Title
Essays in Energy and Development Economics

Author
Burlig, Fiona Elizabeth Wilkes

Publication Date
2017

Peer reviewed|Thesis/dissertation

Essays in Energy and Development Economics

by

Fiona Elizabeth Wilkes Burlig

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Agricultural and Resource Economics

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Catherine Wolfram, Co-chair
Associate Professor Jeremy Magruder, Co-chair
Professor Maximilian Auffhammer
Professor Edward Miguel

Spring 2017

Essays in Energy and Development Economics

Abstract

Essays in Energy and Development Economics

by

Fiona Elizabeth Wilkes Burlig

Doctor of Philosophy in Agricultural and Resource Economics

University of California, Berkeley

Professor Catherine Wolfram, Co-chair

Associate Professor Jeremy Magruder, Co-chair

As demand for electricity grows around the world, so does the need for rigorous evaluation of energy policy interventions. In this dissertation, I use large datasets and modern econometric methods to study two such policies at scale - rural electrification in India and energy efficiency subsidies in California. I find that the benefits associated with these interventions are substantially smaller than previously thought, highlighting the importance of using new techniques for causal inference in these settings. In Chapter 1, I study the impacts of grid-scale rural electrification in India, using a regression discontinuity framework. In Chapter 2, I evaluate energy efficiency upgrades in K-12 schools in California using high-frequency data and novel machine learning methods. In Chapter 3, I develop methods to guide experimental design in the presence of panel data.

In the first chapter, coauthored with Louis Preonas, we study the impacts of energy access in the developing world. Over 1 billion people still lack electricity access. Developing countries are investing billions of dollars in rural electrification, targeting economic growth and poverty reduction, despite limited empirical evidence. We estimate the effects of rural electrification on economic development in the context of India's national electrification program, which reached over 400,000 villages. We use a regression discontinuity design and high-resolution geospatial data to identify medium-run economic impacts of electrification. We find a substantial increase in electricity use, but reject effects larger than 0.26 standard deviations across numerous measures of economic development, suggesting that rural electrification may be less beneficial than previously thought.

In the second chapter, coauthored with Christopher R. Knittel, David Rapson, Mar Reguant, and Catherine Wolfram, we study the impacts of energy efficiency

investments at public K-12 schools in California. We leverage high frequency data – electricity use every 15 minutes – to develop several approaches to estimating counterfactual energy consumption in the absence of the efficiency investments. In particular, We use difference-in-differences approaches with rich sets of fixed effects. We show, however, that these estimates are sensitive to the set of fixed effects included and to the set of schools included as controls. To address these concerns, We develop and implement a novel machine learning approach to predict counterfactual energy consumption at treated schools and validate the approach with non-treated schools. We find that the energy efficiency projects in our sample reduce electricity consumption between 2 to 5% on average, which can result in substantial savings to schools. We also compare our estimates of the energy savings to ex ante engineering estimates. Realized savings are generally less than 50% of ex ante forecasts and quite low for measures other than heating and air-conditioning systems or lighting.

In the third chapter, coauthored with Louis Preonas and Matt Woerman, we seek to answer: 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.

Contents

List of Figures

List of Tables

Acknowledgments

I am deeply grateful to many people for their guidance and support. My advisors, Catherine Wolfram and Jeremy Magruder, have been invaluable. Catherine is a role model and an inspiration, and I am lucky to call myself her student. Jeremy has devoted countless hours to teaching me to think rigorously about difficult problems. Both Catherine and Jeremy simultaneously hold me to an incredibly high standard and teach me how I can make that standard attainable.

I would also like to thank Maximilian Auffhammer, Severin Borenstein, Meredith Fowlie, and Lucas Davis for taking in, putting up with, and striving to improve this wayward development economist, both professionally and personally. I am so lucky to be a part of the Energy Institute family. Aprajit Mahajan, Betty Sadoulet, and Ted Miguel broaden and simultaneously sharpen my thinking.

I have been lucky to work with and learn from a number of other faculty members at Berkeley and elsewhere. Dave Rapson, Chris Knittel, and Mar Reguant, my coauthors on the third chapter of this dissertation, have helped me appreciate the research process. Solomon Hsiang inspires me to ask big questions. Alain de Janvry, Ethan Ligon, Jim Sallee, and Reed Walker provide endless constructive criticism, and have made me a better economist.

I am indebted to Andy Campbell, Casey Hennig, Carmen Karahalios, Karen Notsund, Paula Pedro, and Maggie Smith for their support at Berkeley. I am also grateful to the National Science Foundation for generously funding this work.

Patrick Baylis, Ceren Baysan, Kenny Bell, Susanna Berkouwer, Josh Blonz, Tamma Carleton, Aluma Dembo, Sylvan Herskowitz, Erin Kelley, Dave McLaughlin, Kate Pennington, Louis Preonas, Andrew Stevens, Becca Taylor, and Matt Woerman have allowed me to ask hard questions, learn, fail, succeed, and not take myself too seriously. I could not ask for a better group of friends and compatriots.

Finally, I thank my family, who are and remain an endless source of generosity, support, and love.

Chapter 1

Out of the Darkness and Into the Light? Development Effects of Rural Electrification[1]

1.1 Introduction

Approximately 1.1 billion people around the world still lack access to electricity. These people are overwhelmingly rural, and live almost exclusively in Sub-Saharan Africa and Asia. In recent years, developing countries have made large investments to extend the electricity grid to the rural poor. The International Energy Agency estimates that approximately $9 billion was spent on electrification in 2009, which it expects to rise to $14 billion per year by 2030 (International Energy Agency (2011)). This is not surprising, given that electrification is widely touted as an essential tool to help alleviate poverty and spur economic progress; universal energy access is one of the UN's Sustainable Development Goals (UNDP (2015), World Bank (2015)). While access to electricity is highly correlated with GDP at the national level, there exists limited evidence on the causal effects of electricity access on rural economies.

Recovering causal estimates of the effects of electrification is challenging, since energy infrastructure projects target relatively wealthy or quickly-growing regions. Selection of this kind would bias econometric estimates of treatment effects toward finding large economic impacts. Previous work has relied on instrumental variables strategies to circumvent this problem, and has tended to find large positive effects of

[1]The material in this chapter is from Energy Institute at Haas Working Paper #268, coauthored with Louis Preonas. The original version can be found online at `https://ei.haas.berkeley.edu/research/papers/WP268.pdf`.

electrification. Posited mechanisms for these gains include structural transformation, which in turn changes employment opportunities (Rud (2012)); female empowerment (Dinkelman (2011)); increased agricultural productivity (Chakravorty, Emerick, and Ravago (2016)); health improvements as households switch from kerosene and coal to electricity (Barron and Torero (2016)); and greater educational attainment (Lipscomb, Mobarak, and Barham (2013)).

This paper documents that while large-scale rural electrification causes a substantial increase in energy access and power consumption, it leads at best to small changes in economic outcomes in the medium term. We exploit quasi-experimental variation in electrification generated by a population-based eligibility cutoff in India's massive national rural electrification program, Rajiv Gandhi Grameen Vidyutikaran Yojana (RGGVY). The "Prime Minister's Rural Electrification Program" was launched in 2005 to expand electricity access in over 400,000 rural Indian villages across 27 states. In order to cap program costs, the Central Government introduced a population-based eligibility cutoff based on the size of village neighborhoods ("habitations").[2] When the program was introduced, only villages with constituent habitations larger than 300 people were eligible for electrification under RGGVY.

We pair detailed geospatial information with rich administrative data on the universe of Indian villages and use a regression discontinuity (RD) design to test for the village-level effects of RGGVY eligibility on employment, asset ownership, household wealth, village-wide outcomes, and education. This design relies on relatively weak identifying assumptions, and we provide evidence that these assumptions are satisfied below. We estimate effects using a main sample of nearly 30,000 villages across 22 states. We demonstrate that RGGVY led to statistically significant and economically meaningful increases in electric power availability and consumption that is visible from space. We then show that despite these gains, electrification led to at most modest changes in economic outcomes. More specifically, we are able to reject even small changes, of 0.26 of a standard deviation, across a range of outcomes, including employment, asset ownership, the housing stock, village-wide outcomes, household wealth, and school enrollment. Taken together, these results suggest that the causal impact of large-scale rural electrification on economic development may be substantially smaller than previously thought.

We show that these small effects do not simply reflect issues with the timing or quality of RGGVY project implementation. Our results are quantitatively similar for villages electrified near the beginning and near the end of our sample period, meaning

[2]In the 2001 Indian Census, the village was the lowest-level administrative unit. Villages are composed of habitations (or "hamlets"), which correspond to the inhabited areas of a village. South Asian villages typically have one or more inhabited regions surrounded by agricultural land. India's 600,000 villages contain approximately 1.6 million unique habitations.

that any confounding rollout effects are unlikely. Likewise, we find quantitatively similar results for the subset of states with above-average power supply reliability, which suggests that even in places with relatively infrequent power outages, the economic impacts remain quite small. We also employ an alternative identification strategy, difference-in-differences (DD), which reveals that our RD results appear to generalize to villages far from our 300-person population cutoff. Using this DD approach, we find treatment effects that are broadly consistent with our RD strategy, across the full support of Indian village populations. Our main RD results also stand up to a battery of placebo tests, falsification exercises, and robustness checks.

This paper makes three key contributions to the existing literature. First, our results contrast starkly with the large economic impacts of electrification found in earlier work. They apply directly to rural villages across 27 states in India, representing the world's largest unelectrified population. Perhaps more importantly, we use a regression discontinuity design to quantify the effects of electrification; this necessitates substantially weaker identifying assumptions than the instrumental variables approaches of the prior literature. Second, we add to the knowledge on the economic effects of infrastructure in developing countries. Existing work in this area has tended to find large positive impacts of infrastructure investments.[3] We provide evidence that electricity infrastructure may not necessarily spur large-scale economic growth. Third, our results contribute to a small but growing literature on energy use in the developing world.[4] We demonstrate that while electrified villages are consuming power, this energy use does not appear to be transforming rural economies.

The remainder of the paper proceeds as follows: Sections 1.2, 1.3, and 1.4 describe rural electrification in India, our empirical strategy, and the data used in our analysis. Section 1.5 presents our main empirical results, which we discuss and interpret in Section 1.6. Section 3.7 concludes.

1.2 RGGVY

At the time of its independence in 1947, only 1,500 of India's villages had access to electricity (Tsujita (2014)). By March 2014, that number had risen to 576,554 out of 597,464 total villages. This massive technological achievement is largely attributable to a series of national electrification programs, the first of which began in the 1950s. The flagship program of India's modern electrification efforts was Ra-

[3]See, for example, Donaldson (forthcoming) on the effects of railroads on trade costs and welfare in India and Banerjee, Duflo, and Qian (2012) and Faber (2014) on roads in China.

[4]See Gertler et al. (2016) on income growth and energy demand, Allcott, Collard-Wexler, and O'Connell (2016) on power outages, and McRae (2015) on energy infrastructure.

jiv Gandhi Grameen Vidyutikaran Yojana (RGGVY), or the Prime Minister's Rural Electrification Plan. Prior to RGGVY, over 125,000 (21 percent) of rural villages had no access to power whatsoever. Many of the remaining villages had extremely limited power access; 57 percent of all rural households lacked access to electricity, with the majority of unelectrified households falling below the poverty line. Substantial opportunities remained to expand electricity access in rural communities.

RGGVY was launched in 2005 with the goal of extending power access to over 100,000 unelectrified rural villages in 27 Indian states. The program also set out to provide more intensive electrification to over 300,000 "under-electrified" villages. RGGVY's primary mandate was to install and upgrade electricity infrastructure — specifically transmission lines, distribution lines, and transformers — in order to support electric irrigation pumps, small-to-medium industries, cold chains, healthcare, schooling, and information technology applications. Such infrastructure investments aimed to "facilitate overall rural development, employment generation, and poverty alleviation" (Ministry of Power (2005)). RGGVY also extended electric connections to public places, including schools, health clinics, and local government offices. While the program focused on providing electricity infrastructure to support growing village economies, RGGVY was also charged with extending household electricity access by offering free grid connections to all households below the poverty line.[5] RGGVY investments occurred primarily on the intensive margin, upgrading existing infrastructure to have the capability to power growing rural economies. The majority of RGGVY works, including new grid connections, occurred in villages with some degree of household electrification prior to 2005.

In order for a village to be electrified under RGGVY, its state government had to submit an implementation proposal to the Rural Electrification Corporation (REC), a public-private financial institution overseen by the national government's Ministry of Power. These district-specific proposals, or Detailed Project Reports (DPRs), were based on village-level surveys carried out by local electric utilities, covering both unelectrified villages and partially electrified villages in need of "intensive electrification." Each DPR proposed a village-by-village implementation plan, which included details on new electricity infrastructure to be installed and the number of households and public places to be connected. The REC reviewed DPR proposals, approved projects, and disbursed funds to states.

[5]Above poverty line households were able to purchase connections. All households were required to pay for their own power consumption. The program did not subsidize the consumption of electricity for any household, but Indian retail electricity tariffs are heavily subsidized, and average 2.4 rupees (4 U.S. cents) per kilowatt-hour.

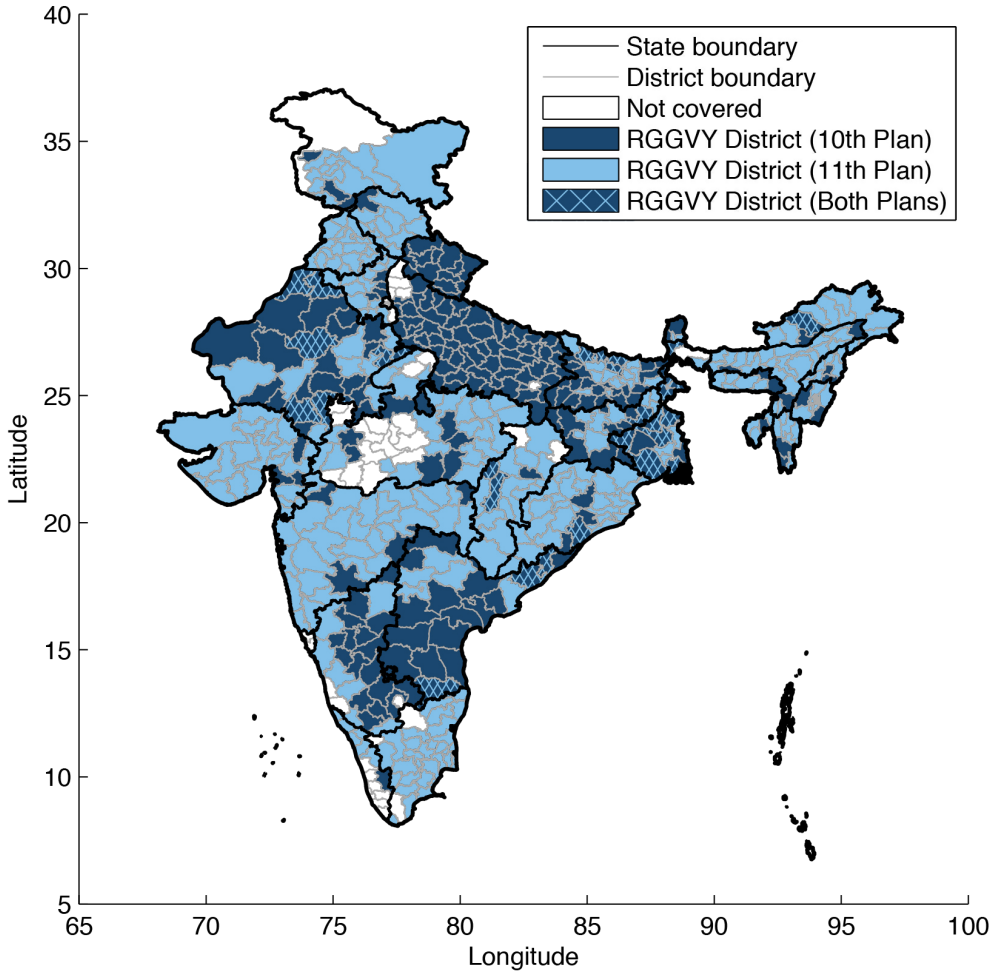
Funding for RGGVY came from India's 10th (2002–2007) and 11th (2007–2012) Five-Year Plans.[6] Districts were sorted into Plans on a first-come, first-serve basis: the first group of approved DPRs were allocated funding under the 10th Plan, and the next group were allocated funding under the 11th Plan. Under the 10th Plan, all villages with habitation populations above 300 were eligible for RGGVY electrification. Under the 11th Plan, this threshold was decreased to 100. Approximately 164,000 (267,000) villages in 229 (331) districts in 25 (25) states were slated for electrification under the 10th (11th) Plan, which also targeted 7.5 million (14.6 million) below-poverty-line households for free connections. Funding for the 10th Plan was disbursed between 2005 and 2010, with over 95 percent of funds released before 2008. The 11th Plan distributed funds between 2008 and 2011.[7]

Figure 1.2.1 shows the spatial distribution of RGGVY districts covered by the 10th and 11th Plans, highlighting the program's broad scope. The vast majority of eligible districts received RGGVY funding under exactly one Five-Year Plan, and 23 out of 27 states contain both 10th- and 11th-Plan districts. We focus our empirical analysis on the districts that received RGGVY funding under the 10th Plan, because electrification in these districts was completed earlier, giving us a longer post-electrification sample period.

