Panel Data and Experimental Design

Fiona Burlig, Louis Preonas, and Matt Woerman

Revised October 2017

Revised version published in
Journal of Development Economics,
144 (102548), 2020

Panel Data and Experimental Design

Fiona Burlig, Louis Preonas, and Matt Woerman*

October 31, 2017

Abstract

How should researchers design experiments with panel data? We derive analytical expressions for the variance of panel estimators under non-i.i.d. error structures, which inform power calculations in panel data settings. Using Monte Carlo simulation, data from a randomized experiment in China, and high-frequency U.S. electricity consumption data, we demonstrate that traditional methods produce experiments that are incorrectly powered with proper inference. Failing to account for serial correlation yields overpowered experiments in short panels and underpowered experiments in long panels. Our theoretical results enable us to achieve correctly powered experiments in both simulated and real data.

Keywords: power, experimental design, panel data, sample size
JEL Codes: B4, C23, C9, O1, Q4

*Burlig: Department of Economics and Energy Policy Institute, University of Chicago. 360 Saieh Hall for Economics, 5757 S. University Avenue, Chicago, IL 60637. Email: burlig@uchicago.edu. Preonas and Woerman: Department of Agricultural and Resource Economics and Energy Institute at Haas. 207 Giannini Hall, UC Berkeley, Berkeley, CA 94720. Emails: lpreonas@berkeley.edu. and matt.woerman@berkeley.edu. We thank 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 the University of California, Berkeley, BITSS, and NEUDC 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. All remaining errors are our own. Our accompanying Stata package, pcpanel, is available from ssc; an R version will follow shortly, which will be available from CRAN.

1 Introduction

Randomized controlled trials (RCTs) are an extremely valuable and increasingly popular tool for causal inference. The number of RCTs published in the top five economics journals has risen substantially over time (Card, DellaVigna, and Malmendier (2011)). As researchers embark on RCTs, they face many challenges in designing experiments: they must choose a sampling frame and sample size, design an intervention, and collect data, all subject to budget constraints. Experiments must have large enough sample sizes to be sufficiently powered, or to be able to statistically distinguish between true and false null hypotheses. At the same time, their sample sizes must be small enough to keep costs down.

Power calculations represent an important tool for calibrating the sample size and design of RCTs. By applying either analytical formulas or simulation-based algorithms, power calculations enable researchers to trade off sample size with the smallest effect an experiment can empirically detect. Bloom (1995) provides an early overview of the power calculation framework.[1] Duflo, Glennerster, and Kremer (2007) and Glennerster and Takavarashi (2013) describe the basics of power calculations and discuss practical considerations. The existing literature on statistical power in economics focuses on single-wave experiments, where units are randomized into a treatment group or a control group, and researchers observe each unit once.[2]

In a widely cited paper based on results from Frison and Pocock (1992), McKenzie (2012) recommends experimental designs that involve panel data, using multiple observations per unit to increase statistical power. This is especially attractive in settings where collecting additional waves of data for one individual is more cost-effective than collecting data on more individuals. In recent years, several prominent papers have employed RCT designs with panel

[1] Cohen (1977) and Murphy, Myors, and Wolach (2014) are classic references.

[2] In economics, researchers often collect two waves of data, but estimate treatment effects using post-treatment data only and controlling for the baseline level of the outcome variable (following McKenzie (2012)). Baird et al. (forthcoming) extends the classic cross-sectional setup to randomized saturation designs, capable of measuring spillover and general equilibrium effects. Athey and Imbens (2016) discusses statistical power using a randomization inference approach.

data.[3] As the costs of data collection fall, panel RCTs are becoming increasingly common, allowing researchers to answer new questions using more flexible empirical strategies.

Panel data also poses challenges in terms of statistical inference. Bertrand, Duflo, and Mullainathan (2004) highlights the notion that units in panel data generally exhibit serial correlation, and that failing to account for this error structure will yield standard errors that are biased towards zero. This dramatically raises the probability of a Type I error. In order to achieve correct false rejection rates, applied econometricians using panel data in quasi-experimental settings generally implement the cluster-robust variance estimator (CRVE), or use "clustered standard errors".[4]

In a panel RCT, it is likewise important to account for serially correlated errors both during *ex post* analysis and in *ex ante* experimental design. If researchers assume that errors are independent and identically distributed (i.i.d.) in *ex ante* power calculations, and then do not adjust their standard errors *ex post*, they will over-reject true null hypotheses if their errors are in fact serially correlated. On the other hand, if researchers adjust their standard errors *ex post* but do not adjust their power calculations *ex ante*, they introduce a fundamental mismatch between *ex ante* and *ex post* assumptions that will yield incorrectly powered experiments in the presence of serial correlation. To the best of our knowledge, there is no existing economics literature on power calculations in panel data that accounts for arbitrary serial correlation.

In this paper, we derive analytical expressions for the variance of panel estimators under non-i.i.d. error structures. We use these expressions to formalize a power calculation formula for difference-in-differences estimators that is robust to serial correlation in panel data settings.[5]

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

[4] See Cameron and Miller (2015) for a practical guide to CRVE standard errors, which were first proposed by White (1984), and popularized by Arellano (1987). Abadie et al. (2017) point out that clustering is important in experiments when treatment assignment is correlated within clusters - as is commonly true in the panel case, where units are typically treated and remain treated throughout the experiment.

[5] Recent experiments published in top economics journals use either the difference-in-differences estimator or the ANCOVA estimator. For analytical tractability, we focus on the difference-in-differences estimator throughout most of this paper. Section 5 provides a discussion of the ANCOVA estimator, where standard power calculation techniques similarly ignore serial correlation in panel data.

We conduct Monte Carlo analysis using both simulated and real data, and demonstrate that standard methods for experimental design yield experiments that are incorrectly powered in the presence of serially correlated errors, even with proper *ex post* inference. Our theoretical results enable us to correct this mismatch between *ex ante* and *ex post* assumptions on the error structure, and our serial-correlation-robust power calculation technique achieves the desired power in both simulated and real data. Ultimately, we provide researchers with both the theoretical insights and practical tools to design well-powered experiments in panel data settings.

We make three main contributions to the literature on experimental design in economics. First, we show that existing power calculation methods for panel data in economics, discussed in McKenzie (2012), fail in the presence of arbitrary serial correlation. We demonstrate this both analytically and via Monte Carlo using real and simulated data. Second, we derive a new expression for the variance of the difference-in-differences estimator under arbitrary serial correlation, which enables us to calibrate panel RCTs to the desired power. Finally, we address practical challenges involved in performing power calculations on panel data real experimental settings.

The paper proceeds as follows. Section 2 provides background on power calculations. Section 3 presents analytical power calculations expressions for panel data, and demonstrates their importance in Monte Carlo simulations with serially correlated errors. Section 4 applies these results to real experimental data. Section 5 discusses practical issues related to power calculations. Section 6 concludes.

2 Background

Randomized controlled trials allow researchers to overcome the fundamental challenge of causal inference highlighted by Rubin (1974): we can never observe the outcome for the same unit i in multiple states of the world simultaneously. RCTs solve this problem in expectation, by randomly assigning treatment to a subset of a population. Comparing the average outcomes of treated and untreated ("control") populations, researchers can identify the average causal effect of treatment. RCTs, and quasi-experimental research designs that attempt to mimic them, have been an important part of the ongoing empirical “credibility revolution” in economics (Angrist and Pischke (2010)).

Designing randomized experiments is challenging, in part because researchers have many degrees of freedom when doing so. They must choose a study location and sampling frame, select a sample size, implement an intervention, and collect data, all subject to partnerships with implementing agencies and to financial constraints. The choice of sample size is of particular importance, as it forces researchers to balance implementation costs and statistical power. Because recruitment and implementation of subjects is costly, an experiment should avoid excessively large samples. At the same time, an experiment that is too small will not be able to statistically distinguish between true and false null hypotheses.

A power calculation computes the smallest effect size that an experiment, with a given sample size and experimental design, will statistically be able to detect. The most general power calculation equation is:

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

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.[6] 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.

Figure 1 illustrates these concepts graphically. The black curve represents the distribution of $\hat{\tau}$ if the null hypothesis is true, and the blue curve represents the distribution of $\hat{\tau}$ if the null hypothesis is false, where τ is instead equal to some value $\tau \neq 0$. Note that the variances of these distributions decrease with the sample size of the experiment. The dashed gray line is the critical value $t^d_{\alpha/2}$. The shaded gray areas represent the likelihood that the

[6] 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.

researcher will reject a true null, and the blue-shaded area represents the statistical power of the test, or the probability that the experiment will correctly reject a false null. Figure 1 displays the case in which $\tau = MDE$, the minimum detectable effect size calibrated to the variance of $\hat{\tau}$, Type I error tolerance α, and desired power κ.

