Mike Brewer[1] / Thomas F. Crossley[2] / Robert Joyce[3]

Inference with Difference-in-Differences Revisited

[1] Institute for Social and Economic Research, University of Essex, Colchester, Essex CO4 3SQ, UK
[2] Department of Economics, University of Essex, Colchester, Essex CO4 3SQ, UK, E-mail: tcross@essex.ac.uk
[3] Institute for Fiscal Studies, London WC1E 7AE, UK

Abstract:
A growing literature on inference in difference-in-differences (DiD) designs has been pessimistic about obtaining hypothesis tests of the correct size, particularly with few groups. We provide Monte Carlo evidence for four points: (i) it is possible to obtain tests of the correct size even with few groups, and in many settings very straightforward methods will achieve this; (ii) the main problem in DiD designs with grouped errors is instead low power to detect real effects; (iii) feasible GLS estimation combined with robust inference can increase power considerably whilst maintaining correct test size – again, even with few groups, and (iv) using OLS with robust inference can lead to a perverse relationship between power and panel length.

Keywords: cluster robust, difference in differences, feasible GLS, hypothesis test, power

JEL classification: C12, C13, C21

DOI: 10.1515/jem-2017-0005

1 Introduction

In labor economics and other empirical, policy-oriented fields, difference-in-differences (DiD) designs are an extremely common way of estimating the effects of policies or programs (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, Duflo, and Mullainathan 2004; Cameron and Miller 2015). 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 (Bertrand, Duflo, and Mullainathan 2004; Cameron, Gelbach, and Miller 2008; Angrist and Pischke 2009).

In this paper we report evidence from Monte Carlo simulations that emphasises a different conclusion. We make four main points. First, our simulations demonstrate that in many typical DiD settings tests of the correct size can 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); and in settings where this works less well, a bootstrap-based approach highlighted by other authors (e.g. Cameron, Gelbach, and Miller 2008; Webb 2013) provides a reliable alternative. All 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 test size. Third, our simulations show that substantial gains in power can be achieved using feasible GLS. Moreover, our simulation results suggest that the the combination of feasible GLS and cluster-robust inference can control test size, even if the parametric assumptions about the error process implicit in feasible GLS estimation are violated, and even with few groups. Fourth, using OLS with robust inference can lead to a perverse relationship between power and panel length. This is another reason to favour feasible GLS. In summary we highlight the need for applied researchers using DiD designs to pay careful attention not just to consistency and test size, but also to the efficiency of their estimators, and we recommend the use of feasible GLS combined with cluster-robust inference as a solution to this problem.

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; Cameron and Miller 2015). Furthermore, treatment status is typically highly serially correlated. In fact, in the most common case on which we focus in this

Thomas F. Crossley is the corresponding author.

paper, treatment is an "absorbing state": once a group is treated, it remains treated in all subsequent periods. This means that serially correlated error terms are likely to have a large impact on the true level of precision with which treatment effects are estimated. In a well-cited Monte Carlo study using US earnings data, Bertrand, Duflo, and Mullainathan (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. So, for example, when evaluating a labor market policy implemented in certain regions from a particular point in time onwards, a researcher should worry both that people in the same region at the same time are affected by common labor market shocks (unrelated to the policy) and that these regional shocks are serially correlated.

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, Duflo, and Mullainathan (2004) and Cameron, Gelbach, and Miller (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.

The crucial finding of Bertrand et al. and others – that inference can go badly wrong in DiD unless one is very careful – is confirmed once again in our experiments. But our simulations also show that a modification to the standard cluster-robust inference procedure described above can dramatically improve test size with few clusters. One can apply a 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. This is straightforward and 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 this procedure by default.[2] When this is done, our simulations show that true test size is within about one percentage point of nominal test size with 50, 20, 10 or 6 groups. This result also holds under a wide range of data generating processes. The key situation in which the method is unreliable is when there is a large imbalance between the numbers of treatment and control groups.

Various alternative techniques for achieving correct test size have been proposed and/or tested (Bertrand, Duflo, and Mullainathan 2004; Donald and Lang 2007; Cameron, Gelbach, and Miller 2008; Bester, Conley, and Hansen 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, Gelbach, and Miller 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. It has also been shown to be quite robust to large imbalances between the numbers of treatment and control groups (Mackinnon and Webb 2017), and hence provides an important alternative to the simpler method described above in such situations. But it is less trivial to implement and computationally more intensive.

Our second point is that, while it is generally not difficult to obtain the correct size, power to detect real effects is a serious concern. When we use the methods above to implement correctly sized hypothesis tests, our simulations suggests that DiD designs can have very low power. This problem is very severe with few groups. 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 5% would be detected with only 17% 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.

More positively, our experiments also demonstrate that substantial gains in power can be achieved using feasible GLS. In particular, with a moderate time series dimension of at least about 10 time periods, one will often be able to increase power by modeling the serial correlation of unobservables inherent in typical DiD designs. A bias-correction for feasible GLS due to Hansen (2007) reduces, but does not eliminate, test size distortion, particularly with small numbers of groups. However, our simulations demonstrate that test size can 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 feasbile GLS estimation, by using the straightforward cluster-robust inference technique described above. This is why we recommend the use of the combination of FGLS (with or without the Hansen correction) and cluster-robust techniques in DiD applications. Furthermore, our simulations also show that OLS estimation is susceptible to a perverse (negative) relationship between power and panel length, and we explain the econometric reason for this. This does not happen with feasible GLS.

The paper proceeds as follows. Section 2 describes the standard econometric setup in DiD designs that we consider, and discusses possible solutions to the inference problems that can arise in this setting. Section 3 details the Monte Carlo design we use to test different inference methods. Section 4 presents and discusses the results of our Monte Carlo simulations. To assist applied researchers who wish to follow the feasbile GLS strategy our experiments support, we provide a STATA ado file that implements this, including the Hansen (2007) bias correction. This is described in Section 5. Finally, Section 6 summarizes and concludes.

2 Approaches to Inference in a Difference-in-Differences Design

We consider the standard linear DiD model

$$y_{igt} = \alpha_g + \delta_t + \beta T_{gt} + \gamma w_{igt} + v_{igt}, \tag{1}$$

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, and v_{igt} is the unobserved individual-level earnings shock.

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). Hence we assume that the OLS DiD estimator based on equation 1 is unbiased, i.e. $E(v_{igt}|\alpha_g, \delta_t, T_{gt}, w_{igt}) = 0$ 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 v_{igt} may not be iid within groups. Some of the variation in v_{igt} may occur at the group-time level, i.e. $v_{igt} = \varepsilon_{gt} + \xi_{igt}$. The DiD estimator is therefore effectively attempting to distinguish between the effects of a group-time level treatment and between-group differences in the evolution of group-time shocks. In addition, the group-time shocks may be serially correlated. The net result is both cross-sectional and serial correlation in within-group shocks. This is highly likely in many DiD applications, including the primary example used in this paper (and much of the previous literature) where groups are US states and the outcome of interest is earnings. The challenge, then, is to quantify accurately the additional uncertainty about β that this causes.