[6]Midway through the 12th Plan, RGGVY was subsumed into Deendayal Upadhyaya Gram Jyoti Yojana (DDUGJY); the remaining projects are slated to be finished by the end of the 13th Plan. As of 2016, all villages are eligible for electrification under DDUGJY, regardless of size.

[7]We downloaded data on RGGVY implementation from `http://www.rggvy.gov.in`, since replaced with `http://www.ddugjy.gov.in`. Appendix A.3 describes the RGGVY program in greater detail, along with additional background on the history of rural electrification in India.

Figure 1.2.1: Indian Districts by RGGVY Implementation Phase



Note. — This map shows 2001 district boundaries, shaded by RGGVY coverage status. Navy districts are covered under the 10th Plan, light blue districts are covered under the 11th Plan, cross-hatched districts were covered under both the 10th and 11th Plans, and white districts are not covered by RGGVY. As of 2001, India had 584 districts across its 28 states and 7 Union Territories. RGGVY covered 530 total districts in 27 states (neither Goa nor the Union Territories were eligible), with 30 districts split between the 10th and 11th Plans.

1.3 Empirical approach

1.3.1 Regression discontinuity design

In this paper, we aim to estimate the causal effect of rural electrification on development. Because energy infrastructure programs are large-scale investments, and because governments allocate funds to specific regions or groups of people in ways that are likely correlated with economic outcomes of interest, it can be challenging to disentangle the impact of electrification from other observed and unobserved factors that affect development. Furthermore, since the electricity grid is spatially integrated, a national-scale rollout of electrification is likely to have different effects than can be observed by a randomized controlled trial that impacts a few hundred rural villages.[8] To overcome these challenges, we implement a regression discontinuity design, allowing us to identify the causal effect of electrification at scale.

Under the RGGVY program rules, villages in 10th-Plan districts were eligible for treatment if they contained habitations with populations of 300 or above. Our RD analysis includes only villages whose districts received funding under the 10th Plan, and we restrict our sample to villages with exactly one habitation. This allows us to use an RD to estimate local average treatment effects for villages with habitation populations close to this 300-person cutoff. In this sharp RD design, eligibility for treatment changes discontinuously from 0 to 1 as village population (our running variable) crosses the 300-person threshold, allowing us to identify the effects of eligibility for RGGVY on both observed changes in electrification and on village-level economic outcomes.[9]

This design necessitates two main identifying assumptions. First, we must assume continuity across the RD threshold for all village covariates and unobservables that might be correlated with our outcome variables. While this assumption is fundamentally untestable, we support it with evidence from several key village characteristics.[10] We know of no other Indian social program with a 300-person eligibility threshold. Second, we assume that our running variable, 2001 Census population, is not manipulable around the threshold. Because our running variable predates the announcement of RGGVY in 2005, we are confident that our population data were

[8] Lee, Miguel, and Wolfram (2016) are implementing a randomized controlled trial of household electrification in 150 rural communities in Western Kenya.

[9] See Imbens and Lemieux (2008) and Lee and Lemieux (2010) for further detail about the formal assumptions underlying RD analysis, and practical issues in applying RD designs.

[10] We find no evidence to suggest that pre-period covariates change discontinuously across the 300-person cutoff. These results are available in Appendix A.2.4.5, as well as in Figure 1.5.5 below.

not influenced by the future existence of RGGVY. Figure 1.4.3 shows no evidence of bunching of villages around this 300-person population cutoff.

Given these assumptions, our RD design provides a consistent estimate of the effect of eligibility for treatment on outcomes of interest for the set of single-habitation villages located in districts that received RGGVY funding under the 10th Plan. Formally, we estimate:

(1.1)
$$Y_{vs}^{2011} = \beta_0 + \beta_1 Z_{vs} + \beta_2(P_{vs} - 300) + \beta_3(P_{vs} - 300) \cdot Z_{vs} + \beta_4 Y_{vs}^{2001} + \eta_s + \varepsilon_{vs}$$
$$\text{for } \; 300 - h \leq P_{vs} \leq 300 + h\,, \quad \text{where } \; Z_{vs} \equiv \mathbf{1}[P_{vs} \geq 300]\,.$$

Y_{vs}^{2011} represents the outcome of interest in village v in state s in 2011, P_{vs} is the 2001 village population, Z_{vs} is the RD indicator equal to one for villages above the cutoff, h is the RD bandwidth, Y_{vs}^{2001} is the 2001 value of the outcome variable, η_s is a state fixed effect, and ε_{vs} is an idiosyncratic error term.[11] We cluster our standard errors at the district level to allow for arbitrary dependence between the errors of villages within the same district. This accommodates both implementer-specific correlations within a district's DPR (RGGVY's unit of project implementation) and natural spatial autocorrelation between nearby villages. We use a preferred RD bandwidth of 150 people on either side of the 300-person cutoff; this allows us to include a large sample of villages, while remaining confident that villages away from the discontinuity are similar to those at the 300-person cutoff.[12]

1.3.2 Economic Outcomes

Economic theory suggests that electrification could impact village economies through several channels. First, as electricity becomes available, we should expect small firms to invest in new capital equipment that uses power. This in turn would raise the marginal product of labor in the non-agricultural sector, drawing workers to new employment opportunities (Rud (2012)). On the other hand, electrification could spur agricultural mechanization, which would improve farm productivity (Chakravorty, Emerick, and Ravago (2016)).[13] This could either increase or decrease employment in agriculture.[14] However, because the marginal product of labor would

[11] Neither the 2001 value of the outcome variable nor the fixed effects are necessary for identification, but they improve the precision of our estimates (see Lee and Lemieux (2010)).

[12] We perform bandwidth sensitivity checks in Appendix A.2.1.2, including calculating the Imbens and Kalyanaraman (2012) optimal bandwidth; our results are not sensitive to bandwidth choice.

[13] In the Indian context, one potential use of electricity in agricultural production is to power irrigation tubewells.

[14] The potential for changes in agricultural employment depends on several factors, including the excess supply of labor, the excess supply of farmland, the degree to which farm mechanization and

unambiguously increase in both the agricultural and non-agricultural sectors, this should increase wages, incomes, and expenditures.

Next, electricity access may lead to gains for women. New employment opportunities, like those described above, could enable more women to work outside the home. Alternatively, newly-electrified households could invest in labor-saving devices, which could decrease the time required for women to complete household duties. This could also lead to increased female employment, either outside the home or in microenterprises. Dinkelman (2011) uses an instrumental variables approach in South Africa, and finds that electrification substantially raises female employment through this latter channel.

Rural electrification may also bring substantial health benefits. Kerosene is widely used throughout the developing world as a fuel for both lighting and cooking, and Indian households also commonly cook with coal and biomass. Combustion of these fuels produces harmful indoor air pollution, which is especially detrimental to young children and infants in utero. Access to electricity may foster investment in electric lights and electric cookstoves, which would likely reduce indoor air pollution and improve child health outcomes (Barron and Torero (2016)). Electrification may also indirectly improve health outcomes, through higher incomes and improved access to health care.

Finally, electrification could impact educational attainment through several channels. On the extensive margin, total school enrollment may increase if electrification leads to income gains, making households less reliant on child labor earnings. On the other hand, rising wages may draw students out of school and into the labor force. Alternatively, we might expect electricity access to change the education production function. Lighting or computing facilities in schools may improve learning in the classroom, and children in homes with electric lighting will likely develop more effective study habits. If electrification improves student performance, it could affect the intensive margin of schooling as students tend to stay in school longer, causing enrollment in upper grades to increase. Using instrumental variables strategies, Lipscomb, Mobarak, and Barham (2013) find that rural electrification increases the number of years that students attend school.

1.4 Data

Our empirical analysis uses data from four main sources. First, we link satellite images of nighttime brightness to village boundary shapefiles, yielding a panel of

agricultural labor are complements or substitutes, and the effect of electricity access on agricultural commodity prices.

village brightness. Next, we use several large administrative datasets published by three different Indian government entities, which contain village populations and a broad set of economic indicators. Armed with a wealth of data on Indian villages, we can test the channels through which we expect electrification under RGGVY to impact economic development.

1.4.1 Nighttime lights data

In order to understand the economic effects of electrification resulting from RGGVY, we must first demonstrate that RGGVY led to a meaningful increase in electricity access and consumption in rural Indian villages. A binary indicator of the presence of electricity infrastructure would be insufficient, since it would mask heterogeneity in power quality, electricity consumption, and connection density. There exists no comprehensive dataset of power consumption at the village level across India, but we are able to construct a measure of electricity consumption using remotely-sensed data.

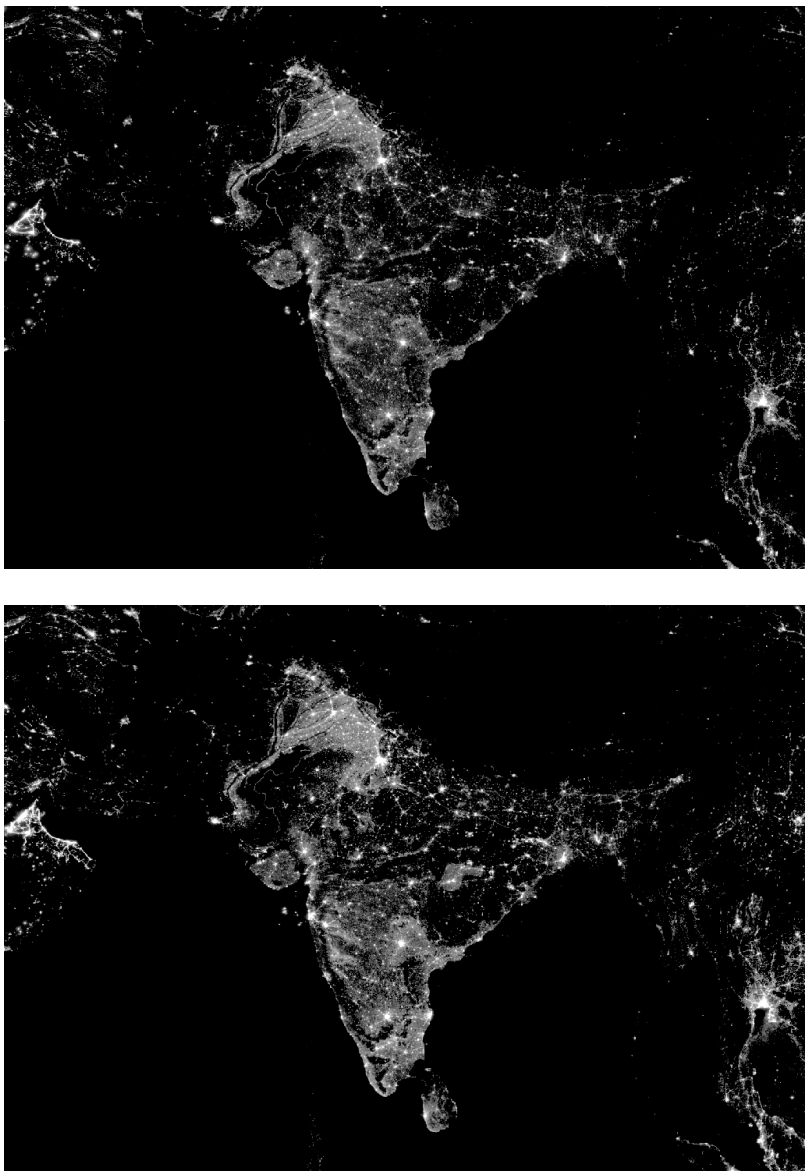
As an indicator of electrification under RGGVY, we use changes in nighttime brightness as observed from space. The National Oceanic and Atmospheric Administration's Defense Meteorological Satellite Program–Operational Line Scan (DMSP-OLS) program collects images from U.S. Air Force satellites, which photograph the earth daily between 8:30pm and 10:00pm local time. After cleaning and processing these images, NOAA averages them across each year and distributes annual composite images online.[15] Each yearly dataset reports light intensity for each 30 arc-second pixel (approximately 1 km^2 at the equator) on a 0–63 scale, which is proportional to average observed luminosity.[16] Figure 1.4.2 shows nighttime brightness in India in 2001 and 2011.

Economists frequently use these nighttime lights data as proxies for economic activity, as popularized by Chen and Nordhaus (2011) and Henderson, Storeygard, and Weil (2012). Existing work demonstrates that nighttime brightness can also be used to detect electrification, even at small spatial scales: Min et al. (2013) find evidence

[15]This cleaning removes any sunlit hours, glare, cloud cover, forest fires, the aurora phenomena, and other irregularities. Nighttime lights data are available for download at `http://ngdc.noaa.gov/eog/dmsp/downloadV4composites.html`. We use the average lights product in our main analysis. See Appendix A.1.3 for further discussion.

[16]Chen and Nordhaus (2011) detail the relationship between physical luminosity and brightness in the nighttime lights images.

Figure 1.4.2: Nighttime Lights in India, 2001 and 2011



Note. — This figure shows the DMSP-OLS nighttime brightness data for India. The top panel shows nighttime lights in 2001, and the bottom panel shows nighttime lights for 2011. The $\approx$1km^2 pixels in this image range in brightness from 0 to 63, covering the full range of the DMSP-OLS data.

of a statistically detectable relationship between NOAA DMSP-OLS brightness and the electrification status of rural villages in Senegal and Mali. Min and Gaba (2014) show that a similar correlation between electrification and nighttime brightness also exists in rural Vietnam. Chand et al. (2009) show a direct relationship between nighttime lights and electric power consumption in India, while Min (2011) finds a strong correlation between brightness and district-level electricity consumption in Uttar Pradesh. We build on this research by using nighttime brightness to demonstrate that RGGVY successfully increased village electricity access, where nighttime lights serve an objective measure of realized energy consumption in these villages.

Importantly, these satellite images represent a lower bound on electricity consumption. While nighttime brightness data record light output (including lighting from houses, public spaces, and outdoor streetlights), they do not directly measure electricity consumed for other purposes. Because all electricity end-uses rely on the same power grid, we treat increases in nighttime brightness as necessary indicators of investments in electricity infrastructure. Likewise, if total electricity consumption increases, we should expect nighttime brightness to increase as well, as more power reaches rural villages. A potential concern with using nighttime lights to proxy for total electricity consumption is that we could mistake new sources of outdoor lighting for increases in electricity access. However, RGGVY's primary mandate was to expand and improve electricity infrastructure, and there is no mention of streetlight installation in 10th-Plan program documentation.[17] Hence, an observed increase in nighttime brightness as a result of RGGVY would very likely be driven not by new streetlights alone, but rather by village-wide increases in access to energy services.

We construct a village-level panel of nighttime brightness by overlaying annual NOAA DMSP-OLS images with 2001 village shapefiles.[18] Our preferred measure of a village's lighting is the maximum brightness of any pixel whose centroid lies within its borders.[19]

We use the brightest pixel because Indian villages are typically organized such that there are centralized populated areas surrounded by fields. This targets our electrification measure at the populated parts of villages, while avoiding measure-

[17]RGGVY 11th-Plan documentation did discuss streetlights in the context of a small carve-out for microgrids targeted at extremely remote villages. Because this carve-out did not exist under the 10th Plan, the 300-person eligibility cutoff did not apply for these villages.

[18]Indian villages have official boundaries, which are recorded by the Census Organization of India. Every square meter in India (excluding bodies of water and forests) is contained in a city, town, or village. We use shapefiles of village boundaries published by ML InfoMap, Ltd.

[19]We calculate this level in ArcGIS, using the standard `Zonal Statistics as Table` operation. For villages too small to contain a pixel's centroid, we assign the brightness value of the pixel at the village centroid.

ment error from brightness averaged across unlit agricultural land.[20] In performing this calculation, we are forced to drop 10 states from our sample. We are missing shapefiles for five states, which represent fewer than 3 percent of the total villages covered by RGGVY. We also exclude five states because we believe these shapefiles to be of extremely low quality: the correlation between the village area implied by the shapefiles and village area recorded by the Indian Census, the entity in charge of defining village boundaries, is below 0.35.[21] We are left with a nighttime lights sample of 370,689 villages across 15 states. We do not impose these sample restrictions for any other outcome variables.

1.4.2 Census of India

We combine several village-level datasets published by the Census of India from the 2001 and 2011 decennial Censuses.[22] The Primary Census Abstract (PCA) contains village population data, and a detailed breakdown of labor allocation by gender and job type. In particular, the PCA reports the number of men and women that are working in agriculture; "household industry workers" (engaged in informal production of goods within the home); and "other workers" that engage in all other types of work.[23] Examples of "other workers" include government servants, municipal employees, teachers, factory workers, and those engaged in trade, commerce, or business. These data allow us to test for sectoral shifts in employment due to RGGVY electrification, either away from agriculture (consistent with structural transformation) or into agriculture (consistent with increased agricultural productivity). We also test for effects on female employment. Because we observe the share of women engaged in economic activity both outside and within the home, these data are well-suited to capture potential impacts of electrification on female labor.

[20] Our results remain largely unchanged if we use the mean lights value rather than the maximum value. We also undertake a procedure to remove measurement error from the nightlights data via linear projection. See Appendix A.1.3 for details.

[21] The five states with missing shapefiles are Arunachal Pradesh, Meghalaya, Mizoram, Nagaland, and Sikkim. The five states with low-quality shapefiles and village areas are Assam, Himachal Pradesh, Jammu and Kashmir, Uttar Pradesh, and Uttarakhand. The remaining states in the sample all have correlations between datasets above 0.6. See Appendix A.1.2 for further discussion.