While α and κ are conventionally set to 0.05 and 0.80, respectively, researcher choices govern the estimator $\hat{\tau}$. The variance of $\hat{\tau}$ depends jointly on the experimental design, the sample size, the model used to estimate $\hat{\tau}$, and the underlying properties of the data. To illustrate this, we first follow Bloom (1995) and Duflo, Glennerster, and Kremer (2007) in considering perhaps the simplest experimental design: a cross-sectional RCT. In this setup, J units are randomly assigned a treatment status D_i, with proportion P in treatment ($D_i = 1$) and proportion $(1 - P)$ in control ($D_i = 0$). We make standard assumptions for randomized trials:

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

$$Y_i = \beta + \tau D_i + \varepsilon_i$$

where ε_i is distributed i.i.d. $\mathcal{N}(0, \sigma_\varepsilon^2)$; and the treatment effect, τ, is homogeneous across all units.

Assumption 2 (Strict exogeneity)**.** $\mathrm{E}[\varepsilon_i \mid \mathbf{X}] = 0$, *where* $\mathbf{X} = [\mathbf{1}\ \mathbf{D}]$. *In practice, this follows from random assignment of D_i.*

Define the OLS estimator of τ to be $\hat{\tau} = (\mathbf{X'X})^{-1}\mathbf{X'Y}$. Under Assumptions 1–2:[7]

$$\begin{aligned} \mathrm{E}[\hat{\tau} \mid \mathbf{X}] &= \tau \\ \mathrm{Var}(\hat{\tau} \mid \mathbf{X}) &= \frac{\sigma_\varepsilon^2}{P(1-P)J} \\ MDE &= \left(t_{1-\kappa}^{J-2} + t_{\alpha/2}^{J-2}\right)\sqrt{\frac{\sigma_\varepsilon^2}{P(1-P)J}} \end{aligned} \tag{2}$$

[7] See Appendix A.1.1 for a full derivation of the variance of $\hat{\tau}$ in this model.

Intuitively, the MDE decreases with sample size J, increases with error variance, σ_ε^2, and is minimized at $P = 0.5$. Given α and κ, larger experiments with less noisy data can statistically reject the null of zero for smaller true treatment effects.

Researchers are not limited to this simple cross-sectional RCT design, however. Alternative designs and estimators may yield similar MDEs at lower cost. McKenzie (2012) highlights the possibility of using multiple waves of data in conjunction with a difference-in-difference (DD) estimator to decrease the number of units required to achieve a given MDE. In this model, P proportion of the J units are again randomized into treatment. The researcher collects the outcome Y_{it} for each unit i, across m pre-treatment time periods and r post-treatment time periods. For units in the treatment group, $D_{it} = 0$ in pre-treatment periods and $D_{it} = 1$ in post-treatment periods; for units in the control group, $D_{it} = 0$ in all $(m + r)$ periods.

Assumption 3 (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 i.i.d. $\mathcal{N}(0, \sigma_\omega^2)$; and the treatment effect, τ, is homogeneous across all units and all time periods.[8]

Assumption 4 (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 again follows from random assignment of D_{it}.*

Assumption 5 (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.*

[8] This is the standard model used in panel RCTs.

The OLS estimator of τ with unit and time fixed effects is $\widehat{\tau} = (\ddot{\mathbf{D}}'\ddot{\mathbf{D}})^{-1}\ddot{\mathbf{D}}'\ddot{\mathbf{Y}}$, where:[9]

$$\ddot{Y}_{it} = Y_{it} - \frac{1}{m+r}\sum_t Y_{it} - \frac{1}{J}\sum_i Y_{it} + \frac{1}{J(m+r)}\sum_i\sum_t Y_{it}$$
$$\ddot{D}_{it} = D_{it} - \frac{1}{m+r}\sum_t D_{it} - \frac{1}{J}\sum_i D_{it} + \frac{1}{J(m+r)}\sum_i\sum_t D_{it}$$

Under Assumptions 3–5:

$$\begin{aligned} \mathrm{E}[\hat{\tau} \mid \mathbf{X}] &= \tau \\ \mathrm{Var}(\hat{\tau} \mid \mathbf{X}) &= \left(\frac{\sigma^2_\omega}{P(1-P)J}\right)\left(\frac{m+r}{mr}\right) \\ MDE &= \left(t^J_{1-\kappa} + t^J_{\alpha/2}\right)\sqrt{\left(\frac{\sigma^2_\omega}{P(1-P)J}\right)\left(\frac{m+r}{mr}\right)} \end{aligned} \tag{3}$$

This is the power calculation equation originally derived by Frison and Pocock (1992) (henceforth FP).[10] The experiment's MDE decreases symmetrically in m and r, because, holding the error variance constant, longer panels decrease the variance of the DD estimator. Note that researchers can potentially trade off J for m and/or r to decrease both MDE and implementation costs.

Importantly, $\sigma^2_\omega \leq \sigma^2_\varepsilon$ by construction, since ω_{it} represents only the idiosyncratic component of the error term. Empirically, the inclusion of fixed effects reduces the error variance to the extent that underlying within-unit and within-time correlations explain Y_{it}. To see this, we can rewrite Equation (3) in terms of σ^2_ε. Let ρ_υ and ρ_δ denote the proportion of the composite variance σ^2_ε contributed by σ^2_υ and σ^2_δ, respectively:

$$\rho_\upsilon \equiv \frac{\sigma^2_\upsilon}{\sigma^2_\upsilon + \sigma^2_\delta + \sigma^2_\omega} = \frac{\sigma^2_\upsilon}{\sigma^2_\varepsilon} \qquad \rho_\delta \equiv \frac{\sigma^2_\delta}{\sigma^2_\upsilon + \sigma^2_\delta + \sigma^2_\omega} = \frac{\sigma^2_\delta}{\sigma^2_\varepsilon}$$

[9] Under the assumption that the researcher knows the true model, random effects is more efficient than fixed effects. In practice, however, this is rarely the case, and researchers use fixed effects instead of random effects. Hence, we consider the fixed effects estimator here.

[10] See Appendix A.2.1 for a full derivation. Here, we use critical values with J degrees of freedom, which is consistent with the assumptions of the CRVE with J clusters. By contrast, the OLS variance estimator would use $J(m+r) - (J+m+r)$ degrees of freedom. Note that the CRVE has been shown to perform poorly with few clusters. In cases where $J < 40$, Cameron, Gelbach, and Miller (2008) recommend using the wild-cluster bootstrap.

Then, (3) can be rewritten as:

$$MDE = \left(t^{J}_{1-\kappa} + t^{J}_{\alpha/2}\right)\sqrt{\left(\frac{\sigma^2_{\varepsilon}}{P(1-P)J}\right)\left(\frac{m+r}{mr}\right)(1-\rho_{\upsilon}-\rho_{\delta})} \tag{4}$$

In this formula, larger unit-level intracluster correlations (i.e., ρ_{υ} closer to 1) or stronger temporal shocks (i.e., ρ_{δ} closer to 1) yield smaller MDEs.[11] Notice that, although Equation (4) includes intracluster correlation coefficient terms, the idiosyncratic component of the error term (ω_{it}) is still assumed to be i.i.d. This highlights an important point: accounting for intracluster correlation is not the same as allowing for arbitrary serial correlation. Indeed, Bertrand, Duflo, and Mullainathan (2004) (henceforth BDM) demonstrate that panel data are likely to exhibit serial correlation within units, meaning that the assumption of i.i.d. errors is unlikely to hold in practice.

