}$.

The first row of table 1 shows the rejection rates obtained assuming iid errors, i.e. by simply forming a t-statistic using the OLS standard error and comparing to standard normal critical values. Rejection rates exceed 40%, more than eight times the nominal test size. This essentially replicates the result in Bertrand

[12] This result is also robust to violations of the assumption of no inter-cluster correlation, as long as data are weakly dependent and some regularity conditions are satisfied. In the context of spatial data where clusters are geographic regions, this implies robustness to the fact that there will be some clustering between observations which are spatially close but put into different clusters by the researcher. The intuition is that cluster size tending to infinity would mean that most observations per cluster are far from other clusters, and hence cluster averages will be approximately independent.

[13] The general requirement is that the variance of the regressor matrix, and of the score, needs to be the same across groups.

[14] See Cameron et al (2008) for full details of this bootstrap. We use 200 bootstrap replications, which is sufficient in this context as bootstrap simulation error will average out across Monte Carlo replications.

et al (2004).

Forming CRSEs using unscaled OLS residuals and comparing the resulting t-statistic to standard normal critical values results in rejection rates that are too high, particularly with small G. Using t_{G-1} rather than the standard normal as the reference distribution is enough to achieve approximately the correct test size when $G \geq 20$, but not with 6 or 10 groups.

The $\sqrt{G/(G-1)}$ residual correction, combined with t_{G-1} critical values, achieves a test size that deviates by less than 1 percentage point from the nominal test size when G ranges between 6 and 50. The same residual correction combined with standard normal critical values also works well for moderate G but, as expected, these critical values result in over-rejection when G is small.

The final row of table 1 shows rejection rates obtained using the wild cluster bootstrap-t procedure. We essentially replicate previous findings: it performs well relative to most tested alternatives, but with over-rejection when $G = 10$ and particularly when $G = 6$.[15]

In summary, table 1 suggests that tests of the correct size can be obtained using very straightforward methods even with very few groups. In particular, this is achieved by computing a t-statistic with CRSEs that use residuals scaled by $\sqrt{\frac{G(N-1)}{(G-1)(N-K)}} \approx \sqrt{G/(G-1)}$, and using critical values from a t distribution with $(G-1)$ degrees of freedom. This is trivial to implement with statistical software. In fact, if one uses a cluster-robust variance matrix in STATA by specifying the "vce(cluster *clustvar*)" option, the confidence intervals and p-values returned are based upon precisely this procedure by default.[16] The best-performing alternative - the wild cluster bootstrap-t procedure highlighted by Cameron et al (2008) - is less trivial to implement, more computationally intensive and still suffers from slight over-rejection with small numbers of groups.

In table 2 we present results from an analagous set of Monte Carlo simulations using employment status rather than earnings as the dependent variable.[17] The performance of different inference methods in data containing varying numbers of groups is essentially the same as in Table 1. In particular, CRSEs formed using the $\sqrt{G/(G-1)}$ residual correction combined with t_{G-1} critical values perform best, with rejection rates always within 1 percentage point of the nominal test size. Again, the wild cluster bootstrap-t also performs well relative to most alternatives but slightly over-rejects with few groups.

[15] This over-rejection was not emphasised in Cameron et al, because it was not statistically significant in that study despite very similar point estimates (we have more Monte Carlo replications and hence lower simulation standard errors).

[16] This is true at the time of writing (STATA version 12.1) and has been the case since at least STATA 6.

[17] Given that we collapse the data to the state-time level in a first stage, this means that Y_{gt} now represents state-time employment rates rather than mean state-time earnings.

Table 1: Rejection rates for tests of nominal 5% size with placebo treatments in log(earnings) data

	G=50	G=20	G=10	G=6
Assume iid	0.422	0.420	0.404	0.412
	(0.007)	(0.007)	(0.007)	(0.007)
CRSE, t(G-1) critical values	0.039	0.042	0.048	0.047
	(0.003)	(0.003)	(0.003)	(0.003)
CRSE, N(0,1) critical values	0.044	0.056	0.075	0.104
	(0.003)	(0.003)	(0.004)	(0.004)
$\sqrt{G/(G-1)}$ residuals, t(G-1) critical values	0.042	0.046	0.050	0.049
	(0.003)	(0.003)	(0.003)	(0.003)
$\sqrt{G/(G-1)}$ residuals, N(0,1) critical values	0.048	0.061	0.079	0.107
	(0.003)	(0.003)	(0.004)	(0.004)
Wild cluster bootstrap-t	0.054	0.055	0.059	0.062
	(0.003)	(0.003)	(0.003)	(0.003)

Simulation standard errors in parentheses
Based on 5,000 Monte Carlo replications

Table 2: Rejection rates for tests of nominal 5% size with placebo treatments in employment data

	G=50	G=20	G=10	G=6
Assume iid	0.357	0.375	0.361	0.375
	(0.007)	(0.007)	(0.007)	(0.007)
CRSE, t(G-1) critical values	0.041	0.041	0.044	0.055
	(0.003)	(0.003)	(0.003)	(0.003)
CRSE, N(0,1) critical values	0.045	0.055	0.070	0.114
	(0.003)	(0.003)	(0.004)	(0.005)
$\sqrt{G/(G-1)}$ residuals, t(G-1) critical values	0.044	0.043	0.047	0.059
	(0.003)	(0.003)	(0.003)	(0.003)
$\sqrt{G/(G-1)}$ residuals, N(0,1) critical values	0.049	0.058	0.075	0.121
	(0.003)	(0.003)	(0.004)	(0.005)
Wild cluster bootstrap-t	0.051	0.052	0.056	0.072
	(0.003)	(0.003)	(0.003)	(0.004)

Simulation standard errors in parentheses
Based on 5,000 Monte Carlo replications

Robustness checks

The CPS data provide specific data generating processes with which to test the performance of computing t-statistics with CRSEs that use $\sqrt{G/(G-1)}$ -scaled residuals and using a t_{G-1} reference distribution. We have shown that our conclusions so far apply both when the outcome of interest is earnings and when it is a binary employment status indicator - two very different data generating processes. But one might still worry how general these findings are.