[22] These data are all publicly available at `http://www.censusindia.gov.in`. Because our research design relies on observing a large number of villages with populations around 300, we are unable to use additional Indian survey datasets such as the NSS or ASI. These datasets do not include a sufficient number of small villages to support our RD analysis, and are not designed to be representative below the district level.

[23] The agriculture category is decomposed further into "cultivators" (on their own land) and "agricultural laborers" (on others' land).

The Houselisting Primary Census Abstract (HPCA) provides extensive data on living conditions, household size, physical household characteristics, and asset ownership. These data report the fraction of households that own a variety of assets, including radios, mobile phones, bicycles, motorcycles, and televisions. RGGVY may have contributed both directly and indirectly to asset ownership, if households purchased electric appliances to take advantage of improved power availability, or if potential income games from electrification enabled increased household expenditures on durable goods. Physical housing characteristics such as floor and roof materials are indicators of household wealth. If RGGVY spurred increases in household expenditures, we expect to observe medium-run investments to improve the housing stock. The HPCA also allows us to examine the health channel, as this dataset reports the fraction of households that cook with electricity and that use kerosene as a main source of lighting.

Finally, the Village Directory (VD), another Census dataset, contains detailed information on village amenities.[24] In particular, the VD includes data on the presence of education and medical facilities; banking facilities and agricultural credit societies; the existence and quality of road network connections and the presence of bus services; and communications access, including postal services and mobile phone networks. We use these data to test for the effects of RGGVY on village amenities. The VD also includes information on village electrification, in the form of binary indicators of electric power availability in each village, separately for the agricultural, domestic, and commercial sectors. These indicators are coded as "1" if *any* electric power was available for a given end use anywhere in the village, and as "0" otherwise. Two-thirds of RGGVY 10th-Plan villages met this criterion at baseline (i.e. were coded as "1" for electric power availability), making these variables particularly poorly suited to analyze the effects of RGGVY. The main goals of RGGVY were to upgrade energy infrastructure and increase the penetration of electricity access within each village. The VD data contain no information on the intensity of electrification within a village, and therefore do not reflect the vast majority of RGGVY works.[25] We instead turn to the nighttime lights data, which allow us to track intensive-margin changes in energy consumption.

We combine the PCA, HPCA, and VD data into a two-wave village-level panel. The 2001 PCA also reports the official 2001 population of each village, which was the population of record for the RGGVY program, and which we use as our RD running

[24] In 2001, the VD was a separate Census product. In 2011, it was bundled into the District Census Handbook (DCHB).

[25] The 2011 Village Directory also reports the average hours of electricity available per day, by sector. Because electricity is distributed over an integrated grid, it is unlikely that RGGVY's infrastructure upgrades would have any effect on these measures of electricity access.

variable.[26] However, RGGVY implementing agencies were instructed to determine eligibility based on 2001 *habitation* (sub-village neighborhood) populations. To the best of our knowledge, the only nation-wide habitation census in existence was conducted by the National Rural Drinking Water Program.[27] We use a fuzzy matching algorithm, modified from Asher and Novosad (2016), to link this habitation census to our village panel and identify the 50 percent of villages with exactly one habitation.[28] For these single-habitation villages, habitation populations are equivalent to village populations—meaning that 2001 village population should exactly correspond to the population that determined RGGVY eligibility for these villages.

The main dataset for our analysis contains the 2001–2011 Census, nighttime brightness, RGGVY program implementation details, and the number of habitations in each village. The subsample of single-habitation, 10th-Plan villages comprises 20 percent of Indian villages.[29] After restricting this 20 percent sample to our preferred RD bandwidth of 150 people above and below the 300-person threshold, we are left with 29,765 10th-Plan single-habitation villages from 22 states.[30] The left panel of Figure 1.4.3 displays a histogram of village populations, showing that the modal village lies within our RD window of 150–450 people. The right panel demonstrates how our two sample restrictions reduce the size of our RD sample, and shows that our running variable, 2001 village population, is smooth across the RD threshold.

[26] RGGVY ledgers we observed in Rajasthan were pre-printed with 2001 Census populations.

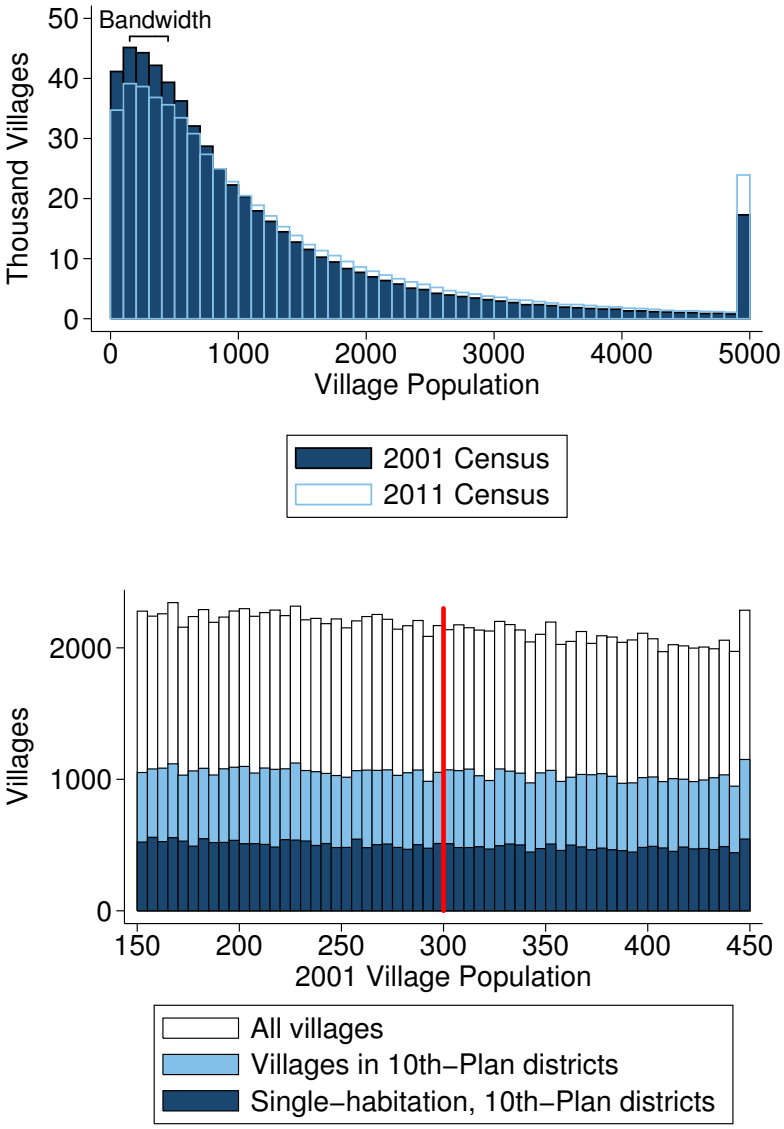
[27] Administered by the Ministry of Drinking Water and Sanitation, this census of habitations was collected in 2003 and 2009, and is available at `http://indiawater.gov.in`.

[28] We thank the authors for sharing their code. Appendix A.1.5 details our matching algorithm.

[29] 50 percent of villages are in districts eligible under RGGVY's 10th Plan, 86 percent of villages match to the habitation census, and 52 percent of matched villages in 10th-Plan districts have one habitation. Our analysis excludes villages that match to the habitation census but have populations that disagree by over 20 percent across datasets, as these matches are likely erroneous. In Appendix A.2, we show that including these villages slightly attenuates our RD point estimates as expected, yet they remain statistically significant.

[30] Three small states with 10th-Plan districts (Manipur, Kerala, and Tripura) are excluded from our final regression because they have no villages that meet these criteria.

Figure 1.4.3: Density of RD Running Variable



Note. — This figure summarizes the distribution of Indian village populations. The top panel shows the population distribution of villages in India in 2001 (solid navy) and 2011 (hollow blue). The bottom panel zooms in on the set of villages used in our RD analysis, within a 150–450 population window around the 300-person cutoff. Our RD sample of single-habitation 10th-Plan villages is shown in navy, relative to all Indian villages (white) and all villages in 10th-Plan districts (light blue).

Table 1.4.1 reports 2001 summary statistics for three sets of villages with populations between 150 and 450: all Indian villages, all villages in 10th-Plan districts, and all villages in 10th-Plan districts that have only one habitation. On average, villages in 10th-Plan districts are geographically smaller and less electrified than the national average, but similar across a range of other covariates. 10th-Plan villages with only one habitation are very similar on observables to average 10th-Plan villages.

Table 1.4.1: Summary Statistics – Villages with Populations Between 150 and 450

2001 Village Characteristics	All Districts	10th-Plan Districts	10th-Plan Districts Single-Habitation
Village area (hectares)	199.74	177.98	173.53
	(462.39)	(561.29)	(661.57)
Share of area irrigated	0.23	0.30	0.35
	(0.30)	(0.33)	(0.34)
Agricultural workers / all workers	0.39	0.37	0.37
	(0.16)	(0.16)	(0.15)
Other workers / all workers	0.06	0.06	0.06
	(0.08)	(0.08)	(0.08)
Employment rate	0.46	0.44	0.44
	(0.14)	(0.14)	(0.14)
Literacy rate	0.45	0.44	0.45
	(0.18)	(0.17)	(0.17)
Education facilities (0/1)	0.66	0.58	0.58
	(0.47)	(0.49)	(0.49)
Medical facilities (0/1)	0.13	0.12	0.12
	(0.34)	(0.32)	(0.32)
Banking facilities (0/1)	0.01	0.01	0.01
	(0.11)	(0.11)	(0.10)
Agricultural credit societies (0/1)	0.03	0.03	0.03
	(0.18)	(0.16)	(0.16)
Electric power (0/1)	0.68	0.62	0.64
	(0.46)	(0.49)	(0.48)
Share households with indoor water	0.21	0.21	0.25
	(0.17)	(0.17)	(0.19)
Share households with thatched roofs	0.27	0.27	0.28
	(0.27)	(0.24)	(0.24)
Share households with mud floors	0.78	0.79	0.77
	(0.17)	(0.16)	(0.17)
Average household size	5.36	5.53	5.56
	(0.58)	(0.61)	(0.60)
Number of villages	129,438	62,638	29,765

Note. — This table shows village-level summary statistics from the 2001 Census, for three sets of villages with 2001 populations between 150 and 450: all villages, villages in 10th-Plan districts, and single-habitation villages in 10th-Plan districts. This third group corresponds to the sample of villages used in our RD analysis. We present workers by sector as the share of total workers in the village; "other" workers are classified as non-agricultural, non-household workers. The employment rate divides the number of workers by village population. Binary variables are labeled (0/1). Standard deviations in parentheses.

1.4.3 Socioeconomic and Caste Census

We draw on individual-level microdata from the Socioeconomic and Caste Census (SECC) for measures of income and alternative employment data. The SECC was collected between 2011 and 2012, with the goal of enumerating the full population of India. We obtained a subset of these data from the Ministry of Petroleum and Natural Gas, whose liquid petroleum gas subsidy program, Pradhan Mantri Ujjwala Yojana, uses SECC data to determine eligibility.[31] As a result, we observe the universe of rural individuals that are eligible for this fuel subsidy program. This includes all individuals living in households that satisfied at least one of seven poverty indicators, and that did not meet any of fourteen affluence criteria.[32] This yields a dataset of data of 332 million individuals from 81 million households, representing roughly half of all households in rural India.

For this selected sample, we observe individual-level data on age, gender, employment, caste, and marital status; and household-level data on the housing stock, land ownership, asset ownership, and income sources. We use the SECC to test for the effects of RGGVY on wealth, using three main indicators. First, we test for the fraction of households with at least one poverty indicator (and no affluence indicators), as measured by the fraction of 2011 Census households that appear in our SECC dataset. Next, the SECC contains an indicator for whether the main income earner in each household earns at least 5,000 rupees per month.[33] This represents the highest-resolution measure of household income in a large-scale Indian dataset, enabling us to directly, albeit coarsely, test the effect of electrification on income. We also use SECC data to test for the effects of RGGVY on the fraction of households that own land or have at least one salaried laborer, two additional wealth indicators. Finally, we construct SECC employment variables that are analogous to the Census's village-wide measures, allowing us to test for distributional employment effects among the subset of households with poverty indicators.

[31] The Ministry of Rural Development, who collected the SECC, are in the process of making the full dataset publicly available. As of now, only district-level aggregates are posted at `http://secc.gov.in/welcome`. We downloaded our data in Excel format from `http://lpgdedupe.nic.in/secc/secc_data.html`.

[32] The sample also excludes the less than 1 percent of the population that met one of five destitution indicators. See Appendix A.1.6 for more details on the inclusion and exclusion criteria. We are missing data from six rural districts, which represent less than 1 percent of Indian villages.

[33] All households whose primary earner made over 10,000 rupees per month were ineligible for the fuel subsidy program, and are not included in our SECC dataset.

1.4.4 District Information System on Education

In order to estimate the effects of electrification on education, we include data on the universe of Indian primary and upper primary schools from the 2005–2006 school year through the 2014–2015 school year.[34] These data come from the District Information System on Education (DISE), which reports annual school-level snapshots on a variety of student, teacher, and school building characteristics. We collected these data at the school level and construct a 10-year panel dataset containing information from 1.68 million unique schools.[35] This panel is strongly unbalanced, and the average school appears in 7 out of a possible 10 years. Given that the reporting of school characteristics varies considerably across years, we focus our analysis on village-wide enrollment counts, which are consistently reported by gender and grade level. We test for effects of RGGVY on total enrollment, enrollment by gender, and enrollment by grade level, which allows us to measure how electrification impacted both the extensive and intensive margins of schooling.

1.5 Regression discontinuity results

1.5.1 Electrification

In order to demonstrate that RGGVY had a meaningful effect on electrification in eligible villages, we examine the effects of eligibility for RGGVY on nighttime brightness. Specifically, we use Equation (1.1) to estimate the effect of having a 2001 population above the RGGVY cutoff on village brightness in 2011. After removing states with low-quality or missing shapefiles, we are left with a sample of 18,686 single-habitation villages, in RGGVY 10th-Plan districts across 12 states, with populations in our RD bandwidth of 150–450 people.

Figure 1.5.4 presents the results from our preferred RD specification graphically, while Table 1.5.2 reports the corresponding numerical results. We find that 2011 nighttime brightness increased discontinuously at the 300-person threshold by 0.15 units of brightness. This jump is statistically significant at the 5 percent level, with

[34] While we use the full time series to match DISE schools to Census villages, we restrict our analysis to the 2010–11 school year, for consistency with our other outcome variables.

[35] We downloaded these data from `http://schoolreportcards.in/SRC-New/`. See Appendix A.1.7 for details.

a p-value of 0.015.[36] Appendix A.2.1 demonstrates that this is robust to a range of alternative bandwidths, functional forms, and specifications.

Though this point estimate might seem small, these results in fact demonstrate that RGGVY eligibility led to a substantial increase in brightness for barely-eligible villages as compared to barely-ineligible villages. To interpret these effects, we turn to the remote sensing literature. The magnitude of the effect we observe is consistent with ground-truthed estimates by Min et al. (2013), who find that electrification is associated with a 0.36-unit increase in nighttime brightness in rural villages in Senegal.[37] Our point estimate of 0.15 is on the same order of magnitude but smaller, which is to be expected, given that villages in our RD bandwidth are significantly smaller than the villages studied in Senegal. In a similar exercise, Min and Gaba (2014) find that a 1-unit increase in brightness corresponds to 60 public streetlights or 240–270 electrified homes in Vietnamese villages.[38]

Extrapolating these results to the Indian context, our estimated 0.15-unit increase translates to roughly 9 additional streetlights per village. This represents a substantial increase in nighttime luminosity, especially considering that RGGVY did not install streetlights. Alternatively, if we extrapolate the (weaker) household relationship to our setting, a 0.15-unit increase would translate to roughly 38 newly electrified homes, or 68 percent of households in the average village in our RD sample. These estimates from Senegal and Vietnam suggest that our effect size in India is consistent with a substantial increase in village electrification under RGGVY, especially given that many electricity end-uses that RGGVY sought to enable are not captured by the nighttime brightness proxy.[39]

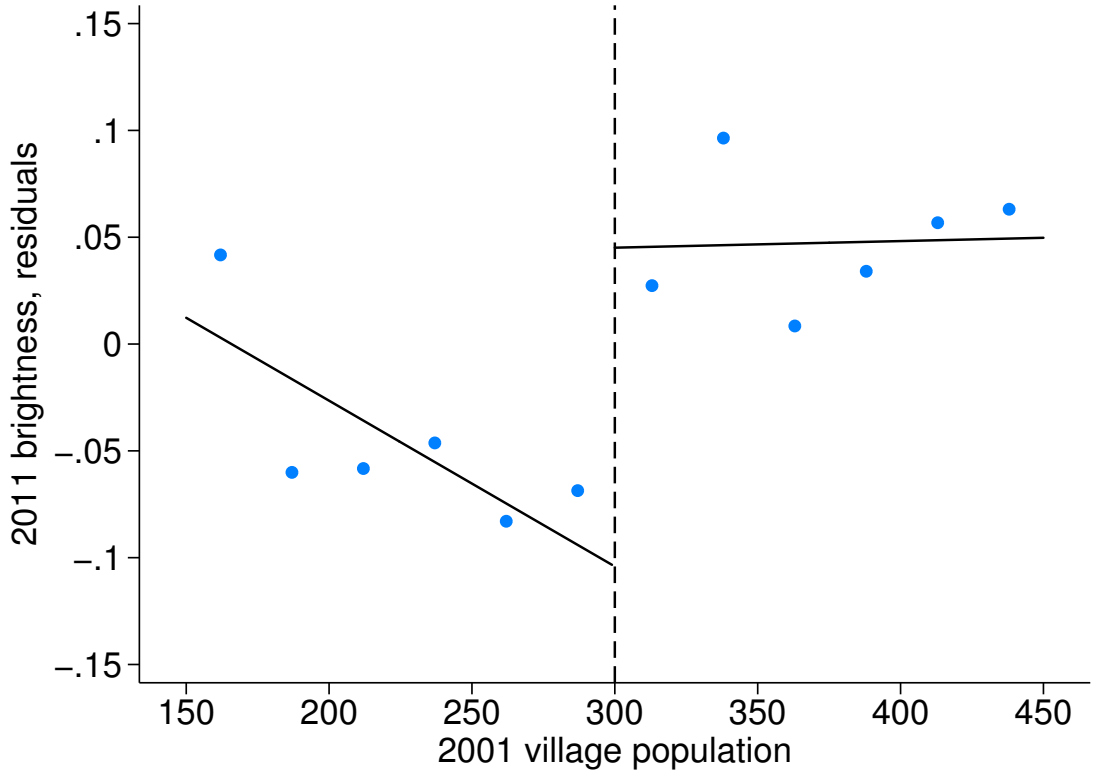
[36]These results include a control for 2001 nighttime brightness. Due to substantial cross-sectional heterogeneity, conditioning on the pre-period level dramatically improves the signal-to-noise ratio. This is common practice with remote sensing data (see also Jayachandran et al. (2016)). If we restrict the RD sample to include only villages that had electric power availability, according to the 2001 Census, we recover a nearly identical result ($\hat{\beta}_1 = 0.16$ with a p-value of 0.046). This suggests that the Census's 1/0 indicator variable for electric power availability masks substantial changes in electricity access under RGGVY, which we are able to detect using nighttime lights.