Given the setup described, the computation of $\hat{\beta}^{OLS}$ from a micro-data regression using equation 1 is equivalent to a two-step procedure. First, run a regression using the micro-data of y_{igt} on w_{igt}, and take the mean residual within each group-time cell. Denote these estimated covariate-adjusted group-time means as $\hat{Y}_{gt}$.[3] Then, since

$$\hat{Y}_{gt} = \alpha_g + \delta_t + \beta T_{gt} + \varepsilon_{gt} + (\hat{Y}_{gt} - Y_{gt}), \tag{2}$$

$\hat{\beta}^{OLS}$ can be obtained from a second-stage regression of $\hat{Y}_{gt}$ on group effects, time effects and the (group-time level) treatment indicator. If state-time cell sizes are large, then estimation error in $\hat{Y}_{gt}$ can essentially be ignored: the composite error term in equation 2 is approximately equal to ε_{gt}, the group-time shock. Equation 2 highlights that, in that case, the true precision of $\hat{\beta}^{OLS}$ depends almost entirely on the number of group-time cells rather than the number of individual-level observations.[4]

As we explain fully in the next section, we first aggregate the data to the state-time level in this way and ignore any estimation error (i.e. we proceed as though $\hat{Y}_{gt} = Y_{gt}$). We then estimate equation 2 and perform inference about β. As the first-stage aggregation accounts for cross-sectional error correlation within states, the key remaining issues for inference are the fact that the state-time shocks may be serially correlated and that there are a finite number of states. A number of methods have been proposed to account for these two issues. We describe them below, as well as some modest proposals of our own.

2.1 Standard Cluster-Robust Inference

Our starting point is the standard OLS estimator of the standard error of $\hat{\beta}$, comparing the resulting t-statistic to standard normal critical values. This effectively assumes that the ε_{gt} in equation 2 are iid, i.e. it ignores serial correlation.

We then look at several ways of performing 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 estimator 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, model over-fitting means that residuals will tend to be smaller in magnitude and less correlated within clusters than the true errors, meaning that CRSEs calculated using equation 3 will tend to be biased downwards. Any small-G bias is larger when the distribution of regressors is skewed: in the DiD context considered here, this is when there is an imbalance between the numbers of treatment and control groups (see Mackinnon and Webb 2017).

2.2 Bias Corrections for Cluster-Robust Inference with Few Clusters

A typical way of attempting to reduce small-G bias (or, under special circumstances, to eliminate it) 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.[5] 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, but we use the additional $\sqrt{(N-1)/(N-k)}$ degrees of freedom adjustment so that our results can be taken as an exact test of how STATA's default performs). This scaling of residuals leads to an unbiased CRSE estimator only under very special circumstances (see Bell and McCaffrey 2002) and so should be viewed generally as a bias-*reducing* correction. The same applies to a second, data-dependent scaling of u_g proposed in Bell and McCaffrey (2002),[6] and extended in Imbens and Kolesár (2012) and Cameron and Miller (2015) investigate the Imbens and Kolesár (2012) adjustment in a set-up that is very similar to the one in this paper, and they show that the degree of freedom adjustment is of minimal importance in balanced DiD designs.

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 $G \to \infty$, 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, in general it does not have an exact theoretical justification.[7] However, recent work by Bester, Conley, and Hansen (2011) provides theoretical justification in certain small-G 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.[8] This result relies on homogeneity requirements, including the need for regressor matrices to converge to the same limit within each group. This would be violated in the canonical DiD setup with a binary treatment indicator where some control groups are never treated.[9] But whether the Bester et al. approach extends well (in terms of getting the test size right) to the standard DiD case in practice is an empirical question which our simulations shed light on.

2.3 Bootstraps

With few groups, an alternative to relying on asymptotic results (such as normality of the t-statistic) or on small sample corrections is to recover the distribution of the test statistic empirically via a bootstrap. Cameron, Gelbach, and Miller (2008) found the wild cluster bootstrap-t procedure to be the best (in terms of test size) of a large number of inference techniques in settings with few groups. In their implementation of the bootstrap (which we follow), they resampled clusters of residuals obtained from regressions which impose the null hypothesis, and scaled the resampled residuals by a constant drawn from a 2-point distribution: 1 and −1, each with probability 0.5. This outperformed 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 it took critical values from the standard normal distribution, rather than from the t distribution.

2.4 Modeling the Error Process Using GLS

The final approach to dealing with the serial correlation in the group-time shocks is to use feasible Generalized Least Squares (GLS): this effectively exploits knowledge of this feature of the data to increase efficiency. 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 lag 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 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 this correction may not work well with small G. Second, one may be worried about misspecification of the error process.

However, neither of these issues affect the consistency of the FGLS estimator. And it is likely that FGLS would still be more efficient than OLS: a weighting matrix based on an incorrect parametrisation of the serial correlation process will often still be closer to the optimal GLS weighting matrix than the identity matrix used by standard OLS.

On the other hand, test size would generally be compromised, because the ordinary formula for the FGLS standard error depends upon the weighting matrix. But robust inference may offer a way to control test size. As noted more generally 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 realizing efficiency gains without compromising test size.[10] One simply plugs the FGLS residuals, rather than OLS residuals, into the formula for a cluster-robust variance matrix.

Hansen (2007) considers combining CRSEs with his FGLS procedure, using bias-corrected estimates of the parameters underlying a group-time error process assumed to be AR, for the case where $G = 50$. The concern with fewer groups would be that test size can not be made robust to misspecification of the error process in this way, because cluster-robust inference relies on having many clusters.This leads us back to the question discussed above and explored in our simulations: how well can bias corrections for cluster-robust inference with few clusters do in controlling test size?

3 Experimental Design

We follow Bertrand, Duflo, and Mullainathan (2004), Cameron, Gelbach, and Miller (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).[11] We focus on log(earnings) as the dependent variable.