3 Theory

3.1 Correlated errors in panel models

Power calculation formulas such as the standard cross-sectional model (Equation (2)) and the FP model (Equation (3)) assume that the error structure in data results from an i.i.d. process. Real data rarely exhibit i.i.d. errors, and researchers frequently apply the CRVE to allow for correlated errors. In cross-sectional models, when treatment is randomly assigned, the OLS variance estimator is an unbiased estimator of the *ex ante* expected variance, even in the presence of cross-sectional error correlations:

Lemma 1. *In a cross-sectional model with random assignment to treatment,* $\frac{\sigma^2_{\varepsilon}}{P(1-P)J}$ *is an unbiased estimator of the expectation of* $\mathrm{Var}(\hat{\tau} \mid \mathbf{X})$ *even if* $\mathrm{E}[\varepsilon_i \varepsilon_j \mid \mathbf{X}] \neq 0$ *for some* $i \neq j$*. See Appendix A.3 for a proof.*[12]

[11] Replacing $\rho_{\delta} = 0$, $P = 0.5$, and $J = 2n$, Equation (4) is equivalent to the power calculation formula for difference-in-differences derived by FP and discussed in McKenzie (2012). See Appendix A.2.1 for complete derivations. Equation (3) is not identical to the model in FP or McKenzie (2012), because these authors assume that the time disturbance δ_t is deterministic and has no variance. Our model allows for $\sigma^2_{\delta} > 0$, in keeping with assumptions economists typically make about data generating processes. Hence, Equation (3) represents a more general version of the FP formula.

[12] Campbell (1977) provides the first version of this proof, which is cited by Moulton (1986), and which imposes a grouped error structure. In Appendix A.3, we provide a proof which allows for arbitrary cross-

This means that in single-wave RCTs, researchers need not adjust standard errors to account for correlation across experimental units.

Cross-sectional randomization does not obviate the need to account for serially correlated errors in panel datasets. When an experimenter collects data from the same cross-sectional units over multiple time periods, each unit's error terms are likely correlated across time.[13] In most DD research designs, once treatment begins, it persists for the remainder of the study, so both a unit's error terms and its treatment status are serially correlated. Hence, researchers still need to account for serial correlation when randomizing treatment at the unit level.[14] BDM demonstrate that serial correlation in DD designs can severely bias conventional standard errors towards zero. This means that failing to account for serially correlated errors can lead to high Type I error rates and substantial over-rejection of true null hypotheses.

Given that a panel DD analysis should account for potential serial correlation *ex post*, what does this imply for the *ex ante* statistical power of such an experiment? BDM find that while applying the CRVE on a serially correlated panel dataset can reduce the Type I error rate to the desired level, this has the effect of increasing the Type II error rate. In other words, correctly accounting for serial correlation will tend to inflate standard errors, which in turn will reduce the rejection rates of both false and true null hypotheses. If a researcher designs a DD experiment using the FP power calculation formula, and then applies the CRVE *ex post*, this suggests that her experiment will likely be underpowered.

sectional error dependence. Athey and Imbens (2016, 2017) still recommend using Eicker-Huber-White standard errors in this case, to allow for heteroskedasticity. We are not aware of any paper that discusses power calculations in the presence of heteroskedastic disturbances.

[13] This is true even in a model with unit and time period fixed effects. These fixed effects control for the average outcome of each unit across all time periods, and the average outcome across all units in each time period. However, if each unit's demeaned outcome realizations evolve non-independently across time, then the resulting "idiosyncratic" error terms (i.e., ω_{it} in Equation (3)) will exhibit some form of correlation that violates the i.i.d. assumption.

[14] As with single-wave RCTs, cross-sectional randomization in panel RCTs eliminates the need to adjust for cross-sectional correlations. Randomizing the timing and duration of treatment within treated units would make the OLS variance estimator unbiased, but would be logistically prohibitive in most settings.

3.2 Power calculations with serial correlation

We derive a more general version of the FP DD power calculation formula in order to accommodate the non-i.i.d. error structures of real-world data, including arbitrary correlations within cross-sectional units over time.[15] Just as in the FP model, there are J units, P proportion of which are randomized into treatment. The researcher again 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.[16]

Assumption 6 (Data generating process)**.** *The data are generated according to the following model:*[17]

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

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 i.i.d.) $\mathcal{N}(0, \sigma_\omega^2)$; and the treatment effect, τ, is homogeneous across all units and all time periods.

Assumption 7 (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 again follows from random assignment of D_{it}.*

Assumption 8 (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.*

[15] We do not consider cross-sectional correlations, because we consider a treatment that is randomized at the unit level. For a full version of this model incorporating both arbitrary serial and cross-sectional correlations, see Appendix A.2.3.

[16] 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 the standard setup in panel experiments in economics.

[17] We follow the previous literature in assuming a homogeneous treatment effect across units and time periods. While this is somewhat restrictive, researchers may relax this assumption by performing power calculations via simulation (see Section 5 below), or by computing multiple power calculations (e.g., for "early" vs. "late" treatment effects, which parallels *ex post* estimation strategies of many existing RCTs). Furthermore, one could imagine combining our formula with that of Baird et al. (forthcoming) to accommodate stratified experimental designs. Our simulation-based software allows stratified randomization, but it is beyond the analytical scope of this paper.

Assumption 9 (Independence across units)**.** $\mathrm{E}[\omega_{it}\omega_{js} \mid \mathbf{X}] = 0, \ \forall \ i \neq j, \ \forall \ t, s.$

Assumption 10 (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^B_T, ψ^A_T, and ψ^X_C analogously, where we consider only the PJ treated units; also define ψ^B_C, ψ^A_C, and ψ^X_C analogously, where we consider only the $(1-P)J$ control units. Using these definitions, assume that $\psi^B = \psi^B_T = \psi^B_C$; $\psi^A = \psi^A_T = \psi^A_C$; and $\psi^X = \psi^X_T = \psi^X_C$.[18]

The OLS estimator with unit and time fixed effects remains $\hat{\tau} = (\ddot{\mathbf{D}}'\ddot{\mathbf{D}})^{-1}\ddot{\mathbf{D}}'\ddot{\mathbf{Y}}$ and again, $\mathrm{E}[\hat{\tau} \mid \mathbf{X}] = \tau$. However, Assumptions 6–10 extend FP to a more general power calculation formula that incorporates arbitrary within-unit correlations:[19]

$$\mathrm{Var}(\hat{\tau} \mid \mathbf{X}) = \left(\frac{1}{P(1-P)J}\right) \left[\left(\frac{m+r}{mr}\right)\sigma^2_\omega + \left(\frac{m-1}{m}\right)\psi^B + \left(\frac{r-1}{r}\right)\psi^A - 2\psi^X\right]$$

$$MDE = (t^J_{1-\kappa} + t^J_{\alpha/2})\sqrt{\left(\tfrac{1}{P(1-P)J}\right)\left[\left(\tfrac{m+r}{mr}\right)\sigma^2_\omega + \left(\tfrac{m-1}{m}\right)\psi^B + \left(\tfrac{r-1}{r}\right)\psi^A - 2\psi^X\right]} \quad (5)$$

Throughout the remainder of the paper, we refer to Equation (5) as the "serial-correlation-robust" (SCR) power calculation formula. Note that under cross-sectional randomization,

[18] We choose the letters "B" to indicate the Before-treatment period, and "A" to indicate the 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.

[19] 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 (5). Applying this formula in practice requires parameterizing three additional values — ψ^B, ψ^A, ψ^X — which may prove difficult without detailed pre-experimental data. However, the standard FP formula implicitly assumes $\psi^B = \psi^A = \psi^X = 0$, which is likely unrealistic, and may lead to substantial errors in the necessary sample size. We discuss practical issues related to power calculations in Section 5.

this expression for the variance of $\hat{\tau}$ still holds in expectation, even in the presence of within-period error correlations across units:

Lemma 2. *In a panel difference-in-differences model with treatment randomly assigned at the unit level,* $\left(\frac{1}{P(1-P)J}\right)\left[\left(\frac{m+r}{mr}\right)\sigma_\omega^2+\left(\frac{m-1}{m}\right)\psi^B+\left(\frac{r-1}{r}\right)\psi^A-2\psi^X\right]$ *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 9–10.*

To illustrate the difference between the FP and SCR models, consider two cross-sectional units (indexed $\{i,j\}$) and four time periods (indexed $\{0,1,2,3\}$). The vector of errors, $\boldsymbol{\omega}$, and the corresponding variance-covariance matrix, $\boldsymbol{\Omega}$, can be represented as follows:[20]

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

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, the FP model assumes that these off-diagonal covariance elements are all zero.

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

[20] We show only the lower diagonal of the variance-covariance matrix because the full matrix is symmetric. Note further that we do not show the cross-unit covariance terms for notational convenience, as these terms are assumed to be zero.

supposing treatment is administered beginning at $t = 2$, these three 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:

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

The SCR power calculation formula above generalizes this structure to a model with J units across m pre-treatment periods and r post-treatment periods. In this model, 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 the MDE formula 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 power calculation formula positively, while ψ^X enters negatively, serial correlation may either increase or decrease the MDE relative to the i.i.d. case. Specifically, serial correlation will increase the MDE if and only if:

$$\left(\frac{m-1}{m}\right)\psi^B + \left(\frac{r-1}{r}\right)\psi^A > 2\psi^X \tag{6}$$

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. Together, these two effects imply that for longer panels, the FP model is increasingly likely to yield underpowered experiments. At the same time, using FP with short panels is likely to yield overpowered experiments.