[37]This result uses the same average annual DMSP–OLS product that we use, unlike many of the other results reported in the paper, which rely on monthly composites that are not publicly available. We exclude the Mali results described in Min et al. (2013) because the authors exclude them from their main regression estimates.

[38]The relationship between nighttime brightness and streetlights is predictably stronger than the relationship between nighttime brightness and electrified homes.

[39]While many factors could cause the relationship between household electrification and nighttime brightness to differ between India and West Africa or Vietnam, Min et al. (2013) and Min

Figure 1.5.4: RD – 2011 Nighttime Brightness



Note. — This figure shows RD results using maximum 2011 nighttime brightness as a dependent variable, as reported in Table 1.5.2. Blue dots show average residuals from regressing the 2011 maximum nighttime brightness on 2001 maximum nighttime brightness and state fixed effects. Each dot contains approximately 1,600 villages, averaged in 25-person population bins. Lines are estimated separately on each side of the 300-person threshold, for 18,686 single-habitation villages between 150–450 people, in 10th-Plan districts. The point estimate on the level shift is 0.149, with a p-value of 0.015. Neither slope coefficient is significant at conventional levels.

We perform a series of validity tests in order to demonstrate that this increase in brightness is, in fact, attributable to the RGGVY program. First, we estimate Equation (1.1) using 2005 nighttime brightness as the dependent variable. Because RGGVY was announced in 2005 and nearly all project implementation began in subsequent years, we should not expect to find an immediate effect of program eligibility on brightness. The left panel of Figure 1.5.5 shows no visual evidence of a discontinuity in 2005 brightness at the 300-person threshold. The point estimate in this regression is 0.031, with a standard error of 0.020, and is not statistically significant at conventional level. This demonstrates that nighttime brightness was smooth at the 300-person cutoff prior to RGGVY.[40]

Next, we conduct a placebo test using 801 placebo RD "thresholds" between 151 and 1000.[41] For each threshold, we re-estimate Equation (1.1) and save $\hat{\beta}_1$. We plot the distribution of these placebo coefficients in the center panel of Figure 1.5.5. We also perform a randomization inference exercise, by scrambling the relationship between nighttime brightness and village population 10,000 times.[42] For each iteration, we estimate Equation (1.1), and the right panel of Figure 1.5.5 shows the resulting distribution of RD point estimates. The red lines indicate our estimate of $\hat{\beta}_1$, which falls above the 99th percentile of the placebo distribution and above the 98th percentile of the randomization distribution. This provides evidence that our RD estimates do not simply reflect spurious volatility in the relationship between nighttime lights and village population data.

We also perform a falsification exercise based on the implementation details of the RGGVY program. Our RD sample includes only villages that were eligible for RGGVY under the 10th Plan, for which the relevant eligibility cutoff was 300 people. It also includes only those villages confirmed to have exactly one habitation, for which 2001 village population is the appropriate running variable. We should not find effects at the 300-person cutoff on nighttime brightness for villages eligible under the 11th Plan, for which the relevant eligibility cutoff was moved from 300 to 100 people. Similarly, we should not find any RD effects for villages comprising multiple habitations, because these villages' populations do not correspond to the habitation

and Gaba (2014) provide evidence that the magnitude of our RD point estimate is consistent with what we might expect from a substantial increase in electricity access in these small villages.

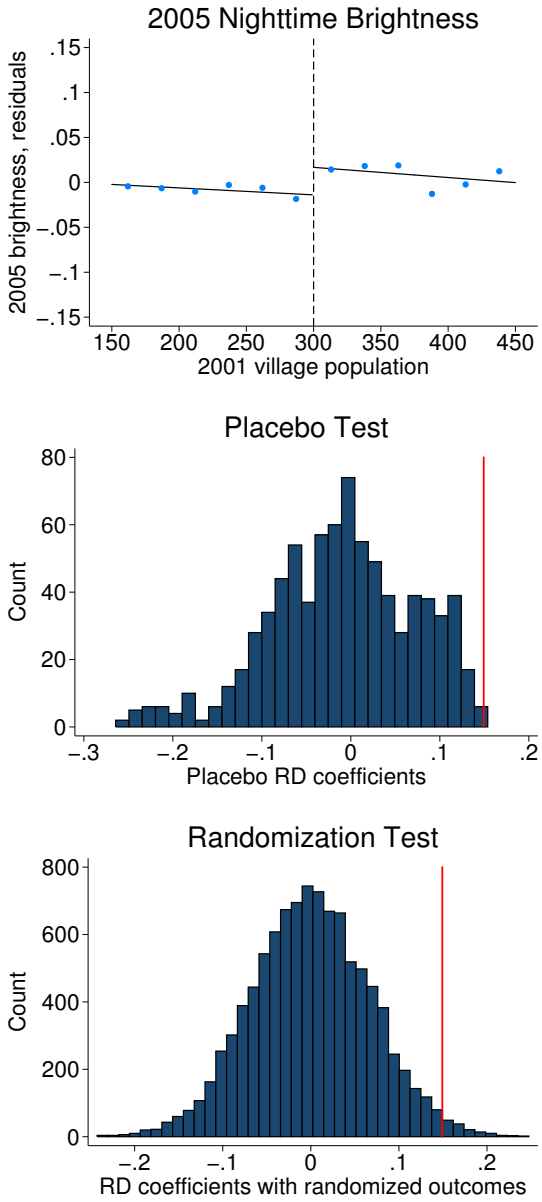
[40] We perform a variety of additional pre-period covariate smoothness checks in Appendix A.2.4.5, and find no evidence of discontinuities prior to RGGVY. Appendix A.2.3 demonstrates that the discontinuity in brightness steadily increases from 2006 onward.

[41] We test all 801 integer values in $[151, 275] \cup [325, 1000]$, which is asymptotically equivalent to simulating placebo draws across this discrete support. We omit thresholds between 275 and 325 to avoid possible contamination of the placebo results with the real threshold. We also avoid placebo thresholds below 151, to ensure positive values across the full 300-person RD window.

[42] We assign lights values to each village by sampling $\{Y_v^{2001}, Y_v^{2011}\}$ pairs without replacement.

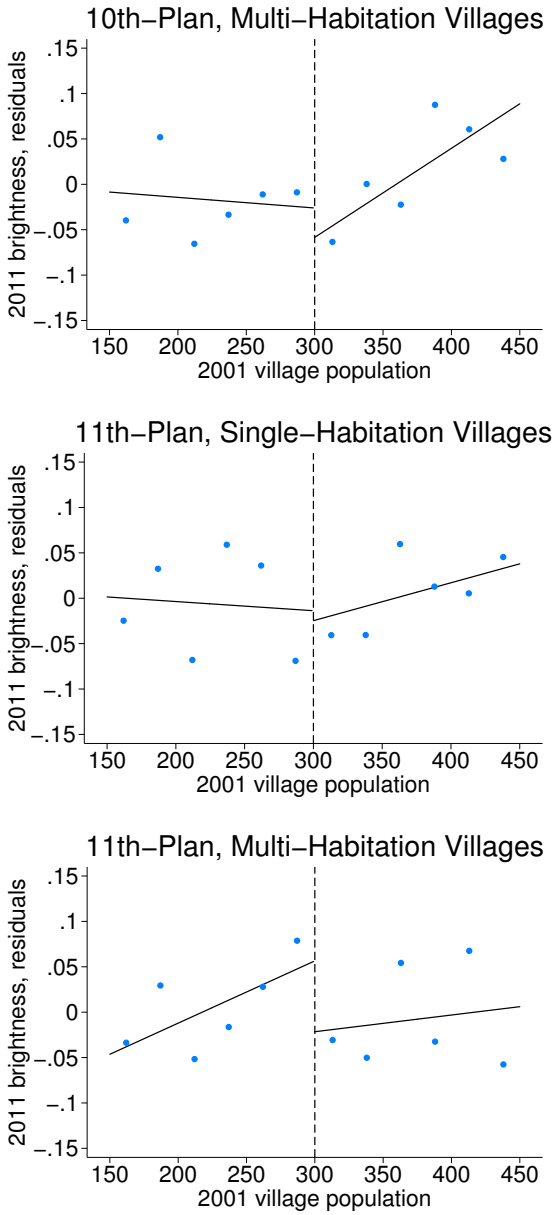
populations that determined RGGVY eligibility. Figure 1.5.6 presents RD results estimated using these alternative samples: as expected, none exhibits evidence of a discontinuity at the 300-person cutoff. This provides strong evidence that RGGVY, rather than spurious effects or other programs, is causing these effects.

Figure 1.5.5: Nighttime Brightness – Validity Tests



Note. — This figure presents results from three RD validity checks. The left panel displays results from estimating our main specification using 2005 brightness as the dependent variable; the point estimate is 0.031 with a standard error of 0.020. The center panel was generated by estimating Equation (1.1) on 801 placebo RD thresholds, representing all integer values in $[151, 275] \cup [325, 1000]$. We omit placebo thresholds within 25 people of the true 300-person threshold to ensure that placebo RDs do not detect the true effects of RGGVY eligibility, and we exclude thresholds below 151 due to our 150-person bandwidth. The right panel was generated by scrambling village brightness 10,000 times and re-estimating Equation (1.1). The red lines represent the RD coefficient from the actual data at the correct 300-person threshold. Our RD point estimate falls above the 99th percentile of the placebo distribution and above the 98th percentile of the randomization distribution.

Figure 1.5.6: Nighttime Brightness – Falsification Tests



Note. — This figure presents three falsification tests for our RD on nighttime brightness. The top and right panels include only villages with multiple habitations, for which the running variable of village population did not determine village eligibility. The center and bottom panels include only villages in districts that became eligible for RGGVY under the 11th Plan, for which the appropriate eligibility cutoff was lowered from 300 to 100 people. Blue dots show average residuals from regressing 2011 nighttime brightness on 2001 brightness and state fixed effects. Each dot contains approximately 900–1,600 villages, averaged in 25-person population bins. Lines are estimated separately on each side of the 300-person threshold, for villages within the 150–450 population bandwidth. Supplementary Table A.2.10 reports the regression results that correspond to these figures.

Table 1.5.2: RD – Nighttime Brightness

	2011 village brightness
$\mathbf{1}$[2001 pop $\geq$ 300]	0.1493**
	(0.0603)
2001 population	−0.0008
	(0.0007)
$\mathbf{1}$[2001 pop $\geq$ 300] $\times$ 2001 pop	0.0008
	(0.0008)
2001 Control	Yes
State FEs	Yes
RD bandwidth	150
Number of observations	18,686
Number of districts	130
Mean of dependent variable	6.370
R^2	0.766

Note. — This table shows results from estimating Equation (1.1), which corresponds to Figure 1.5.4. We define village brightness based on the brightest pixel contained within the village boundary. This regression includes all single-habitation villages in 10th-Plan districts with 2001 populations in the RD bandwidth (a 150-person bandwidth includes villages with 2001 populations between 150 and 450), for the 12 states with available village shapefiles that match to Census village areas with a correlation above 0.35. Standard errors are clustered at the district level. Significance: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

1.5.2 Economic outcomes

We now turn to the effects of RGGVY eligibility on village economies, and test for impacts of electrification via each of the potential channels discussed in Section 1.3.2. We estimate Equation (1.1) using outcome variables from six broad categories: employment, asset ownership, housing stock characteristics, village-wide outcomes, household income, and education. Each RD regression uses a dependent variable from 2011, while controlling for 2001 population as the running variable, state fixed effects, and the 2001 level of the dependent variable (unless otherwise noted).

First, we test for employment effects by estimating Equation (1.1) using the total number of male (female) workers in a given category divided by the total male (female) population of a village as the dependent variable.[43] Figure 1.5.7 summarizes these workforce results graphically for each gender and sector, and Panel B of Table 1.5.3 reports them numerically. We find that eligibility for RGGVY caused a 0.7 percentage point decrease in the share of men working in agriculture, on a mean of 42 percent. In contrast, the percentage of men in non-agricultural, non-household labor increased by 0.5 percentage points, on a mean of 10 percent. While these sectoral shifts are statistically significant and in a direction consistent with the structural transformation hypothesis, these effect sizes are very small: we can reject changes in male labor allocation larger than 1.3 percentage points. We find no statistically significant effects of electrification on the share of women working in any sector, and similarly narrow confidence intervals allow us to reject changes in female employment larger than 1.3 percentage points.[44]

We next test for effects of RGGVY eligibility on the share of households with a variety of different assets and housing stock characteristics. Figure 1.5.8 depicts RD results for the percent of households that own a telephone, own a television, own a motorcycle, have kerosene lighting, have mud floors, and are categorized as "dilapidated" by the Census. We see no strong graphical evidence of discontinuous changes in any of these dependent variables at the 300-person cutoff. Table 3 presents these results numerically in Panels C and D, while also reporting on the share of households with radios, bicycles, and without assets; the share of households cooking with electricity or gas; and the share of households with thatched roofs. Consistent with the graphical evidence in Figure 1.5.8, these results show that RGGVY did

[43] 2011 population does not change discontinuously at the 300-person threshold. See Panel A of Table 1.5.3, where we find that RGGVY caused no meaningful changes in village demographics.

[44] These results focus on the extensive margin of employment (i.e., number of workers). We also test for effects the intensive margin of employment in Appendix A.2.5 (i.e., share of workers working at least six months of the year). We find no evidence of statistically significant or economically meaningful changes on the intensive margin.

not lead to economically meaningful investments in electricity-using assets, non-electricity-using assets, or the housing stock in the medium term. We can reject increases larger than 1 percentage point in all cases. This suggests that RGGVY is unlikely to have contributed to significant increases in household expenditures. The program is also unlikely to have led to meaningful reductions in indoor air pollution, since we see no effects on the share of households with kerosene lighting or electric/gas cooking.

In Panel E of Table 1.5.3, we present RD results for village-level outcomes, including mobile phone coverage, the presence of agricultural credit societies, and the presence of irrigation tubewells, and the share of village area planted and irrigated.[45] These results are not statistically significant, and even the upper bounds on the 95 percent confidence intervals represent economically insignificant changes (smaller than 2 percentage points) in these outcomes. Taken together, these result imply that if RGGVY did lead to increases in agricultural productivity, farmers did not respond by increasing either the scale of irrigation or total farmland.