Our control variables are a quartic in age. As in the aforementioned papers, we first aggregate the data to the state-time level in the way just described and ignore any estimation error from this procedure (i.e. we proceed as though $\hat{Y}_{gt} = Y_{gt}$).[12] We then estimate equation 2. As the first-stage aggregation accounts for cross-sectional error correlation within states, the key remaining issues for inference are the fact that the state-time shocks may be serially correlated and that there are a finite number of states.[13]

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."[14] In treating exactly half of the states, we follow the main approach in Bertrand, Duflo, and Mullainathan (2004) and Cameron, Gelbach, and Miller (2008). This is the most favorable possible choice in terms of the resulting precision of treatment effect estimates, as it maximizes between-group variation in treatment status. For all treated states in each Monte Carlo replication, the placebo treatment is applied in the same randomly chosen year and in all subsequent years.[15] We estimate the "effect" of this placebo treatment by estimating equation 2. We initially use OLS, and 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 also vary the fraction of groups that are treated, to explore robustness to unbalanced designs.

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 then look at power by reporting how often the null of no effect is rejected when there are real treatment effects of various sizes. We do this only for techniques that we have shown to work well in controlling test size. Nevertheless, to ensure that we are comparing the power of hypothesis tests which have *exactly* the same size, we adopt the procedure suggested by Davidson and MacKinnon (1998): the nominal significance level used to determine whether to reject the null hypothesis is that which gives a test of true size 0.05.

We also 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 given probabilities. 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.[16] 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 (see later). We therefore recover the entire relationship between power and MDEs. We do this for DiD designs with varying numbers of groups.

We then explore whether power can be improved by using FGLS rather than OLS estimation. We rerun the Monte Carlo simulations, this time implementing FGLS assuming an AR(2) process for the group-time shocks, homogeneous across states. 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.

Finally, we show how the performance of OLS and FGLS/BC-FGLS compare when the error process assumed in FGLS estimation is misspecified, by simulating state-time shocks according to a AR(2) process that is heterogeneous across states and a (homogeneous) MA(1) process, and when the length of the available panel is shorter, varying T between 10 and 30.

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 at which the null of no effect is rejected when generating placebo treatments, estimating equation 2 by OLS, and using different methods to conduct 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 $se(\hat{r}) = \sqrt{\hat{r}(1-\hat{r})/4999}$.

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	0.366	0.394	0.396	0.398
	(0.007)	(0.007)	(0.007)	(0.007)
CRSE, N(0,1) critical values	0.058	0.078	0.127	0.228
	(0.003)	(0.004)	(0.005)	(0.006)
$\sqrt{G/(G-1)}$-CRSEs, t(G − 1) critical values	0.046	0.039	0.039	0.053
	(0.003)	(0.003)	(0.003)	(0.003)
Wild cluster bootstrap-t	0.038	0.060	0.044	0.060
	(0.001)	(0.002)	(0.003)	(0.003)

The table shows the proportion of the time that the null hypothesis of no treatment effect was rejected in 5000 Monte Carlo simulations. Simulation standard errors are reported in parentheses. The simulations resample groups (US states) from CPS-MORG data, having imposed the sampling restrictions described in the text. The treatment parameter has a true coefficient of zero. G is the number of sampled states. Data from 1979 to 2008 inclusive are sampled (i.e. T = 30). Regressions are run on aggregated state-year data. The different inference methods used are discussed in the text.

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, Duflo, and Mullainathan (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 ≥ 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 no more than about 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 as in Cameron, Gelbach, and Miller (2008).[17] We essentially replicate previous findings: it also performs well relative to most tested alternatives.

In summary, the simulations reported in Table 1 suggest that tests with very little size distortion 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. Happily, this is the default procedure used by STATA if one specifies the "vce(cluster *clustvar*)" option.

The accuracy of robust variance matrix estimators in finite samples depends on the skewness of the distribution of regressors as well as sample size [as illustrated in Monte Carlo work by Imbens and Kolesár (2012)]. With cluster-robust standard errors in the DiD context considered here, a researcher should pay attention not only to the number of groups but also to whether treatment status is skewed, i.e. whether the number of treated groups is similar to the number of controls.

With unbalanced designs, inference based on cluster-robust standard errors should produce tests of correct size less reliably than when the numbers of treatment and control groups are equal (as in the Monte Carlo experiments up to now). This has been illustrated recently by Mackinnon and Webb (2017). Scaling the residuals that are plugged into the standard CRSE formula by $\sqrt{G/(G-1)}$ (and using a t_{G-1} reference distribution) may not be a sufficient small-G correction in the presence of this imbalance.

Table 2 and Table 3 repeat the simulations presented in Table 1 but with varying degrees of imbalance, for the cases with 10 and 50 groups respectively. $G1$ denotes the number of treated groups. The first columns repeat the results for the balanced designs shown in Table 1, and subsequent columns reduce the number of treated groups. When G = 10, the simple method that works well under a wide a wide range of balanced designs – using CRSEs with $\sqrt{G/(G-1)}$-scaled residuals and a t_{G-1} reference distribution – continues to achieve correct test size with $G1 = 4$, but over-rejects a little when $G1$ drops to 3, and more severely when it drops to 2. The wild cluster bootstrap-t continues to work well with $G1 = 3$ but also performs poorly with $G1 = 2$ (with significant *under*-rejection). When G = 50, both methods continue to produce tests of the close to the right size when $G1$ drops as low as 10. With $G1 = 5$ the bootstrap is needed to conduct approximately accurate inference, while the simpler method over-rejects.

Table 2: Rejection Rates for Tests of Nominal 5% Size with Placebo Treatments in Log-Earnings Data with 10 Groups.

	G1 = 5	G1 = 4	G1 = 3	G1 = 2
$\sqrt{G/(G-1)}$-CRSEs, t(G − 1) critical values	0.039	0.046	0.071	0.137
	(0.003)	(0.003)	(0.004)	(0.005)
Wild cluster bootstrap-t	0.044	0.052	0.053	0.024
	(0.003)	(0.003)	(0.003)	(0.002)

The table shows the proportion of the time that the null hypothesis of no treatment effect was rejected in 5000 Monte Carlo simulations. Simulation standard errors are reported in parentheses. The simulations resample groups (US states) from CPS-MORG data, having imposed the sampling restrictions described in the text. The treatment parameter has a true coefficient of zero. G1 denotes the number of treated groups. Data from 1979 to 2008 inclusive are sampled (i.e. T = 30). Regressions are run on aggregated state-year data. The different inference methods used are discussed in the text.

Table 3: Rejection Rates for Tests of Nominal 5% Size with Placebo Treatments in Log-Earnings Data with 50 Groups.

	G1 = 25	G1 = 15	G1 = 10	G1 = 5
$\sqrt{G/(G-1)}$-CRSEs, t(G − 1) critical values	0.046	0.044	0.057	0.109
	(0.003)	(0.003)	(0.003)	(0.004)
Wild cluster bootstrap-t	0.038	0.045	0.039	0.052
	(0.001)	(0.001)	(0.001)	(0.002)

The table shows the proportion of the time that the null hypothesis of no treatment effect was rejected in 5000 Monte Carlo simulations. Simulation standard errors are reported in parentheses. The simulations resample groups (US states) from CPS-MORG data, having imposed the sampling restrictions described in the text. The treatment parameter has a true coefficient of zero. G1 denotes the number of treated groups. Data from 1979 to 2008 inclusive are sampled (i.e. T = 30). Regressions are run on aggregated state-year data. The different inference methods used are discussed in the text.