3.3 Monte Carlo simulations

If a randomized experiment relies on a power calculation that fails to 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 FP 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 both *ex ante* and *ex post*; (ii) experiments that fail to account for serial correlation *ex ante* but apply the CRVE to account for serial correlation *ex post*; and (iii) experiments that both account for serial correlation *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^{FP} equal to the MDE from the FP 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 data generating process:

$$Y_{it} = \beta + \upsilon_i + \delta_t + \omega_{it} \tag{7}$$

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

$$\omega_{it} = \gamma\omega_{i(t-1)} + \xi_{it} \tag{8}$$

Third, we randomly assign treatment, with effect sizes τ^{FP}, 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 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^{FP} 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^2_υ, σ^2_δ, α, and κ fixed across all simulations, and we adjust the variance of the white noise term σ^2_ξ such that every simulation has a fixed idiosyncratic variance σ^2_ω. This allows γ to govern the proportion of σ^2_ω that is serially correlated.[21] 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}.[22] This causes τ^{SCR} to vary both with the degree of serial correlation and panel length, whereas τ^{FP} varies only with panel length.

Figure 2 displays the results of this exercise. The left column shows rejection rates under the FP formula using OLS standard errors, which assumes zero serial correlation 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 *ex post* only. The right column show rejection rates under our SCR formula using CRVE standard errors, which allows for serial correlation both *ex ante* and *ex post.* The top row plots realized power as a function

[21] 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 $\sigma^2_\omega = \frac{\sigma^2_\xi}{1-\gamma^2}$.

[22] We provide these formal derivations in Appendix B.1, along with further details on these Monte Carlo simulations. We also provide additional simulation results that separately vary m and r in Appendix C.

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.

The left column confirms the BDM result that failing to account for serial correlation leads to false rejection rates dramatically higher than $\alpha = 0.05$. Even a modest serial correlation 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 correlation 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 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 broadly 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.[23]

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 (6). For shorter panels, using the FP formula instead of our SCR formula yields dramatically overpowered experiments. While this may seem counterintuitive, (6) is increasingly unlikely to hold as m

[23] As an alternative strategy for correcting false rejection rates in the presence of serial correlation, BDM suggest collapsing data down to one pre-treatment and one post-treatment period. In Appendix A.2.4, we demonstrate that this does not suffice for power calculations. While collapsing data enables the researcher to achieve the correct false rejection rate without knowing the true variance of their estimator, (an estimate of) this variance is required for power calculations.

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.[24]

These results are striking. For even a modest degree of serial correlation, applying the FP power calculation formula will not yield experiments of the desired statistical power. By contrast, the SCR formula achieves the desired 80 percent 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, it stands to reason that such serial correlation can similarly impact the statistical power of experiments if not accounted for *ex ante*.

4 Applications to real-world data

4.1 Bloom et al. (2015) data

In order to understand whether the differences in power demonstrated in the Monte Carlo simulations above are meaningful in practice, we conduct an analogous simulation exercise using a real dataset from an experiment in a developing-country setting. We use data 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.[25] The authors estimate the

[24] 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. See Appendix C for additional short-panel results.

[25] 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.

following equation to derive the central result, reported in Table 2 in the original paper:

$$\text{Performance}_{it} = \alpha \text{Treat}_i \times \text{Experiment}_t + \beta_t + \gamma_i + \varepsilon_{it} \tag{9}$$

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 40 periods ($m = r = 20$). For each simulation panel length, we randomly assign three treatment effect sizes, τ^{FP}, $\tau^{AR(1)}$, and τ^{SCR}, at the individual level and estimate Equation (9) separately for each treatment effect size. We calibrate τ^{FP} using the FP formula that assumes no serial correlation; $\tau^{AR(1)}$ 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, and ψ^X, where the subscript $\hat{\omega}$ denotes the variance/covariance of *residuals* rather than errors.[26]

Figure 3 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 serial correlation leads to experiments that deviate dramatically from 80 percent power, even in the presence of relatively weak serial correlation. For an experiment with 12 pre/post-treatment periods, applying the FP formula with $\kappa = 0.80$ yields an experiment with only 35 percent power. This is consistent with our results from simulated 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.

[26] Appendix D.1 outlines how to estimate these parameters. For all three treatment effects, we estimate the value of $\sigma^2_{\hat{\omega}}$ estimated from residuals.

4.2 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)).[27] Electricity consumption data tend to exhibit high within-household autocorrelation, making them particularly well-suited for this study. Additionally, RCTs using energy consumption data are becoming increasingly common in economics, making our Pecan Street application relevant to this growing literature.[28]

We aggregate these data to four different temporal levels: hourly, daily, weekly, and monthly, each with a different correlation structures and amounts of idiosyncratic variation, which sheds light on the performance of different power calculation approaches over a range of underlying error structures. We conduct Monte Carlo simulations on all four versions of the Pecan Street data to assess the performance of alternative power calculation assumptions. These follow the same procedure as the Bloom et al. (2015) simulations.[29] Figure 4 shows that in all four versions of the Pecan Street data, realized power sharply deviates from the desired 80 percent under both the FP assumption of i.i.d. errors and an assumed 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. This results holds at both high and low temporal frequencies, and is consistent with the Bloom et al. (2015) simulations.

4.3 Power calculations in real data

We can use the SCR formula to perform power calculations, which quantify the tradeoff between MDE and J over different panel lengths. In order to operationalize the SCR formula,

[27] 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 for 699 households over 26,888 hours. Appendix B.3 describes the data, including summary statistics, in further detail.

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

[29] Appendix B.3 provides further details on these simulations.

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 their (future) experimental datasets. In the best case scenario, researchers will have access to data that are representative of what will be collected in the field, and they can simply calculate these variance and covariance terms from this pre-existing dataset.[30] Plugging these estimates into the SCR formula, researchers can evaluate the tradeoffs between different experimental design elements, such as the desired power of the experiment, the number of units to recruit, the number of observations to collect for each unit, and the expected effect size of the intervention.

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 under the assumption of uncorrelated errors and also allowing for arbitrary correlations among the error terms of a household using the FP formula and the SCR formula, respectively. For simplicity, we consider only balanced panels of households with the same number of observations before and after the experimental intervention (i.e. $m = r$). For each experiment length, we estimate the average $\sigma^2_{\hat{\omega}}$ and $\psi_{\hat{\omega}}$ terms from the daily Pecan Street dataset, and we assume constant values for the remaining parameters.[31]

We plot the results of this exercise in Figure 5. The left panel depicts power calculations using the FP formula, which assumes an i.i.d. error structure. The right panel applies the SCR formula to compute the number of units required for the same set of MDEs, 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 an experiment of a particular length, ranging from $m = r = 1$ to $m = r = 12$; the curves plot the number of households required to achieve 80 percent power as a function of the MDE. For each MDE, fewer households are required as the length of the experiment increases. However, the "naive" power calculation always implies a substantially smaller number of households than the SCR power calculation - for example,

[30] 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.

[31] See Appendix B.4 for details.

with $m = r = 12$, the "naive" power calculation suggests the researcher only needs 106 households to achieve an MDE of 5 percent; the SCR power calculation implies that the required sample size is 354 households - over 3 times greater. Hence, if a researcher in this setting applies the CRVE *ex post* but assumes i.i.d. errors *ex ante*, he will likely include too few households to achieve the desired statistical power.

5 Power calculations in practice

5.1 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:

$$\min_{P,J,m,r} MDE(P,J,m,r) \quad \text{s.t.} \quad C(P,J,m,r) \leq B \tag{10}$$

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.[32] This tradeoff also arises from differences in the marginal effects of units and time periods on the MDE. Using the SCR power calculation formula, the "elasticities" of the MDE with respect to number of units and number of time periods are:

[32] 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.

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

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 varies as a function of the error structure and the number of time periods.[33] 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.

Figure 6 illustrates the fact that additional time periods may either increase or decrease statistical power. The left panel plots the MDE of an experiment as a function of the number of pre- and post-treatment waves, holding the number of units J constant. The right panel depicts the tradeoff between additional units and additional time periods by plotting the combinations of J and $m = r$ that yield a given MDE. To analytically construct these curves, we use the SCR formula and assume that the error structure is AR(1) with varying γ values.