To explore this, we now repeat our Monte Carlo simulations using simulated data. The earnings generating process still conforms with equation 2, but we now simulate the state-time shocks ourselves. In doing so we vary their degree of serial correlation and non-normality. The state-time shocks for each state evolve according to the AR(1) process

$$\varepsilon_{gt} = \rho\varepsilon_{g,t-1} + \omega_{gt}, \quad t = 2, ..., 30$$

$$\varepsilon_{g1} = \frac{\omega_{g1}}{\sqrt{1-\rho^2}},$$

where ω_{gt} is iid across groups and time and is drawn from a t distribution with d degrees of freedom. To control the degree of non-normality in the white noise, we vary d between 2 (indicating very high non-normality) and 120 (at which point the t distribution is essentially standard normal). We generate the initial condition (ε_{g1}) such that it has the same variance as state-time shocks in other time periods. To control the degree of serial correlation, we vary ρ. We also examine a scenario in which the data generating process is heterogeneous, by drawing ρ_g separately for each state from a uniform distribution between 0 and 1.

In each Monte Carlo replication, we first resample states with replacement from the CPS data and randomly choose treated states and the year in which the placebo treatment begins, just as before. We then regress Y_{gt} on state and year fixed effects only. For each state-time combination, we simulate the outcome variable by summing the relevant (estimated) state effect, the relevant (estimated) year effect, and the random state-year shock generated as above. We then estimate the DiD model using the transformed outcome and conduct the hypothesis test on β. We use 10,000 replications.

Tables 3 to 6 report rejection rates for various combinations of ρ and d, when varying the number of groups between 50 and 6. They show that our finding is robust to a very wide range of error processes. Rejection rates remain within about a percentage point of the nominal test size under all of the tested combinations of degrees of serial correlation, non-normality in the white noise, and number of groups.

Table 3: Rejection rates for tests of nominal 5% size under various error processes with 50 groups

	$\rho = 0$	$\rho = 0.2$	$\rho = 0.4$	$\rho = 0.6$	$\rho = 0.8$	ρ varies with g
d=2	0.037	0.043	0.040	0.041	0.042	0.043
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
d=4	0.041	0.041	0.042	0.043	0.042	0.043
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
d=20	0.040	0.042	0.043	0.043	0.043	0.042
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
d=60	0.041	0.043	0.045	0.041	0.042	0.043
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
d=120	0.041	0.042	0.044	0.041	0.043	0.043
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)

Simulation standard errors in parentheses
Based on 10,000 Monte Carlo replications

Table 4: Rejection rates for tests of nominal 5% size under various error processes with 20 groups

	$\rho = 0$	$\rho = 0.2$	$\rho = 0.4$	$\rho = 0.6$	$\rho = 0.8$	ρ varies with g
d=2	0.039	0.045	0.047	0.046	0.043	0.047
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
d=4	0.043	0.047	0.045	0.044	0.043	0.047
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
d=20	0.044	0.048	0.046	0.044	0.043	0.047
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
d=60	0.045	0.049	0.045	0.044	0.045	0.046
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
d=120	0.046	0.049	0.044	0.044	0.046	0.045
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)

Simulation standard errors in parentheses
Based on 10,000 Monte Carlo replications

Table 5: Rejection rates for tests of nominal 5% size under various error processes with 10 groups

	$\rho = 0$	$\rho = 0.2$	$\rho = 0.4$	$\rho = 0.6$	$\rho = 0.8$	ρ varies with g
d=2	0.040	0.045	0.048	0.045	0.046	0.048
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
d=4	0.044	0.048	0.047	0.045	0.048	0.047
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
d=20	0.047	0.047	0.047	0.046	0.047	0.046
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
d=60	0.045	0.051	0.046	0.047	0.048	0.046
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
d=120	0.048	0.049	0.044	0.046	0.048	0.047
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)

Simulation standard errors in parentheses
Based on 10,000 Monte Carlo replications

Table 6: Rejection rates for tests of nominal 5% size under various error processes with 6 groups

	$\rho = 0$	$\rho = 0.2$	$\rho = 0.4$	$\rho = 0.6$	$\rho = 0.8$	ρ varies with g
d=2	0.055	0.058	0.058	0.058	0.052	0.051
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
d=4	0.055	0.058	0.056	0.056	0.051	0.054
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
d=20	0.053	0.059	0.057	0.057	0.051	0.054
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
d=60	0.056	0.061	0.058	0.057	0.053	0.053
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
d=120	0.056	0.060	0.057	0.057	0.052	0.052
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)

Simulation standard errors in parentheses
Based on 10,000 Monte Carlo replications

4.2 Power to detect real effects

Given the sobering conclusions reached by the burgeoning literature on inference in DiD designs, we believe our findings in the previous section are important: correct test size in the presence of grouped errors can be achieved with straightforward methods, under a wide range of data generating processes, even with few groups. However, we now show that power to detect real treatment effects with tests of correct size can be extremely low.

