NBER WORKING PAPER SERIES

PANEL DATA AND EXPERIMENTAL DESIGN

Fiona Burlig
Louis Preonas
Matt Woerman

Working Paper 26250
http://www.nber.org/papers/w26250

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
September 2019

We thank Andrew Foster, several anonymous referees, Michael Anderson, Maximilian Auffhammer, Patrick Baylis, Joshua Blonz, Severin Borenstein, Hai-Anh Dang, Aureo de Paula, Meredith Fowlie, James Gillan, Erin Kelley, Jeremy Magruder, Aprajit Mahajan, Edward Miguel, Tavneet Suri, Catherine Wolfram, and seminar participants at UC Berkeley, BITSS, NEUDC, and EMEE for valuable comments. Molly Van Dop provided excellent research assistance. Funding for this research was provided by the Berkeley Initiative for Transparency in the Social Sciences, a program of the Center for Effective Global Action (CEGA), with support from the Laura and John Arnold Foundation. Burlig was generously supported by the National Science Foundation's Graduate Research Fellowship Program under grant DGE-1106400, and Preonas was generously supported by Resources for the Future's Joseph L. Fisher Dissertation Fellowship. All remaining errors are our own. Our accompanying Stata package, pcpanel, is available from ssc. We have no financial relationships that relate to this research. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Panel Data and Experimental Design
Fiona Burlig, Louis Preonas, and Matt Woerman
NBER Working Paper No. 26250
September 2019
JEL No. B4,C23,C9,O1,Q4

ABSTRACT

How should researchers design panel data experiments? We analytically derive the variance of panel estimators, informing power calculations in panel data settings. We generalize Frison and Pocock (1992) to fully arbitrary error structures, thereby extending McKenzie (2012) to allow for non-constant serial correlation. Using Monte Carlo simulations and real world panel data, we demonstrate that failing to account for arbitrary serial correlation ex ante yields experiments that are incorrectly powered under proper inference. By contrast, our “serial-correlation-robust” power calculations achieve correctly powered experiments in both simulated and real data. We discuss the implications of these results, and introduce a new software package to facilitate proper power calculations in practice.

Fiona Burlig
Harris School of Public Policy
University of Chicago
1307 East 60th Street
Chicago, IL 60637
and NBER
burlig@uchicago.edu

Matt Woerman
Resource Economics
Stockbridge Hall
University of Massachusetts Amherst
80 Campus Center Way
Amherst, MA 01003
mwoerman@umass.edu

Louis Preonas
University of Maryland
2104 Symons Hall
7998 Regents Drive
College Park, MD 20742
lpreonas@umd.edu

A data appendix is available at http://www.nber.org/data-appendix/w26250
A Stata package: pcpanel is available at https://econpapers.repec.org/software/bocbocode/s458286.htm

1 Introduction

Randomized controlled trials (RCTs) are an increasingly popular method for applied economics research (Card, DellaVigna, and Malmendier (2011). When designing RCTs, researchers typically use *ex ante* power calculations to evaluate the trade-off between sample size and statistical precision. If the sample is too small, the experiment will be unable to distinguish between true and false null hypotheses; at the same time, overly large samples waste resources.[1] The economics literature on power calculations has focused on single-wave experiments, where units are randomized into treatment and control groups, and researchers observe each unit once.[2] In a widely cited paper based on results from Frison and Pocock (1992), McKenzie (2012) recommends panel data experiments, where multiple observations per unit help to increase statistical power. This is especially attractive in settings where collecting additional waves of data for one individual is cheaper than enrolling more individuals. As a result, panel RCTs have become increasingly common in recent years.[3]

At the same time, panel data pose challenges for statistical inference, due to non-constant within-unit serial correlation. Bertrand, Duflo, and Mullainathan (2004) demonstrate that failing to account for this correlation structure can bias standard errors towards zero, raising the probability of a Type I error. In order to achieve correct false rejection rates, applied econometricians often implement the cluster-robust variance estimator (CRVE), or use “clustered standard errors”, which accommodates arbitrary serial correlation within panel units.[4]

[1] Bloom (1995) provides an early framework for power calculations, while Duflo, Glennerster, and Kremer (2007) and Glennerster and Takavarasha (2013) detail their implementation in practice. Cohen (1977) and Murphy, Myors, and Wolach (2014) are also classic references.

[2] Researchers often collect two waves of data, but estimate treatment effects using post-treatment data only, controlling for the baseline level of the outcome variable (following McKenzie (2012)). Baird et al. (2018) extends the standard cross-sectional setup to randomized saturation designs, capable of measuring spillover and general equilibrium effects. Athey and Imbens (2017) discusses statistical power using a randomization inference approach.

[3] Recent panel RCTs include Bloom et al. (2013); Blattman, Fiala, and Martinez (2014); Jessoe and Rapson (2014); Bloom et al. (2015); Fowlie et al. (2017); Atkin, Khandelwal, and Osman (2017); Atkin et al. (2017); McKenzie (2017); and Fowlie, Greenstone, and Wolfram (2018).

[4] See White (1984), Arellano (1987), and Cameron and Miller (2015) for more details on the CRVE. Abadie et al. (2017) underscore the need to use the CRVE in experiments where treatment assignment is correlated within clusters—as in most panel RCTs, where treated units remain treated throughout the experiment.

Besides affecting the false rejection rate, serial correlation can also impact statistical power. Hence, it is important to properly account for serial correlation during both *ex post* analysis and *ex ante* power calculations.[5] However, McKenzie (2012) only allows for *constant* serial correlation *ex ante* (equivalent to assuming i.i.d. errors after removing unit fixed effects), despite evidence that panel data typically exhibit more complex serial correlation (Bertrand, Duflo, and Mullainathan (2004)). While Frison and Pocock (1992) propose a framework for incorporating non-constant serial correlation into *ex ante* power calculations, they rely on restrictive assumptions—homogeneous error structures across units and deterministic time shocks—which are likely unrealistic in practice. As a result, there is a gap in the existing literature that may preclude researchers from designing properly powered experiments using panel data.

In this paper, we derive analytical expressions for the variance of panel estimators under *arbitrary* non-i.i.d. error structures, while also allowing for random time shocks. We use these expressions to (i) formalize a power calculation formula for difference-in-differences estimators that is robust to arbitrary serial correlation, and (ii) devise a method for estimating the required inputs to this formula from real data.[6]

We conduct Monte Carlo analyses using both simulated and real data, and demonstrate that the methods outlined in McKenzie (2012) yield experiments that are incorrectly powered in the presence of non-constant serial correlation, even with proper *ex post* inference. These methods tend to yield dramatically overpowered experiments in short panels and dramatically underpowered experiments in long panels. By allowing for non-constant serial correlation *ex ante*, our "serial-correlation-robust" power calculation approach achieves the desired power in both simulated and real data. Ultimately, we provide researchers with

[5] If researchers do not adjust their standard errors *ex post* to account for within-unit serial correlation in panel data, they will likely over-reject true null hypotheses. If they adjust their standard errors *ex post* but do not adjust their *ex ante* power calculations to account for within-unit serial correlation, they introduce a mismatch between *ex ante* and *ex post* assumptions that will likely yield incorrectly powered experiments.

[6] Recent experiments published in top economics journals use either the difference-in-differences estimator or the ANCOVA estimator. We discuss ANCOVA in Sections 2.2.2 and 3.1, where standard power calculation techniques similarly ignore non-constant serial correlation within panel units. However, analytically deriving the variance of the ANCOVA estimator necessitates restrictive assumptions on the data generating process, causing analytical ANCOVA power calculations to perform poorly with real data. For this reason, we focus on power calculations for difference-in-differences.

both the theoretical insights and practical tools to design properly powered experiments in panel data settings.

We make three main contributions to the economics literature on experimental design. First, we show that standard power calculation methods for panel RCTs (discussed in McKenzie (2012)) fail in the presence of arbitrary serial correlation. Second, we derive a new power calculation formula for difference-in-differences, allowing for *arbitrary* serial correlation, which extends Frison and Pocock (1992) by accommodating heterogeneous error structures across units and random time shocks. This serial-correlation-robust formula enables researchers to calibrate panel RCTs to the desired power. Finally, we provide guidance for designing panel RCTs in real experimental settings, and introduce an accompanying Stata package (`pcpanel`) to facilitate proper power calculations in practice.

The paper proceeds as follows. Section 2 presents analytical power calculations for panel data with arbitrary serial correlation, and uses Monte Carlo simulations to evaluate the performance of these results. Section 3 extends these simulation results to real experimental data. Section 4 discusses practical issues related to power calculations, and introduces our accompanying software package. Section 5 concludes.

2 Power calculations for panel data

Power calculations provide an *ex ante* estimate of the smallest effect size that an experiment, with a given sample size and experimental design will be able to statistically detect. Most power calculations take the following form:

$$MDE = \left(t^{d}_{1-\kappa} + t^{d}_{\alpha/2}\right)\sqrt{\mathrm{Var}\left(\hat{\tau} \mid \mathbf{X}\right)} \tag{1}$$

where $\mathrm{Var}(\hat{\tau} \mid \mathbf{X})$ is the exact finite sample variance of the treatment effect estimator, conditional on independent variables $\mathbf{X}$; $t^{d}_{\alpha/2}$ is the critical value of a t distribution with d degrees of freedom associated with the probability of a Type I error, α, in a two-sided test against a null hypothesis of $\tau = 0$; and $t^{d}_{1-\kappa}$ is the critical value associated with the probability of correctly rejecting a false null, κ.[7] These parameters determine the minimum detectable effect (MDE), the smallest value $|\tau|> 0$ for which the experiment will (correctly) reject the null $\tau = 0$ with probability κ at the significance level α.

This paper's core contribution is our "serial-correlation-robust" (SCR) power calculation formula for designing experiments with panel data. This section outlines our model and resulting SCR formula for the difference-in-differences (DD) estimator, which extends Frison and Pocock (1992) and McKenzie (2012) by incorporating arbitrary (non-constant) serial correlation. Using simulations, we demonstrate the importance of accounting for this correlation *ex ante* in order to achieve the desired statistical power *ex post*. We also consider three sensitivities: short panels, which are traditionally of interest in development economics; alternative treatment effect estimators; and alternative assumptions on the data generating process (DGP).

2.1 Serial-correlation-robust power calculations

We begin with a model in which there are J units, P proportion of which are randomized into treatment. The researcher collects outcome data Y_{it} for each unit i, across m pre-treatment time periods and r post-treatment time periods. For treated units, $D_{it} = 0$ in pre-treatment periods and $D_{it} = 1$ in post-treatment periods; for control units, $D_{it} = 0$ in all periods.[8]

Assumption 1 (Data generating process)**.** *The data are generated according to the following model:*

$$Y_{it} = \beta + \tau D_{it} + \upsilon_i + \delta_t + \omega_{it}$$

where υ_i is a unit-specific disturbance distributed i.i.d. $\mathcal{N}(0, \sigma_\upsilon^2)$; δ_t is a time-specific disturbance distributed i.i.d. $\mathcal{N}(0, \sigma_\delta^2)$; ω_{it} is an idiosyncratic error term distributed (not necessarily

[7] For one-sided tests, $t^d_{\alpha/2}$ can be replaced with t^d_{α}. $1 - \kappa$ gives the probability of a false rejection, or a Type II error. The degrees of freedom, d, will depend on the dimensions of $\mathbf{X}$ and the treatment effect estimator in question.

[8] Put differently, we assume that there is a control group of units that is never treated in the sample period, and a treatment group of units for which treatment turns on in a particular time period (and persists through all subsequent periods). This is a standard design for panel RCTs in economics.

i.i.d.) $\mathcal{N}(0,\sigma_\omega^2)$*; and* τ *is the treatment effect, which is assumed to be homogeneous across all units and all time periods.*[9]

Assumption 2 (Strict exogeneity)**.** $\mathrm{E}[\omega_{it} \mid \mathbf{X}] = 0$*, where* $\mathbf{X}$ *is a full rank matrix of regressors, including a constant, the treatment indicator* $\mathbf{D}$*,* $J-1$ *unit dummies, and* $(m+r)-1$ *time dummies. This follows from random assignment of* D_{it}*.*

Assumption 3 (Balanced panel)**.** *The number of pre-treatment observations, m, and post-treatment observations, r, is the same for each unit, and all units are observed in every time period.*

Assumption 4 (Independence across units)**.** $\mathrm{E}[\omega_{it}\omega_{js} \mid \mathbf{X}] = 0, \ \forall\ i \neq j, \ \forall\ t, s.$

Assumption 5 (Symmetric covariance structures)**.** *Define:*

$$\psi^B \equiv \frac{2}{Jm(m-1)} \sum_{i=1}^{J} \sum_{t=-m+1}^{-1} \sum_{s=t+1}^{0} \mathrm{Cov}\left(\omega_{it}, \omega_{is} \mid \mathbf{X}\right)$$
$$\psi^A \equiv \frac{2}{Jr(r-1)} \sum_{i=1}^{J} \sum_{t=1}^{r-1} \sum_{s=t+1}^{r} \mathrm{Cov}\left(\omega_{it}, \omega_{is} \mid \mathbf{X}\right)$$
$$\psi^X \equiv \frac{1}{Jmr} \sum_{i=1}^{J} \sum_{t=-m+1}^{0} \sum_{s=1}^{r} \mathrm{Cov}\left(\omega_{it}, \omega_{is} \mid \mathbf{X}\right)$$

to be the average pre-treatment, post-treatment, and across-period covariance between different error terms of the same unit, respectively. Define ψ_T^B*,* ψ_T^A*, and* ψ_T^X *analogously, where we consider only the PJ treated units; also define* ψ_C^B*,* ψ_C^A*, and* ψ_C^X *analogously, where we consider only the* $(1-P)J$ *control units. Using these definitions, assume that* $\psi^B = \psi_T^B = \psi_C^B$*;* $\psi^A = \psi_T^A = \psi_C^A$*; and* $\psi^X = \psi_T^X = \psi_C^X$*.*[10]

[9] While we assume a homogeneous treatment effect, our derivations also hold if τ varies across time periods (as in Frison and Pocock (1992), and McKenzie (2012)). In addition, this data generating process is quite general and nests versions with no random unit or time shocks; reducing σ_υ^2 (σ_δ^2) to zero is equivalent to removing random unit (time) shocks from the data generating process. Section 2.2.3 explores data generating processes where υ_i and δ_t are deterministic, rather than random shocks. In Section 3.1 below, we explore the performance of our serial-correlation-robust power calculations in a real data setting where the true data generating process is unknown.