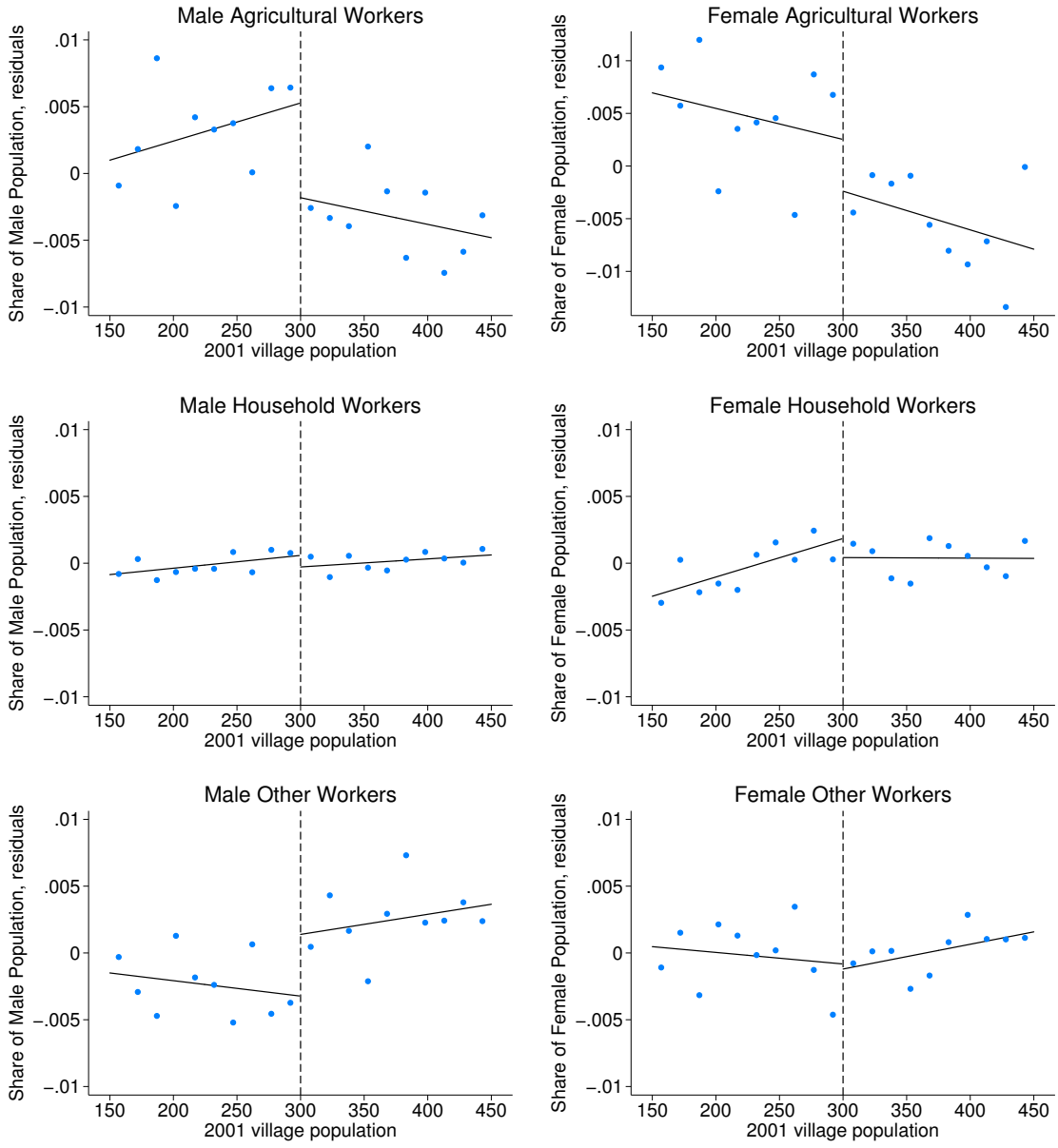
Next, we test for effects of RGGVY eligibility on economic outcomes among households with at least one poverty indicator. We estimate Equation (1.1) using the fraction of households with at least one poverty indicator (and zero affluence indicators) as the dependent variable. We also test for effects on the fraction of this subset of households for which the main income earner earns at least 5,000 rupees per month. The top row of Figure 1.5.9 presents these results graphically, revealing no evidence that RGGVY led to changes in these outcomes. Panel A of Table 1.5.4 reports the corresponding regression results, along with RD estimates for the fraction of households that report salaried employment and that own land. For each outcome, we can reject increases larger than 1.6 percentage points, at 95 percent confidence, suggesting that eligibility for RGGVY did not have economically meaningful effects on household poverty or wealth.

Using the SECC dataset, we also can test whether RGGVY eligibility had different employment impacts among individuals of lower socioeconomic status. We construct sector-specific labor shares that are analogous to Panel B of Table 1.5.3, except that they include only adults living in households with at least one poverty indicator.[46] We report these results graphically in the bottom row of Figure 1.5.9 and

[45] Tubewells are deep wells used for groundwater extraction, which are a common means of irrigation throughout rural India. Electric pumps improve the efficiency of tubewells.

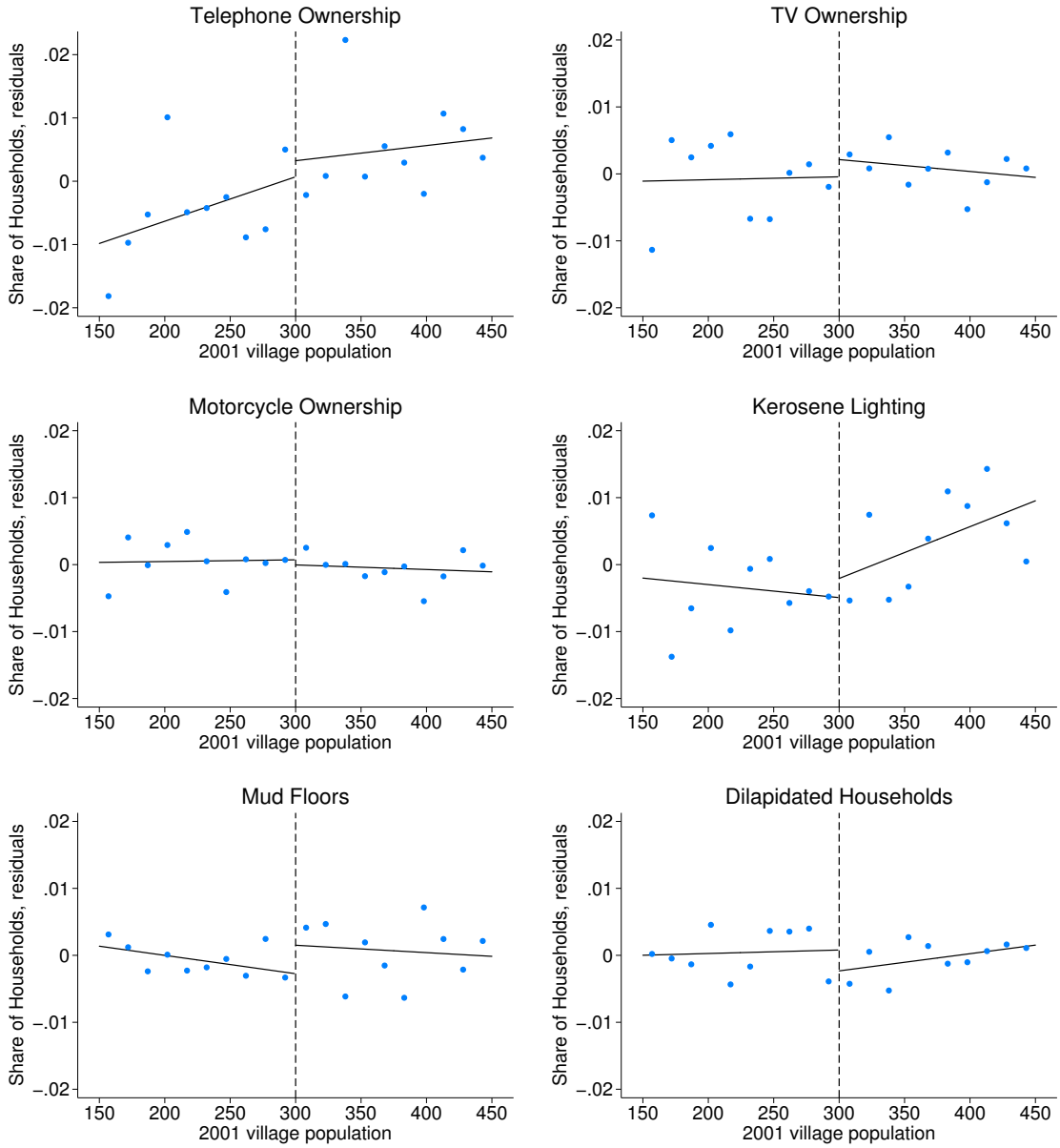
[46] See Appendix A.1.6 for further information on how we constructed these categories from the SECC data.

Figure 1.5.7: RD – Labor Outcomes



Note. — This figure shows the results from our preferred RD specification (Equation (1.1)), as reported numerically in Table 1.5.3. Blue dots show average residuals from regressing the 2011 percentage of the male/female population classified in each labor category on the corresponding 2001 percentage and state fixed effects. Each dot contains approximately 1,500 villages, averaged in 15-person population bins. Lines are estimated separately on each side of the 300-person threshold, for all 29,765 single-habitation villages between 150 and 450 people, in 10th-Plan districts.

Figure 1.5.8: RD – Housing and Asset Ownership



Note. — This figure shows the results from our preferred RD specification (Equation (1.1)), as reported numerically in Table 1.5.3. Blue dots show average residuals from regressing the 2011 percentage of households owning each asset (or with each characteristic) on the corresponding 2001 percentage and state fixed effects. Each dot contains approximately 1,500 villages, averaged in 15-person population bins. Lines are estimated separately on each side of the 300-person threshold, for all 29,765 single-habitation villages between 150 and 450 people, in 10th-Plan districts.

Table 1.5.3: RD – Census Outcomes

2011 Outcome Variable	RD Coefficient	Standard Error	95 Percent Confidence	Mean of Outcome
A. Demographic outcomes				
Total population	−0.8647	(2.528)	[−5.820, 4.091]	271.09
0–6 cohort / total population	0.0009	(0.001)	[−0.001, 0.002]	0.14
Average household size	−0.0051	(0.013)	[−0.030, 0.020]	5.13
Literacy rate	−0.0025	(0.002)	[−0.007, 0.002]	0.57
B. Labor outcomes				
Male agricultural workers / male pop	−0.0071**	(0.003)	[−0.013, −0.002]	0.42
Female agricultural workers / female pop	−0.0049	(0.004)	[−0.013, 0.003]	0.29
Male household workers / male pop	−0.0009	(0.001)	[−0.002, 0.000]	0.01
Female household workers / female pop	−0.0014	(0.001)	[−0.004, 0.001]	0.01
Male other workers / male pop	0.0046**	(0.002)	[0.001, 0.008]	0.10
Female other workers / female pop	−0.0004	(0.002)	[−0.004, 0.004]	0.05
C. Asset ownership				
Share of households with telephone	0.0025	(0.006)	[−0.008, 0.013]	0.54
Share of households with TV	0.0026	(0.004)	[−0.005, 0.010]	0.26
Share of households with bicycle	−0.0015	(0.004)	[−0.010, 0.007]	0.50
Share of households with motorcycle	−0.0008	(0.003)	[−0.006, 0.004]	0.13
Share of households without assets	0.0039	(0.004)	[−0.004, 0.012]	0.22
D. Housing stock				
Share of households with elec/gas cooking	0.0005	(0.003)	[−0.005, 0.006]	0.07
Share of households with kerosene lighting	0.0029	(0.006)	[−0.009, 0.015]	0.48
Share of households with mud floors	0.0043	(0.004)	[−0.003, 0.012]	0.73
Share of households with thatched roof	−0.0034	(0.005)	[−0.013, 0.007]	0.23
Share of households dilapidated	−0.0031	(0.003)	[−0.009, 0.002]	0.07
E. Village-wide outcomes				
1/0 Mobile phone coverage in village	−0.0008	(0.011)	[−0.023, 0.021]	0.75
1/0 Post office in village	0.0018	(0.004)	[−0.005, 0.009]	0.03
1/0 Ag credit societies in village	0.0013	(0.004)	[−0.006, 0.009]	0.02
1/0 Water from tubewell in village	−0.0023	(0.011)	[−0.024, 0.019]	0.44
Share of village area irrigated	−0.0057	(0.005)	[−0.016, 0.004]	0.35
Share of village area planted	0.0015	(0.006)	[−0.010, 0.013]	0.58

Note. — Each row represents a separate regression estimating Equation (1.1) on the outcome variable. The RD bandwidth includes 29,765 villages with 2001 populations between 150 and 450, across 225 districts. The second column shows the RD point estimate ($\hat{\beta}_1$) for each regression. All specifications control for the 2001 level of the outcome variable, except for share of village area planted (where 2001 values are not available) and 1/0 indicator variables. All specifications also include state fixed effects. Standard errors are clustered at the district level, which we use to calculate 95 percent confidence intervals in the fourth column. The fifth column reports the mean of the dependent variable for each RD regression. Significance: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

numerically in Panel B of Table 1.5.4. While the SECC sample differs notably from the village averages in the PCA, these results are broadly consistent with our main labor results, and visual evidence suggests a small decrease (increase) in agricultural (other) employment for adult men. We can reject 2 percentage point shifts across all six labor categories, which suggests that the average employment effects of RGGVY were similar to the effects on less wealthy households.

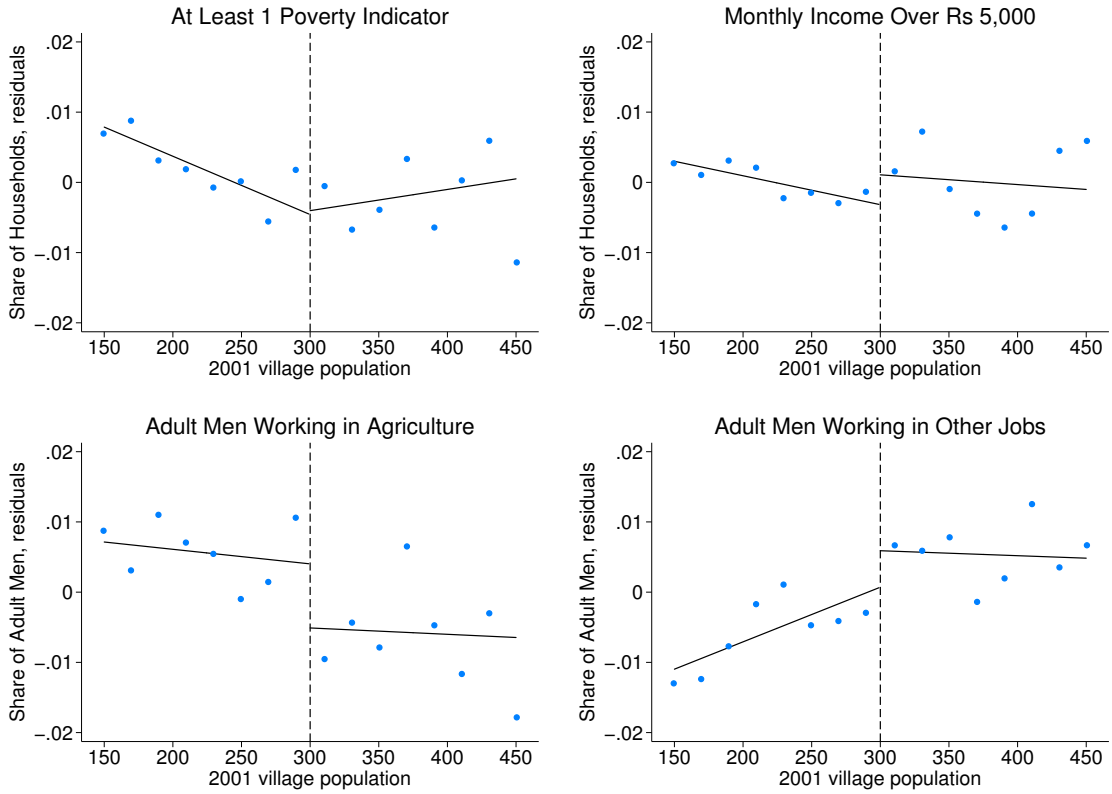
Table 1.5.4: RD – SECC Village-Level Outcomes

2011 Outcome	RD Coefficient	Standard Error	95 Percent Confidence	Mean of Outcome
A. Share of households				
At least one poverty indicator	0.0006	(0.006)	$[-0.011, 0.012]$	0.48
Monthly income greater than Rs 5,000	0.0043	(0.004)	$[-0.004, 0.013]$	0.08
One member holding salaried job	0.0030	(0.002)	$[-0.002, 0.008]$	0.02
Owning any land	−0.0005	(0.008)	$[-0.017, 0.016]$	0.44
B. Adult employment				
Male agricultural workers / adult men	−0.0091*	(0.005)	$[-0.019, 0.001]$	0.29
Female agricultural workers / adult women	−0.0039	(0.005)	$[-0.013, 0.006]$	0.08
Male household workers / adult men	0.0008	(0.001)	$[-0.002, 0.004]$	0.01
Female household workers / adult women	−0.0015	(0.008)	$[-0.016, 0.013]$	0.51
Male other workers / adult men	0.0052	(0.006)	$[-0.007, 0.017]$	0.42
Female other workers / adult women	0.0054	(0.005)	$[-0.005, 0.016]$	0.16

Note. — Each row represents a separate regression estimating Equation (1.1) on a different SECC village-level outcome. The first row of Panel A is coded as the share of *total* households in the village with at least one poverty indicator. Other outcomes in Panel A are coded as the proportion of this *subset* of households (with poverty indicators) that meet each criterion. Panel B outcomes are coded as the share of adult men (women) with an occupation in each subcategory, for the sample of adults in households with at least one poverty indicator. (We treat all invididuals over 16 years of age as adults.) The second column shows the RD point estimate ($\hat{\beta}_1$) for each regression. All specifications include state fixed effects, but they do not include any additional baseline control variables. The RD bandwidth includes 25,942 villages with 2001 populations between 150 and 450. These regressions contain fewer villages than regressions in Table 1.5.3 because only 87 percent of 10th-Plan, single-habitation, 150–450 villages match to the SECC dataset. Standard errors are clustered at the district level with 222 clusters, which we use to calculate 95 percent confidence intervals in the fourth column. The fifth column reports the mean of the dependent variable for each RD regression. Significance: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

Finally, we test for the effects of RGGVY eligibility on education. We estimate Equation (1.1) using village-wide enrollment for grades 1–8, both pooled and separately by gender, as the dependent variable. We also test for separate effects for