At low to moderate levels of serial correlation, increasing the panel length always reduces the MDE for a given J and, likewise, reduces the J required to achieve a given MDE. However, at higher levels of serial correlation, this relationship is no longer monotonic. For

[33] 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 10).

$\gamma \geq 0.6$, marginally increasing m or r in a relatively short panel increases the MDE for a given J and, likewise, increases the J required to achieve a given MDE. This suggests that for experiments using highly correlated data, such as any of the four Pecan Street datasets above, additional periods of data might *decrease* statistical power if the panel is not sufficiently long.[34]

5.2 ANCOVA

Many published and pre-registered ongoing experiments in economics estimate treatment effects using analysis of covariance (ANCOVA) methods.[35] To do this, the econometrician estimates the following specification using post-treatment data only:

$$Y_{it} = \alpha + \tau D_i + \theta \bar{Y}_i^B + \varepsilon_{it} \tag{11}$$

where $\bar{Y}_i^B = \sum_{t=-m+1}^{0} Y_{it}$ 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 what has become the standard formula for ANCOVA power calculations (henceforth FP ANCOVA), which we adapt to our notation:

$$MDE \approx (t_{1-k}^J + t_{\alpha/2}^J)\sqrt{\left(\frac{1}{P(1-P)J}\right)\left[(1-\theta)^2\sigma_v^2 + \left(\frac{\theta^2}{m} + \frac{1}{r}\right)\sigma_\omega^2\right]} \tag{12}$$

[34] McKenzie (2012) argues that stronger unit-specific shocks (i.e. higher σ_v^2) can erode the benefits of collecting additional waves of data; this result extends 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.

[35] 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) comment thats, with i.i.d. error structures, ANCOVA is always more efficient than the DD model with the same number of periods, but that these gains are eroded as the intracluster correlation coefficient increases. Neither paper 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.

where $\theta = \frac{m\sigma_\upsilon^2}{m\sigma_\upsilon^2 + \sigma_\omega^2}$.[36] Importantly, deriving this formula under serial correlation necessitates an additional simplifying assumption for analytical tractability: we must assume away time shocks.[37] Under this strong assumption, we also derive the variance of the ANCOVA allowing for serial correlation, along with the corresponding serial-correlation-robust power calculation formula (henceforth SCR ANCOVA):

$$\text{Var}(\hat{\tau} \mid \mathbf{X}) \approx \frac{1}{P(1-P)J}\left[(1-\theta)^2\sigma_\upsilon^2 + \left(\frac{\theta^2}{m} + \frac{1}{r}\right)\sigma_\omega^2 + \frac{\theta^2(m-1)}{m}\psi^B + \frac{r-1}{r}\psi^A - 2\theta\psi^X\right] \tag{13}$$

$$MDE \approx (t_{1-\kappa}^J + t_{\alpha/2}^J)\times \sqrt{\frac{1}{P(1-P)J}\left[(1-\theta)^2\sigma_\upsilon^2 + \left(\frac{\theta^2}{m} + \frac{1}{r}\right)\sigma_\omega^2 + \frac{\theta^2(m-1)}{m}\psi^B + \frac{r-1}{r}\psi^A - 2\theta\psi^X\right]} \tag{14}$$

where $\theta = \frac{m\sigma_\upsilon^2 + m\psi^X}{m\sigma_\upsilon^2 + \sigma_\omega^2 + (m-1)\psi^B}$.[38]

As in Section 3, we compare the FP ANCOVA formula to the SCR ANCOVA formula by simulating data with an AR(1) error term and varying panel lengths. For each level of the AR(1) parameter and panel length, we compute the treatment effect size τ implied from the FP ANCOVA or SCR ANCOVA formulas. Next, we simulate 10,000 experiments, randomly assigning units to treatment with the appropriate effect size.[39] For each simulated

[36] The FP ANCOVA formula presented here differs slightly from that in Frison and Pocock (1992) and McKenzie (2012) because previous derivations have assumed that the true data generating process follows Equation (12), where outcomes are determined in part by pre-treatment values of the outcome. We instead assume that that there are unit-specific random effects, as in Assumption 6. The FP ANCOVA model assumes that time shocks are deterministic and have 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), this formula is approximate because we ignore sampling error in the estimation of θ, which approaches zero as the number of units increases.

[37] 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 $\bar{Y}_i^B$ 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.

[38] 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.

[39] In keeping with the assumptions of these models, we do not include time shocks in the data generating process.

experiment, we estimate $\hat{\tau}$ using the ANCOVA estimator, and record whether $\hat{\tau}$ is statistically different from zero with clustered standard errors. We calculate realized power as the fraction of experiments that reject the null hypothesis of $\tau = 0$. Figure 7 presents the results of this exercise. While the FP ANCOVA formula produces properly-powered experiments only when errors are i.i.d., the SCR ANCOVA is robust to all levels of AR(1) serial correlation.

Even though the SCR ANCOVA formula outperforms the FP ANCOVA formula in the presence of serial correlation, we do not recommend that researchers use this formula in real-world applications. Time shocks with non-zero variance are a common feature of panel data, and assuming them away may result in improperly powered experiments.[40] For this reason, we suggest that researchers either conduct power calculations by simulation (as discussed below) or perform power calculations using the SCR DD formula, which yields correctly powered DD experiments using real-world data. Power calculations based on *ex ante* assumptions of the DD estimator will understate the realized power from estimating a more-efficient ANCOVA model *ex post*.[41]

5.3 A simulation-based approach

In this paper, we develop an analytical framework for performing power calculations in a DD model with panel data that have a non-i.i.d. error structure. Shifting from the i.i.d. model to a non-i.i.d. model increases the number of parameters required to calibrate a DD power calculation formula. This reveals a fundamental challenge of analytical power calculations: more complex experimental designs and data generating processes require more complex treatment effect estimators, which in turn have analytic variance expressions that are increasingly difficult to derive and parameterize. For example, if we relax the assumption of randomization and instead consider a quasi-experimental DD design, where cross-sectional correlations remain important, the expression for $\text{Var}(\hat{\tau})$ includes 13 separate ψ parameters

[40] See Appendix C where we apply the SCR ANCOVA formula to the Bloom et al. (2015) data. We do not achieve the desired power in simulated experiments due to the unaccounted-for time shocks that exist in this real-world dataset.

[41] This will always be true with i.i.d. errors and with serial correlation in the absence of time shocks; the full result for the ANCOVA model including time shocks has yet to be proven.

— each of which would need to be non-parametrically estimated to fully characterize the error structure of the data and conduct an analytical power calculation.[42]

In light of these challenges, we recommend performing power calculations via simulation rather than by using analytic formulas, in cases where researchers have access to a pre-existing dataset that is representative of their future experimental data. Simulation-based power calculations follow a straightforward Monte Carlo process, where each simulation implements the proposed estimating equation over a range of assumed treatment effect sizes (τ), numbers of units (J), proportion of units treated (P), and panel lengths (m and r).[43] Calculating the average rejection rates of the null hypothesis $\tau = 0$ over all simulations will reveal the statistical power of each parameterization, and researchers can compare power across parameterizations to find their preferred values of P, J, m, and r. Importantly, simulation-based power calculations do not require researchers to estimate $\text{Var}(\hat{\tau})$ as a function of the underlying error structure in the data. This allows for greater flexibility in selecting research designs, and easily facilitates comparisons across alternative estimating equations.[44]

Simulation-based power calculations allow researchers to leverage any representative pre-existing data that may exist, and our analytical results provide key intuition for interpreting this simulation output. Given that power calculations via simulation are computationally intensive and necessitate a grid search over the full space of candidate design parameters, researchers may begin by using analytical power calculation formulas to narrow the range of plausible parameter values. In the absence of representative data *ex ante*, researchers may apply analytical techniques (with appropriate sensitivity analyses) to inform experimental design. It may still be possible to calibrate the variance-covariance parameters in the serial-correlation-robust power calculation formula, even if the ideal pre-existing dataset is not available.

[42] See Appendix A.2.3 for the full derivation and resulting power calculation equation.

[43] For each simulation, the researcher re-randomizes PJ units into treatment, adds τ to treated units' outcomes for all post-treatment periods, and estimates $\hat{\tau}$ using her preferred variance estimator. We provide further guidance on simulation-based power calculations in Appendix D.2.