[10] These ψ terms come from the derivation of the variance of the DD estimator (see Appendix A.2.2). Each ψ averages covariances over units with potentially heterogeneous error structures (ω_{it} from the underlying DGP). In addition, the ψ terms depend both on the error structure and on the length of the experiment (m and r). We choose the letters "B" to indicate the Before-treatment period, and "A" to indicate the

Under these assumptions, the OLS estimator with unit and time fixed effects is $\widehat{\tau} = (\ddot{\mathbf{D}}'\ddot{\mathbf{D}})^{-1}\ddot{\mathbf{D}}'\ddot{\mathbf{Y}}$, with $\mathrm{E}[\hat{\tau}] = \tau$. Assumptions 1–5 yield a power calculation formula that is robust to arbitrary serial correlation:[11]

$$MDE = (t^J_{1-\kappa} + t^J_{\alpha/2})\sqrt{\underbrace{\left(\tfrac{1}{P(1-P)J}\right)\left[\left(\tfrac{m+r}{mr}\right)\sigma^2_\omega + \left(\tfrac{m-1}{m}\right)\psi^B + \left(\tfrac{r-1}{r}\right)\psi^A - 2\psi^X\right]}_{\mathrm{Var}(\hat{\tau}|\mathbf{X})}} \qquad (2)$$

Throughout the remainder of the paper, we refer to Equation (2) as the "serial-correlation-robust" (SCR) power calculation formula. Note that under cross-sectional randomization, this expression for the variance of $\hat{\tau}$ still holds in expectation, even in the presence of within-period error correlations across units:

Lemma 1. *In a panel difference-in-differences model with treatment randomly assigned at the unit level,* $\left(\frac{1}{P(1-P)J}\right)\left[\left(\frac{m+r}{mr}\right)\sigma^2_\omega + \left(\frac{m-1}{m}\right)\psi^B + \left(\frac{r-1}{r}\right)\psi^A - 2\psi^X\right]$ *is an unbiased estimator of the expectation of* $\mathrm{Var}(\hat{\tau} \mid \mathbf{X})$*, even in the presence of arbitrary within-period cross-sectional correlations. See Appendix A.3 for a proof, and see Appendix A.2.3 for a more general model that relaxes Assumptions 4–5.*

2.1.1 Accounting for serial correlation

Our SCR power calculation formula generalizes the Frison and Pocock (1992) (henceforth FP) difference-in-differences formula to accommodate fully arbitrary correlation structures. Whereas FP assume that all units share a homogeneous correlation structure, our SCR formula accommodates arbitrary heterogeneous correlation structures across units. We also allow for random time shocks across all units, which are characteristic of most economic datasets.[12] As a result, our SCR formula is able to flexibly accommodate more realistic panel data structures.

After-treatment period. We index the m pre-treatment periods $\{-m+1, \ldots, 0\}$, and the r post-treatment periods $\{1, \ldots, r\}$. In a randomized setting, $\mathrm{E}\left[\psi^B\right] = \mathrm{E}\left[\psi^B_T\right] = \mathrm{E}\left[\psi^B_C\right]$, $\mathrm{E}\left[\psi^A\right] = \mathrm{E}\left[\psi^A_T\right] = \mathrm{E}\left[\psi^A_C\right]$, and $\mathrm{E}\left[\psi^X\right] = \mathrm{E}\left[\psi^X_T\right] = \mathrm{E}\left[\psi^X_C\right]$, making this a reasonable assumption *ex ante*. However, it is possible for treatment to alter the covariance structure of treated units only.

[11] We present the formal derivation of this formula in Appendix A.2.2. Note that if $m = 1$ (or $r = 1$), ψ^B (or ψ^A) is not defined and is multiplied by 0 in Equation (2).

[12] If we impose FP's assumptions, our SCR formula collapses to the FP formula for the difference-in-differences model with unequal correlations.

McKenzie (2012) is the most widely cited reference for power calculations using panel data in economics. McKenzie's results follow from FP's original derivations, yet they impose a more restrictive assumption on the data generating process: constant serial correlation between any two time periods, within each cross-sectional unit.[13] Using our notation, McKenzie's difference-in-differences power calculation formula becomes:[14]

$$MDE = \left(t^J_{1-\kappa} + t^J_{\alpha/2}\right) \sqrt{\left(\tfrac{\sigma^2_\omega}{P(1-P)J}\right)\left(\tfrac{m+r}{mr}\right)} \tag{3}$$

This is equivalent to the SCR formula when ψ^A, ψ^B, and ψ^X are all equal to zero—allowing only *constant* serial correlation of the *composite* error term (ε_{it}, in McKenzie's notation), or an i.i.d. idiosyncratic error term (ω_{it}, in our notation). As highlighted by Bertrand, Duflo, and Mullainathan (2004, henceforth BDM), this assumption is likely unrealistic because most panel datasets in economics exhibit non-constant serial correlation.[15]

To further illustrate the difference between the McKenzie and SCR models, consider two cross-sectional units (indexed $\{i, j\}$) and four time periods (indexed $\{0, 1, 2, 3\}$), with the data generating process described in Assumption 1. The vector of idiosyncratic errors, $\boldsymbol{\omega}$, and the corresponding variance-covariance matrix, $\boldsymbol{\Omega}$, can be represented as follows:[16]

[13] The preceding paragraph describes the most general FP model, reported on page 1701. McKenzie draws from the more restrictive FP model, reported on p. 1693.

[14] See Appendix A.2.1 for the derivation. This formula is analogous to McKenzie's theoretical formula. McKenzie (2012, p. 215) suggests an alternative approach for empirical applications with non-constant serial correlation, which we explore in Appendix C.4 using both simulated and real data. We find that this approach, while effective in panels with only 1 pre- and 1 post-treatment period, delivers improperly powered experiments in longer panels. We also find this approach to be less effective as the degree of autocorrelation increases. In Appendices D.1 and E, we derive a method to use real pre-existing data to perform power calculations using the SCR method, which is effective even in settings where the true DGP is unknown.

[15] Our SCR formula differs from that of McKenzie in one additional way: McKenzie's empirical specification does not include a unit fixed effect, whereas the estimator underlying the SCR model does. This fixed effect absorbs any *constant* serial correlation of composite errors, leaving no remaining serial correlation between the idiosyncratic error terms of our SCR model. Thus, constant serial correlation in McKenzie's framework translates to *no* serial correlation in our SCR framework. Section 2.2.3 discusses the implications of removing *constant* serial correlation from the true DGP, and Section 2.2.2 considers alternative treatment effect estimators.

[16] We show only the lower diagonal of the variance-covariance matrix because $\boldsymbol{\Omega}$ is symmetric. We also omit the cross-unit covariance terms for notational convenience, which are zero under Assumption 4.

$$\boldsymbol{\omega} = \begin{bmatrix} \omega_{i0} \\ \omega_{i1} \\ \omega_{i2} \\ \omega_{i3} \\ \omega_{j0} \\ \omega_{j1} \\ \omega_{j2} \\ \omega_{j3} \end{bmatrix} \qquad \mathbf{\Omega} = \begin{bmatrix} \sigma^2_{i0} & & & & & & & \\ \sigma_{i0,i1} & \sigma^2_{i1} & & & & & & \\ \sigma_{i0,i2} & \sigma_{i1,i2} & \sigma^2_{i2} & & & & & \\ \sigma_{i0,i3} & \sigma_{i1,i3} & \sigma_{i2,i3} & \sigma^2_{i3} & & & & \\ & & & & \sigma^2_{j0} & & & \\ & & & & \sigma_{j0,j1} & \sigma^2_{j1} & & \\ & & & & \sigma_{j0,j2} & \sigma_{j1,j2} & \sigma^2_{j2} & \\ & & & & \sigma_{j0,j3} & \sigma_{j1,j3} & \sigma_{j2,j3} & \sigma^2_{j3} \end{bmatrix}$$

Serial correlation within each unit is represented by the (potentially non-zero) covariance terms $\sigma_{it,is}$ and $\sigma_{jt,js}$, for all $t \neq s$. In contrast, McKenzie's model allows only for constant serial correlation—that is, $v_i \neq 0$ and ω_{it} i.i.d., in our notation. The assumption of i.i.d. idiosyncratic errors would force all off-diagonal covariance elements in $\mathbf{\Omega}$ to equal zero, thereby precluding the types of non-constant serial correlation that BDM highlight (e.g., autoregressive processes).

The magnitudes of these off-diagonal covariance terms directly affect the variance of the DD estimator. The three ψ terms defined above, along with the error variance and experimental design parameters, are sufficient to fully characterize the true variance of the treatment effect estimator in this model. To fix ideas, using the four-period model above and supposing treatment is administered beginning at $t = 2$, these covariance parameters are:

$$\begin{aligned} \psi^B &= \frac{\sigma_{i0,i1} + \sigma_{j0,j1}}{2} \\ \psi^A &= \frac{\sigma_{i2,i3} + \sigma_{j2,j3}}{2} \\ \psi^X &= \frac{\sigma_{i0,i2} + \sigma_{i1,i2} + \sigma_{i0,i3} + \sigma_{i1,i3} + \sigma_{j0,j2} + \sigma_{j1,j2} + \sigma_{j0,j3} + \sigma_{j1,j3}}{8} \end{aligned}$$

Alternatively, if treatment is administered beginning at $t = 1$, these covariance terms become:

$$\begin{aligned}
\psi^B &= \text{(not defined for only 1 pre-treatment period)} \\
\psi^A &= \frac{\sigma_{i1,i2} + \sigma_{i1,i3} + \sigma_{i2,i3} + \sigma_{j1,j2} + \sigma_{j1,j3} + \sigma_{j2,j3}}{6} \\
\psi^X &= \frac{\sigma_{i0,i1} + \sigma_{i0,i2} + \sigma_{i0,i3} + \sigma_{j0,j1} + \sigma_{j0,j2} + \sigma_{j0,j3}}{6}
\end{aligned}$$

Assumption 5 generalizes this structure to a model with J units across m pre-treatment periods and r post-treatment periods. Equation (2) shows that greater average covariance in the pre- or post-treatment periods (ψ^B or ψ^A) increases the MDE. Intuitively, as errors for treated and control units are more serially correlated, the benefits of collecting multiple waves of pre- and post-treatment data are eroded. However, cross-period covariance (ψ^X) enters Equation (2) negatively. This highlights a key property of the DD estimator: because DD identifies the treatment effect off of differences between post- and pre-treatment outcomes, greater serial correlation between pre- and post-treatment observations makes differences caused by treatment easier to detect.

Assuming that the within-unit correlation structure does not vary systematically across time periods, positively correlated errors will imply positive ψ^B, ψ^A, and ψ^X. Because ψ^B and ψ^A enter the SCR formula positively, while ψ^X enters negatively, serial correlation may either increase or decrease the MDE relative to the McKenzie i.i.d. case. Specifically, serial correlation will increase the MDE if and only if:

$$\left(\frac{m-1}{m}\right)\psi^B + \left(\frac{r-1}{r}\right)\psi^A > 2\psi^X \tag{4}$$

This inequality is more likely to hold in longer panels, for two reasons. First, as the number of pre- and post-treatment periods increases, $\left(\frac{m-1}{m}\right)$ and $\left(\frac{r-1}{r}\right)$ approach one. Second, the covariance terms contributing to ψ^X lie farther away from the diagonal of the variance-covariance matrix than the covariance terms contributing to ψ^B and ψ^A. Because errors from non-adjacent time periods are likely to be less correlated than errors from adjacent time periods, and because the number of far-off-diagonal covariances increases relatively more quickly for ψ^X as the panel becomes longer, ψ^X is increasingly likely to be smaller than ψ^B and ψ^A in longer panels.[17]

2.1.2 Monte Carlo simulations

If a randomized experiment relies on a power calculation that fails to properly account for serial correlation *ex ante*, its realized power may be different from the desired κ. To understand the extent to which this matters in practice, we conduct a series of Monte Carlo simulations comparing the McKenzie model and the SCR model over a range of panel lengths and error correlations. We simulate three cases and compute the Type I error rate and the statistical power for each: (i) experiments that fail to account for serial correlation in ω_{it} (i.e. the *idiosyncratic* component of error term) both *ex ante* and *ex post*; (ii) experiments that fail to account for serial correlation in ω_{it} *ex ante* but apply the CRVE to account for serial correlation *ex post*; and (iii) experiments that both account for serial correlation in ω_{it} *ex ante* and apply the CRVE *ex post.*

For each set of parameter values characterizing both a data generating process and an experimental design, we first calculate two treatment effect sizes: τ^{McK} equal to the MDE from the McKenzie formula, and τ^{SCR} equal to the MDE from our SCR formula. Second, we use these parameter values to create a panel dataset from the following DGP:

$$Y_{it} = \beta + \upsilon_i + \delta_t + \omega_{it} \tag{5}$$

where ω_{it} follows an AR(1) process:

$$\omega_{it} = \gamma \omega_{i(t-1)} + \xi_{it} \tag{6}$$

Third, we randomly assign treatment, with effect sizes τ^{McK}, τ^{SCR}, and $\tau^0 = 0$ at the unit level, to create three separate outcome variables. Fourth, we regress each of these outcome variables on their respective treatment indicators and include unit fixed effects and time fixed

[17] This analytical result illustrates how the correlation structure impacts the variance of estimators that use both pre- and post-treatment data. Figure 10 below demonstrates that, in cases with strong serial correlation, adding time periods can actually increase the MDE. This is not specific to the DD estimator: Appendix Figure C3 shows the same pattern for power calculations using the ANCOVA estimator.

effects. Fifth, we compute both OLS standard errors and CRVE standard errors clustered at the unit level, for all three regressions. We repeat steps two through five 10,000 times for each set of parameters, calculating rejection rates of the null hypothesis $\tau = 0$ across all simulations. For τ^{McK} and τ^{SCR}, this rate represents the realized power of the experiment. For the placebo τ^0, it represents the realized false rejection rate.