Rejection rates in table 7 indicate power to detect treatment effects of approximately 2% on earnings (precisely, effects of +0.02 on log(earnings)), as in Bertrand et al (2004) and Cameron et al (2008).[18] This is based upon the same Monte Carlo replications as table 1, but with a transformed dependent variable: we add $0.02T_{gt}$ to Y_{gt}.

The test that we have shown to have correct size detects a 2% effect with a probability of only about 1 in 5 even with all 50 US states. Power declines further with G and is just 7% when the data are sampled from 6 states - barely greater than the size of the test. The power of other tests is no better, allowing for the fact that they have larger (incorrect) size.

Table 7: Rejection rates for tests of nominal 5% size with a treatment effect of +0.02 in log(earnings) data

	G=50	G=20	G=10	G=6
Assume iid	0.678	0.552	0.473	0.454
	(0.007)	(0.007)	(0.007)	(0.007)
CRSE, t(G-1) critical values	0.212	0.108	0.076	0.070
	(0.006)	(0.004)	(0.004)	(0.004)
CRSE, N(0,1) critical values	0.225	0.140	0.116	0.136
	(0.006)	(0.005)	(0.005)	(0.005)
$\sqrt{G/(G-1)}$ residuals, t(G-1) critical values	0.220	0.118	0.079	0.073
	(0.006)	(0.005)	(0.004)	(0.004)
$\sqrt{G/(G-1)}$ residuals, N(0,1) critical values	0.236	0.148	0.121	0.142
	(0.006)	(0.005)	(0.005)	(0.005)
Wild cluster bootstrap-t	0.255	0.142	0.099	0.091
	(0.006)	(0.005)	(0.004)	(0.004)

Simulation standard errors in parentheses
Based on 5,000 Monte Carlo replications

We now consider the minimum effects that would be detected (i.e. that would lead to a rejection of the null of no effect) with given probabilities - a way of assessing statistical power first outlined by Bloom (1995).

[18] As with the Monte Carlos looking at rejection rates under the null, we are not considering the same set of inference methods as those authors. The techniques that we include and which neither of those papers included are those that use CRSEs formed using $\sqrt{G/(G-1)}$ -scaled residuals, and all variants which involve comparing t-statistics to critical values from a t distribution (i.e. rows 2 and 4-6 of table 7).

We vary power between 1% and 99% and compute the minimum detectable effects (MDEs) in each case. We continue just with the hypothesis test that we have demonstrated to have approximately correct size for G between 50 and 6: using CRSEs with $\sqrt{G/(G-1)}$ residuals and t_{G-1} critical values.

For a given level of power, x, the MDE is

$$MDE(x) = \hat{se(\hat{\beta})} \left[c_u - p^t_{1-x}\right], \tag{4}$$

where $\hat{se(\hat{\beta})}$ is the $\sqrt{G/(G-1)}$-corrected CRSE estimate, c_u is the upper critical value (the 97.5th percentile of the t(G-1) distribution), and p^t_{1-x} is the $(1-x)th$ percentile of the t-statistic under the null hypothesis of no treatment effect.

We proceed with the same Monte Carlo design underlying the results in Table 1. Monte Carlo replications provide us with an estimate of the distribution of the t-statistic under the null. They also provide repeated estimates of the $\sqrt{G/(G-1)}$-corrected CRSE: we plug each of those estimates into equation 4 in turn, and take the average. Due to the low computational intensity of this approach, we are able to use 100,000 Monte Carlo replications so that simulation error is negligible. We use equation 4 to compute MDEs for power ranging from 1% to 99%.

Figure 1 plots MDEs against power when the number of groups is 50, 20, 10 and 6. Even with earnings data on the entire US population (50 states), one would need a treatment effect of about 5% in order to detect it with 80% probability and 3.5% to have even a 50-50 chance. Power declines much further as G gets small. With a sample from 6 US states - by no means an extreme example in the applied DiD literature - the MDE on earnings is about 16% for 80% power and 11% for 50% power. In other words, for treatment effects of typically realistic magnitude it is unlikely that one would detect effects using a correctly sized test, particularly when the number of states is small.

The results using a binary employment indicator as the dependent variable lead to similar conclusions, as shown in figure 2. For 80% power, the MDE on the employment rate with data from all 50 states is about 2 percentage points, and this rises to 6.5 percentage points with 6 states.[19]

4.3 Increasing power with feasible GLS

The previous subsection argued that lack of power is a key problem in typical DiD designs. This suggests that there may be large gains from efforts to improve the efficiency of estimation. The serial correlation 'problem' inherent in a typical DiD study also suggests one way to go about this: exploit this feature of the data to increase efficiency using feasible GLS.

[19] The baseline employment rate in the sample is 67%.