[44] For example, a simulation-based power calculation for a proposed experiment using hourly electricity consumption data could compare the standard DD specification with individual and time fixed effects to an alternative specification that also includes group-specific time trends, without needing to formally derive an expression for $\text{Var}(\hat{\tau})$ under this alternative model.

6 Conclusion

Randomized experiments are costly, and it is important that researchers avoid underpowered experiments that are not informative, as well as dramatically overpowered experiments that waste resources. Statistical power calculations help researchers to calibrate the sample size of experiments *ex ante*, such that they are likely to collect enough data to detect treatment effects of a meaningful size, while also unlikely to collect excessive amounts of costly data. As data collection becomes easier and cheaper, panel data are becoming increasingly common in randomized experiments. Temporally disaggregated data allow researchers to ask new questions, and to apply a wider range of empirical methods to answer these questions (McKenzie (2012)).

In this paper, we identify a fundamental mismatch between existing *ex ante* power calculation techniques and *ex post* inference in panel data settings. We develop new tools to incorporate serial correlation into the design of panel RCTs. We derive the variance of a panel difference-in-differences estimator allowing for arbitrary within-unit correlation, which we use to update the conventional differences-in-differences power calculation formula derived by Frison and Pocock (1992). This new "serial-correlation-robust" formula is consistent with the CRVE variance estimator, which has become standard practice in *ex post* analysis of panel RCTs. We use Monte Carlo analyses to demonstrate that our updated power calculation formula achieves the desired statistical power, whereas the conventional formula is likely to produce either dramatically underpowered or overpowered experiments in the presence of serially correlated errors. These results hold in real data from a panel RCT in China, and for household electricity consumption data similar to that used in panel RCTs in the energy economics literature.

Our work highlights the need to carefully consider the assumptions that will enter *ex post* analysis when calibrating the design of experiments *ex ante*. The serial-correlation-robust power calculation framework allows researchers to conduct power calculations that correctly account for within-unit correlation, and provides intuition about these calculations under non-i.i.d. errors. We extend the main results by providing researchers with a framework for trading off units for effect sizes that takes cost into account; discussing the

ANCOVA estimator in the presence of serial correlation; and discussing implementation of power calculations in practice.[45] Ultimately, this paper serves as a useful starting point for power calculations with serial correlation. Future research should seek to extend this framework to accommodate heterogeneous effects, cluster-randomized trials, and ANCOVA models with time shocks.

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.

Angrist, Joshua D., and Jorn-Steffen Pischke. 2010. "The Credibility Revolution in Empirical Economics: How Better Research Design is Taking the Con out of Econometrics." *Journal of Economic Perspectives* 24 (2): 3–30.

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 Imbens. 2017. "The State of Applied Econometrics: Causality and Policy Evaluation." *Journal of Economic Perspectives* 31 (2): 3–32.

Athey, Susan, and Guido W. Imbens. 2016. "The Econometrics of Randomized Experiments." Working Paper.

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 ´Ozler. forthcoming. "Optimal Design of Experiments in the Presence of Interference." *Review of Economics and Statistics.*

[45] We have an accompanying software package, `pcpanel`, which makes our power calculation method easily accessible and user-friendly. The Stata package `pcpanel` is currently is available from `ssc`, with an `R` version to follow shortly, which will be available from `CRAN`.

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, Jonah B. Gelbach, and Douglas L. Miller. 2008. "Bootstrap-Based Improvements for Inference with Clustered Errors." *Review of Economics and Statistics* 90 (3): 414–427.

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

Campbell, Cathy. 1977. "Properties of Ordinary and Weighted Least Square Estimators of Regression Coefficients for Two-Stage Samples." *Proceedings of the Social Statistics Section, American Statistical Association:* 800–805.

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. Forthcoming. "Do Energy Efficiency Investments Deliver? Evidence from the Weatherization Assistance Program."

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 Takavarashi. 2013. *Running Randomized Evaluations: A Practical Guide.* Princeton, NJ: Princeton University Press.

Ito, Koichiro, Takanori Ida, and Makoto Tanaka. Forthcoming. “Moral Suasion and Economic Incentives: Field Experimental Evidence from Energy Demand.” *American Economic Journal: Economic Policy.*

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.

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

Moulton, Brent. 1986. “Random group effects and the precision of regression estimates.” *Journal of Econometrics* 32 (3): 385–397.

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/`.

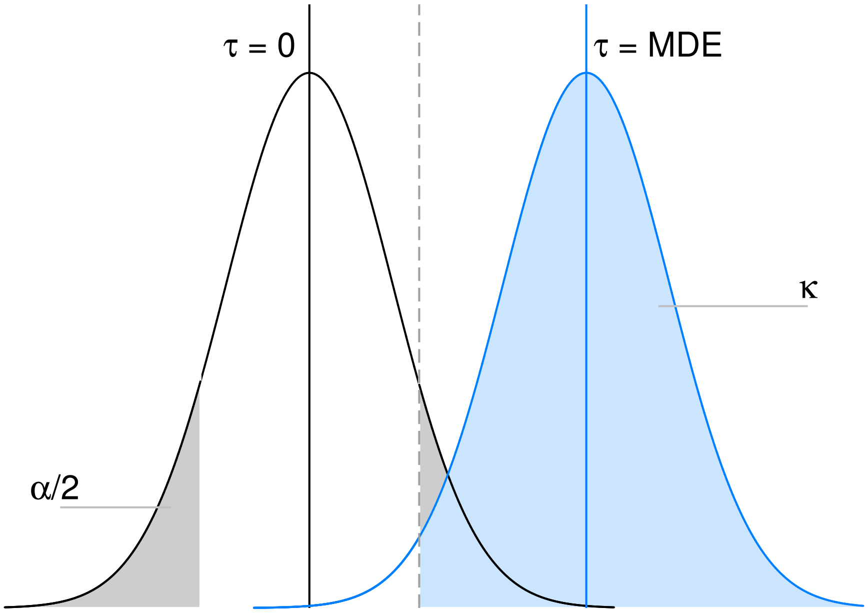
Rubin, Donald B. 1974. “Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies.” *Journal of Educational Psychology* 66 (5): 688–701.

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.