Figure 1.5.9: RD – SECC Village-Level Outcomes

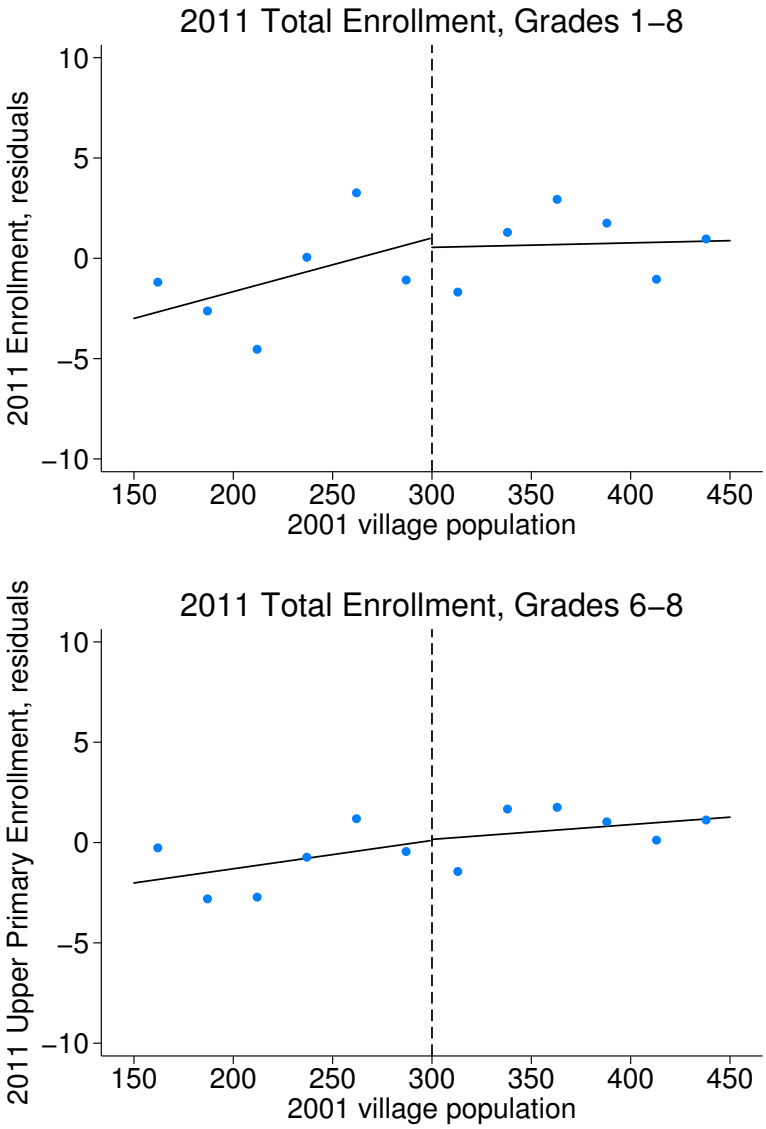


Note. — This figure shows the results from our preferred RD specification (Equation (1.1)), as reported numerically in Table 1.5.4. The upper-left panel reports the proportion of total village households with at least one poverty indicator in 2011, while the upper-right panel reports the proportion of households with a poverty indicator that had a maximum monthly income over Rs 5,000 in 2011. The lower panels report the share of adult men in households with a poverty indicator with occupations in each category. Blue dots show average residuals from regressing the 2011 share of households on state fixed effects. Each dot contains approximately 1,600 villages, averaged in 20-person population bins. Lines are estimated separately on each side of the 300-person threshold for 25,942 villages, i.e. all 10th-Plan single-habitation villages within our 150–450 population RD bandwidth, that match to the SECC dataset.

primary (grades 1–5) and upper primary (grades 6–8) enrollment, where the latter reflects changes on the intensive margin of schooling. We report these results in Figure 1.5.10 and Table 1.5.5, which show no statistically significant changes in enrollment at the 300-person threshold.[47] As with our other results, our 95 percent confidence intervals can reject even moderate changes in enrollment on either the intensive or extensive margins.

[47]These regressions control for the 2005 level of the outcome variable, which is the earliest year of enrollment data available. Also, we not that because these village-level enrollment regressions aggregate enrollment across all schools in each village, they might confound changes in within-school attendance with changes in enrollment due to new school construction over time. Appendix A.2.8 repeats these same regressions using school-level enrollment observations, while conducting additional sensitivity analysis.

Figure 1.5.10: RD – School Enrollment



Note. — This figure shows the results from our preferred RD specification (Equation (1.1)), as reported numerically in the first and last rows of Table 1.5.5. Blue dots show average residuals from regressing the 2011 number of (total, grades 6–8 only) students on the corresponding 2005 enrollment counts and state fixed effects. Each dot contains approximately 1,000 villages, averaged in 25-person population bins. Lines are estimated separately on each side of the 300-person threshold, for 12,251 single-habitation villages between 150 and 450 people, in 10th-Plan districts, with school-village matches and nonmissing 2005 and 2011 enrollment data.

Table 1.5.5: RD – School Enrollment

2011 Outcome Variable	RD Coefficient	Standard Error	95 Percent Confidence	Mean of Outcome
Total enrollment, grades 1–8	−0.472	(3.93)	[−8.18, 7.24]	74.05
Male enrollment, grades 1–8	0.197	(2.00)	[−3.72, 4.11]	37.60
Female enrollment, grades 1–8	−0.650	(2.02)	[−4.61, 3.31]	36.45
Total enrollment, grades 1–5	−0.408	(2.95)	[−6.19, 5.37]	60.58
Total enrollment, grades 6–8	0.051	(1.50)	[−2.89, 2.99]	13.47

Note. — Each row represents a separate regression estimating Equation (1.1) on a different enrollment count, aggregating schools enrollment up to village-level observations. The second column shows the RD point estimate ($\hat{\beta}_1$) for each regression. All specifications control for the 2005 level of the outcome variable and state fixed effects. The RD bandwidth includes 12,251 village observations with 2001 populations between 150 and 450, with a single habitation, in RGGVY 10th-Plan districts. These regressions contain fewer villages than regressions in Table 1.5.3 because only 51 percent of 10th-Plan, single-habitation, 150–450 person villages match to a school, and only 76 percent of these matched villages contain schools that report nonmissing enrollment values for 2011 and 2005. Standard errors are clustered at the district level, with 215 clusters, which we use to calculate 95 percent confidence intervals in the fourth column. The fifth column reports the mean of the dependent variable for each RD regression. Significance: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

Taking these results together, we conclude that while the provision and consumption of electricity substantially increased as a result of RGGVY eligibility, we detect no economically meaningful changes in labor outcomes, asset ownership, the housing stock, village-level outcomes, household income, or school attendance. Our RD results are precisely estimated, enabling us to rule out even modest effect sizes for these outcomes. This suggests that eligibility for RGGVY did not lead to structural transformation, increased agricultural productivity, female empowerment, reductions in indoor air pollution, improved education, or poverty reductions.

1.6 Interpretations and Extensions

1.6.1 Scaling

The above regressions recover intent-to-treat estimates: they show the effect of being *eligible* for RGGVY on our outcomes of interest. In order to compute average treatment effects, we need to scale these estimates such that we recover the effect of electrification on development.[48]

[48] We do not scale via two-stage least squares because we do not have access to a binary "RGGVY electrification" variable, nor would this variable capture different levels of energy access and consumption across villages treated under RGGVY, as discussed above.

We propose several methods of scaling our estimates. First, we consider inflating our outcomes based on the proportion of villages within our bandwidth that RGGVY claims to have treated. This is akin to the scale factor we would apply with a traditional instrumental variables estimator. RGGVY's district-level aggregate data suggest that between 56 and 82 percent of eligible villages were treated by the program.[49] This implies that our estimates should be inflated by approximately a factor of 1.5 in order to recover the causal effects of treatment under RGGVY.

Alternatively, we can calibrate a scaling factor to the magnitude of the increase in nighttime brightness, which we estimate to be 0.15 units of brightness. Min et al. (2013) suggest that when villages in Senegal were electrified, they experienced increases of approximately 0.4 nighttime brightness points. If, alternatively, we apply Min and Gaba (2014)'s estimates of a 1-unit increase in brightness corresponding to 240–270 electrified households, then full electrification of the average village in our RD sample with 56 households would imply an increase of 0.2 brightness points.[50] This suggests that our RD estimates should be inflated by a factor of between 1.3 and 3 to recover the average effect of RGGVY electrification.[51]

Scaling the point estimates reported in Tables 1.5.3–1.5.5 by a factor of 3 does not yield adjusted estimates that are economically meaningful. For the vast majority of outcomes, we see no visual evidence of a discontinuity, suggesting that these upper bounds are quite conservative. Even after inflating the 95 percent confidence intervals by these factors, we can still reject 4 percentage point changes in labor outcomes, 4 percentage point changes in asset ownership, 5 percentage point changes in the housing stock, and 8 percentage point changes in village-level outcomes. We can also reject 6 percentage point changes in outcomes in Table 1.5.4, as well as 21 student (30 percent) increases in total school enrollment. Scaling by a factor of 3, we can rule out effects larger than 0.26 of one standard deviation in all outcomes presented in Tables 1.5.3–1.5.5.

Even if we were to scale our estimates by an extremely conservative factor of 10, we can still reject effect sizes consistent with the previous literature.[52] Dinkelman

[49]RGGVY's aggregate village counts in 10th-Plan districts sum to 56 percent of the total number of villages in these districts, and 82 percent of villages with 2001 populations over 300.

[50]These increases of 0.4 and 0.2 are internally consistent; the average villages in Min et al. (2013) and Min and Gaba (2014) are larger than the villages in our RD bandwidth.

[51]We do not propose a scale factor based on Min and Gaba (2014)'s streetlights estimate, since we do not have data on the number of streetlights per village, and because RGGVY did not install streetlights.

[52]In order to arrive at factor of 10, which we believe to be the most conservative interpretation of our results, we assume that RGGVY *only* impacted household electricity end-uses. Our nighttime brightness effect of 0.15 is comparable to the change in brightness associated with a 10 percentage point increase in the share of households with electric lighting, a proxy for household power con-

(2011) finds that electrification caused 9–9.5 percentage point increases in female employment; we can reject 2 percentage point increases in total female employment.[53] Lipscomb, Mobarak, and Barham (2013) likewise find large effects of electrification on total employment rates; we can reject 1 percentage point increases in the village-wide employment rate even after applying a conservative scaling factor of 10.[54] Chakravorty, Emerick, and Ravago (2016) find that rural electrification leads to a 56 percent decrease in a deprivation index, and a 38 percent increase in household expenditures; scaling our Table 1.5.4 results by 10, we can reject an 11 percentage point decrease in the share of households with at least one poverty indicator, and a 13 percentage point increase in the share of households (with at least one poverty indicator) with monthly incomes greater than 5,000 rupees.

1.6.2 Heterogeneous effects

It is possible that our results mask heterogeneity in the quality of energy services experienced by RGGVY villages. In particular, India faces major electricity shortages, which vary across locations (Allcott, Collard-Wexler, and O'Connell (2016)). If half of the villages in our sample experienced frequent power outages while the other half received consistent power, our average intent-to-treat estimate across both groups would be small even if RGGVY led to large economic effects in places with high-quality energy supply. We test for this by re-estimating all of our RD results using the subset of states with above-average power availability (Central Electricity Authority (2011)).[55] In this subsample, our estimated RD coefficient on nighttime brightness increases from 0.15 to 0.25, statistically significant at the 1 percent level. However, the results for labor, asset ownership, the housing stock, village-level outcomes, and household wealth are quantitatively similar to those estimated using the full RD sample.[56]

sumption, at the mean of our RD sample. This suggests a scaling factor of 10 to translate this into an increase from 0 to 100 percent of households.

[53]We estimate Equation (1) pooling female employment across all three sectors, resulting in an RD point estimate of -0.0067 with the upper end of our 95 percent confidence interval of 0.0015, which we multiply by 10. We can similarly reject increases of 3 percentage points in female agricultural employment, 1 percentage point in female household employment, and 4 percentage points in female other employment.

[54]If we pool all six labor outcomes in Panel B of Table 1.5.3, the resulting RD point estimate is -0.0053 with an upper 95 percent confidence interval of 0.0002.

[55]These seven states are (in decreasing order of 2011 power quality): Chhattisgarh, Orissa, Karnataka, West Bengal, Gujarat, Haryana, and Rajasthan.

[56]Appendix A.2.10 reports regression results for both split-sample exercises discussed in this section. The schooling results are qualitatively similar, but somewhat less robust.

This suggests that poor power quality in a subset of states is not attenuating our estimate of the average effect across the full sample. Moreover, our main RD results reflect the realized implementation of a large-scale national rural electrification program in the developing world. Even if we had found substantial positive effects for a subset of states, the overall treatment effect would be indicative of the degree to which future rural electrification programs might be limited by the supply reliability.

It is also possible that we do not detect large effects because the benefits of electrification take many years to accrue. While we cannot rule this possibility out completely, our 2011 outcome data were collected between three and five years after 95 percent of villages in our sample received RGGVY funding.[57] Even if there were significant delays in implementation, this is much longer than the time span over which development interventions are typically studied. Nevertheless, we recover quantitatively similar RD point estimates when we restrict our RD sample to districts with early RGGVY funding. Therefore, it is unlikely that our small results are driven by villages that failed to take advantage of the full set of possible medium-run benefits of electric power before being surveyed by the 2011 Census.

1.6.3 Difference-in-differences

Finally, we might be concerned that villages close to the 300-person RD threshold stand little to gain from electrification. Perhaps these small villages are simply too poor, too credit-constrained, or too economically isolated to translate increased electricity access into new employment or income-generating opportunities. We employ a second identification strategy, difference-in-differences (DD), to test for the effects of RGGVY eligibility on larger villages far from our RD threshold. Recall that there were two major phases of RGGVY implementation: the 10th-Plan phase and the 11th-Plan phase. The majority of 11th-Plan electrification projects had not been completed before the 2011 Census. We can therefore use 10th-Plan districts as a "treated" group and 11th-Plan districts as a "control" group in a DD framework.[58] We estimate the following fixed effects specification on our two-decade village panel:

$$Y_{vst} = \gamma_0 + \sum_b \gamma_1^b \mathbf{1}\left[\text{10th} \times \text{Post}\right]_{vt} \times \mathbf{1}\left[P_v \in \text{Bin}_b\right] + \delta_t + \eta_v + \varepsilon_{vt} \tag{1.2}$$

where $\mathbf{1}\left[\text{10th} \times \text{Post}\right]_{vt}$ is an indicator equal to one if village v was eligible for RGGVY under the 10th Plan and the year t is 2011, $\mathbf{1}\left[P_v \in \text{Bin}_b\right]$ are 2001 vil-

[57] Over 70 percent of villages in our RD sample are in districts that received RGGVY funding before the end of 2006. See Appendix Table A.2.26.

[58] Selection into the different plans was non-random. It is plausible that 10th-Plan districts were more administratively capable than 11th-Plan districts, likely biasing our DD estimates upward.

lage population bins along the full support of populations (shown in Figure 1.4.3), δ_t are year fixed effects, and η_v are village fixed effects. This necessitates stronger identifying assumptions than our RD specification, namely that villages in 10th-Plan districts were trending in parallel to 11th-Plan villages prior to RGGVY. Village-level data are not available for the 1991 Census, therefore we are unable to directly test this assumption.[59]

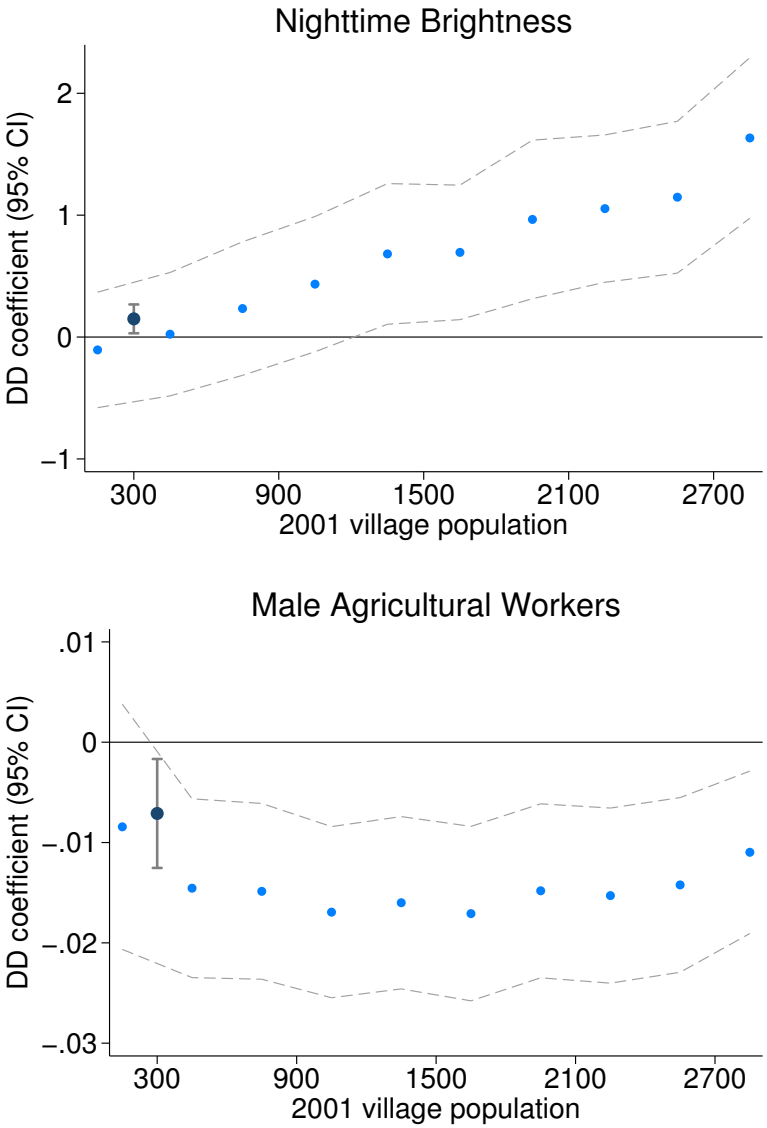
Figure 1.6.11 compares our main RD results with DD results from estimating Equation (1.2) with 300-person population bins, for nighttime brightness and male agricultural workers. For both outcomes, the RD point estimates lie within the DD confidence intervals. Moreover, the DD effect of RGGVY on nighttime lights increases nearly monotonically in population, while the DD effect for male agricultural labor is close to constant as population increases. This suggests that our small RD results for male agricultural employment are likely to be externally valid outside of our RD bandwidth. Other economic outcome variables show similarly constant DD coefficients across small and large villages.[60]

These DD results are broadly consistent with our RD results, despite using a much larger population of villages (from 10th- and 11th-Plan districts, including multi-habitation villages) and using 11th-Plan villages as counterfactuals (as opposed to barely ineligible 10th-Plan villages). Beyond allowing us to extend our RD results to larger villages, the DD results are encouragingly similar to the RD. Relying on alternative identifying assumptions on a different sample of villages, we again demonstrate that RGGVY caused nighttime brightness to increase, but has not meaningfully improved the economic outcomes that we observe.