In summary these simulations confirm that, if the imbalance between treatment and control groups is large enough, using CRSEs with $\sqrt{G/(G-1)}$ -scaled residuals and a t_{G-1} reference distribution can lead to over-rejection. But as emphasised in Mackinnon and Webb (2017), the wild cluster bootstrap-t is relatively robust to this imbalance. Hence, although a slightly less trivial procedure is necessary to get test size right with very unbalanced designs, it can be done even with few groups.[18]

4.1.1 Robustness

We undertook a number of further experiments to explore the robustness of these findings. First, we considered the case where a binary employment indicator is the dependent variable (in a linear probability model.) Together with log(earnings), this covers the two most common outcomes of interest in DiD studies, according to a survey of the applied literature in Bertrand, Duflo, and Mullainathan (2004). The results for employment are very similar to those reported below for earnings, and the full results are available in an earlier working paper (Brewer, Crossley, and Joyce 2013).

Second, to ensure that our findings are not specific to the CPS data generating process we have also conducted our Monte Carlo experiments using simulated data (returning to a balanced design). The earnings generating process still conformed with equation 2, but we simulated the state-time shocks ourselves. In doing so we varied their degree of serial correlation and non-normality. In particular, we assumed that the state-time shocks for each state evolve according to the AR(1) process

$$\varepsilon_{gt} = \rho\varepsilon_{g,t-1} + \sqrt{\frac{0.004(1-0.4^2)(d-2)}{d}}\omega_{gt}, t = 2, \dots, 30$$

$$\varepsilon_{g1} = \sqrt{\frac{0.004d}{d-2}}\omega_{g1},$$

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 varied d between 4 (very high non-normality) and 120 (at which point the t distribution is essentially standard normal). To control the degree of serial correlation, we varied ρ. We also examined a scenario in which the data generating process is heterogeneous, by drawing ρ separately for each state from a uniform distribution between 0 and 1. The scaling applied to ω_{gt} ensures that, when $\rho = 0.4$, the variance of ε_{gt} is equal to 0.004 – approximately the empirical variance of the residuals in the CPS data. This means that the degree of serial correlation is allowed to affect the stationary variance of ε_{gt}, but the distribution of the white noise is not. We generated the initial condition (ε_{g1}) such that its variance matched the stationary variance of state-time shocks in other time periods.

In each Monte Carlo replication, we first resampled 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 regressed Y_{gt} on state and year fixed effects only. For each state-time combination, we simulated 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.

Table 4 and Table 5 report rejection rates for various combinations of ρ and d, when the number of groups are, respectively 50 and 10. They show that our main 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 4: Rejection Rates for Tests of Nominal 5% Size Using $\sqrt{G/(G-1)}$-CRSEs and t_{G-1} Critical Values with 50 Groups (Simulated Data).

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

d = 120	0.045	0.047	0.044	0.044	0.045	0.046
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)

The table shows the proportion of the time that the null hypothesis of no treatment effect was rejected in 10,000 Monte Carlo simulations. Simulation standard errors are reported in parentheses. The simulations resample groups (US states) from CPS-MORG data, having imposed the sampling restrictions described in the text. The treatment parameter has a true coefficient of zero. Data from 1979 to 2008 inclusive are sampled (i.e. T = 30). Regressions are run on aggregated state-year data. Simulated log-earnings are generated by effectively replacing the empirical regression residuals with a simulated error term generated according to an AR(1) process. Each cell in the table represents a different AR(1) process. ρ denotes the AR(1) parameter. In the final column the AR(1) parameter is drawn separately for each group, from a uniform distribution between 0 and 1. d denotes the degrees of freedom of the scaled t distribution from which the white noise is drawn (hence it controls the degree of non-normality). See text for full details.

Table 5: Rejection Rates for Tests of Nominal 5% Size Using $\sqrt{G/(G-1)}$-CRSEs and t_{G-1} Critical Values with 10 Groups (Simulated Data).

	ρ = 0	**ρ = 0.2**	**ρ = 0.4**	**ρ = 0.6**	**ρ = 0.8**	**ρ varies with g**
d = 4	0.049	0.053	0.049	0.053	0.047	0.050
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
d = 20	0.055	0.052	0.050	0.053	0.050	0.048
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
d = 60	0.055	0.053	0.051	0.055	0.052	0.050
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
d = 120	0.055	0.048	0.050	0.051	0.054	0.052
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)

Notes as for Table 4.

4.2 Power to Detect Real Effects

Our findings in the previous section indicate that controlling test size need not be a major concern in DiD designs. However, we now show that power to detect real treatment effects with tests of correct size can be extremely low.

Rejection rates in Table 6 indicate power to detect treatment effects on earnings of approximately 2% (precisely, 0.02 log-points), 5%, 10% and 15%. This is based upon the same Monte Carlo replications as Table 1, except we transform the dependent variable: for example, to look at power to detect a 5% effect we add $0.05T_{gt}$ to Y_{gt}.

Table 6: Rejection Rates for Tests of True 5% Size with Different Treatment Effects (β) in Log-Earnings Data.

	G = 50	**G = 20**	**G = 10**	**G = 6**
β = 0.02: $\sqrt{G/(G-1)}$-CRSEs, t(G − 1) critical values	0.378	0.162	0.095	0.072
	(0.007)	(0.005)	(0.004)	(0.004)
β = 0.02: wild cluster bootstrap-t	0.405	0.131	0.095	0.078
	(0.004)	(0.003)	(0.004)	(0.004)
β = 0.05: $\sqrt{G/(G-1)}$-CRSEs, t(G − 1) critical values	0.954	0.625	0.338	0.178
	(0.003)	(0.007)	(0.007)	(0.005)
β = 0.05: wild cluster bootstrap-t	0.954	0.510	0.309	0.176
	(0.002)	(0.005)	(0.006)	(0.005)
β = 0.10: $\sqrt{G/(G-1)}$-CRSEs, t(G − 1) critical values	1.000	0.963	0.782	0.482
	(.)	(0.003)	(0.006)	(0.007)
β = 0.10: wild cluster bootstrap-t	1.000	0.902	0.737	0.439
	(.)	(0.003)	(0.006)	(0.007)
β = 0.15: $\sqrt{G/(G-1)}$-CRSEs, t(G − 1) critical values	1.000	0.999	0.933	0.776
	(.)	(0.000)	(0.004)	(0.006)
β = 0.15: wild cluster bootstrap-t	1.000	0.992	0.906	0.707
	(.)	(0.001)	(0.004)	(0.006)

The table shows the proportion of the time that the null hypothesis of no treatment effect was rejected in 5000 Monte Carlo simulations. Simulation standard errors are reported in parentheses. The simulations resample states from CPS-MORG data, having imposed the sampling restrictions described in the text. β is the true value of the treatment parameter. G is the number of sampled states. Data from 1979 to 2008 inclusive are sampled (i.e. T = 30). Regressions are run on aggregated state-year data. The inference methods used are discussed in the text. We adjust for test size when making power comparisons using the procedure outlined by Davidson and MacKinnon (1998). See text for details.