We test five levels of the AR(1) parameter: $\gamma \in \{0, 0.3, 0.5, 0.7, 0.9\}$. For each γ, we simulate symmetric panels with an equal number of pre-treatment and post-treatment periods, with panel lengths ranging from 2 periods ($m = r = 1$) to 40 periods ($m = r = 20$). We hold J, P, β, σ_v^2, σ_δ^2, α, and κ fixed across all simulations, and we adjust the variance of the white noise term σ_ξ^2 such that every simulation has a fixed idiosyncratic variance σ_ω^2.[18] This allows γ to govern the proportion of σ_ω^2 that is serially correlated.[19] The covariance terms ψ^B, ψ^A, and ψ^X have closed-form expressions under the AR(1) structure, and we use these expressions to calculate τ^{SCR}.[20] This causes τ^{SCR} to vary both with the degree of serial correlation and panel length, whereas τ^{McK} varies only with panel length.

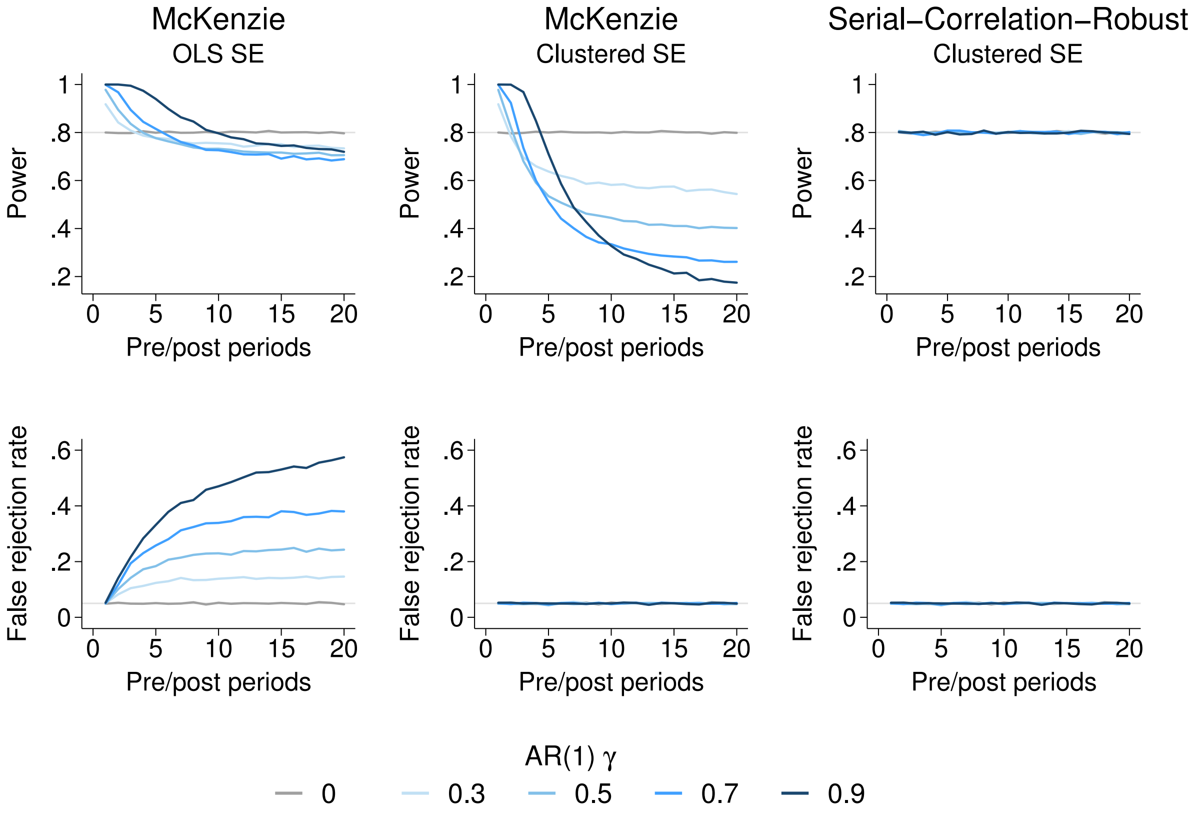
Figure 1 displays the results of this exercise. The left column shows rejection rates under the McKenzie formula using OLS standard errors, which assumes zero serial correlation in ω_{it} both *ex ante* and *ex post*. The middle column shows rejection rates under the FP formula using CRVE standard errors, which accounts for serial correlation in ω_{it} *ex post* only. The right column show rejection rates under our SCR formula using CRVE standard errors, which allows for serial correlation in ω_{it} both *ex ante* and *ex post*. The top row plots realized power as a function of the number of pre/post-treatment periods, which should equal $\kappa = 0.80$ in a properly designed experiment. The bottom row plots the corresponding realized false rejection rates, which should equal the desired $\alpha = 0.05$. Only the SCR formula, in conjunction with CRVE standard errors, achieves the desired 0.80 and 0.05 across all panel lengths and AR(1) parameters.

[18] Note that σ_ω^2 is a parameter of the DGP, as described in Assumption 1; it is not a function of a particular realization of data, desired experiment length, or estimating equation. In this section, we ignore the practicalities of estimating σ_ω^2 (and other DGP parameters) because we know their true underlying values. See Appendix D.1 for a detailed discussion of how to estimate these parameters from data.

[19] In an AR(1) model, the relationship between the variance of the AR(1) process and the variance of the white noise disturbance depends on γ, with $\sigma_\omega^2 = \frac{\sigma_\xi^2}{1-\gamma^2}$.

[20] Appendix B.1 provides these derivations, along with further details on these Monte Carlo simulations.

Figure 1: Standard methods result in improperly powered experiments in AR(1) data



Notes: This figure displays power and false rejection rates from performing power calculations with three different sets of assumptions, using simulated data with AR(1) processes (with differing levels of serial correlation) and differing panel lengths (ranging from 2 ($m = r = 1$) to 40 ($m = r = 20$) periods). In the left column, we apply the standard McKenzie formula (Equation (3), which assumes away non-constant serial correlation), and use OLS standard errors *ex post*, in line with the assumptions of this formula. In the middle column, we again apply the McKenzie formula, but cluster standard errors *ex post*—which is inconsistent with the *ex ante* formula, but corrects for within-unit serial correlation following Bertrand, Duflo, and Mullainathan (2004). In the right column, we apply our serial-correlation-robust formula (Equation (2), which accounts for non-i.i.d. errors *ex ante*), and cluster standard errors *ex post*. As expected, this third set of simulations achieves the desired 80 percent power and 5 percent false rejection rate.

The left column confirms the BDM result that failing to appropriately account for serial correlation leads to false rejection rates dramatically higher than $\alpha = 0.05$. Even a modest AR(1) parameter of $\gamma = 0.5$ yields a 20 percent probability of a Type I error, for panels with $m = r > 5$. This underscores the fact that randomization cannot correct serial correlation in panel settings, and experiments that collect multiple waves of data from the same cross-sectional units should account for within-unit correlations over time. By contrast, the middle and right columns apply the CRVE and reject placebo effects at the desired rate of $\alpha = 0.05$.

The middle column shows how failing to properly account for serial correlation *ex ante* can yield dramatically overpowered or underpowered experiments. Particularly for longer panels with $m = r > 5$, performing power calculations via Equation (3) may actually produce experiments with less than 50 percent power, even though researchers intended to achieve power of 80 percent (i.e., $\kappa = 0.80$). For a relatively high serial correlation of $\gamma = 0.7$, simulations based on the conventional power calculation formula yield power less than 32 percent for $m = r > 10$. This is consistent with the BDM finding that applying the CRVE reduces statistical power, even though doing so achieves the desired Type I error rate. By contrast, the right column applies both the SCR power calculation formula and the CRVE, and these simulations achieve the desired power of $\kappa = 0.80$ for each value of γ.

The middle column also highlights how failing to account for serial correlation *ex ante* may either increase *or* decrease statistical power, as shown in Equation (4). For shorter panels, using the FP formula instead of our SCR formula yields dramatically overpowered experiments. While this may seem counterintuitive, Equation (4) is increasingly unlikely to hold as m and r decrease to 1. In the extreme case where $m = r = 1$, ψ^B and ψ^A do not enter, and the only covariance term in the SCR formula is ψ^X, which enters negatively. These simulations reveal that just as higher γ yields more dramatically underpowered experiments for longer panels, higher γ yields more dramatically *over*powered experiments for shorter panels.[21]

These results are striking. For even a modest degree of AR(1) serial correlation, applying the McKenzie power calculation formula will not yield experiments of the desired statistical power. By contrast, the SCR formula achieves the desired power for all panel lengths and AR(1) parameters. While AR(1) is a relatively simple correlation structure, it serves as a reasonable first-approximation for more complex forms of serial correlation. Given that real-world panel datasets exhibit enough serial correlation to produce high Type I error rates,

[21] Intuitively, serial correlation has two opposite effects on the statistical power of a DD estimator. It decreases power by reducing the effective number of observations for each cross-sectional unit, and it increases power by increasing the signal in estimating treatment effects off of a post−pre difference. In shorter panels, this second effect tends to dominate.

it stands to reason that such serial correlation can similarly impact the statistical power of experiments if not properly accounted for *ex ante*.

2.2 Sensitivities

2.2.1 Short panels

While experiments with long panels have become increasingly common, short-panel experiments remain a development economics staple. Our SCR formula generalizes to panels as short as 2 periods ($m = r = 1$), where Equation (2) simplifies to:

$$MDE = (t^J_{1-\kappa} + t^J_{\alpha/2})\sqrt{\left(\tfrac{1}{P(1-P)J}\right)[2\sigma^2_\omega - 2\psi^X]} \tag{7}$$

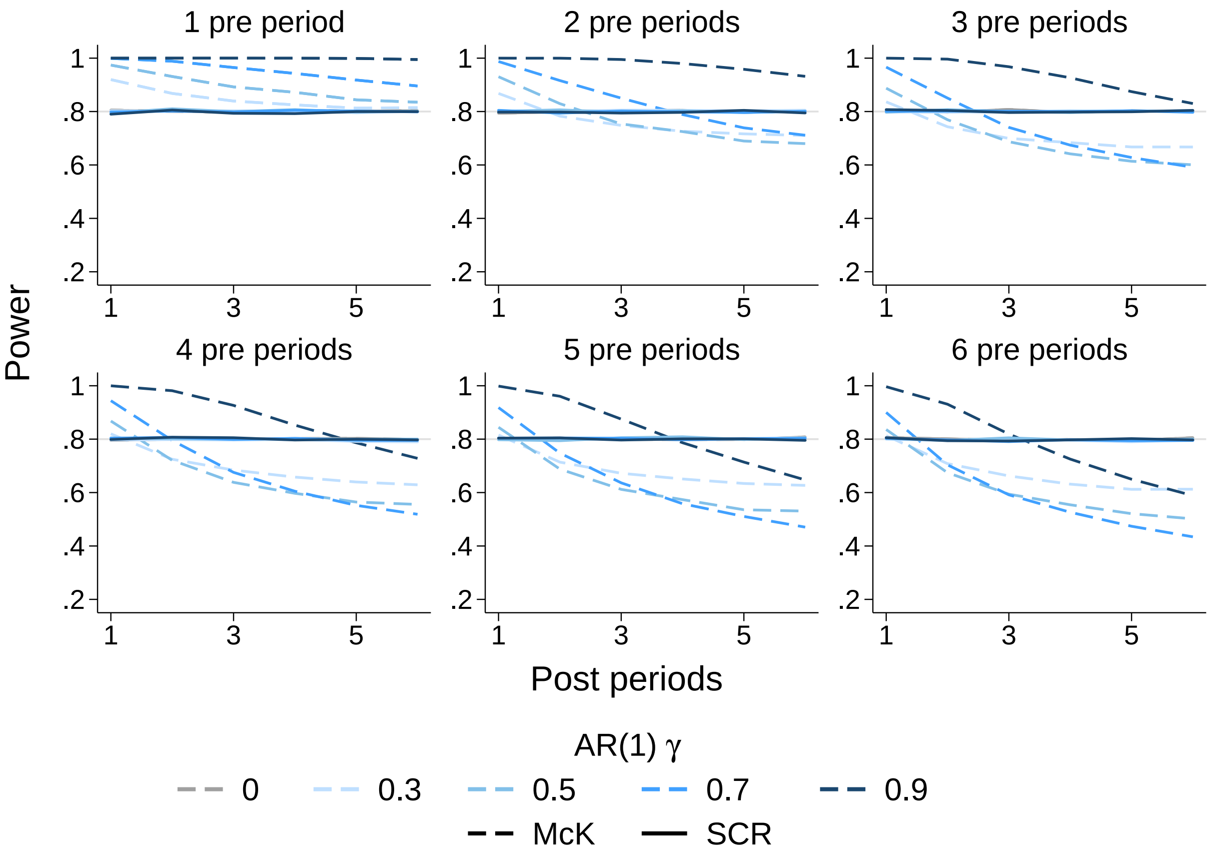
Notably, ψ^B and ψ^A are no longer defined for $m = r = 1$. However, ψ^X remains, meaning that the wedge between the SCR and McKenzie formulas also remains. By omitting ψ^X (which enters Equation (7) negatively), the McKenzie formula will tend to yield over-powered two-period experiments.