[59] In Appendix A.2.11, we test for differential pre-trends using district-level data. These trends are not statistically zero, suggesting that our DD results should be interpreted with some caution.

[60] We report additional DD results in Appendix A.2.11.

Figure 1.6.11: Difference-in-Differences Results



Note. — This figure compares the reduced form effects from our preferred RD specification (Equation (1.1)) to the results from our DD specification (Equation (1.2)), using 300-person population bins. Navy blue dots show the RD coefficients, with whiskers indicating 95 percent confidence intervals. Light blue dots and dashed lines show the binned DD point estimates and 95 percent confidence intervals. The top panel shows the effects for nighttime lights, as measured by maximum village brightness. The bottom panel shows the effects for male agricultural workers. The RD results are statistically significant at the 5 percent level and the 1 percent level, respectively. The pooled DD point estimates are 0.45 and −0.008; both are statistically significant at the 10 percent level (Appendix A.2.11 reports these results in a regression table). DD regressions for lights and labor include 629,778 and 994,802 village-year observations, respectively.

1.6.4 Costs and benefits

We do not have direct estimates of village-level program costs, incomes, or expenditures.[61] However, we perform several back-of-the-envelope calculations based on our RD results. This enables us to better understand the overall economics of RGGVY, while also quantifying the costs and benefits of electrification.

First, we consider the per-village costs of RGGVY implementation. In 2005, RGGVY was expected to cost 634.2 billion rupees, or approximately $17.2 billion.[62] Given the stated scope of the program detailed in Section 1.2, this suggests a cost per village of approximately 1,470,000 rupees, or $36,000 in 2015 USD.[63]

We can apply average Indian rural wage rates to estimate the income differential that might have resulted from the (small) sectoral shift from agricultural to non-agricultural employment we observe under RGGVY. According to India's National Sample Survey Office, the average 2011 wage for male (female) non-agricultural workers was 196 (116) rupees per day, which was 26 (0.9) percent higher than the average agricultural wage of 155 (115) rupees per day. To compute the average increase in village-level income, we scale the lower bound of our confidence interval for male (female) agricultural labor from Table 1.5.3, -0.013 (-0.013), by a factor of 3. This converts our intent-to-treat estimate into an average treatment effect, and it implies a maximum shift out of male and female agricultural employment of 3.9 percentage points.[64] If all of these men (women) shifted from agriculture into non-agriculture employment, then total daily male (female) village wage earnings would have increased by approximately 293 (7) rupees. If each employed person worked 365 days per year, this would translate into a total annual village income increase of approximately 109,000 rupees, or an upper bound of 1.4 percent.

Alternatively, we can use our RD estimates on household asset ownership to infer changes in expenditures resulting from electrification. Scaling the upper confidence intervals in Panel C of Table 1.5.3 by 3, we can reject increases in asset ownership of greater than 3.9 percent for mobile phones, 3.0 percent for televisions, 2.1 percent for bicycles, and 1.2 percent for motorcycles. Monetizing these upper bounds using asset prices from ICRISAT's Village Dynamics in South Asia dataset, this implies a

[61] The SECC income data indicate whether households' main income earners earned more or less than 5,000 rupees per month, and comes from a selected subset of households. Hence, we exclude these data from the subsequent cost-benefit analysis.

[62] We use the 2005 exchange rate of 44 rupees per dollar, and convert to 2015 USD.

[63] This is comparable to Chakravorty, Emerick, and Ravago (2016), who report average electrification costs of $42,000 per village in the Philippines.

[64] In keeping with Section 1.6.1, we apply a scaling factor of 3 throughout this section.

maximum average household expenditure of 572 rupees.[65] Supposing that only 10 percent of RGGVY-driven expenditure increases were spent on these four durable goods implies a maximum increase in per-household expenditure of 5,720 rupees, or a total village-wide increase of around 398,000 rupees. These asset purchases occurred during the 3–6 year period after electrification; if we conservatively assume that they all occurred within 3 years of electrification, this would represent at best a 2.1 percent increase in annual village expenditures.[66]

Our back-of-the-envelope estimates suggest that annual village income increased by a maximum of 109,000 rupees, that annual village expenditures increased by a maximum of 133,000 rupees, and that RGGVY electrification came at a cost of approximately 1,470,000 rupees per village. These results are quite conservative: though we do not measure all possible benefits from electrification, the benefits we do use in performing this calculation come almost entirely from regression estimates where we cannot reject zero; our assumptions in performing this calculation also make it biased towards finding large effects. Using the larger expenditure estimate and applying a conservative 3 percent discount rate, this translates into a payback period of approximately 12 years.[67]

At best, we find that RGGVY increased annual incomes by 1.4 percent and annual expenditure by 2.1 percent, despite causing a substantial shift in nighttime lights. This suggests exercising caution when using nighttime brightness as a proxy for income or expenditures. The DMSP-OLS dataset measures light emissions. Because brightness relates directly to energy consumption through lighting, it serves as a useful indicator of electrification. Since electrification should lead to increased brightness even absent a corresponding increase in incomes, we do not use the DMSP-OLS data as a proxy for income/expenditures, and caution others against doing so when evaluating programs that directly increase light emissions.

Importantly, our results do not speak directly to the effects of RGGVY on welfare. It is quite possible that electrification has dramatically increased average quality of life for rural Indians. Indeed, since villagers are using more power as a result of RGGVY, revealed preference suggests that they benefit from the program. Even though we measure a wide range of outcome variables which are typical of large-scale administrative datasets, there may be important utility benefits that we cannot mea-

[65] The average prices for durables commonly purchased after electrification are: Rs 2,796 for cell phones, Rs 4,166 for televisions, Rs 1,259 for bicycles, and Rs 25,922 for motorcycles.

[66] India's average rural monthly per capita expenditures were 1,430 rupees for 2011–2012.

[67] This starkly contrasts with Chakravorty, Emerick, and Ravago (2016), who find a payback period of approximately 1 year; however, it is corroborated by evidence from Lee, Miguel, and Wolfram (2016), who use revealed preference results to suggest that the costs of electrification are much larger than the benefits.

sure. Our results highlight the need to incorporate additional non-market measures into future administrative data collection efforts.

1.7 Conclusion

In this paper, we evaluate the medium-run effects of electrification on development using a regression discontinuity (RD) design which exploits a population eligibility threshold in India's national rural electrification program, RGGVY. We find that eligibility for RGGVY led to substantial changes in nighttime brightness and power availability. Despite this increase in energy access, we find that electrification did not have economically meaningful impacts on a range of development outcomes.

These results hold when we rescale our reduced form estimates to account for the proportion of eligible villages that underwent treatment. We see similar effects on development among states with high and low average reliability of electricity supply. We also find similar effects when we restrict our analysis to the earliest districts to obtain RGGVY funding, suggesting that our results do not depend on the timing of our post-intervention data. Finally, we apply a difference-in-differences strategy, which relies on alternative identifying assumptions and includes a larger sample of villages well outside our RD bandwidth. These results support the main conclusions from our RD analysis that while nighttime lights, and therefore power consumption, increased substantially with RGGVY electrification, other development outcomes that we observe did not. Our cost-benefit calculations suggests a much longer payback period than previously estimated.

These results are the first to suggest that electrifying rural villages may not cause sizable economic gains in the medium term. Our regression discontinuity strategy relies on much less stringent identifying assumptions than the instrumental variables approaches of previous work, allowing us to measure effects of a natural rollout of rural electrification, at scale. In contrast to the existing literature, we find that electrification did not yield even modest changes in labor, income, household wealth, asset ownership and expenditures, village-level outcomes, and education. These null results come from the world's largest unelectrified population, and appear to generalize to over 400,000 villages across rural India.

Nevertheless, electrification may lead to large economic benefits in certain contexts, and may have important positive effects on human well-being that we are unable to quantify. An important direction for future work will be to understand when, where, and after how long electricity access and power availability have the greatest economic impact. For example, electrification may lead to substantial gains in economic productivity in urban settings, or in regions with budding local indus-

tries. There may also be substantial long-run effects of electrification, and more research is necessary to identify these benefits. Finally, we encourage future research on quantifying the non-market benefits from electrification that frequently go unmeasured.

Chapter 2

Machine Learning from Schools About Energy Efficiency[1]

2.1 Introduction

Energy efficiency is a cornerstone of global greenhouse gas (GHG) abatement efforts. For example, worldwide proposed climate mitigation plans rely on energy efficiency to deliver 42 percent of emissions reductions (IEA (2015)). The appeal of energy efficiency investments is straightforward: they may pay for themselves by lowering future energy bills. At the same time, lower energy consumption reduces reliance on fossil fuel energy sources, providing the desired GHG reductions. A number of public policies – including efficiency standards, utility-sponsored rebate programs, and information provision requirements – aim to encourage more investment in energy efficiency.

Policymakers are likely drawn to energy efficiency because a number of analyses point to substantial unexploited opportunities for cost-effective investments (see, e.g., McKinsey & Company (2009)). These analyses are almost universally based on engineering estimates of the potential energy savings over time rather than field evidence of actual savings. One strand of the economics literature has attempted to explain why consumers might fail to avail themselves of profitable investment opportunities (see, e.g., Allcott and Greenstone (2012), Gillingham and Palmer (2014), and Gerarden, Newell, and Stavins (2015)). The most popular explanations have emphasized the possibility of market failures, such as imperfect information, capital market failures, split incentive problems, and behavioral biases, including myopia,

[1]The material in this chapter is from an unpublished working paper, coauthored with Christopher R. Knittel, David Rapson, Mar Reguant, and Catherine Wolfram.

inattentiveness, prospect theory and reference-point phenomena. The literature also points to the possibility that engineering estimates understate the costs consumers face.

Economists have also pointed out that accurately measuring the returns to energy efficiency investments is difficult as it requires constructing a counterfactual energy consumption path from which reductions caused by the efficiency investments can be measured (Joskow and Marron (1992)). Recent studies use both experimental (e.g., Fowlie, Greenstone, and Wolfram (2015a)) and quasiexperimental (e.g., Levinson (2016), Myers (2014), Davis, Fuchs, and Gertler (2014)) approaches to developing this counterfactual. Many of these papers suggest that the energy savings from efficiency investments are lower than expected, and the results cast doubt on the extent to which emissions reductions can be achieved through energy efficiency. At the same time, energy efficiency proponents have criticized these findings as only pertaining to specific contexts (Kushler (2015)), and are skeptical that experimental and quasi-experimental estimation approaches can be applied broadly. In general, there is a clear need for techniques to estimate returns to energy efficiency programs that can be applied in a wide set of contexts.

The proliferation of high-frequency data in electricity markets provides a promising opportunity to estimate treatment effects associated with energy efficiency investments wherever advanced metering infrastructure (AMI, or "smart metering") is installed. In this paper, we use high-frequency data to implement several approaches to estimating counterfactual energy consumption absent the energy efficiency investments. We begin with difference-in-difference approaches that employ rich sets of fixed effects. We show, however, that the resulting estimates are sensitive to the set of observations included as controls as well as to the fixed effects included in the specification. To address these concerns, we develop and implement a novel machine-learning approach to predict counterfactual energy consumption at treated schools and validate the approach with non-treated schools.

Specifically, we match hourly electricity consumption data from public K-12 schools in California to energy efficiency installation records, and exploit temporal and cross-sectional variation to estimate the causal effect of the energy efficiency investments on energy use. Our data span 2008 to 2014, although only 20 percent of schools had smart meter data in 2008 and half have entered our data set by 2011. Our methodology includes traditional regression-based approaches, which allow us to non-parametrically control for a wide array of potential confounders. We implement a novel machine learning approach, using model selection and forecasting methods to construct school-specific counterfactual electricity usage using only pre-treatment data. We compare the results of this approach to the traditional

difference-in-difference methods, and find, consistent with Monte Carlo simulations, that our machine learning approach yields slightly larger treatment effects.

Our contributions to the literature are both policy-relevant and methodological. From a policy perspective, this paper departs from much of the previous academic literature on energy efficiency by examining energy efficiency outside the residential sector. 37 percent of electricity use in the United States in 2014 was residential, and over half is attributable to commercial and industrial uses (EIA (2015)). A more complete view of what energy efficiency opportunities are cost-effective requires more evidence from a variety of settings, which, in turn, requires an informed understanding of the costs and benefits of investment in settings that have traditionally been difficult to study.

Our results demonstrate that energy efficiency investments can lead to substantial energy savings in schools. Across all types of investments, energy efficiency appears to deliver between 2 and 5 percent reductions in electricity use. We also look at the two most prevalent upgrade categories in our sample, lighting, which makes up 22 percent of upgrades; and heating, ventilation, and air conditioning (HVAC), which makes up 51 percent of upgrades. During daytime hours, investments in energy efficient lighting lead to 3 to 7 percent reductions in electricity use and smaller reductions at night. Investments in energy efficient HVAC systems produce a 2-4 percent reduction in the daytime hours, when temperatures are highest. These results translate into a significant amount of overall energy savings, representing about 60 kWh in daily electricity savings per school, which we estimate using both regression and machine learning techniques. When we compare these ex post effects to ex ante savings, however, we find that these ex-post saving estimates appear to deliver less than 50 percent of expected ex-ante savings.

From a methodological perspective, high frequency data provides large benefits, but also presents new challenges. Using 15-minute interval electricity consumption data allows us to incorporate a rich set of controls and fixed effects in order to non-parametrically separate the causal effect of energy efficiency upgrades from other confounding factors. However, over-saturation is a concern; fixed effects estimators that absorb too much identifying variation can spuriously detect "treatment effects" that are simply artifacts of measurement problems in the data (Fisher et al. (2012)). The machine learning method that we develop in this paper uses LASSO, a form of regularized regression, to generate a model of counterfactual electricity consumption that leverages our high-frequency data while avoiding overfitting. This allows us to optimally saturate the regression (trading off precision with bias). Furthermore, we only use pre-treatment data to train the LASSO model, meaning that the school-specific prediction models we generate do not risk absorbing part of the treatment, which is not present in the data used to build the model.

To our knowledge, this is the first paper in economics to incorporate machine learning methods into a selection-on-unobservables design in order to conduct causal inference.[2] We validate the machine learning predictions at our control schools, finding tightly-estimated zero effects for non-treated schools. We further embed our machine learning predictions into a difference-in-difference framework to account for the possibility of systematic trends in the prediction errors. Using a series of Monte Carlo simulations, we demonstrate that the machine learning approach performs as expected relatively to the regressions. We believe our methodology may be of interest to empiricists in a broad class of settings.

The remainder of the study proceeds by describing our empirical setting and datasets (Section 2.2). We then describe the baseline difference-in-differences methodology and estimate results using these standard tools (Section 2.3). Section 2.4 introduces our machine learning methodology, and presents the results. In Section 2.5, we compare our savings estimates to the ex ante engineering projections to calculate realization rates. Section 2.6 concludes.

2.2 Context and Data

Existing ex ante estimates suggest that commercial buildings, including schools, may present important opportunities to increase energy efficiency. For example, McKinsey & Company, who developed the iconic global abatement cost curve (see McKinsey & Company (2009)), note that buildings account for 18 percent of global emissions and as much as 30 percent in many developed countries. In turn, commercial buildings account for 32 percent of building emissions, with residential buildings making up the balance. Opportunities to improve commercial building efficiency primarily revolve around lighting, office equipment, and HVAC systems.

Commercial buildings such as schools, which are not operated by profit-maximizing agents, may be less likely to take advantage of cost-effective investments in energy efficiency, so they may yield particularly high returns to targeted programs that encourage energy efficiency investments. On the other hand, schools are open fewer hours than many commercial buildings, so the returns may be lower. Energy efficiency retrofits for schools gained prominence in California with Proposition 39, which voters passed in November 2012. The proposition closed a corporate tax loophole and devoted half of the revenues to reducing the amount public schools spend on energy, largely through energy-efficiency retrofits. Over the first three fiscal years of the program, the California legislature appropriated $1 billion to the program (CEC

[2]In a recent NBER working paper, Cicala (2017) implements a variant on this methodology, using random forests rather than LASSO, in the context of electricity market integration.