Figure 1: Minimum detectable effects on log(earnings) using tests of size 0.05

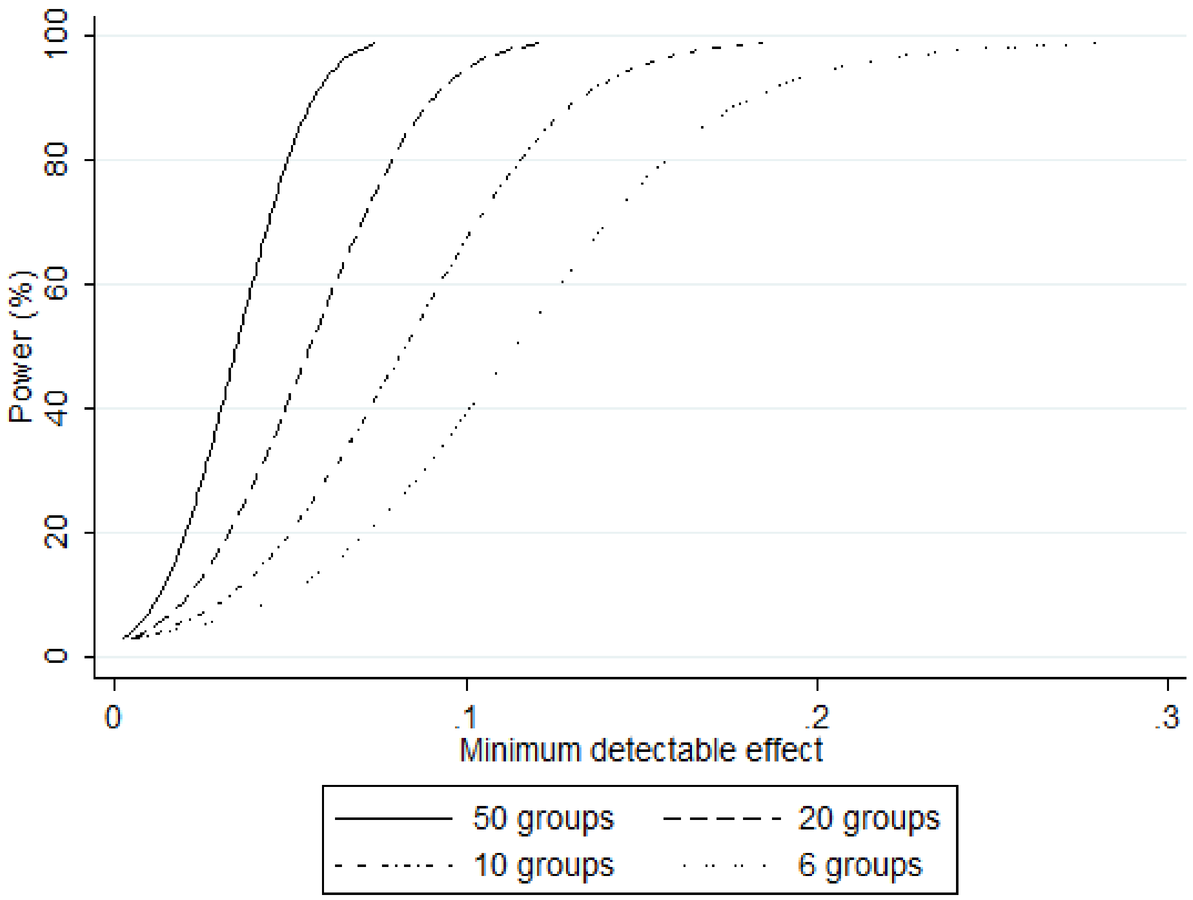
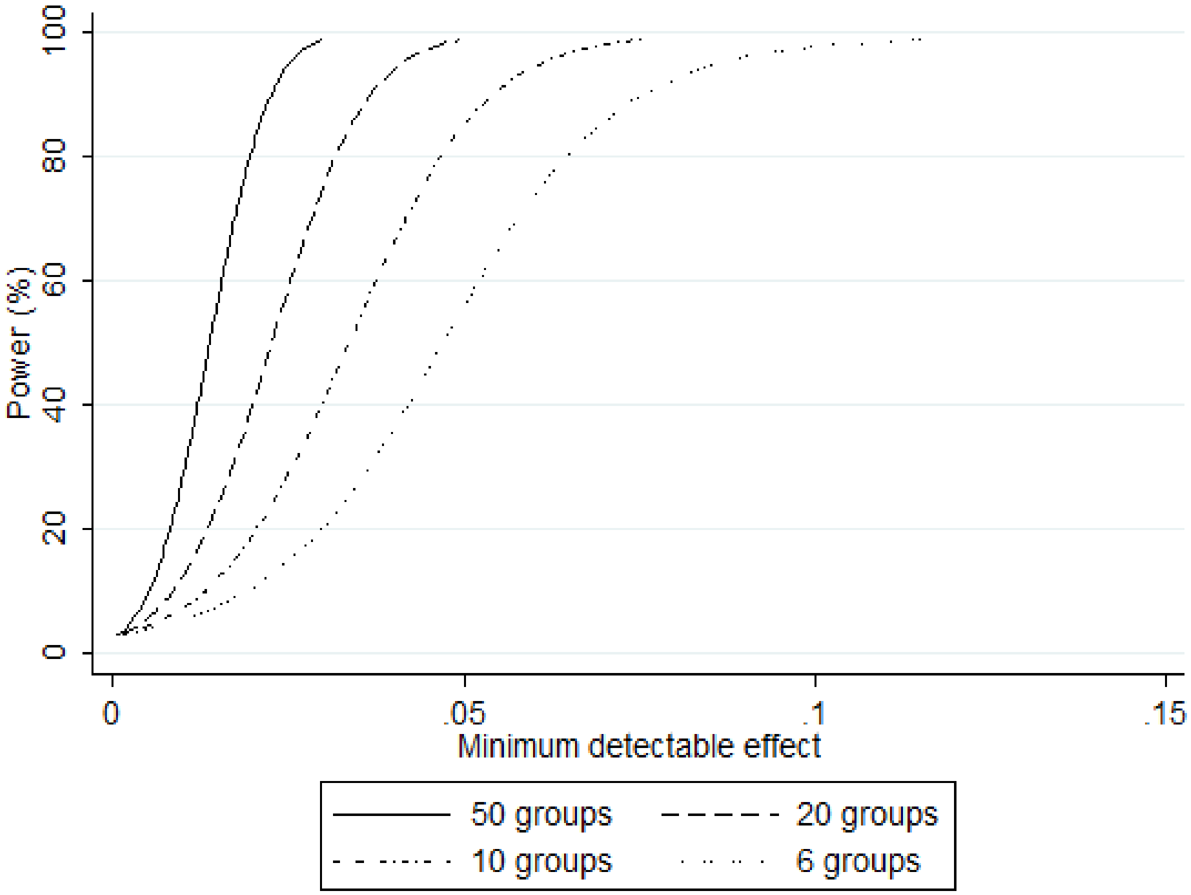