It may be unintuitive that serial correlation still matters in a two-period panel, which is isomorphic to a cross-sectional first-differences model, and does not require clustered standard errors to achieve the desired Type I error rate. However, unlike the Type I error rate, statistical power (i.e. 1 minus the Type II error rate) is a function of the variance of the treatment effect estimator, which depends on ψ^X. Hence, ignoring ψ^X from Equation (7) will bias MDE upward, likely leading researchers to choose an unnecessarily large J.[22]

To illustrate the difference between the SCR and McKenzie formulas for short panel experiments, we extend Figure 1 for panels with between 1 and 6 pre/post-treatment periods. We again simulate data with serially correlated errors of varying AR(1) parameters γ, and calibrate treatment effect sizes using each power calculation formula. Figure 2 displays realized power for panels with $1 \leq m \leq 6$ and $1 \leq r \leq 6$, varying m and r independently. Across all levels of non-zero serial correlation ($\gamma > 0$), the McKenzie formula yields over-powered experiments for panels with either 1 pre-treatment or 1 post-treatment period. Consistent

[22] Appendix A.2.4 mathematically demonstrates how power depends on ψ^X even in a two-period panel, despite the fact that the false rejection rate in a two-period panel is independent of ψ^X.

Figure 2: Power in short panels – AR(1) data



Notes: This figure repeats the same simulation exercise as Figure 1, except that we separately vary the number of pre-treatment and post-treatment periods. Each panel conducts simulations with a distinct number of pre-treatment periods ($1 \leq m \leq 6$), and each horizontal axis varies the number of post-treatment periods ($1 \leq r \leq 6$). Dotted lines report realized power for experiments calibrated using the McKenzie formula (i.e. the top-middle panel of Figure 1), while solid lines report realized power using the SCR formula (i.e. the top-right panel of Figure 1). For all cases where $m = 1$ or $r = 1$, the McKenzie formula yields over-powered experiments across the full range of positive AR(1) parameters. This shows that traditional "one baseline, one follow-up" experiments will calibrate to excessively large sample sizes if they ignore non-constant serial correlation *ex ante*. The SCR formula is properly powered in all cases.

with Figure 1, realized power under the McKenzie formula decreases monotonically as panel length increases; the longer the panel, the more likely the McKenzie formula is to deliver underpowered experiments. While the McKenzie formula yields overpowered short panel experiments using these simulated data, it can yield *underpowered* experiments in real data, even for short panels.[23] By contrast, experiments calibrated using the SCR formula achieve the desired power regardless of panel length.

[23] See Appendix Figure C1.

2.2.2 Alternative estimators

Our SCR power calculation formula assumes that researchers will employ a DD estimator with unit and time fixed effects. However, particularly in experiments with short panels, researchers often choose to use one of two alternative approaches: a simplified differences-in-differences estimator or an analysis of covariance (ANCOVA) estimator.

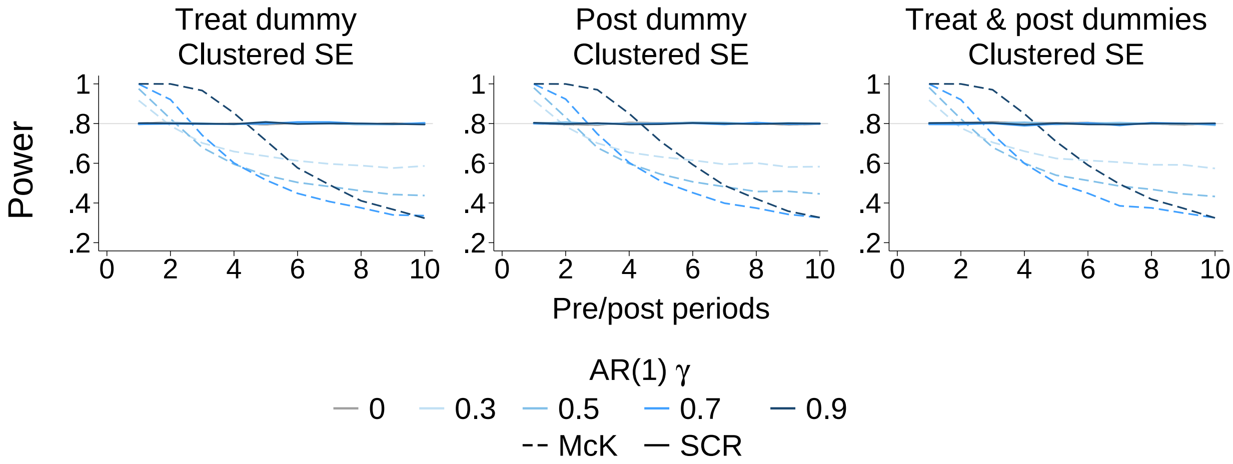
Simplified differences-in-differences Researchers often estimate DD models that replace a full set of unit and time fixed effects with a treatment group dummy and/or post-period dummy, respectively. The resulting estimating equation takes the following form:

$$Y_{it} = \beta + \tau\left[\text{Treat}_i \times \text{Post}_t\right] + \upsilon\left[\text{Treat}_i\right] + \delta\left[\text{Post}_t\right] + \varepsilon_{it} \tag{8}$$

where $\text{Treat}_i = 1$ if unit i is in the treatment group, and $\text{Post}_t = 1$ for all post-treatment periods. Figure 3 explores how the McKenzie and SCR formulas perform under three alternative *ex post* DD estimating equations, which alter our original DD estimating equation by: (i) replacing unit fixed effects with a Treat_i dummy, (ii) replacing time fixed effects with a Post_t dummy, and (iii) replacing both sets of fixed effects with dummies (Equation (8)). In all cases, we use the same DGP as in Assumption 1, which includes idiosyncratic unit and time shocks; as before, we simulate data following Equation (5), with an idiosyncratic error term following an AR(1) process.

As in Figure 1, the SCR formula achieves correctly powered experiments at all levels of serial correlation, for all three DD estimators in Figure 3. This is because all three yield the same treatment effect estimator—and therefore have the same variance—as our original DD estimating equation with unit and time fixed effects. However, these different estimating equations have varying levels of serial correlation in the *error term.* Using OLS standard errors would produce different estimates of the variance, even though the true underlying variance is the same across these estimating equations. On the other hand, with randomized treatment, the CRVE (clustered at the unit level) correctly estimates the (same) variance in all cases, because it accounts for serial correlation within unit, regardless of whether that

Figure 3: SCR outperforms traditional methods with alternative estimators



Notes: This figure repeats the same simulation exercise as Figure 1, except that each panel alters the *ex post* DD estimating equation. The left panel replaces unit fixed effects with a Treat$_i$ dummy. The middle panel replaces time fixed effects with a Post$_t$ dummy. The right panel replaces both sets of fixed effects with Treat$_i$ and Post$_t$ dummies. Dotted lines report realized power for experiments calibrated using the McKenzie formula (i.e. the top-middle panel of Figure 1), while solid lines report realized power using the SCR formula (i.e. the top-right panel of Figure 1). All three DD estimating equations yield power that is identical to Figure 1 (estimated with unit and time fixed effects). This is because the variance of these estimators is a function of the underlying error structure, and the CRVE correctly accounts for unmodeled serial correlation in each case. Hence, the performance gap between McKenzie vs. SCR formulas remains for all DD estimators, provided that researchers account for serial correlation when conducting *ex post* inference.

correlation enters through the underlying error structure or the omission of fixed effects.[24] Realized power under the McKenzie formula is likewise equivalent across all three alternative estimators, yielding overpowered experiments in short panels and underpowered experiments in long panels. In Figure 5 below, we extend this analysis to also consider alternative DGPs.

ANCOVA Another common strategy, particularly in short panels, is to employ the analysis of covariance (ANCOVA) estimator.[25] To do this, the econometrician estimates the

[24] Appendix A.4 mathematically shows that the DD estimator in Equation (8) has the same variance as the DD estimator with unit and time fixed effects.

[25] In a randomized setting where unit fixed effects are not needed for identification, this method may be preferred to DD because it more efficiently estimates $\hat{\tau}$ (Frison and Pocock (1992)). McKenzie (2012) notes that under constant serial correlation (i.i.d. ω_{it} in our SCR framework), ANCOVA is always more efficient than the DD model with the same number of time periods, but that these gains are eroded as the intracluster correlation coefficient increases. These gains are also eroded as the number of pre-treatment periods increases. Neither Frison and Pocock (1992) nor McKenzie (2012) handles the fully general case of arbitrary serial correlation. Teerenstra et al. (2012) begins with a similar setup for the ANCOVA framework, but considers the $m = r = 1$ case only, obviating the need to address the CRVE-related issues raised here.

following specification using post-treatment data only:

$$Y_{it} = \alpha + \tau D_i + \theta \overline{Y}_i^B + \varepsilon_{it} \tag{9}$$

where $\overline{Y}_i^B = \sum_{t=-m+1}^{0} Y_{it}$ is the pre-treatment average value of the dependent variable for unit i. This estimator has become popular in economics, as it is more efficient than the DD model with the same number of periods (McKenzie (2012)).

Frison and Pocock (1992) also derive a power calculation formula for the ANCOVA estimator, based on the same assumptions as their DD formula. As with the DD estimator, McKenzie draws on a more restrictive FP formula, which assumes constant serial correlation within each panel unit. McKenzie's formula (henceforth McKenzie ANCOVA) has become the standard for ANCOVA power calculations, which we adapt to our notation:

$$MDE \approx (t_{1-k}^J + t_{\alpha/2}^J)\sqrt{\left(\tfrac{1}{P(1-P)J}\right)\left[(1-\theta)^2\sigma_v^2 + \left(\tfrac{\theta^2}{m} + \tfrac{1}{r}\right)\sigma_\omega^2\right]} \tag{10}$$

where $\theta = \frac{m\sigma_v^2}{m\sigma_v^2+\sigma_\omega^2}$.[26]

Importantly, deriving this formula under non-constant within-unit serial correlation necessitates an additional simplifying assumption for analytical tractability: we must assume away time shocks.[27] Under this assumption, we derive the variance of the ANCOVA estimator allowing for arbitrary serial correlation, along with the corresponding serial-correlation-

[26] Our McKenzie ANCOVA formula differs slightly from that in McKenzie (2012) and Frison and Pocock (1992, page 1693). These previous derivations have assumed that the true data generating process follows Equation (9), where post-treatment outcomes are determined in part by pre-treatment outcomes. We instead assume unit-specific random effects, as in Assumption 1. FP and McKenzie assumed deterministic time shocks with no variance, or $\sigma_\delta^2 = 0$. We instead must assume that there are no time shocks for analytical tractability. While both data generating processes yield identical treatment effect estimators, they imply different variances of this estimator. As in Frison and Pocock (1992) and McKenzie (2012), our McKenzie ANCOVA formula is approximate because we ignore sampling error in the estimation of θ, which approaches zero as the number of units increases.

[27] Frison and Pocock (1992) and McKenzie (2012) both assume away random time shocks. A critical step in the derivation of the ANCOVA model with time shocks and arbitrary serial correlation requires us to calculate a conditional expectation that depends on the error term ε_{it} and the pre-period mean $\overline{Y}_i^B$ of *every* unit in the experiment, which becomes analytically intractable for any reasonable number of experimental units. See Appendix A.2.5 for more details. By contrast, the variance of the DD estimator depends on the distribution of errors conditional on only the treatment indicator, which is orthogonal to the error terms by randomization.

robust power calculation formula (henceforth SCR ANCOVA):

$$MDE \approx (t_{1-\kappa}^{J} + t_{\alpha/2}^{J}) \times \sqrt{\frac{1}{P(1-P)J}\left[(1-\theta)^2\sigma_\upsilon^2 + \left(\frac{\theta^2}{m} + \frac{1}{r}\right)\sigma_\omega^2 + \frac{\theta^2(m-1)}{m}\psi^B + \frac{r-1}{r}\psi^A - 2\theta\psi^X\right]} \tag{11}$$

where $\theta = \frac{m\sigma_\upsilon^2 + m\psi^X}{m\sigma_\upsilon^2 + \sigma_\omega^2 + (m-1)\psi^B}$.[28]

Figure 4 compares the McKenzie vs. SCR ANCOVA formulas, using simulations analogous to Figure 1. For each level of the AR(1) parameter and panel length, we compute the treatment effect size τ implied by each ANCOVA formula. Unlike Figures 1–3, these simulations use a data-generating process *without* random time shocks (i.e. $\sigma_\delta^2 = 0$), for consistency with the assumptions underlying Equations (10)–(11). The McKenzie ANCOVA formula produces properly-powered experiments only when idiosyncratic errors are i.i.d., while the SCR ANCOVA is robust to all levels of AR(1) serial correlation.