We focus on the two methods that we have shown above to produce approximately correctly sized hypothesis tests even when the number of groups is small: the $\sqrt{G/(G-1)}$ residual scaling combined with t_{G-1} critical values, and the wild cluster bootstrap-t. Nevertheless, to ensure that we are comparing the power of hypothesis tests which have *exactly* the same size, we adopt the procedure suggested by Davidson and MacKinnon (1998): the nominal significance level used to determine whether to reject the null hypothesis is that which gives a test of true size 0.05. This nominal significance level is obtained from the 5th percentile of the empirical distribution of p-values from Monte Carlo simulations under a true null (i.e. the simulations underlying the results in Table 1). As the results in Table 1 suggest, for both of these methods this is a number very close (but not generally identical) to 0.05. All results reported in Table 6 use this "size-adjusted" measure of power.

The results indicate that power is a serious issue in these designs. A 2% effect would be detected with a probability of less than 1 in 4, even with data from all 50 US states. To detect a 5% effect with a probability of about 80% – a conventional benchmark for power – one would need data on all 50 US states. Power declines much further with G. With 6 states, a researcher would have less than a 50−50 chance of detecting even a 10% effect, a 17% chance of detecting a 5% effect, and a 7% chance of detecting a 2% effect (power barely greater than the size of the test). In other words, it is unlikely that one would detect effects of a typically realistic magnitude using a correctly sized test, and highly unlikely when the number of groups is small.

A comparison of the two inference methods suggests that their power is similar for all combinations of number of groups and size of treatment effect considered in these simulation experiments. If anything, the simpler $\sqrt{G/(G-1)}$ residual scaling combined with t_{G-1} critical values tends to have slightly higher size-adjusted power than the wild cluster bootstrap-t.

Figure 1 documents power more comprehensively by showing 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). 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 uses CRSEs with $\sqrt{G/(G-1)}$-scaled residuals and t_{G-1} critical values.

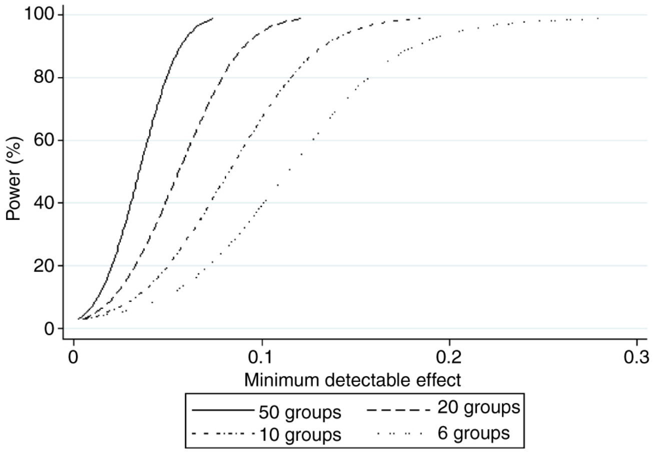

Figure 1: Minimum Detectable Effects on Log-Earnings Using $\sqrt{G/(G-1)}$-CRSEs and t_{G-1} Critical Values and Tests of Size 0.05. The figure shows the proportion of the time that the null hypothesis of no treatment effect is rejected when the treatment parameter has a true coefficient ranging from 0 to 0.3. Numbers are computed using the results of 100,000 Monte Carlo simulations combined with equation 4, as described in the text. The simulations resample states from CPS-MORG data, having imposed the sampling restrictions described in the text. Data from 1979 to 2008 inclusive are sampled (i.e. T = 30). Regressions are run on aggregated state-year data.

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

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

where $\widehat{se(\hat{\beta})}$ 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. With earnings data on the entire US population (50 states), one would need a treatment effect of about 3.5% to have even a 50—50 chance of detecting it. 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.We have also calculated analogous results using a binary employment indicator as the dependent variable. This leads to similar conclusions. For 80% power, the MDE on the employment rate with data from all 50 states is about 2 percentage points, rising 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 knowledge of this feature of the data using feasible GLS. Here we implement the FGLS estimation procedure suggested by Hansen (2007), with and without the straightforward robust inference procedure that our simulations have shown to produce correctly sized tests even with few groups (see Section 2 for details). Our simulation experiments indicate that, in combination with our recommended robust inference technique, FGLS can improve power considerably whilst maintaining correctly sized tests, in a way that is robust to misspecification (or mis-estimation) of the error process, even with few groups.[20]

The first, third and fifth columns of Table 7 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 three of Table 1). The second row shows that FGLS without the bias correction and without robust inference gives tests which over-reject when G is small. Note, however, that even this size distortion is considerably smaller than when using 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 greater than the nominal test size without robust inference (fourth row), particularly when G is small. This is what we would expect, because the bias correction is consistent as $G \to \infty$. 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 7: Rejection Rates for Tests of True 5% Size with Treatment Effects of Zero and +0.05 in Log-Earnings Data.

	G = 50		G = 20		G = 6	
	effect = 0	**effect = 0.05**	**effect = 0**	**effect = 0.05**	**effect = 0**	**effect = 0.05**
OLS, $\sqrt{G/(G-1)}$-CRSEs, t(G − 1) critical values	0.046	0.954	0.039	0.625	0.053	0.178
	(0.003)	(0.003)	(0.003)	(0.007)	(0.003)	(0.005)
FGLS	0.060	0.996	0.092	0.800	0.114	0.292
	(0.003)	(0.001)	(0.004)	(0.006)	(0.004)	(0.006)
FGLS, $\sqrt{G/(G-1)}$-CRSEs, t(G − 1) critical values	0.046	0.994	0.044	0.799	0.058	0.272
	(0.003)	(0.001)	(0.003)	(0.006)	(0.003)	(0.006)
BC-FGLS	0.042	0.995	0.064	0.798	0.089	0.291
	(0.003)	(0.001)	(0.003)	(0.006)	(0.004)	(0.006)
BC-FGLS, $\sqrt{G/(G-1)}$-CRSEs, t(G − 1) critical values	0.045	0.991	0.051	0.784	0.063	0.267
	(0.003)	(0.001)	(0.003)	(0.006)	(0.003)	(0.006)

The table shows the proportion of the time that the null hypothesis of no treatment effect was rejected in 5000 Monte Carlo simulations. Simulation standard errors are reported in parentheses. The simulations resample states from CPS-MORG data, having imposed the sampling restrictions described in the text. G is the number of sampled states. Data from 1979 to 2008 inclusive are sampled (i.e. T = 30). Regressions are run on aggregated state-year data. The inference methods used are discussed in the text. We adjust for test size when making power comparisons using the procedure outlined by Davidson and MacKinnon (1998). See text for details.