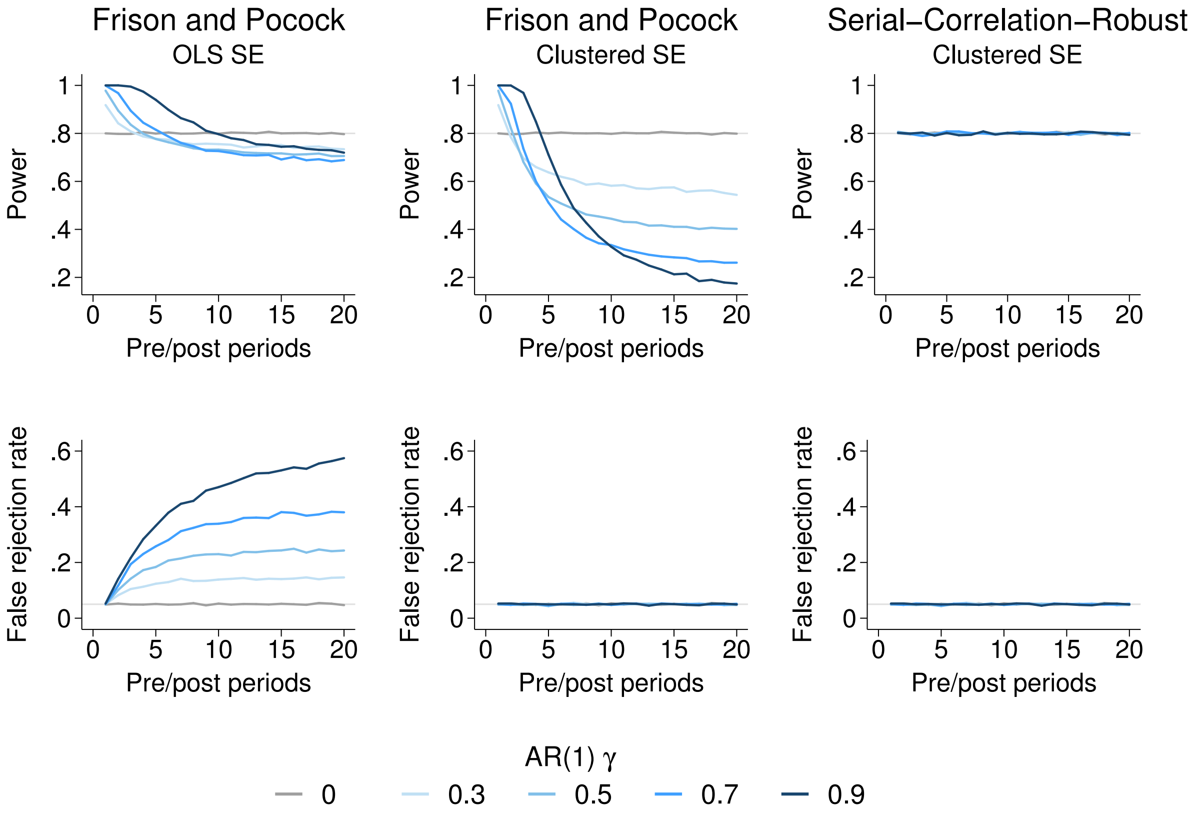
Tables and Figures

Figure 1: Hypothesis testing framework



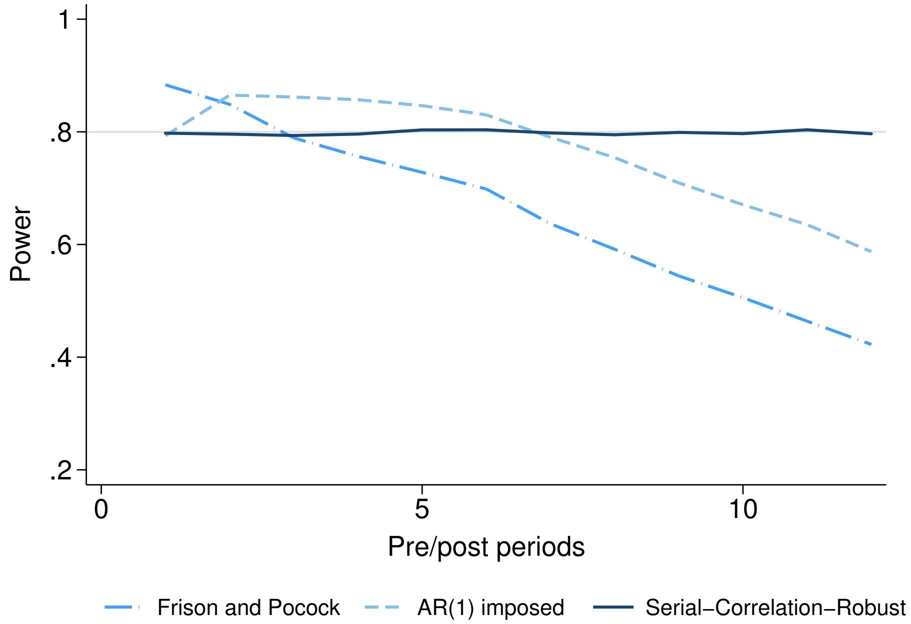
Notes: This figure displays the theoretical underpinnings of statistical power calculations. The black curve represents the distribution of the treatment effect estimator $\hat{\tau}$ under the null hypothesis of a zero effect. For a chosen significance level α, we will reject this null if $\hat{\tau}$ lies above (below) the $1 - \alpha/2$ ($\alpha/2$) percentile of the distribution. The gray-shaded area represents the likelihood of a Type I error. The blue curve is the distribution of $\hat{\tau}$ under the hypothesis that τ is equal to some other value, where this value is the minimum detectable effect (MDE) that yields a statistical power of κ. Given that $\tau = MDE$ and given the sample size J, the shaded blue area is the power of this test. The unshaded area to the left of the critical value (the dashed gray line) and under the blue distribution represents the likelihood of committing a Type II error.

Figure 2: Traditional methods result in improperly powered experiments in AR(1) data



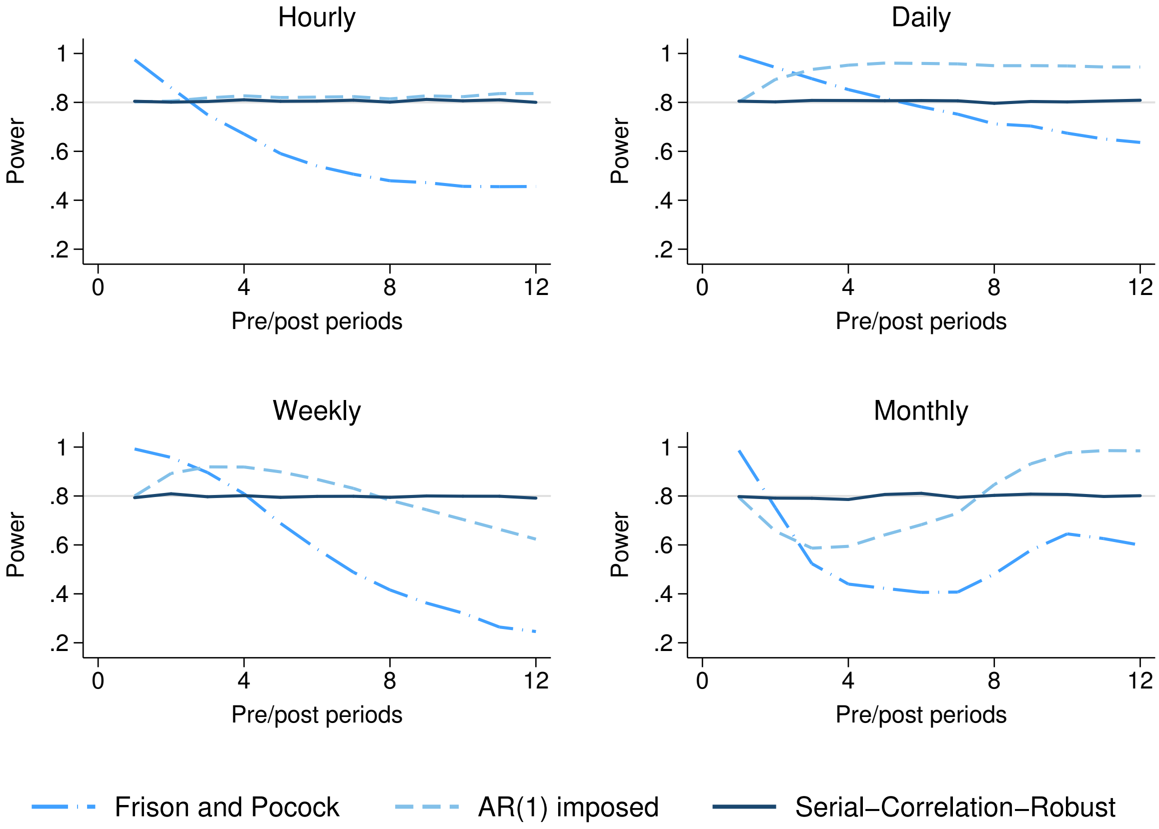
Notes: This figure displays power and rejection rates from performing power calculations with three different sets of assumptions on data generated with AR(1) processes with differing levels of serial correlation and differing panel lengths (ranging from 2, $m = r = 1$, to 40, $m = r = 20$). In the left column, we apply the formula of Frison and Pocock (1992) (Equation (3)), and use OLS standard errors *ex post*, in line with the assumptions of this formula. In the middle column, we again apply calibrate power calculations using Frison and Pocock (1992)'s 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 the serial-correlation-robust power calculation formula to account for non-i.i.d. errors *ex ante*, and we cluster standard errors at the individual level *ex post*. As expected, this third set of simulations achieves the desired 80 percent power and 5 percent false rejection rate.