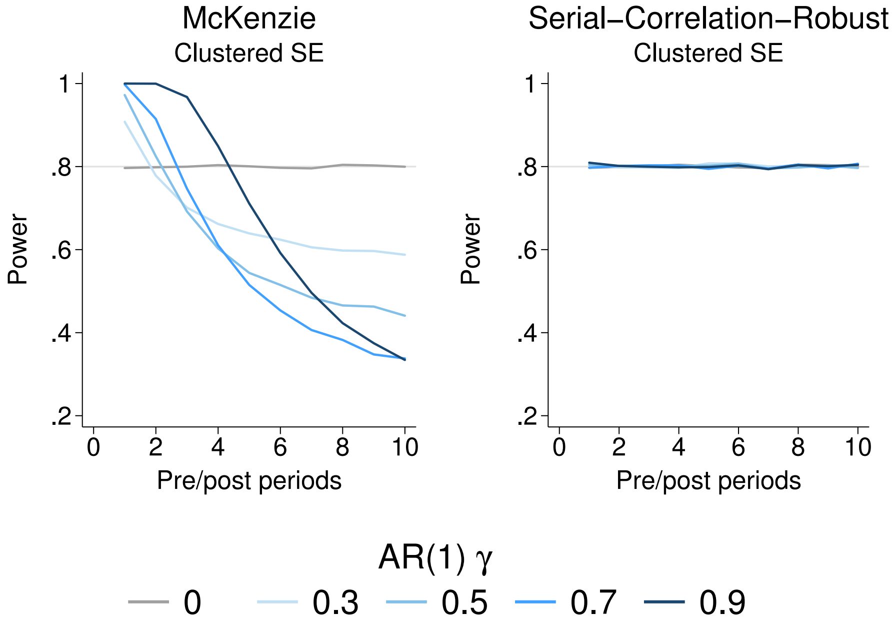
Even though the SCR ANCOVA formula outperforms the McKenzie ANCOVA formula in the presence of AR(1) errors, we do not recommend that researchers use this formula in real-world applications (even in short panels). Time shocks with non-zero variance are a common feature of panel data, and assuming them away may result in improperly powered experiments. We discuss this further in Section 3 below, where we apply our SCR methods to real data. If researchers plan to use the ANCOVA estimator, we recommend that they perform power calculations by simulation, as discussed in Section 4.2.2 below.

2.2.3 Alternative data generating processes

In addition to considering alternative estimators, we also investigate the effectiveness of the SCR power calculation formula under alternative DGPs. We consider three cases: (i) deterministic unit effects ($\sigma_\upsilon^2 = 0$), (ii) deterministic time effects ($\sigma_\delta^2 = 0$), and (iii), deterministic unit *and* time effects ($\sigma_\upsilon^2 = \sigma_\delta^2 = 0$). Given that each is simply a special case of the DGP

[28] We present the formal derivation of the SCR ANCOVA formula in Appendix A.2.5. For analytical tractability, we assume that the ψ parameters are uniform across all units. We also ignore sampling error in the estimation of θ, which approaches zero as the number of units increases; Frison and Pocock (1992) and McKenzie (2012) also make this simplification. Through additional Monte Carlo simulations, we confirm that neither of these assumptions is likely to affect statistical power.

Figure 4: Standard ANCOVA methods fail under serial correlation

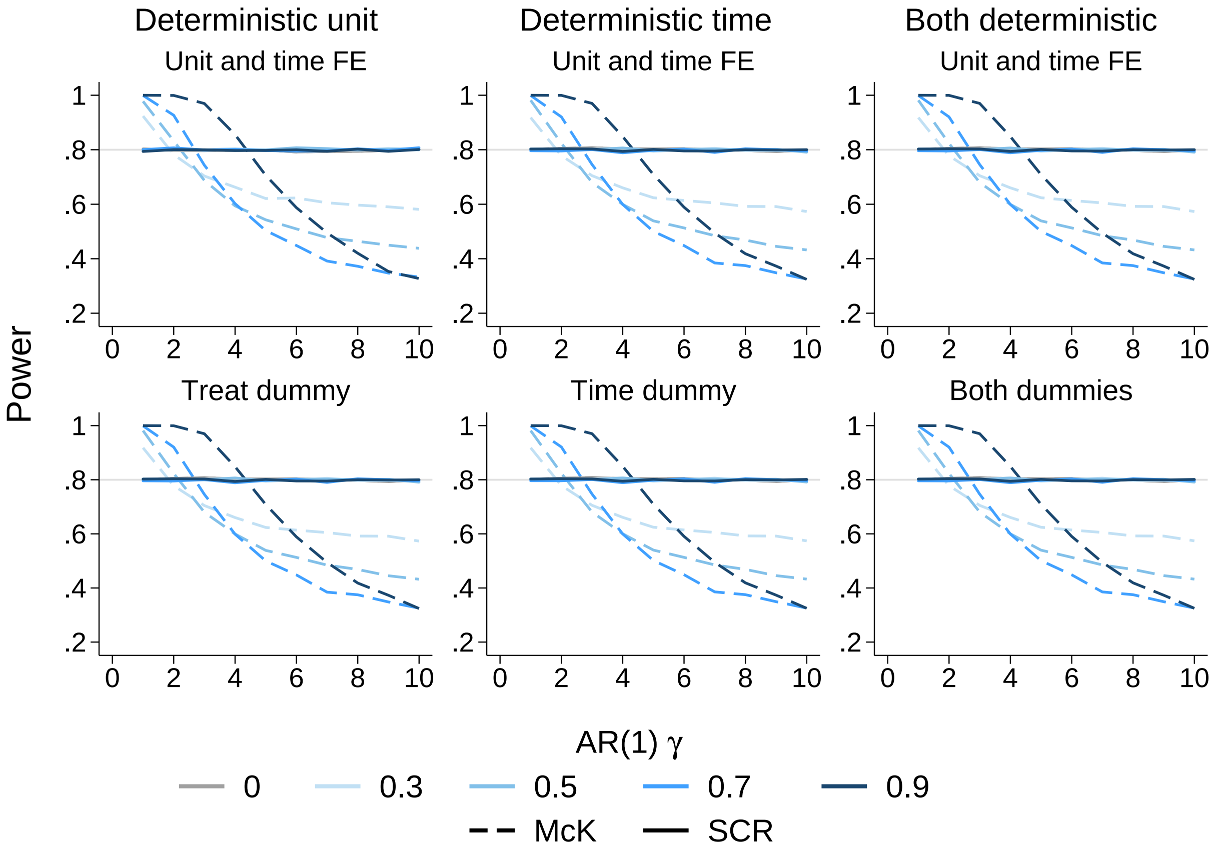


Notes: This figure repeats the same simulation exercise as Figure 1, using ANCOVA power calculation formulas *ex ante* and an ANCOVA estimating equation *ex post*. The left panel calibrates MDE using the McKenzie ANCOVA formula (Equation (10), and is analogous to the top-middle panel of Figure 1. The right panel calibrates MDE using the SCR ANCOVA formula (Equation (11), and is analogous to the top-right panel of Figure 1. Both set of simulations estimate Equation (9) *ex post*, cluster standard errors by unit, and simulate DGPs without random time shocks ($\sigma^2_\delta = 0$). As in the DD simulations, the McKenzie ANCOVA formula fails to generate correctly-powered experiments, whereas the SCR formula is properly powered across all panel lengths and degrees of AR(1) correlation.

in Assumption 1, we expect the SCR to continue to achieve the desired power. However, these cases provide more favorable conditions to test the McKenzie formula: in the absence of random unit or time shocks, the variance of the *composite* error term is the same as the variance of the *idiosyncratic* error term (ω_{it}).

We test the relative performance of the SCR and McKenzie formulas in Figure 5. The top row presents simulations using the three alternative DGPs, while holding the fixed effects estimator fixed. In the bottom row, we match each DGP with its "matched" estimator (e.g., for the DGP with deterministic unit effects, we replace unit fixed effects with a Treat_i dummy). The SCR formula yields correctly powered experiments, even under alternative DGPs with misspecified *ex post* estimators. In contrast, the McKenzie formula generates improperly powered experiments even in the absence of unit and time shocks. Next, we test

Figure 5: SCR outperforms traditional methods with alternative DGPs



Notes: This figure repeats the same simulation exercise as Figure 1, for alternative DGPs and DD estimating equations. The left column removes random unit effects from the DGP ($\sigma_\upsilon^2 = 0$); the middle column removes random time effects from the DGP ($\sigma_\delta^2 = 0$); and the right column removes both random unit effects and random time effects ($\sigma_\upsilon^2 = \sigma_\delta^2 = 0$). The top row uses an *ex post* estimator with unit and time fixed effects (as in Figure 1, albeit now misspecified); the bottom rows replaces fixed effects with dummy variables (Treat_i and/or Post_t, as in Figure 3) to align with each DGP, respectively. Dotted lines report realized power for experiments calibrated using the McKenzie formula (i.e. the top-middle panel of Figure 1), while solid lines report realized power using the SCR formula (i.e. the top-right panel of Figure 1). In all cases, our results are identical to Figure 1 (DGP with random unit and time effects; DD estimator with unit and time fixed effects). This show that our SCR formula is robust to alternative DGPs, which may not match the DD estimator. By contrast, the McKenzie formula continues to perform poorly in the presence of non-constant serial correlation—even after removing intracluster correlation, random time shocks, and fixed effects.

the performance of the SCR formula when we do not control the DGP, using real-world panel data.

3 Applications to real-world data

3.1 Bloom et al. (2015) data

In this section, we conduct an analogous simulation exercise using a real dataset from an experiment in a developing-country setting. These data come from Bloom et al. (2015), in which Chinese call center employees were randomly assigned to work either from home or from the office for a nine-month period.[29] The authors estimate the following equation to derive the central result, reported in Table 2 in the original paper:

$$\text{Performance}_{it} = \alpha \text{Treat}_i \times \text{Experiment}_t + \beta_t + \gamma_i + \varepsilon_{it} \tag{12}$$

This is a standard DD estimating equation with fixed effects for individual i and week t. From this model's residuals, we estimate an AR(1) parameter of $\hat{\gamma} = 0.233$, which is highly statistically significant and indicates that these worker performance data exhibit weak serial correlation.

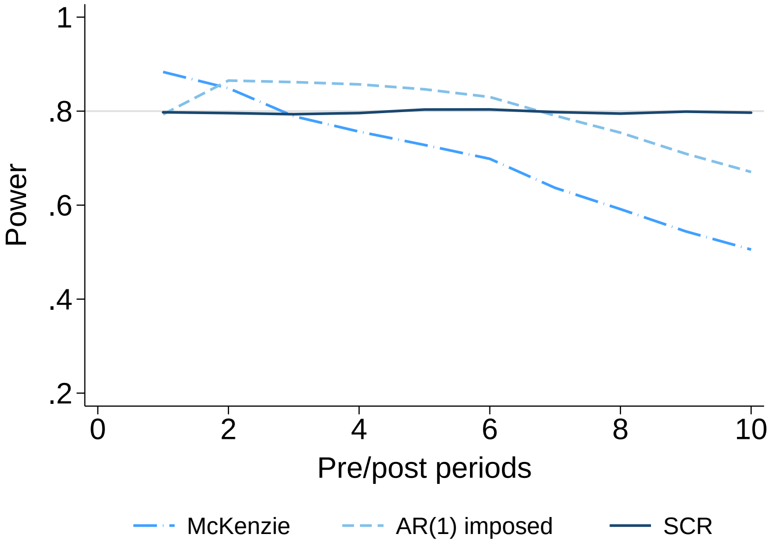
We perform Monte Carlo simulations on this dataset that are analogous to those presented above. We subset consecutive periods of the Bloom et al. (2015) dataset to create panels ranging in length from 2 periods ($m = r = 1$) to 20 periods ($m = r = 10$). For each simulation panel length, we randomly assign three treatment effect sizes, τ^{McK}, $\tau^{AR(1)}$, and τ^{SCR}, at the individual level and estimate Equation (12) separately for each treatment effect size. We calibrate τ^{McK} using the McKenzie formula that assumes no serial correlation;[30]

$\tau^{AR(1)}$ using the SCR formula with ψ parameters consistent with an AR(1) error structure of $\gamma = 0.233$; and τ^{SCR} using the SCR formula with non-parametrically estimated $\psi_{\hat{\omega}}$ parameters. We define $\sigma^2_{\hat{\omega}}$, $\psi^B_{\hat{\omega}}$, $\psi^A_{\hat{\omega}}$ and $\psi^X_{\hat{\omega}}$ to be the estimated analogues of σ^2_{ω}, ψ^B, ψ^A,

[29] This dataset consists of weekly performance measures for the 249 workers enrolled in the experiment between January 2010 and August 2011. We keep only those individuals who have non-missing performance data for the entire pre-treatment period, leaving us with a balanced panel of 79 individuals over 48 pre-treatment weeks (a different sample from that in the paper). Our purpose with this exercise is not to comment on the statistical power of the original paper, but rather to investigate the importance of accounting for serial correlation *ex ante* in real experimental data. Appendix B.2 provides more information on this simulation dataset, including summary statistics.

[30] In Appendix C.4, we apply McKenzie's alternative empirical approach: calculating $\tau^{McK-Avg}$ by parameterizing his power calculation formula using the average autocorrelation for panels of length $(m + r)$. In both this dataset and the Pecan Street data described in Section 3.2, this alternative approach yields underpowered experiments, except in two-period panels where $m = r = 1$.