Figure 2: Minimum detectable effects on employment rates using tests of size 0.05



A natural way to proceed is to assume an AR(k) process for the group-time shocks. FGLS can then be implemented by estimating equation 2 using OLS, as before; estimating the k AR parameters using the OLS regression residuals; using those estimates to apply the standard GLS linear transformations to the variables entering equation 2; and estimating the analog of equation 2 on the transformed variables via OLS.

Two issues arise. First, estimates of the AR(k) parameters obtained by regressing OLS residuals on k lags are inconsistent with T fixed, due to the presence of fixed group effects (Nickell, 1981; Solon, 1984). Hansen (2007) derives a bias correction which is consistent as $G \to \infty$, and develops the asymptotic properties of a FGLS estimator which uses it. But with small G, this correction may not work well. Second, one may be worried about misspecification of the error process.

Neither of these issues affect the unbiasedness or consistency of the FGLS estimator. And it is highly likely that FGLS would still be more efficient than OLS - a Mahalanobis weighting matrix based on an incorrect parameterisation of the serial correlation process will normally still be closer to the optimal GLS weighting matrix than the identity matrix (which is what ordinary OLS uses). But test size will generally be compromised, because the ordinary formula for the FGLS standard error depends upon the weighting matrix.

However, test size might still be controlled using robust inference. As noted by Wooldridge (2006), the combination of FGLS estimation and robust inference is used relatively little in practice, but will often be a sensible way of realising efficiency gains without compromising test size. Hansen (2007) also considers this approach in the context of his FGLS procedure using bias-corrected estimates of the AR(k) parameters underlying the group-time error process. One simply plugs the FGLS residuals, rather than OLS residuals, into the formula for a cluster-robust variance matrix.

In the DiD context, the prevailing view is that the limitation of using cluster-robust inference is that its validity depends on having lots of groups. But one of the contributions of this paper has been to show that simple modifications to standard cluster-robust inference enable test size to be controlled, even with few groups. This suggests that it may be possible to use FGLS to improve power in DiD, whilst maintaining correctly sized tests, in a way that is robust to misspecification (or mis-estimation) of the error process, even with a small number of groups.

We now confirm this with our final set of simulations. We rerun the Monte Carlo simulations, this time implementing FGLS (rather than OLS). We assume an AR(2) process for the group-time shocks. We estimate the 2 AR parameters in two ways. First, we simply regress the residuals from OLS estimation of equation 2 on two lags. With fixed T and fixed group effects, this produces inconsistent estimates of the AR parameters. Second, we apply to these estimates the bias correction derived by Hansen (2007). This correction is consistent as G goes to infinity. We label these “FGLS” and “BC-FGLS” respectively. In both cases, we explore what happens when the estimator is used with and without cluster-robust inference. We use the cluster-robsut technique that we have shown to work well even when G is small: using CRSEs with $\sqrt{G/(G-1)}$ residuals and t_{G-1} critical values.

Table 8 shows rejection rates under a true null hypothesis (no treatment effect). The first row reiterates the good size properties of OLS estimation combined with CRSEs that use $\sqrt{G/(G-1)}$-scaled residuals and t_{G-1} critical values (i.e. it repeats row four of Table 1). The second row shows that FGLS without the bias correction and without robust inference gives tests at least double the nominal size. Note however that even this size distortion is considerably smaller than with OLS without robust inference. Hansen's bias correction for the estimated parameters of the AR process reduces this size distortion, though still returns rejection rates that are too high without robust inference (fourth row), particularly when G is small. This is what we would expect, because the bias correction is consistent as $G \rightarrow \infty$. But, as with OLS, the size of the test can be controlled using robust inference, even with few groups, using the methods described earlier in this paper: when doing this, the rejection rate remains within about 1 percentage point of the nominal test size. This is true both for FGLS and BC-FGLS (third and fifth rows).

Table 9 turns attention to power, showing rejection rates when there is a 2% treatment effect on earnings. Again, the first row reiterates the earlier finding that OLS estimation combined with a correctly sized test provides very low power (i.e. it repeats the fourth row of Table 7 . As Hansen (2007) showed in the case where G = 50, the FGLS procedures deliver substantial improvements in power. Combined with robust inference which delivers the correct test size, BC-FGLS detects the treatment effect with 37% probability, whereas OLS detects it with 22% probability. Table 9 shows that FGLS also delivers substantial proportionate power gains relative to OLS with smaller G, though from a much lower base: with G = 6, power is 7.3% using OLS and 10.3% using FGLS. Using Hansen's bias correction delivers a little more power than 'ordinary' FGLS, and more so when G is larger (again, as expected theoretically).