Figure 3: Power simulations for Bloom et al. (2015) data



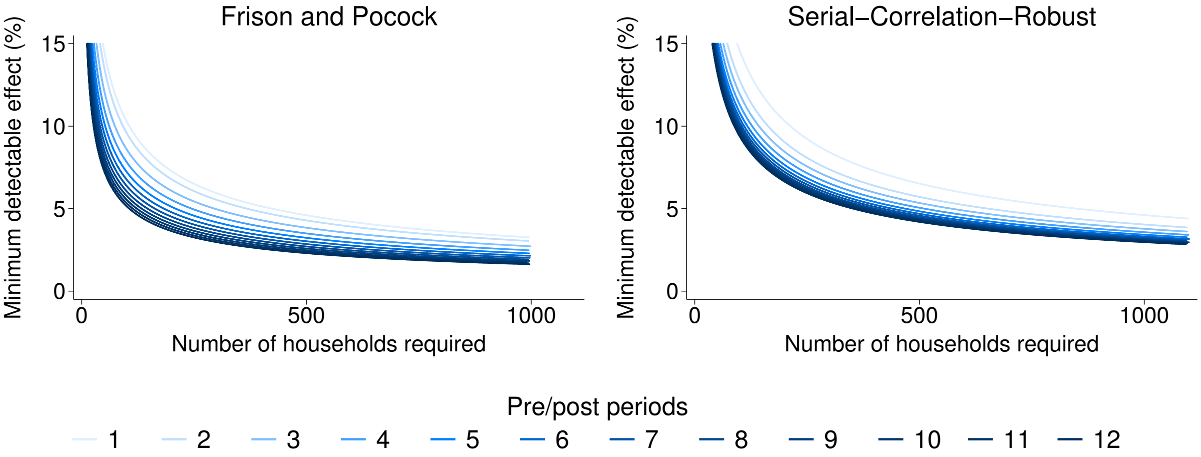
Notes: This figure shows results from Monte Carlo simulations using Bloom et al. (2015) data. Each curve displays the relationship between realized power and the number of pre/post periods used, applying different *ex ante* assumptions. The long-dashed line uses the Frison and Pocock (1992) formula. The short-dashed line uses the serial-correlation-robust formula, under the assumption that the error structure is AR(1). We estimate the AR(1) parameter via Equation (8). The solid line applies the serial-correlation-robust formula, where we non-parametrically generate estimates of $\psi^B_{\hat{\omega}}$, $\psi^A_{\hat{\omega}}$, and $\psi^X_{\hat{\omega}}$ terms using the Bloom et al. (2015) dataset. All three sets of simulations apply the CRVE *ex post*, clustering at the individual level. Only the serial-correlation-robust power calculation formula achieves the desired power of 80 percent, even though the Bloom et al. (2015) data exhibit relatively weak serial correlation.

Figure 4: Power simulations for Pecan Street data



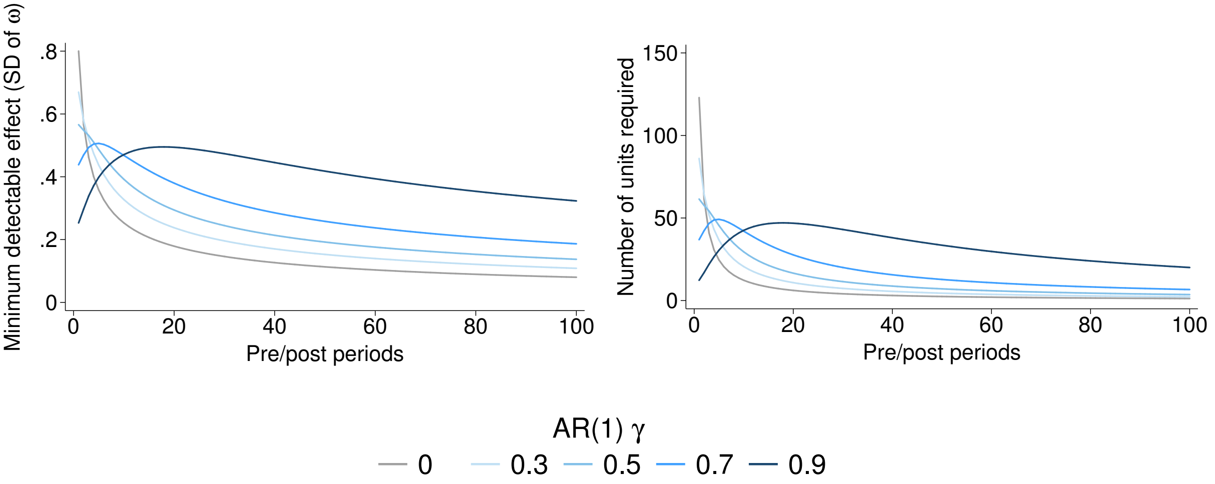
Notes: This figure shows results from Monte Carlo simulations using the Pecan Street electricity data, collapsed to different levels of aggregation. Each curve displays the relationship between power and the number of pre/post periods used, applying different *ex ante* assumptions. The long-dashed lines use the Frison and Pocock (1992) formula (Equation (3)). The short-dashed lines use Equation (5) assuming an AR(1) error structure, where we estimate the AR(1) parameters by estimating Equation (8) separately for each dataset. The solid lines apply Equation (5), by estimating separate covariance components of $\psi^B_{\hat{\omega}}$, $\psi^A_{\hat{\omega}}$, and $\psi^X_{\hat{\omega}}$ terms using residuals from each Pecan Street dataset. All simulations apply the CRVE *ex post*, clustering at the household level. For each temporal resolution, only the serial-correlation-robust power calculation formula achieves desired power of 80 percent.

Figure 5: Analytical power calculations – daily Pecan Street dataset



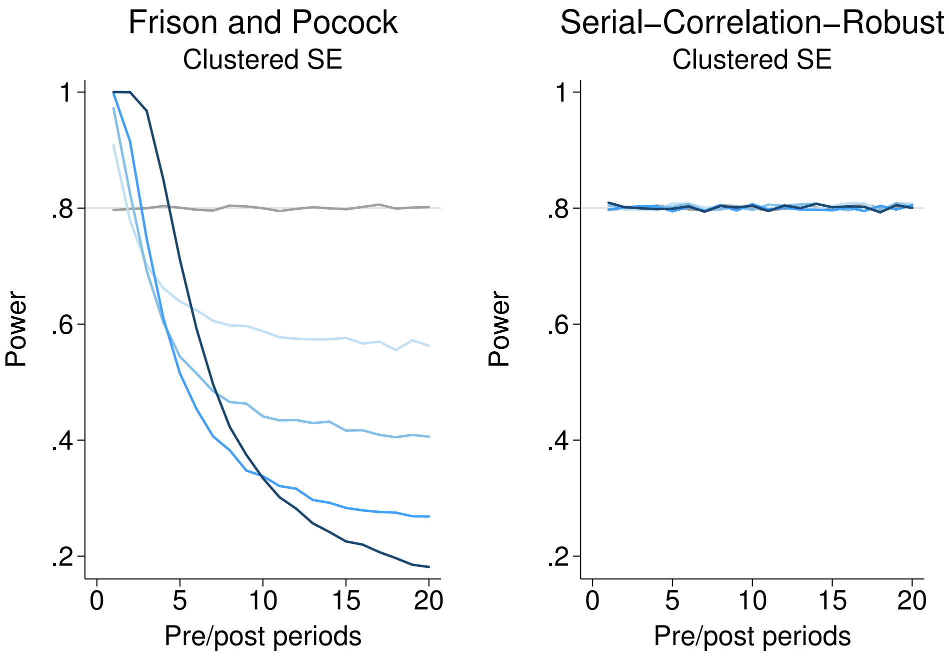
Notes: This figure shows the result of analytic power calculations on Pecan Street electricity data, collapsed to the daily level. Each curve displays the number of units required to detect a given minimum detectable effect with 80 percent power. The individual lines are 80 percent power curves for datasets with varying panel lengths, with the shortest panel (1 pre-period, 1 post-period) in light blue, and the longest panel (12 pre-periods, 12 post-periods) in navy. The left panel shows a power calculation using the Frison and Pocock (1992) formula, which assumes an i.i.d. error structure. The right panel applies the serial-correlation-robust formula, which accounts for the real error structure of the data. Note that, as discussed above, failing to account for the full covariance structure will lead a researcher to dramatically underestimate the sample size required to detect a given effect.

Figure 6: Analytical power calculations with increasing panel length



Notes: This figure displays the results of analytical power calculations using the serial-correlation-robust 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, the MDE declines monotonically in m and r. However, for higher γ, increasing m and r actually *increases* the MDE when $m = r$ is relatively small and *decreases* the 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.

Figure 7: Traditional ANCOVA methods fail under serial correlation



Notes: This figure displays power from performing power calculations for the ANCOVA estimator under two sets of assumptions on data generated with AR(1) processes with differing levels of serial correlation and differing panel lengths (ranging from 2, $m = r = 1$, to 40, $m = r = 20$). In the left panel, we apply the Frison and Pocock (1992) ANCOVA power calculation formula. In the right panel, we use the serial-correlation-robust ANCOVA power calculation. In both cases, we cluster our standard errors *ex post*, and achieve the desired 5 percent false rejection rate (not shown). As in the diffence-in-difference models above, the FP ANCOVA formula fails to generate correctly-powered experiments, whereas the serial-correlation-robust formula is properly powered across all panel lengths and degrees of serial correlation.