Figure 6: Power simulations for Bloom et al. (2015) data



Notes: This figure shows results from Monte Carlo simulations using Bloom et al. (2015) data. Each curve plots realized power for a DD experiment with a certain number of pre/post-treatment periods (from $m = r = 1$ to $m = r = 10$), using different *ex ante* assumptions. The long-dashed line applies the McKenzie formula (Equation (3)), which assumes away non-constant serial correlation. The short-dashed line applies the SCR formula, under the assumption of AR(1) serial correlation (using Equation (6) to estimate an AR(1) parameter). The solid line applies the SCR formula (Equation (2)), where we non-parametrically estimate $\psi^B_{\hat{\omega}}$, $\psi^A_{\hat{\omega}}$, and $\psi^X_{\hat{\omega}}$ terms from the (residualized) Bloom et al. (2015) dataset. All simulations apply the CRVE *ex post*, clustering at the individual level. Only the SCR formula achieves the desired 80 percent power, even though the Bloom et al. (2015) data exhibit relatively weak serial correlation.

and ψ^X, where the subscript $\hat{\omega}$ denotes the variance/covariance of *residuals* rather than errors. Unlike Section 2.1.2 above, where the simulated DGP is known, real datasets require researchers to estimate these residual-based parameters to properly implement our SCR formula.[31]

Figure 6 reports the results of this exercise, demonstrating that only the SCR formula achieves the desired statistical power in the Bloom et al. (2015) data. Failing to account for non-constant serial correlation leads to experiments that deviate dramatically from 80 percent power, even when that serial correlation is relatively weak. For an experiment with 10 pre/post-treatment periods, applying the McKenzie formula with $\kappa = 0.80$ yields an experiment with only 51 percent power. This is consistent with our results from simulated

[31] Appendix D.1 explains how to estimate these residual-based parameters, which is not trivial and requires small-sample corrections. Appendix E shows how to correct for estimation bias in $\sigma^2_{\hat{\omega}}$, $\psi^B_{\hat{\omega}}$, $\psi^A_{\hat{\omega}}$ and $\psi^X_{\hat{\omega}}$, in order to calculate an unbiased MDE. In real-world settings, researchers will need to estimate parameters of an unobserved DGP, as we do here.

data, demonstrating that researchers can calibrate a panel RCT to 80 percent power if the *ex ante* formula properly accounts for the within-unit correlation structure of the data.[32]

ANCOVA in real data The ANCOVA estimator is more efficient than the DD estimator. As discussed in Section 2.2.2, for DGPs with no time shocks, our SCR ANCOVA formula can achieve properly powered experiments; however, an SCR ANCOVA formula that accommodates random time shocks is not analytically tractable. Here, we test McKenzie and SCR ANCOVA formulas using the Bloom et al. (2015) data, where the true DGP is unknown, and may contain time shocks. Figure 7 displays the results of these power calculations in simulated experiments of varying length, with $m \in \{1, 5, 10\}$ and $1 \leq r \leq 10$. Neither formula consistently yields properly powered experiments. The McKenzie formula comes closest with one pre-treatment period and many post-treatment periods, but performs poorly for $m > 1$ or $r < 4$. On the other hand, the SCR ANCOVA performs reasonably well with 10 pre-treatment periods, but yields overpowered experiments with $m \leq 5$ or $r < 5$.

Given that neither formula delivers properly-powered experiments with real data, we caution against using ANCOVA power calculation formulas in practice. When researchers intend to estimate an ANCOVA model *ex post*, we suggest they conduct *ex ante* power calculations using simulation-based methods (see Section 4.2.2 below).[33]

3.2 Pecan Street data

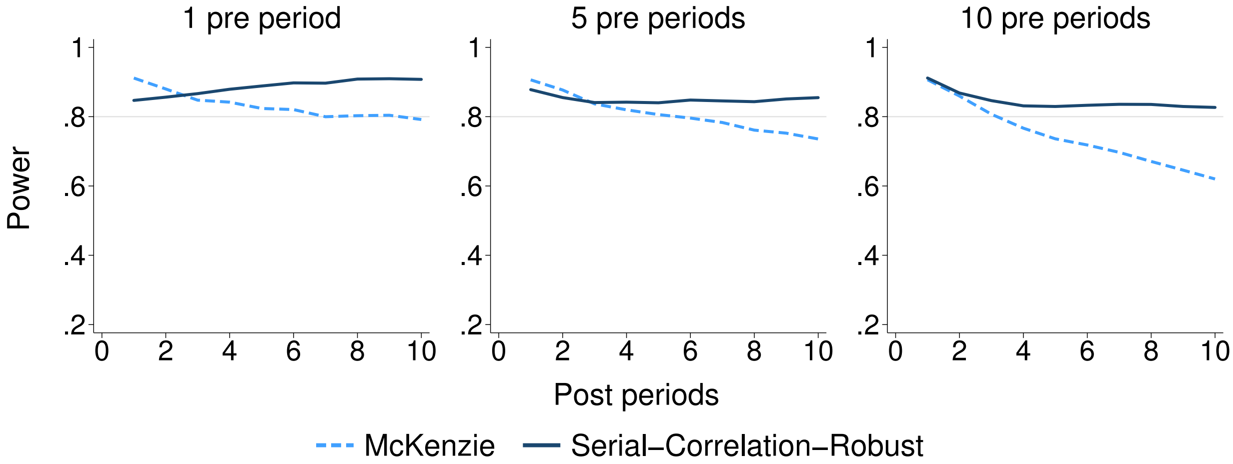
Having demonstrated the importance of properly accounting for serial correlation using data from an RCT in a developing country setting, we now turn to a much higher-frequency dataset with higher serial correlation: household electricity consumption in the United States (Pecan Street (2016)).[34] Electricity consumption data tend to exhibit high within-household

[32] Appendix C.1 replicates the sensitivity analyses from Section 2.2 using the Bloom et al. (2015) data (and the Pecan Street data discussed below). We again find that the SCR formula achieves 80 percent power for short panels and DD estimators without fixed effects, while the McKenzie formula fails in both cases.

[33] If researchers lack access to pre-existing data for performing simulations, one option is to apply the SCR DD power calculation formula to calibrate the sample size *ex ante*, but then estimate ANCOVA *ex post*. Since ANCOVA is (weakly) more efficient than DD, this will tend to yield overpowered experiments.

[34] Pecan Street is a research organization, based at the University of Texas at Austin, that makes high-resolution energy usage data available to academic researchers. The raw data, which are available with a login from `https://dataport.pecanstreet.org/data/interactive`, consist of hourly electricity consumption

Figure 7: ANCOVA power calculations in real data



Notes: This figure repeats the same exercise as Figure 6, for ANCOVA experiments simulated on the Bloom et al. (2015) dataset. Each panel simulates experiments with a certain number of pre-treatment periods ($m \in \{1, 5, 10\}$), and horizontal axes vary the number of post-treatment periods ($1 \leq r \leq 10$). Dotted lines apply the McKenzie ANCOVA formula *ex ante* (Equation (10)), which assumes away non-constant serial correlation. Solid lines apply the SCR ANCOVA formula *ex ante* (Equation (11)), where we non-parametrically estimate $\psi^B_{\hat{\omega}}$, $\psi^A_{\hat{\omega}}$, and $\psi^X_{\hat{\omega}}$ terms from the (residualized) Bloom et al. (2015) dataset. All cases estimate ANCOVA *ex post* (Equation (9)), clustering standard errors by individual. Both formulas generate improperly powered experiments with real data that likely have time shocks, which neither formula can account for.

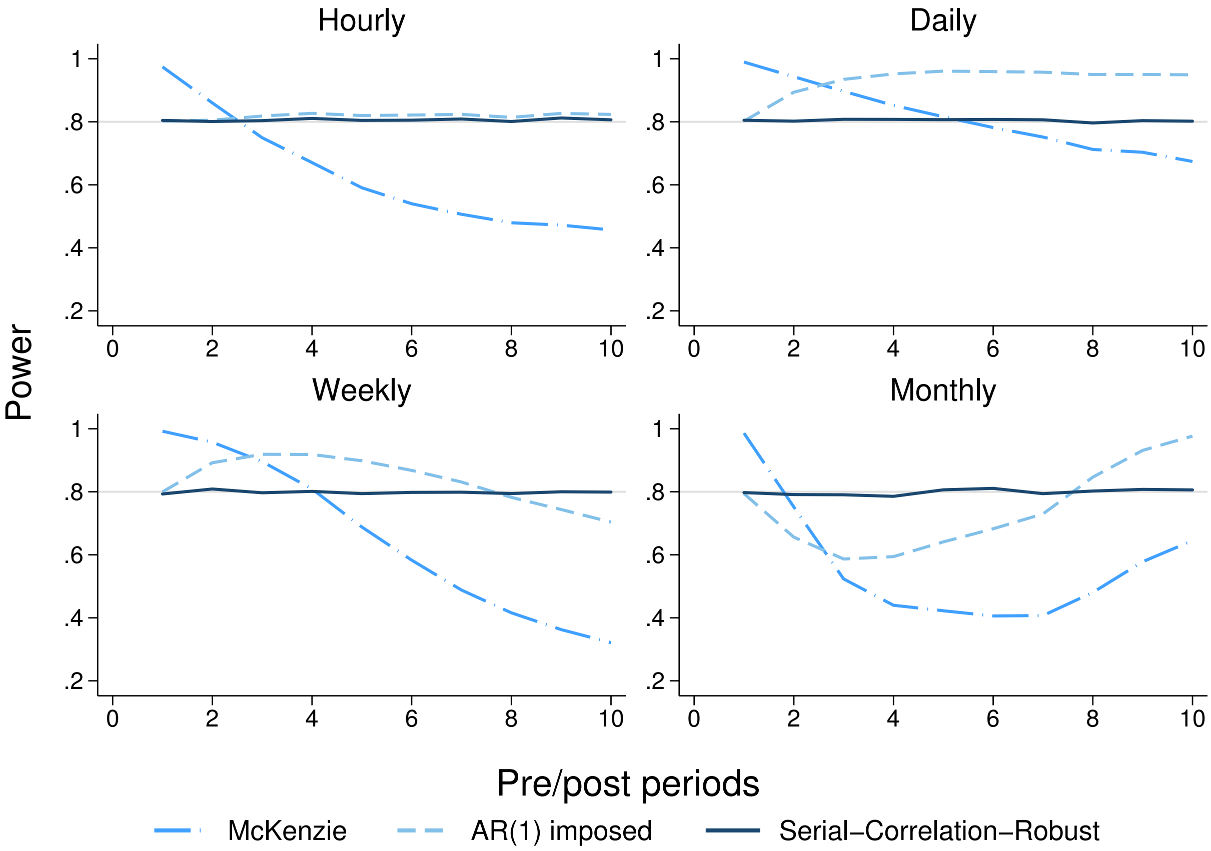
autocorrelation, making them particularly well-suited for this analysis. Additionally, RCTs using energy consumption data are becoming increasingly common in economics, making our Pecan Street application relevant to this growing literature.[35]

We aggregate these data to four different temporal frequencies: hourly, daily, weekly, and monthly, each with a different correlation structures and amounts of idiosyncratic variation. This allows us to compare the McKenzie vs. SCR power calculation formula over a range of underlying error structures. We conduct Monte Carlo simulations on all four temporal frequencies, following the same procedure as the Bloom et al. (2015) simulations. Figure 8 shows that in all four cases, realized power sharply deviates from the desired 80 percent under both the McKenzie assumption of i.i.d. idiosyncratic errors and an assumed

for 699 households over 26,888 hours. Appendix B.3 provides further detail on both these data and the ensuing simulations.

[35] For example, see Allcott (2011); Jessoe and Rapson (2014); Ito, Ida, and Tanaka (2018); Fowlie, Greenstone, and Wolfram (2018); Fowlie et al. (2017); Allcott and Greenstone (2017); and Jack and Smith (2019). There is also a large quasi-experimental literature that uses energy consumption data.

Figure 8: Power simulations for Pecan Street data



Notes: This figure conducts simulations that are identical to Figure 6, using four Pecan Street datasets collapsed to different levels of temporal frequency. Each curve plots realized power for a DD experiment with a certain number of pre/post-treatment periods (from $m = r = 1$ to $m = r = 10$), using different *ex ante* assumptions. The long-dashed lines apply the McKenzie formula (Equation (3)), which assumes away non-constant serial correlation. The short-dashed lines apply the SCR formula, under the assumption of AR(1) serial correlation (using Equation (6) to estimate an AR(1) parameter). The solid lines apply the SCR formula (Equation (2)), where we non-parametrically estimate $\psi^B_{\hat{\omega}}$, $\psi^A_{\hat{\omega}}$, and $\psi^X_{\hat{\omega}}$ terms from the (residualized) Pecan Street data. All simulations apply the CRVE *ex post*, clustering at the individual level. While each temporal frequencies exhibits a unique correlation structure, only the SCR power calculation formula achieves the desired power in each case.

AR(1) structure. We achieve correctly powered experimental designs only by applying the SCR method, which accounts for the full covariance structure of the Pecan Street data.[36]

[36] As with the Bloom et al. (2015) dataset, we estimate parameters of the unobserved DGP for each temporal frequency. Appendix C.4 shows again that the McKenzie (2012, p. 215) approach yields underpowered experiments for all four temporal frequencies for panels longer than 2 periods.

3.3 Power calculations in real data

To operationalize our SCR power calculation formula in practice, researchers must assume values for σ^2_ω, ψ^B, ψ^A, and ψ^X that reflect the error structure likely to be present in (future) experimental datasets. In the best case scenario, researchers have access to data that are representative of what will be collected in the field, and they can estimate these variance and covariance terms from this pre-existing dataset.[37] Plugging these estimates into the SCR formula, researchers can evaluate the tradeoffs between desired power (κ), number of units (J), pre- and post-treatment observations per unit (m, r, respectively), proportion of the population treated (P), and expected effect size (MDE).