Figure 3 illustrates the power gains more comprehensively, plotting MDEs against power and comparing the results for OLS and BC-FGLS estimation with varying numbers of groups (always combined with cluster-robust inference, so that test size is correct). The power gains from BC-FGLS are substantial.

Tables 10 and 11 and Figure 4 repeat this analysis for the case where the outcome of interest is a binary employment status indicator. This confirms that these conclusions all hold qualitatively in that very different generating process: FGLS delivers substantial power gains over OLS, and this can be done whilst controlling test size, even with few groups. We also re-ran the earlier robustness checks where we simulate our own state-time earnings shocks according to an AR(1) process, varying the degree of serial correlation and non-normality in the white noise. Again, the results were robust.[20]

[20] These are not shown in the paper but are available from the authors on request.

Table 8: Rejection rates for tests of nominal 5% size with placebo treatments in log(earnings) data

	G=50	G=20	G=10	G=6
OLS, $\sqrt{G/(G-1)}$-CRSEs, t(G-1) critical values	0.042	0.046	0.050	0.049
	(0.003)	(0.003)	(0.003)	(0.003)
FGLS	0.100	0.106	0.112	0.126
	(0.004)	(0.004)	(0.004)	(0.005)
FGLS, $\sqrt{G/(G-1)}$-CRSEs, t(G-1) critical values	0.047	0.053	0.057	0.061
	(0.003)	(0.003)	(0.003)	(0.003)
BC-FGLS	0.068	0.077	0.088	0.099
	(0.004)	(0.004)	(0.004)	(0.004)
BC-FGLS, $\sqrt{G/(G-1)}$-CRSEs, t(G-1) critical values	0.049	0.057	0.057	0.064
	(0.003)	(0.003)	(0.003)	(0.003)

Simulation standard errors in parentheses

Based on 5,000 Monte Carlo replications

Table 9: Rejection rates for tests of nominal 5% size with a treatment effect of +0.02 in log(earnings) data

	G=50	G=20	G=10	G=6
OLS, $\sqrt{G/(G-1)}$-CRSEs, t(G-1) critical values	0.220	0.118	0.079	0.073
	(0.006)	(0.005)	(0.004)	(0.004)
FGLS	0.460	0.275	0.206	0.191
	(0.007)	(0.006)	(0.006)	(0.006)
FGLS, $\sqrt{G/(G-1)}$-CRSEs, t(G-1) critical values	0.348	0.175	0.110	0.096
	(0.007)	(0.005)	(0.004)	(0.004)
BC-FGLS	0.395	0.224	0.163	0.150
	(0.007)	(0.006)	(0.005)	(0.005)
BC-FGLS, $\sqrt{G/(G-1)}$-CRSEs, t(G-1) critical values	0.365	0.187	0.118	0.103
	(0.007)	(0.006)	(0.005)	(0.004)