The second, fourth and sixth columns of Table 7 turn attention to power, showing rejection rates when there is a 5treatment effect on earnings. Again, the first row reiterates the earlier finding (i.e. see Table 6) that OLS estimation combined with a correctly sized test provides low power, particularly with few groups. 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 96% probability, whereas OLS detects it with 80% probability. Table 7 shows that FGLS also delivers very substantial proportionate power gains relative to OLS with smaller G: with G = 6, power is 18% using OLS and 29% using FGLS. Using Hansen's bias correction delivers a little more power than "ordinary" FGLS.

Figure 2 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.

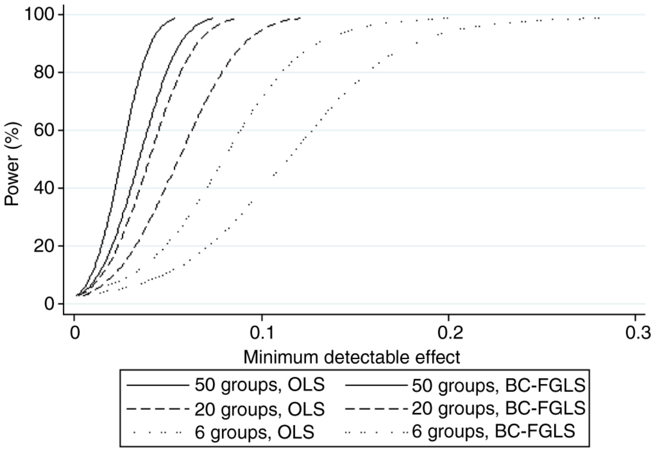

Figure 2: Minimum Detectable Effects on Log-Earnings Using $\sqrt{G/(G-1)}$-CRSEs and t_{G-1} Critical Values and Tests of Size 0.05. The figure shows the proportion of the time that the null hypothesis of no treatment effect is rejected when the treatment parameter has a true coefficient ranging from 0 to 0.3. Numbers are computed using the results of 100,000 Monte Carlo simulations combined with equation 4, as described in the text. The simulations resample groups (US states) from CPS-MORG data, having imposed the sampling restrictions described in the text. Data from 1979 to 2008 inclusive are sampled (i.e. T = 30). Estimation of the treatment effect is conducted on aggregated state-year data either by OLS (reproducing part of Figure 1), or by feasible GLS assuming a AR(2) error process (homogeneous across states) and using bias-corrected AR parameter estimates as in Hansen (2007) (denoted "BC-FGLS"). See text for full details.

4.3.1 Robustness

As we did for the analysis of test size in Section 4.1, we have repeated this analysis for the case where the outcome of interest is a binary employment status indicator, and the results, which are available in in Brewer, Crossley, and Joyce (2013), confirm that the findings reported here hold in that case: FGLS delivers substantial power gains over OLS, and this can be done whilst controlling test size, even with few groups.

We now further explore the robustness of these results to different data settings. Of particular interest are cases where the parametric assumptions about the serial correlation process inherent in the FGLS procedure are incorrect. In such cases, can test size still be reliably controlled (even with few groups), and what power gains (if any) can FGLS offer?

We continue to use FGLS estimation based on the assumption of an AR(2) process for the state-time shocks which is the same for all states. We explore its properties under two forms of misspecification of the error process. First, we simulate an AR(2) process which is heterogeneous across states. Second, we simulate an MA(1) process with parameter 0.5.[21] In each case, we effectively replace the empirical log-earnings residuals in the CPS (from a regression of state-year earnings on state and year fixed effects) with our simulated error terms, as in the robustness checks reported above.

Columns one and three of Table 8 shows the results of these simulations under the null hypothesis of no treatment effect using tests of nominal size 0.05, for the case where G = 10. The results show that the previous results on test size hold under these alternative processes: without robust inference, Hansen's bias correction for the AR parameter estimates brings true test size closer to the nominal size when using FGLS estimation (although there is still some over-rejection); but test size can be controlled reliably using our suggested robust inference technique, whether estimation is carried out using OLS, FGLS or BC-FGLS.[22] Columns two and four of Table 8 report power to detect a treatment effect of 0.05 log-points on earnings, again for the case where $G = 10$.[23] The second column shows that, using correctly sized hypothesis tests (i.e. those that use our suggested robust inference technique), FGLS estimation does offer substantial power gains over OLS even when based upon the incorrect assumption that the AR(2) error process is homogeneous across states.[24] The fourth column shows that, where the true error process is MA(1) rather than AR(2), there is no power gain from FGLS. Intuitively this makes sense: where the parametric assumptions about the serial correlation in the error terms are a very poor approximation to the true process, FGLS does not offer efficiency gains; where the assumptions are a better approximation, FGLS does offer efficiency gains over the OLS estimator, which does not exploit any knowledge of the nature of the error process.

Table 8: Rejection Rates with 5%-Level Tests and Treatment Effects of Zero and +0.05 in Simulated Log-Earnings Data (10 Groups).

	Heterogenous AR(2)		**MA(1)**	
	effect = 0	**effect = 0.05**	**effect = 0**	**effect = 0.05**
OLS, $\sqrt{G/(G-1)}$-CRSEs, t(G − 1) critical values	0.056	0.520	0.056	0.619
	(0.003)	(0.007)	(0.003)	(0.007)
FGLS	0.106	0.782	0.088	0.713
	(0.004)	(0.006)	(0.004)	(0.006)
FGLS, $\sqrt{G/(G-1)}$-CRSEs, t(G − 1) critical values	0.065	0.712	0.061	0.606
	(0.003)	(0.006)	(0.003)	(0.007)
BC-FGLS	0.069	0.807	0.072	0.710
	(0.004)	(0.006)	(0.004)	(0.006)
BC-FGLS, $\sqrt{G/(G-1)}$-CRSEs, t(G − 1) critical values	0.063	0.727	0.059	0.603
	(0.003)	(0.006)	(0.003)	(0.007)

The table shows the proportion of the time that the null hypothesis of no treatment effect was rejected in 5000 Monte Carlo simulations. Simulation standard errors are reported in parentheses. The simulations resample states from CPS-MORG data, having imposed the sampling restrictions described in the text. Data from 1979 to 2008 inclusive are sampled (i.e. T = 30). Regressions are run on aggregated state-year data. The underlying data effectively replaces empirical CPS regression residuals with a simulated error term, generated according to an AR(2) process which varies across groups and an MA(1) process respectively. The inference methods used are discussed in the text. We adjust for test size when making power comparisons using the procedure outlined by Davidson and MacKinnon (1998). See text for details.