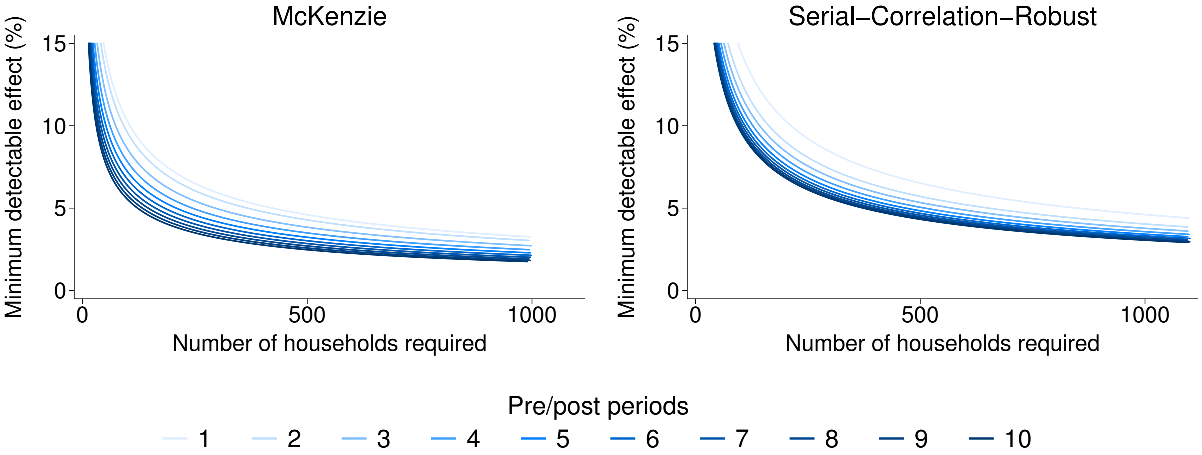
We perform this procedure on the daily Pecan Street dataset to imitate the design of an experiment that affects household electricity consumption. We do so both assuming i.i.d. idiosyncratic errors (using the McKenzie formula) and allowing for arbitrary serial correlation (using the SCR formula). For simplicity, we consider only balanced panels of households with the same number of observations before and after treatment (i.e. $m = r$). For each panel length, we re-estimate $\sigma^2_{\hat{\omega}}$ and $\psi_{\hat{\omega}}$ terms from the daily Pecan Street data.[38]

We plot the results of this exercise in Figure 9. The left panel applies the McKenzie formula and assumes i.i.d. idiosyncratic errors. The right panel applies the SCR formula, using our non-parametric estimates of $\psi^B_{\hat{\omega}}$, $\psi^A_{\hat{\omega}}$, and $\psi^X_{\hat{\omega}}$ to reflect the real error structure of the data. Each curve corresponds to experiments of a particular length (ranging from $m = r = 1$ to $m = r = 10$), and plots the number of households (J) required to achieve 80 percent power as a function of MDE. In all cases, longer panels lower J needed to achieve a given MDE. However, the McKenzie formula always calls for substantially fewer households than the SCR formula. For example, with $m = r = 10$ and an MDE of 5 percent, the McKenzie formula solves for 123 households, while the SCR formula solves for 374 households—over 3 times greater sample size. Hence, if a researcher in this setting

[37] Appendix D.1 provides details on how to estimate $\sigma^2_{\hat{\omega}}$, $\psi^B_{\hat{\omega}}$, $\psi^A_{\hat{\omega}}$, and $\psi^X_{\hat{\omega}}$ from pre-existing data, and Appendix E proves that power calculations using estimated parameters recover the same MDE in expectation as those using true parameters. The plausibility of estimating these parameters will vary across settings. Researchers with implementing partners that have access to large amounts of historical data may use these data to estimate $\sigma^2_{\hat{\omega}}$, $\psi^B_{\hat{\omega}}$, $\psi^A_{\hat{\omega}}$, and $\psi^X_{\hat{\omega}}$. On the other hand, this may not be possible for experiments in completely unstudied settings. See Appendix D.3 for more details on how to overcome a lack of pre-experimental data.

[38] We fix $P = 0.5$, $\kappa = 0.80$, and $\alpha = 0.05$. See Appendix B.4 for details.

Figure 9: Analytical power calculations – daily Pecan Street dataset



Notes: This figure shows the result of analytic power calculations on Pecan Street electricity data, collapsed to the daily level. Each curve displays the number of units required to detect a given minimum detectable effect with 80 percent power. Each iso-power curve corresponds to a particular panel length, with the shortest panel (1 pre-period, 1 post-period) in light blue, and the longest panel (10 pre-periods, 10 post-periods) in navy. The left panel shows a power calculation using the standard McKenzie (2012) formula, which does not accommodate non-constant serial correlation. The right panel applies the serial-correlation-robust formula, which accounts for the real error structure of the data. In this setting, failing to account for the full covariance structure will dramatically understate the sample size required to detect a given effect.

applied the CRVE *ex post* but assumes i.i.d. idiosyncratic errors *ex ante*, he would likely include too few households to achieve the desired statistical power.

4 Power calculations in practice

4.1 Trading off units and time periods

Recruiting participants, administering treatment, and collecting data are all costly, and these implementation costs are often the limiting factor in study size. We can use the power calculation framework to conceptualize the optimal design of a panel RCT given a budget, by couching it in a simple constrained optimization problem of the following form:

$$\min_{P,J,m,r} MDE(P,J,m,r) \quad \text{s.t.} \quad C(P,J,m,r) \leq B \tag{13}$$

where $C(P,J,m,r)$ is the cost of conducting an experiment and B is the experiment's budget.

The budget constraint creates a fundamental tradeoff between including additional units and including additional time periods in the experiment, since each comes at a cost.[39] This tradeoff also arises from differences in the marginal effects of units and time periods on the MDE. Using our SCR formula, the "elasticities" of the MDE with respect to number of units and number of time periods are:

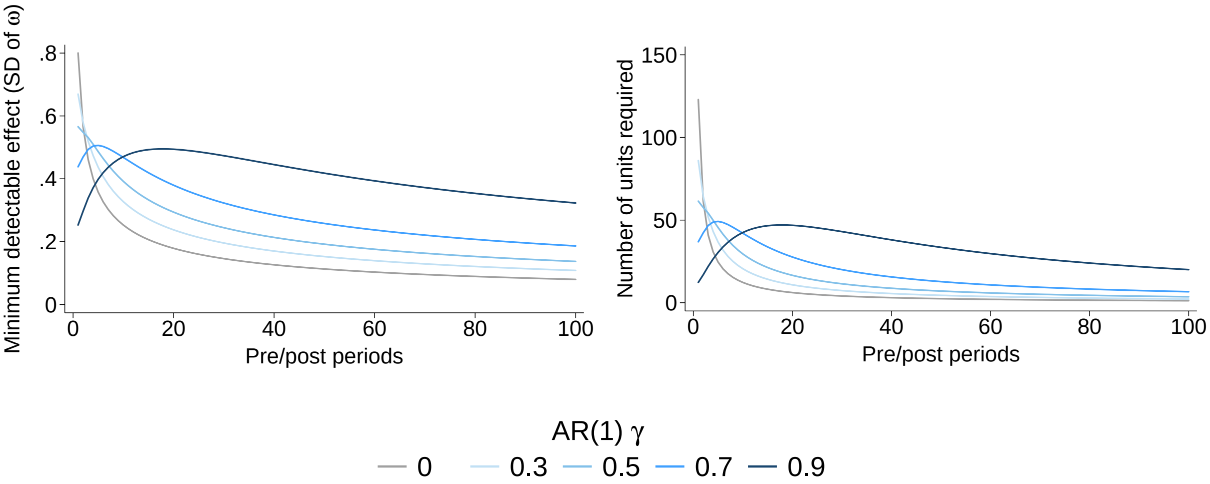
$$\begin{aligned}
\frac{\partial MDE/MDE}{\partial J/J} &= -\frac{1}{2} \\
\frac{\partial MDE/MDE}{\partial m/m} &= -\frac{1}{2}\left[\frac{\frac{\sigma_\omega^2}{m} - \frac{\psi^B}{m} - (m-1)\frac{\partial\psi^B}{\partial m} + 2m\frac{\partial\psi^X}{\partial m}}{\left(\frac{m+r}{mr}\right)\sigma_\omega^2 + \left(\frac{m-1}{m}\right)\psi^B + \left(\frac{r-1}{r}\right)\psi^A - 2\psi^X}\right] \\
\frac{\partial MDE/MDE}{\partial r/r} &= -\frac{1}{2}\left[\frac{\frac{\sigma_\omega^2}{r} - \frac{\psi^A}{r} - (r-1)\frac{\partial\psi^A}{\partial r} + 2r\frac{\partial\psi^X}{\partial r}}{\left(\frac{m+r}{mr}\right)\sigma_\omega^2 + \left(\frac{m-1}{m}\right)\psi^B + \left(\frac{r-1}{r}\right)\psi^A - 2\psi^X}\right]
\end{aligned}$$

There is a constant elasticity of MDE with respect to J of -0.5, meaning that a 1 percent increase in the number of units always yields a 0.5 percent reduction in the MDE. However, the elasticity of MDE with respect to m and r depends on the error structure and panel length.[40] For some parameter values, this elasticity can be positive, such that increasing the length of the experiment would actually *increase* the MDE. This may seem counter-intuitive, but adding time periods can reduce the average covariance between pre- and post-treatment observations (ψ^X), which introduces more noise in the estimation of pre- vs. post-treatment difference. For relatively short panels with errors that exhibit strong serial correlation, this effect can dominate the benefits of collecting more time periods.

[39] Researchers may also adjust P to make an experimental design more cost effective. An RCT will have the lowest MDE at $P = 0.5$, but if control units are cheap compared to treatment units, the same power may be achieved at lower cost by decreasing P and increasing J. See Duflo, Glennerster, and Kremer (2007) for more details. We also typically consider α and κ to be fixed "by convention." While α is the product of research norms, and therefore relatively inflexible, researchers may want to adjust κ. $1 - \kappa$ is the probability of being unable to distinguish a true effect from 0. In lab experiments which are cheaply replicated, researchers may accept $\kappa < 0.80$, whereas in large, expensive field experiments that can only be conducted once, researchers may instead wish to set $\kappa > 0.80$. Researchers may also choose to size their experiments such that they achieve a power of 80 percent for the smallest economically meaningful effect, even if they expect the true MDE to be larger.

[40] Note that J, m, and r must all be integer-valued, hence these derivatives serve as continuous approximations of discrete changes in these parameters. Likewise, the partial derivatives of ψ^B, ψ^A, and ψ^X with respect to m and r are not technically defined, as these covariance terms are averaged over discrete numbers of periods (as shown in Assumption 5).

Figure 10: Analytical power calculations with increasing panel length



Notes: This figure displays the results of analytical power calculations using the SCR formula for varying AR(1) parameters. The left panel shows the tradeoff between the minimum detectable effect (MDE) and the number of time periods ($m = r$) for varying levels of serial correlation, holding the number of units fixed at $J = 100$ and normalizing MDE by the standard deviation of ω_{it}. At low levels of γ, MDE declines monotonically in m and r. However, for higher γ, increasing m and r actually *increases* MDE when $m = r$ is relatively small, and *decreases* MDE when $m = r$ is relatively large. The right panel shows the relationship between the number of units (J) and number of pre/post periods ($m = r$) required to detect an MDE equal to one standard deviation of ω_{it}. Similarly, for low levels of serial correlation, the trade-off between J and $m = r$ is monotonic. However, as γ increases, adding periods in short panels necessitates a greater number of units to achieve the same MDE, while adding periods in longer panels means that fewer units are required to achieve the same MDE. Appendix Figure C3 replicates this result using the SCR ANCOVA formula.

Figure 10 illustrates how adding time periods may either increase or decrease power. The left panel plots the MDE of an experiment as a function of the number of pre- and post-treatment periods, holding the number of units J constant. The right panel depicts the tradeoff between adding units vs. time periods by plotting the combinations of J and $m = r$ that yield a given MDE. We use the SCR formula to analytically construct these curve, assuming AR(1) idiosyncratic errors with varying γ values.

At low to moderate levels of serial correlation, increasing the panel length always reduces MDE given J, and vice versa. However, at higher levels of serial correlation, this relationship is no longer monotonic. For $\gamma \geq 0.6$, marginally increasing m or r in a relatively short panel increases MDE for a given J, and likewise increases J required to achieve a given MDE.

This suggests that in settings with strong non-constant serial correlation, adding periods of data might *decrease* statistical power if the panel is not sufficiently long.[41]

4.2 Power calculations in Stata

To facilitate implementation of *ex ante* power calculations in practice, we have developed the Stata package `pcpanel`.[42] This software implements panel data power calculations that properly account for serial correlation, a feature not present in Stata's built-in `power` command.[43] `pcpanel` operationalizes our SCR formula based on *either* user-input assumptions on the correlation structure *or* nonparametric estimates of variance/covariance parameters from a pre-existing dataset. Our package also executes power calculations by simulation, accommodating a range of experimental designs outside the scope of this paper's analytical framework.