Simulation standard errors in parentheses

Based on 5,000 Monte Carlo replications

Figure 3: Minimum detectable effects on log(earnings) using tests of size 0.05

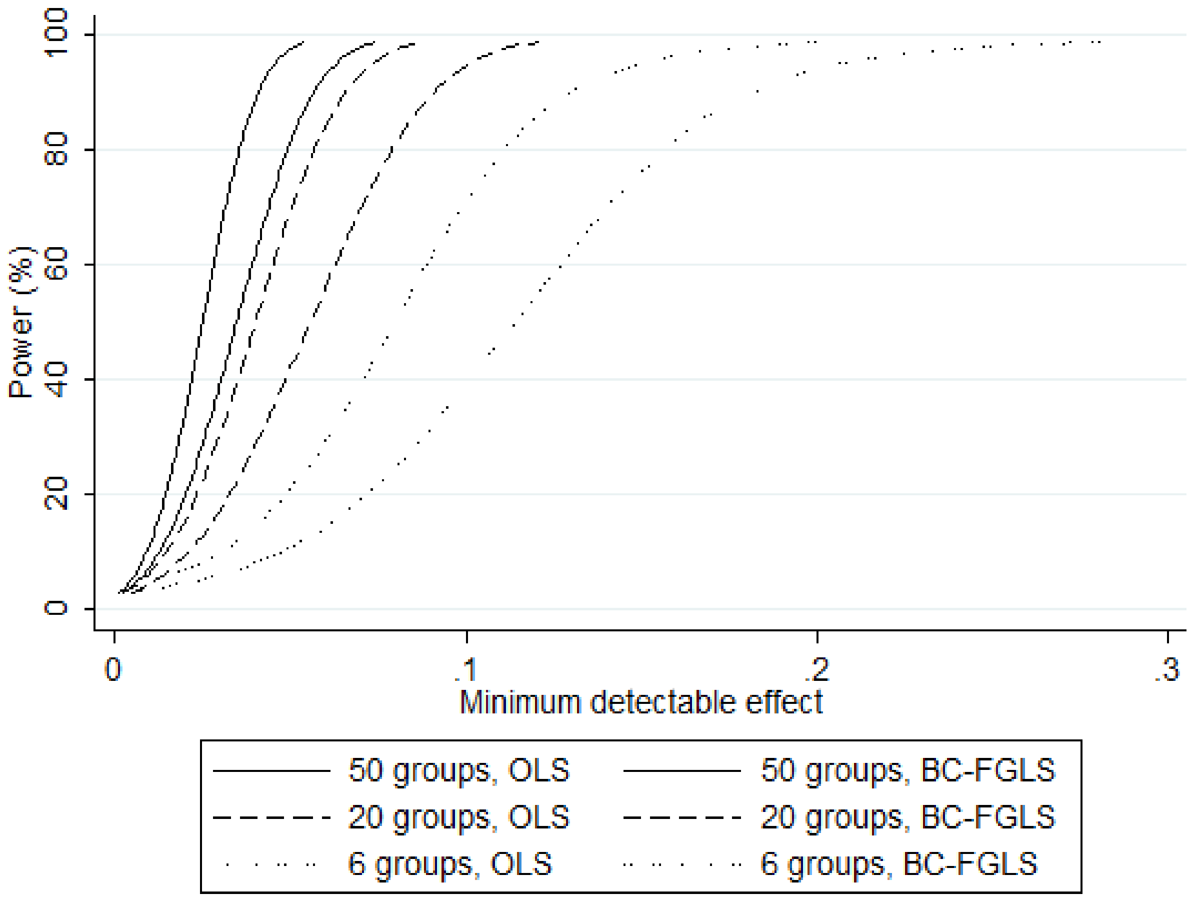


Table 10: Rejection rates for tests of nominal 5% size with placebo treatments in employment data

	G=50	G=20	G=10	G=6
OLS, $\sqrt{G/(G-1)}$-CRSEs, t(G-1) critical values	0.044 (0.003)	0.043 (0.003)	0.047 (0.003)	0.059 (0.003)
FGLS	0.165 (0.005)	0.181 (0.005)	0.182 (0.005)	0.213 (0.006)
FGLS, $\sqrt{G/(G-1)}$-CRSEs, t(G-1) critical values	0.041 (0.003)	0.046 (0.003)	0.048 (0.003)	0.061 (0.003)
BC-FGLS	0.124 (0.005)	0.140 (0.005)	0.146 (0.005)	0.179 (0.005)
BC-FGLS, $\sqrt{G/(G-1)}$-CRSEs, t(G-1) critical values	0.042 (0.003)	0.046 (0.003)	0.049 (0.003)	0.063 (0.003)

Simulation standard errors in parentheses

Based on 5,000 Monte Carlo replications

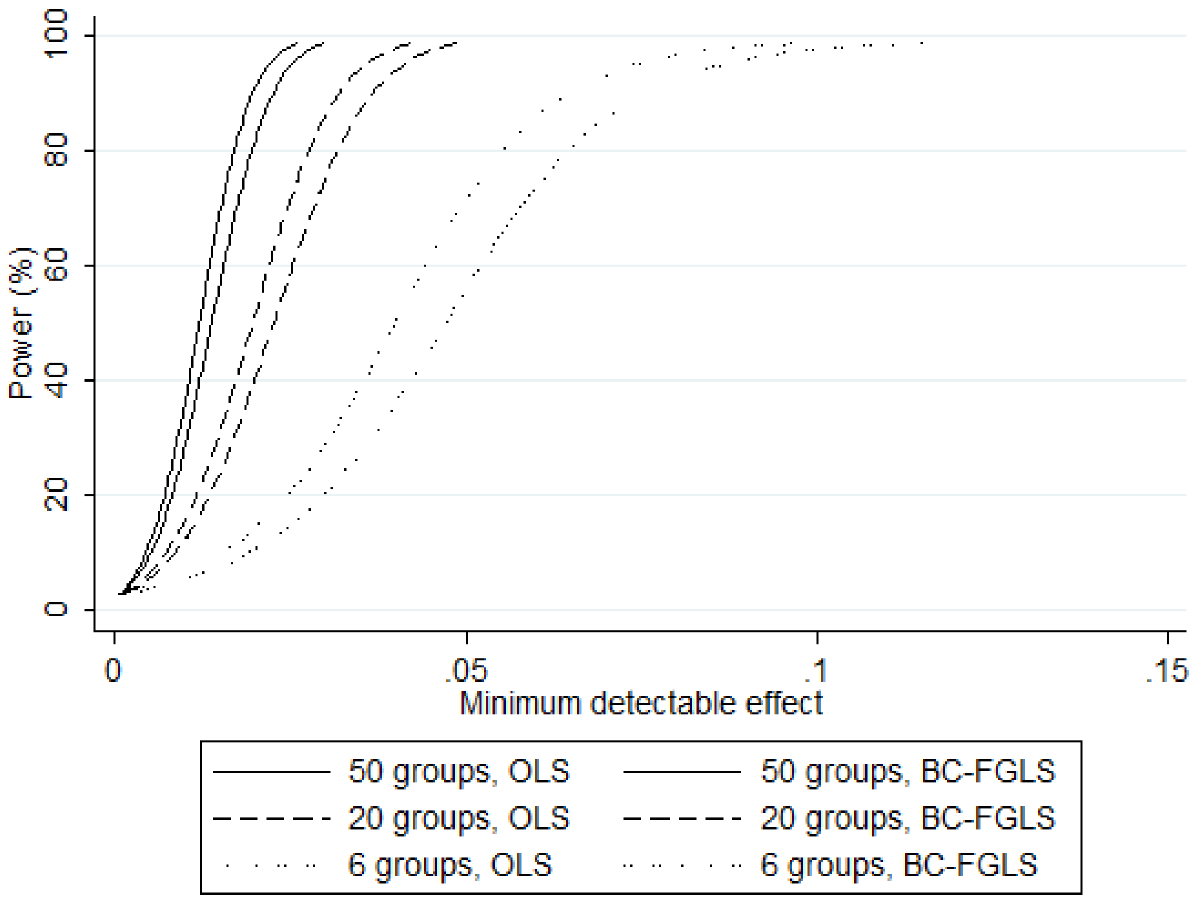
Table 11: Rejection rates for tests of nominal 5% size with a treatment effect of +0.02 in employment data