Finally, we show how these results vary with the number of time periods available. Table 9 repeats the simulations from Table 8 for the cases where $T = 20$ and $T = 10$. This shows that test size can still be controlled using our recommended robust inference approach with few time periods, whether estimation is based on OLS or FGLS. It also shows, however, that the power gains from FGLS diminish with decreasing T, and with $T = 10$ the power of OLS and FGLS (when combined with inference which provides a correctly sized test) are essentially the same. An interesting feature of Table 8 is that with OLS and robust inference, power actually declines with T.[25] Indeed, with these data, the growing power advantage of FGLS with increasing T comes about because the power of OLS declines and not because the power of the FGLS improves. Further inspection of the underlying simulations suggests that there is a genuine increase in the precision of OLS estimation as T falls: the estimated policy effects become more tightly distributed around their true value.

Table 9: Rejection Rates with 5%-Level Tests and Treatment Effects of Zero and +0.05 in Log-Earnings Data (10 Groups).

	T = 30		**T = 20**		**T = 10**	
	effect = 0	**effect = 0.05**	**effect = 0**	**effect = 0.05**	**effect = 0**	**effect = 0.05**
OLS, $\sqrt{G/(G-1)}$-CRSEs, t(G − 1) critical values	0.039	0.338	0.034	0.380	0.037	0.447
	(0.003)	(0.007)	(0.003)	(0.007)	(0.003)	(0.007)
FGLS	0.101	0.449	0.107	0.436	0.091	0.415
	(0.004)	(0.007)	(0.004)	(0.007)	(0.004)	(0.007)

FGLS, $\sqrt{G/(G-1)}$-CRSEs, t(G − 1) critical values	0.044	0.484	0.039	0.478	0.040	0.479
	(0.003)	(0.007)	(0.003)	(0.007)	(0.003)	(0.007)
BC-FGLS	0.078	0.453	0.078	0.431	0.078	0.413
	(0.004)	(0.007)	(0.004)	(0.007)	(0.004)	(0.007)
BC-FGLS, $\sqrt{G/(G-1)}$-CRSEs, t(G − 1) critical values	0.049	0.479	0.043	0.481	0.038	0.486
	(0.003)	(0.007)	(0.003)	(0.007)	(0.003)	(0.007)

The table shows the proportion of the time that the null hypothesis of no treatment effect was rejected in 5000 Monte Carlo simulations. Simulation standard errors are reported in parentheses. The simulations resample states from CPS-MORG data, having imposed the sampling restrictions described in the text. T is the number of (consecutive) time periods (years). The first year of data is chosen from a uniform distribution between 1979 and (2009-T) in each Monte Carlo simulation. Regressions are run on aggregated state-year data. The inference methods used are discussed in the text. We adjust for test size when making power comparisons using the procedure outlined by Davidson and MacKinnon (1998). See text for details.

The reason why this is possible in this context is as follows.[26] DiD regressions effectively estimate the difference in mean outcomes between post-treatment periods and pre-treatment periods (and then compare these differences across treatment and control groups). The variance of this difference is decreasing in the covariance between the error terms pre and post treatment (intuitively, if error terms pre and post treatment covaried perfectly then they would not add any noise to the difference between pre and post treatment outcomes because they would cancel out). If serial correlation between observations decays with time, then the error term from an additional time period pre (post) treatment will covary less strongly with error terms in the post (pre) treatment period than the error terms already present. Hence the "covariance effect" acts to increase the variance of the DiD estimate of the treatment effect when you add another time period to the data. On the other hand, of course, the variance is also increasing in the variance of the average error terms both pre and post treatment, and adding more time periods will reduce this variance. But this reduction will be small if serial correlation is high, and hence it can be dominated by the covariance effect.

An additional lesson, then, is that researchers who use OLS to perform DiD estimation should be very cautious about blindly using as many time periods as possible. With serially correlated error terms, this can result in a reduction in power. This issue does not arise with FGLS, which effectively transforms the data in a way that removes the serial correlation from the error terms. This seems to us another reason to favour a FGLS approach.

In summary, our simulation experiments suggest that as long as the number of time periods is not too small (approximately 10 or less), FGLS combined with robust inference is likely to offer substantial power gains; and even if it does not – because the error process is badly misspecified – test size can still be reliably controlled, even with few groups. We therefore recommend this approach is used more routinely by applied researchers doing DiD estimation.

5 STATA Code for FGLS

The replication files provided with this paper include the STATA ado file Hansen.ado which implements the the FGLS estimators used in this paper, with or without the iterative bias-correction procedure set out in Hansen (2007). It also allows a user to combine GLS with CRSEs as illustrated above. This implementation assumes an AR error process homogeneous across groups. It is designed to look and feel like the "regress" command in STATA in terms of syntax and outputs, but with options to control the FGLS estimation. Full details are at the top of the file. This package can be used by applied researchers to implement the empirical strategies we recommend.

6 Summary and Conclusion

This paper contributes to a growing literature on inference in difference-in-differences designs with grouped errors. The literature has emphasized difficulties in obtaining correctly sized hypothesis tests, particularly with few groups, but our results suggest this is not the key challenge.

Using Monte Carlo evidence, we have made four points. First, it is possible to obtain tests of the correct size, even with few groups, and in many settings this is possible using methods that are very straightforward to implement. 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 also show that with OLS and serially correlated errors, power can actually fall as panel length grows. This seems to us another reason to prefer a FGLS approach.

We therefore recommend that applied researchers adopt FGLS estimation combined with robust inference methods in practice. We also suggest that future research could usefully focus on improving power, as well as on getting test size correct, when using difference-in-difference designs.

Acknowledgement

We would like to thank a large number of colleagues and seminar and conference participants for helpful comments. All remaining errors are our own. This work was supported by the ESRC through Programme Evaluation for Policy Analysis, a former Node of the National Centre for Research Methods, through the Research Centre on Micro-Social Change (MiSoC) at the University of Essex (ES/L009153/1), and through the Centre for the Microeconomic Analysis of Public Policy (CPP) at the IFS (RES-544-28-5001).