4.2.1 Analytical DD power calculations

The program `pc_dd_analytic` conducts analytic power calculations for DD experiments using our SCR formula (Equation (2)). `pc_dd_analytic` takes (exactly two of) sample size (J), a desired MDE, and a desired power (κ), and returns the third. It behaves similarly to the legacy Stata function `sampsi`, except that `pc_dd_analytic` accepts the *idiosyncratic* residual variance σ_ω^2, equal to $\sigma_\epsilon^2(1-\rho)$ in the notation of McKenzie (2012).[44]

[41] McKenzie (2012) argues that stronger unit-specific shocks (i.e. higher σ_v^2) can erode the benefits of collecting additional waves of data. Here, we extend that argument to within-unit serial correlation, demonstrating that higher autocorrelation in the idiosyncratic error term can similarly erode—and even reverse—the benefits of increased panel length. This result reflects the analytical properties of estimators that leverage both pre- and post-treatment data, and does not reflect the DD estimator over-controlling for pre-period data. Appendix Figure C3 replicates this figure using the SCR ANCOVA formula and finds the same result.

[42] `pcpanel` is available for download from SSC. Both Appendix D.4 and the `pcpanel` help file describe the software package in further detail.

[43] While `power` does allow for repeated-measures ANOVA, it cannot accommodate standard panel research designs used in economics, including DD and ANCOVA. The previous Stata power calculations command, `sampsi`, did allow for this, but it was depreciated as of 2013 and is no longer supported.

[44] For example, `-sampsi 0 10, n1(150) n2(150) pre(3) post(5) sd(50) r1(0.3) m(change)-` becomes `-pc_dd_analytic, mde(10) n(300) p(0.5) pre(3) post(5) var(1750)-`, as $1750 = 50^2{*}(1-0.3)$. Both of these commands yield a power of 0.81. For very small samples, `sampsi` and `pc_dd_analytic` will yield slightly different results, as `pc_dd_analytic` uses a finite-sample t distribution to calculate critical values, whille `sampsi` uses a Normal distribution.

Users have two options for incorporating non-constant serial correlation. First, they may allow the subprogram `pc_dd_covar` to nonparametrically estimate the average covariance structure of a pre-existing dataset, using the same procedure that generated the solid lines in Figures 6 and 8. By characterizing the relevant components of the data's actual correlation structure, `pc_dd_covar` enables users with representative pre-existing data to accurately estimate $\sigma^2_{\hat{\omega}}$, $\psi^B_{\hat{\omega}}$, $\psi^A_{\hat{\omega}}$, and $\psi^X_{\hat{\omega}}$, in order to perform accurate *ex ante* power calculations (as in Section 3.1 above).[45]

Second, in the absence of pre-existing data, users may instead input assumed within-unit AR(1) correlations (γ), assumed average idiosyncratic covariances (ψ^B, ψ^A, ψ^X), or assumed average idiosyncratic correlations (ψ^B/σ^2_ω, ψ^A/σ^2_ω, ψ^X/σ^2_ω).[46]

4.2.2 Power calculations by simulation

Simulation-based power calculations are the most robust, flexible strategy for designing experiments *ex ante*. Hence, `pcpanel` includes the program `pc_simulate`, which enables researchers to implement our simulation approach from Section 3.[47] `pc_simulate` recovers the *ex ante* power of a given experimental design and MDE—by simulating the *ex post* estimating equation on subsamples of a pre-existing dataset.[48] `pc_simulate` enables power calculations via simulation for four standard treatment effect estimators: cross-sectional ("one-shot"), repeated cross-sections (post-treatment periods only), DD, and ANCOVA. Users may

[45] The option `-depvar(y)-` tells `pc_dd_covar` to nonparametrically estimate the average covariance structure of the variable `y`, given the number of pre- and post-treatment periods. `pc_dd_covar` implements the estimation approach detailed in Appendix D.1, and `pc_dd_analytic` then applies the bias-correction factors derived in Appendix E. Implementing this procedure correctly is not trivial, and we encourage users with pre-existing data to use our software to do so.

[46] Alternatively, adding the option `-ar1(0.4)-` to `-pc_dd_analytic, mde(10) n(300) p(0.5) pre(3) post(5) var(1750)-` reduces power from 0.81 to 0.64. Figures 6 and 8 illustrate how AR(1) imperfectly approximates the correlation structures present in real data. However, we provide this option for `pc_dd_analytic` to let users without pre-existing data examine the sensitivity of their power calculations to differing levels of serial correlation.

[47] To construct the solid lines in Figures 6 and 8, we perform *both* analytical and simulation-based power calculations. For each panel length, we calculate the MDE that achieves 80 percent power, based on nonparametric estimates of the covariance structure (as in, run `pc_dd_analytic`, calling subprogram `pc_dd_covar`). Then, we simulate (as in, run `pc_simulate`) to confirm that realized power is indeed 80 percent.

[48] Each iteration re-randomizes PJ units into treatment and adds MDE to treated units' outcomes for all post-treatment periods. Appendix D.2 outlines this simulation algorithm in more detail. Users may also provide simulated data based on a plausible DGP. Appendix D.3 provides guidance on conducting simulation-based power calculations in the absence of a representative pre-existing dataset.

flexibly condition on pre-determined covariates (e.g. household size) or more detailed fixed effects (e.g. unit interacted with month-of-year), both common strategies to increase an experiment's power. `pc_simulate` also accommodates datasets in which a longer panel has been collapsed to a single pre-treatment and post-treatment period, as recommended by Bertrand, Duflo, and Mullainathan (2004) as a way of controlling the false rejection rate.[49]

Finally, `pc_simulate` supports two additional randomization techniques: stratified randomization and cluster randomization. For stratified randomization, users identify one or multiple categorical covariates (e.g. gender), and `pc_simulate` selects PJ treated units and $(1 - P)J$ control units within each stratification group. This randomization approach ensures balance across the treated and control groups. For cluster randomization, users input a group identifier (e.g. village) and `pc_simulate` randomizes *groups* of units into treatment (e.g. selecting treated villages, rather than treated households). Users may alter the number of units per group, and also vary the intensity of unit-level treatment within treated groups.[50]

5 Conclusion

Randomized experiments are costly, and researchers should avoid both underpowered experiments that are uninformative and overpowered experiments that waste resources. *Ex ante* power calculations enable researchers to design experiments with sample sizes that are sufficient, but not excessive. As multi-wave data collection becomes cheaper, panel RCTs are becoming increasingly common. Temporally disaggregated data allow researchers to ask new questions, applying a wider range of empirical methods (McKenzie (2012)).

In this paper, we develop a framework for panel data power calculations, generalizing the standard power calculation formula for difference-in-differences (originally derived by Frison and Pocock (1992)) to accommodate fully arbitrary correlations in the error structure.

[49] Collapsing to cross-sectional unit-specific averages (or two-period pre/post averages for DD estimators) obviates the need to adjust the Type I error rate using the CRVE. However, collapsing data does *not* obviate the need to account for serial correlation in power calculations (see Appendix A.2.4).

[50] `pc_simulate` does not allow for treatment spillovers, which would require (essentially arbitrary) *ex ante* assumptions on the strength and direction of these spillovers, which are beyond the scope of this work. Researchers interested in randomized saturation designs should consult Baird et al. (2018), which considers these designs in the cross-section. Appendix C.3 uses `pc_simulate` to compare cluster randomized designs to unit-level randomization in simulated data.

Unlike the power calculation formulas highlighted in McKenzie (2012), our "serial-correlation-robust" formula achieves the desired power in settings with arbitrary serial correlation. These results hold in Monte Carlo simulations, real data from a panel RCT in China, and household electricity consumption data similar to that used in panel RCTs in the energy economics literature.

Our new method is robust to alternative difference-in-differences estimators, to ANCOVA with deterministic time shocks, and to real-world data generating processes. We also provide a framework for trading off minimum detectable effects vs. sample sizes, while discussing practical issues in designing panel RCTs. Our accompanying STATA package `pcpanel` executes both analytical and simulation-based power calculations, and we recommend simulation-based methods when researchers have access to representative pre-existing data *ex ante*. A productive avenue for future work would be extending the treatment interference framework in Baird et al. (2018) to panel data experiments.

References

Abadie, Alberto, Susan Athey, Guido W. Imbens, and Jeffrey Wooldridge. 2017. "When Should You Adjust Standard Errors for Clustering?" ArXiv Working Paper No. 1710.02926.

Allcott, Hunt. 2011. "Social Norms and Energy Conservation." *Journal of Public Economics* 95 (9): 1082–1095.

Allcott, Hunt, and Michael Greenstone. 2017. "Measuring the Welfare Effects of Residential Energy Efficiency Programs." National Bureau of Economic Research Working Paper No. 23386.

Arellano, Manuel. 1987. "Computing Robust Standard Errors for Within-Group Estimators." *Oxford Bulletin of Economics and Statistics* 49 (4): 431–34.

Athey, Susan, and Guido W. Imbens. 2017. "The Econometrics of Randomized Experiments." *Handbook of Economic Field Experiments* 1:73–140.

Atkin, Azam, David aand Chaudhry, Shamyla Chaudry, Amit K. Khandelwal, and Eric Verhoogen. 2017. “Organization Barriers to Technology Adoption: Evidence from Soccer-Ball Producers in Pakistan.” *The Quarterly Journal of Economics* 132 (3): 1101–1164.

Atkin, David, Amit K. Khandelwal, and Adam Osman. 2017. “Exporting and Firm Performance: Evidence from a Randomized Experiment.” *The Quarterly Journal of Economics* 132 (2): 551–615.

Baird, Sarah, J. Aislinn Bohren, Craig McIntosh, and Berk Özler. 2018. “Optimal Design of Experiments in the Presence of Interference.” *Review of Economics and Statistics* 100 (5): 844–860.

Bertrand, Marianne, Esther Duflo, and Sendhil Mullainathan. 2004. “How Much Should We Trust Differences-In-Differences Estimates?” *The Quarterly Journal of Economics* 119 (1): 249–275.

Blattman, Christopher, Nathan Fiala, and Sebastian Martinez. 2014. “Generating Skilled Self-Employment in Developing Countries: Experimental Evidence from Uganda.” *The Quarterly Journal of Economics* 129 (2): 697–752.

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

Bloom, Nicholas, Benn Eifert, Aprajit Mahajan, David McKenzie, and John Roberts. 2013. “Does Management Matter? Evidence from India.” *The Quarterly Journal of Economics* 128 (1): 1–51.

Bloom, Nicholas, James Liang, John Roberts, and Zhichun Jenny Ying. 2015. “Does Working from Home Work? Evidence from a Chinese Experiment.” *The Quarterly Journal of Economics* 130 (1): 165–218.

Cameron, A. Colin, and Douglas L. Miller. 2015. “A Practitioner’s Guide to Cluster-Robust Inference.” *Journal of Human Resources* 50 (2): 317–372.

Card, David, Stefano DellaVigna, and Ulrike Malmendier. 2011. “The Role of Theory in Field Experiments.” *Journal of Economic Perspectives* 25 (3): 39–62.

Cohen, Jacob. 1977. *Statistical Power Analysis for the Behavioral Sciences.* New York, NY: Academic Press.

Duflo, Esther, Rachel Glennerster, and Michael Kremer. 2007. "Using Randomization in Development Economics Research: A Toolkit." Chap. 61 in *Handbook of Development Economics,* edited by Paul T. Schultz and John A. Strauss, 3895–3962. Volume 4. Oxford, UK: Elsevier.

Fowlie, Meredith, Michael Greenstone, and Catherine Wolfram. 2018. "Do Energy Efficiency Investments Deliver? Evidence from the Weatherization Assistance Program." *The Quarterly Journal of Economics* 133 (3): 1597–1664.

Fowlie, Meredith, Catherine Wolfram, C. Anna Spurlock, Annika Todd, Patrick Baylis, and Peter Cappers. 2017. "Default Effects and Follow-on Behavior: Evidence from an Electricity Pricing Program." National Bureau of Economic Research Working Paper No. 23553.

Frison, L., and S. J. Pocock. 1992. "Repeated Measures in Clinical Trials: Analysis Using Mean Summary Statistics and its Implications for Design." *Statistics in Medicine* 11 (13): 1685–1704.

Glennerster, Rachel, and Kudzai Takavarasha. 2013. *Running Randomized Evaluations: A Practical Guide.* Princeton University Press.

Ito, Koichiro, Takanori Ida, and Makoto Tanaka. 2018. "Moral Suasion and Economic Incentives: Field Experimental Evidence from Energy Demand." *American Economic Journal: Economic Policy* 10 (1): 240–267.

Jack, B. Kelsey, and Grant Smith. 2019. "Charging ahead: Prepaid electricity metering in South Africa." *American Economic Journal: Applied Economics.* Forthcoming.

Jessoe, Katrina, and David Rapson. 2014. "Knowledge Is (Less) Power: Experimental Evidence from Residential Energy Use." *American Economic Review* 104 (4): 1417–1438.

McKenzie, David. 2012. "Beyond Baseline and Follow-up: The Case for More T in Experiments." *Journal of Development Economics* 99 (2): 210–221.

McKenzie, David. 2017. “Identifying and Spurring High-Growth Entrepreneurship: Experimental Evidence from a Business Plan Competition.” *American Economic Review* 107 (8): 2278–2307.

Murphy, Kevin, Brett Myors, and Allen Wolach. 2014. *Statistical Power Analysis: A Simple and General Model for Traditional and Modern Hypothesis Tests.* 4th ed. New York, NY: Routledge.

Pecan Street. 2016. “Dataport.” `https://dataport.pecanstreet.org/`.

Teerenstra, Steven, Sandra Eldridge, Maud Graff, Esther de Hoop, and George F. Borm. 2012. “A simple sample size formula for analysis of covariance in cluster randomized trials.” *Statistics in Medicine* 31 (20): 2169–2178.

White, Halbert. 1984. *Asymptotic Theory for Econometricians.* 1st ed. San Diego, CA: Academic Press.