	G=50	G=20	G=10	G=6
OLS, $\sqrt{G/(G-1)}$-CRSEs, t(G-1) critical values	0.824	0.417	0.222	0.160
	(0.005)	(0.007)	(0.006)	(0.005)
FGLS	0.975	0.770	0.552	0.463
	(0.002)	(0.006)	(0.007)	(0.007)
FGLS, $\sqrt{G/(G-1)}$-CRSEs, t(G-1) critical values	0.895	0.509	0.264	0.189
	(0.004)	(0.007)	(0.006)	(0.006)
BC-FGLS	0.970	0.738	0.511	0.420
	(0.002)	(0.006)	(0.007)	(0.007)
BC-FGLS, $\sqrt{G/(G-1)}$-CRSEs, t(G-1) critical values	0.904	0.531	0.279	0.198
	(0.004)	(0.007)	(0.006)	(0.006)

Simulation standard errors in parentheses

Based on 5,000 Monte Carlo replications

Figure 4: Minimum detectable effects on employment rates using tests of size 0.05



5 Summary and Conclusion

This paper has contributed to a growing literature on inference in difference-in-differences designs with grouped errors. The literature has emphasised difficulties in obtaining correctly sized hypothesis tests, particularly with few groups.

Using Monte Carlo evidence, we have made three main points. First, it is possible to obtain tests of the correct size, even with few groups, using methods that are very straightforward to implement. In fact, when one uses cluster-robust inference in STATA, these methods are implemented by default. Second, the main problem in difference-in-differences designs with grouped errors is instead low power to detect real effects. Third, feasible GLS estimation combined with robust inference methods can increase power considerably whilst maintaining correct test size - again, even with few groups. These findings have proven robust to a wide range of data generating processes.

We therefore recommend that applied researchers think seriously about efficiency, rather than just consistency and test size, when using difference-in-differences designs.

References

[1] Angrist, J. and J. Pischke, *Mostly Harmless Econometrics: An Empiricist's Companion*, Princeton University Press, Princeton, NJ (2009).

[2] Barrios,T. R. Diamond, G.W. Imbens and M. Kolesár, "Clustering, Spatial Correlations, and Randomization Inference", *Journal of the American Statistical Association* 107:498 (2012), 578-591.

[3] Bell, R. M., and D. F. McCaffrey, "Bias Reduction in Standard Errors for Linear Regression with Multi-Stage Samples", *Survey Methodology* 28:2 (2002), 169–179.

[4] Bertrand, M., E. Duflo, and S. Mullainathan, "How Much Should We Trust Differences-in-Differences Estimates?" *Quarterly Journal of Economics* 119 (2004), 249–275.

[5] Bester, A.C., T.G. Conley and C.B. Hansen, "Inference with dependent data using cluster covariance estimators", *Journal of Econometrics* 165:2 (2011), 137-151.

[6] Bloom, H. S. "Minimum Detectable Effects: A simple way to Report the Statistical Power of Experimental Designs", *Evaluation review* 19:5 (1995), 547-556.

[7] Cameron, A. C., J. G. Gelbach, and D. L. Miller, "Bootstrap-Based Improvements for inference with clustered errors", *The Review of Economics and Statistics* 90:3 (2008), 414-427.

[8] Donald, S. G., and K. Lang, "Inference with Difference-in-Differences and Other Panel Data", *The Review of Economics and Statistics* 89:2 (2007), 221–233.

[9] Hansen, C., "Generalized Least Squares Inference in Panel and Multilevel Models with Serial Correlation and Fixed Effects," *Journal of Econometrics* 140:2 (2007), 670-694.

[10] Imbens, G. and M. Kolesár, "Robust standard errors in small samples: some practical advice", National Bureau of Economic Research Working Paper 18478 (2012).

[11] Liang, K.-Y., and S. L. Zeger, "Longitudinal Data Analysis Using Generalized Linear Models", *Biometrika* 73 (1986), 13–22.

[12] Moulton, B. R., "An Illustration of a Pitfall in Estimating the Effects of Aggregate Variables on Micro Units", *The Review of Economics and Statistics* 72 (1990), 334–338.

[13] Nickell, S., "Biases in dynamic models with fixed effects", Econometrica 49:6 (1981), 1417–1426.

[14] Solon, G., “Estimating autocorrelations in fixed effects models”, National Bureau of Economic Research Technical Working Paper 32 (1984).

[15] Wooldridge, J. M., “Cluster-Sample Methods in Applied Econometrics”, *American Economic Review* 93 (2003), 133–138.

[16] Wooldridge, J.M., “Cluster-Sample Methods in Applied Econometrics: An Extended Analysis,” mimeograph (2006), Michigan State University Department of Economics.