Notes

[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 issues for inference are dealing with a finite number of groups and any serial correlation in group-time shocks.

[2] This is true at the time of writing (STATA version 14.1) and has been the case since at least STATA 6.

[3] Equivalently, one could include a full set of group-time dummies in this first regression (and omit the constant). The $\hat{Y}_{gt}$ are the estimated coefficients on those dummies.

[4] 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 compute a minimum distance estimator of β. One can then test the over-identifying restrictions. This is asymptotically valid as group-time cell sizes tend to infinity.

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

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

[7] Donald and Lang (2007) show that a similar reference distribution - t_{G-2} - would provide tests of exactly the right size in the special case where the ε_{gt} were normal, homoscedastic and independent (i.e. serially uncorrelated).

[8] 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.

[9] The asymptotic variance of the score also needs to be the same across groups.

[10] Romano and Wolf (2015) make this point in the more specific context where errors are heteroskedastic but independent (unlike in the typical DiD context we consider here, where serial correlation is a key issue), arguing that weighted least squares should be used to realise efficiency gains in the face of heteroskedasticity in combination with robust inference to control test size.

[11] 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.

[12] Given large state-time cell sizes, aggregation should average out the individual-level shock component precisely. Mean cell sizes are 500 and 780 when the dependent variables are log(earnings) and employment status respectively.

[13] We recommend the first-step aggregation 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, Gelbach, and Miller (2008) and is noted in Hansen (2007).

[14] In the particular example we use here where groups are geographical units, the assumption of no inter-group error correlation is not likely to be reasonable close to the groups' boundaries. This is an advantage of generating placebo treatments randomly in the experiment: Barrios et al. (2012) show that, as long as there is no *cross-cluster* spatial correlation in treatment status, correct test size is robust to some correlation in the error terms across clusters, as long as the data are weakly dependent so that error correlation decays with distance. Hence, we will not be confusing the impacts on test size of inadequately accounting for grouped errors with the impacts of (incorrectly) assuming that the earnings shocks of people in geographical proximity but in different states are independent.

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

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

[17] We use 199 bootstrap replications, which is sufficient in this context as bootstrap simulation error will average out across Monte Carlo replications. We note that, as pointed out by Webb (2013), p-values are not point identified when the number of groups is very small. For example, with $G = 6$ there are only $2^G = 64$ potential unique bootstrap samples and $2^{G-1} = 32$ possible t-statistics (in absolute value).

[18] For the extreme case – not considered here – where there is just one treated group, see Conley and Taber (2011). They develop a method for inference which is valid in this design when there are many control groups.

[19] The baseline employment rate in the sample is 67%. Full results are available from the authors.
[20] The results shown use the Cochrane-Orcutt transformation. The Prais-Winsten proceedure gave very similar results.
[21] For the heterogeneous AR(2) process, the coefficient on the first lag (α_1^g) is drawn from a uniform distribution between zero and one for each state. The coefficient on the second lag is set equal to $0.5*min(\alpha_1^g, 1-\alpha_1^g)$, which ensures stationarity. For both the heterogeneous AR(2) and (homogeneous) MA(1) processes, the white noise in the process is normally distributed. Its variance is chosen so that the stationary variance of the simulated error term matches the empirical variance of the log-earnings residuals in the CPS (0.04).
[22] We showed in Section 4.1 that this inference technique is not reliable if there is a large imbalance between the numbers of treatment and control groups; but that the wild cluster boostrap-t procedure is relatively robust in such settings. FGLS estimation combined with bootstrap-based inference would be a sensible alternative in those situations.
[23] We have also conducted this analysis with $G = 50$. Conclusions are qualitatively the same, although of course the power of all procedures is higher with more groups.
[24] We also re-ran the earlier robustness checks where state-time earnings shocks evolve according to an AR(1) process, with varying degrees of serial correlation and varying degrees of non-normality in the white noise. The same qualitative conclusions about the size and power of FGLS and OLS combined with robust inference continued to hold. These results are available from the authors on request.
[25] The same qualitative result is reported without comment in Tables 3 to 5 of Hansen (2007).
[26] We are extremely grateful to Joao Santos Silva for pointing out this mechanism to us.

References

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

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

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

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

Bester, A. C., T. G. Conley, and C. B. Hansen. 2011. "Inference with Dependent Data Using Cluster Covariance Estimators." *Journal of Econometrics* 165 (2): 137–151.

Bloom, H. S. 1995. "Minimum Detectable Effects: A Simple Way to Report the Statistical Power of Experimental Designs." *Evaluation Review* 19 (5): 547–556.

Brewer, M., T. F. Crossley, and Robert Joyce. 2013. "Inference with Differences-in-Differences Revisited." IZA DP No. 7742.

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

Cameron, A. C., and D. L. Miller. 2015. "A Practitioner's Guide to Cluster-Robust Inference." *Journal of Human Resources* 50 (2): 317–372.

Conley, T., and C. Taber. 2011. "Inference with 'Difference in differences' with a Small Number of Policy Changes." *The Review of Economics and Statistics* 93 (1): 113–125.

Davidson, R., and J. G. MacKinnon. 1998. "Graphical Methods for Investigating the Size and Power of Test Statistics." *The Manchester School* 66: 1–26.

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

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

Imbens, G., and M. Kolesár. 2012. "Robust Standard Errors in Small Samples: Some Practical Advice." National Bureau of Economic Research Working Paper 18478.

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

Mackinnon, J., and M. Webb. 2017. "Wild Bootstrap Inference for Wildly Different Cluster Sizes." *Journal of Applied Econometrics* 32 (2): 233–254.

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

Nickell, S. 1981. "Biases in Dynamic Models with Fixed Effects." *Econometrica* 49 (6): 1417–1426.

Romano, J., and M. Wolf. 2015. "Resurrecting Weighted Least Squares (August 2015)." University of Zurich, Department of Economics, Working Paper No. 172.

Solon, G. 1984. "Estimating Autocorrelations in Fixed Effects Models." National Bureau of Economic Research Technical Working Paper 32.

Webb, M. 2013. "Reworking Wild Bootstrap Based Inference for Clustered Errors." Queen's Economics Department Working Paper No. 1315.

Wooldridge, J. M. 2003. "Cluster-Sample Methods in Applied Econometrics." *American Economic Review* 93: 133–138.

Wooldridge, J. M. 2006. "Cluster-Sample Methods in Applied Econometrics: An Extended Analysis." Mimeograph, Michigan State University Department of Economics.

Supplemental Material: The online version of this article offers supplementary material (DOI:https://-doi.org/10.1515/jem-2017-0005).