(2017)). To put this in perspective, it represents about one-third of what California is currently spending on *all* utility-funded energy efficiency programs (ranging from low-interest financing to light bulb subsidies to complex industrial programs) and about 5 percent of what utilities nationwide spend on energy efficiency (Barbose et al. (2013)). Though our time period precedes most investments financed through Proposition 39, our results are relevant to expected energy savings from this large public program.

Methodologically, schools provide a convenient laboratory to isolate the impacts of energy efficiency. School buildings are all engaged in nearly the same activities, subject to relatively similar trends in education, and are clustered within distinct neighborhoods and towns. Other commercial buildings, by contrast, can house anything from an energy intensive data center that operates around the clock to a church that operates very few hours per week. Finally, given the public nature of schools, we are able to assemble relatively detailed data on building characteristics and recent investments.

Most of the existing empirical work on energy efficiency focuses on the residential sector. There is little existing work on energy efficiency in commercial buildings. Kahn, Kok, and Quigley (2014) provide descriptive evidence on differences in energy consumption across one utility's commercial buildings as a function of various observables, including incentives embedded in the occupants' leases, age and other physical attributes of the buildings. In other work, Kok and co-authors analyze the financial returns to energy efficiency attributes, though many of the attributes were part of the building's original construction and not part of deliberate retrofits, which are the focus of our work (Kok and Jennen (2012) and Eichholtz, Kok, and Quigley (2013)).

2.2.1 Data Sources

This project merges data from several sources. We combine high-frequency electricity consumption and account information with data on energy efficiency upgrades, school characteristics, community demographics, and weather.

We obtained 15-minute interval electricity metering data for the universe of public K-12 schools in Northern California served by Pacific Gas and Electric Company (PG&E). The data begin in January 2008, or the first month after the school's smart meter was installed, whichever comes later. 20 percent of the schools in the sample appear in 2008; the median year schools enter the sample is 2011. The data

series runs through 2014. To speed computation time, we aggregate these 15-minute observations to three-hourly "blocks."[3]

In general, PG&E's databases link meters to customers for billing purposes. For schools, this creates a unique challenge: in general, school bills are paid at the district, rather than individual school site, level. In order to estimate the effect of energy efficiency investments on electricity consumption, we required a concordance between meters and school sites. We developed a meter matching process in parallel with PG&E. The final algorithm that was used to match meters to schools was implemented as follows: first, PG&E retrieved all meters associated with "education" customers.[4] Next, they used GPS coordinates attached to each meter to match meters from this universe to school sites, using school location data from the California Department of Education. This results in a good but imperfect match between meters and schools. In some cases, multiple school sites match to one or more meters. This can often be resolved by hand, and was wherever possible, but several "clusters" remain. We use only school-meter matches that did not need to be aggregated. Robustness tests suggest that the results presented here do not change substantively when we include these "clusters." Our final sample includes 2,094 schools.

The PG&E data also describe energy efficiency upgrades at the schools as long as the school applied for rebates from the utility.[5] A total of 6,971 upgrades occurred at 1,039 schools between January 2008 and December 2014. For each energy efficiency measure installed, our data include the measure code, the measure description[6], a technology family (e.g., "HVAC", "Lighting", "Food service technology"), the number of units installed, the installation date, the expected lifetime of the project, the engineering-estimate-based expected annual kWh savings, the incremental measure cost, and the PG&E upgrade incentive received by the school.[7]

Many schools undertake multiple upgrades, either within or across categories. We include all upgrades in our analysis, and break out results for the two most common upgrade categories: HVAC and lighting. Together these two categories make up over 70 percent of the total upgrades, and over 60 percent of the total projected savings.

[3]Robustness checks suggest that our results are similar if we aggregate to hourly blocks.

[4]PG&E records a NAICS code for most customers in its system; this list of education customers was based on the customer NAICS code.

[5]Anecdotally, the upgrades in our database are likely to make up a large share of energy efficiency upgrades undertaken by schools. PG&E reports making concerted marketing efforts to reach out to schools to induce them to make these investments; schools often lack funds to devote to energy efficiency upgrades in the absence of such rebates.

[6]One example lighting measure description from our data: "PREMIUM T-8/T-5 28W ELEC BALLAST REPLACE T12 40W MAGN BALLAST-4 FT 2 LAMP"

[7]We have opted not to use the cost data as we were unable to obtain a consistent definition of the variables.

We obtained school and school-by-year information from the California Department of Education on academic performance, number of students, the demographic composition of each school's students, the type of school (i.e., elementary, middle school, high school or other) and location. We also matched schools and school districts to Census blocks in order to incorporate additional neighborhood demographic information, such as racial composition and income. Finally, we obtained information on whether school district voters had approved facilities bonds in the two to five years before retrofits began at treated schools.[8]

We obtained hourly temperature data from 2008 to 2014 from over 4,500 weather stations across California from MesoWest, a weather data aggregation project hosted by the University of Utah.[9] We matched school GPS coordinates provided by the Department of Education with weather station locations from MesoWest to pair each school with its closest weather station to create a school-specific hourly temperature record.

2.2.2 Summary Statistics

Table 2.2.1 displays summary statistics for the data described above, across schools with and without energy efficiency projects. Of the 2,094 schools in the sample, 1,039 received some type of energy efficiency upgrade. 628 received only HVAC upgrades and 493 received only lighting upgrades. There are 1,055 schools that received no upgrade. Our main variable of interest is electricity consumption, which we observe every 15 minutes but summarize in Table 2.2.1 at the 3-hourly "block" level for practical purposes. We observe electricity consumption data for the average school for a three-year period.

For schools that are treated, expected energy savings are almost 30,000 kWh, which is approximately 5 percent of average electricity consumption. Savings are a slightly larger share of consumption for schools with lighting interventions.[10]

The first three columns of Table 2.2.1 highlight measurable differences between treated and non-treated schools. Treated schools use over 50 percent more electricity, have 30 percent more enrolled students, different student demographics and

[8]Bond data are from EdSource (edsource.org).

[9]We performed our own sample cleaning procedure on the data from these stations, dropping observations with unreasonably large fluctuations in temperature, and dropping stations with more than 10% missing or bad observations.

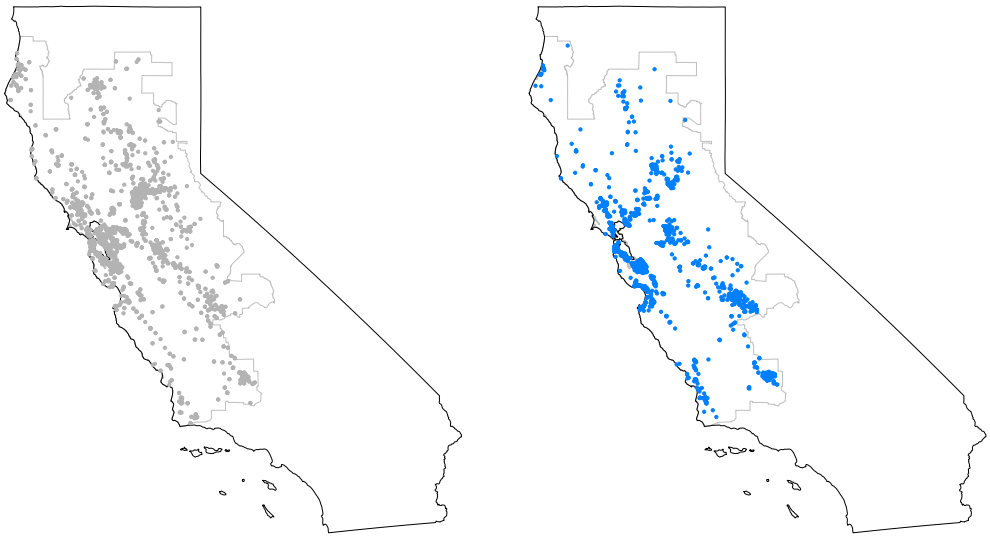
[10]We do not summarize projected savings since it is zero for the control schools.

Table 2.2.1: Average characteristics of schools in the sample

Category	Untreated	Any Intervention		HVAC Interventions		Lighting Interventions	
		Treated	T-U	Treated	T-U	Treated	T-U
Hourly energy cons. (kWh)	34.2	57.1	22.9	63.5	29.4	61.6	27.5
	[15.3, 39.1]	[24.7, 59.5]	(2.5)	[28.7, 63.9]	(2.9)	[27.3, 60.8]	(3.1)
First year in sample	2012	2009.9	-2.0	2009.5	-2.3	2009.8	-2.1
	[2011.0, 2013.0]	[2008.0, 2012.0]	(0.1)	[2008.0, 2011.0]	(0.1)	[2008.0, 2011.0]	(0.1)
Total enrollment	553.5	722.3	168.8	765.9	212.4	747.7	194.2
	[355.4, 658.9]	[462.5, 808.3]	(19.9)	[496.8, 832.0]	(22.8)	[479.8, 812.4]	(24.5)
API	789.4	793.7	4.3	792.8	3.4	786.5	-2.9
	[734.8, 855.3]	[737.9, 854.6]	(4.2)	[734.0, 856.0]	(4.9)	[736.5, 836.0]	(5.1)
Bond passed – 2 yrs	0.3	0.3	0.0	0.3	-0.0	0.3	-0.1
	[0.0, 1.0]	[0.0, 1.0]	(0.0)	[0.0, 1.0]	(0.0)	[0.0, 1.0]	(0.0)
Bond passed – 5 yrs	0.4	0.4	0.0	0.4	-0.0	0.4	-0.1
	[0.0, 1.0]	[0.0, 1.0]	(0.0)	[0.0, 1.0]	(0.0)	[0.0, 1.0]	(0.0)
High School Graduates	23.5	23.3	-0.2	23.7	0.1	24.1	0.6
	[14.4, 32.0]	[15.3, 31.8]	(0.5)	[15.3, 32.5]	(0.6)	[17.5, 31.5]	(0.7)
College Graduates	20.1	20.3	0.2	19.4	-0.6	19.7	-0.4
	[9.5, 29.8]	[9.8, 29.4]	(0.5)	[9.0, 28.8]	(0.6)	[9.3, 28.5]	(0.7)
Percent single mothers	20.1	19.4	-0.7	20.0	-0.1	19.9	-0.2
	[4.8, 30.4]	[4.9, 28.8]	(0.9)	[4.9, 30.4]	(1.0)	[4.8, 31.1]	(1.1)
Percent African-American	5.7	6.2	0.5	5.5	-0.3	5.9	0.1
	[1.0, 6.3]	[1.3, 7.8]	(0.4)	[1.3, 7.3]	(0.4)	[1.3, 8.1]	(0.5)
Percent White	34.3	30.7	-3.6	29.8	-4.5	29.7	-4.6
	[8.5, 56.5]	[8.8, 51.9]	(1.1)	[8.3, 49.3]	(1.3)	[8.5, 48.0]	(1.4)
Percent 2+ races	2.4	2.0	-0.4	1.6	-0.8	1.9	-0.5
	[0.3, 3.5]	[0.3, 2.8]	(0.1)	[0.0, 2.0]	(0.1)	[0.3, 2.8]	(0.2)
Temperature (F)	60.2	60.7	0.5	61.2	1.0	60.9	0.7
	[57.5, 62.6]	[58.1, 63.0]	(0.2)	[58.4, 63.7]	(0.2)	[58.2, 63.2]	(0.2)
Latitude	37.7	37.5	-0.2	37.4	-0.3	37.4	-0.2
	[37.2, 38.2]	[36.9, 38.0]	(0.0)	[36.8, 38.0]	(0.1)	[36.8, 38.0]	(0.1)
Longitude	-121.6	-121.3	0.4	-121.0	0.6	-121.2	0.4
	[-122.2, -121.1]	[-122.0, -120.4]	(0.0)	[-122.0, -119.8]	(0.1)	[-122.0, -120.0]	(0.1)
Number of schools	1055	1039		628		493	

Notes: This table displays characteristics of the treated and untreated schools in our sample, by type of intervention. Interquantile range in brackets; p-value on differences between treated and untreated schools in parentheses.

Figure 2.2.1: Locations of Treated and Untreated Schools



Notes: This figure displays the locations of schools in our sample. The left panel shows "untreated" schools that did not undertake any energy efficiency upgrades during our sample period, and the right panel shows "treated" schools that had at least one upgrade during our sample. There is substantial overlap in the locations of treated and untreated schools. The light gray outline shows the PG&E service territory.

are generally further south and further east. Figure 2.2.1 shows the spatial distribution of treatment and control schools. Schools that received HVAC and/or lighting upgrades look different across an array of observable characteristics from schools that did not receive these upgrades. Schools receiving lighting upgrades perform less well academically than non-upgrading schools, but this difference disappears when comparing schools that did and did not receive HVAC upgrades. Because these schools are different on a range of observable characteristics, and because these indicators may be correlated with electricity usage, it is important that we consider selection into treatment as a possible threat to econometric identification in this setting.

2.3 Regression Analysis

We first present regression results that estimate the treatment effect of the energy efficiency interventions using a difference-in-difference framework. Specifically, we

compare schools with and without energy efficiency interventions, using data that spans the period before and after schools invested.

2.3.1 Difference-in-Difference Approach

In order for a difference-in-difference analysis to be valid, the identifying assumption that treated and untreated schools were trending similarly prior to the treatment must hold. While this assumption is fundamentally untestable, we provide evidence in support of the assumption by using an event study framework. In this event study, we compare treated and untreated schools in event time (i.e., in the years prior to and after treatment), and demonstrate that these groups of schools were not trending differently before the treated schools installed energy efficiency upgrades. Because we observe upgrades, including HVAC investments, that may be expected to deliver savings during some parts of the year and not others, we estimate our event studies at the annual level. Figure 2.3.2 displays the results of this event study using a parsimonious specification, including only school and block fixed effects.[11]

In the years prior to treatment, we are unable to statistically distinguish the difference between treated group and untreated group outcomes from zero. However, after treatment, we observe a statistically significant reduction in electricity consumption. This suggests that the difference-in-difference design is likely to be valid in this context.

Moving into the full difference-in-difference design, as our base specifications, we estimate regressions similar to the following:

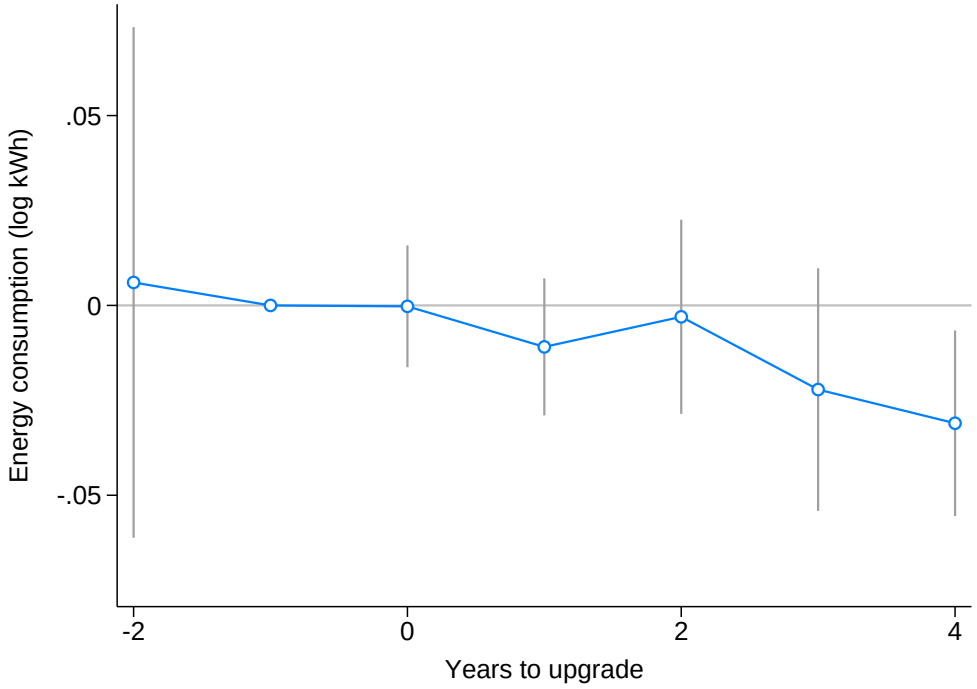
$$Y_{ith} = \beta_h D_{it} + \alpha_i + \kappa_h + \gamma_t + \varepsilon_{ith}, \tag{2.1}$$

where Y_{ith} represents the log of electricity consumption in kWh at school i on date t during hour-block h. D_{it} is a treatment variable equal to the cumulative fraction of upgrades installed in school i by date t, as measured by ex ante expected kWh savings.[12] Therefore, this coefficient can be interpreted as the effect of going from

[11] To generate the coefficients displayed in Figure 2.3.2, we randomly assign a treatment date to untreated schools, and regress the log of energy consumption on a set of years-to-treatment dummies (excluding the year before treatment), a set of years-to-treatment × treated-school dummies, school fixed effects, and block fixed effects. Figure 2.3.2 reports the coefficients on the years-to-treat × treated-school dummies.

[12] Specifically, an untreated school will have $D_{it} = 0$ in all periods. At a treated school that undergoes two upgrades with projected savings of 10 kWh each, D_{it} will be zero prior to the first upgrade; after the first upgrade, D_{it} will be 0.5; and after the second upgrade, D_{it} will be 1.

Figure 2.3.2: Energy efficiency upgrades: Event study



Notes: This event study figure displays the effects of installing any energy efficiency upgrade over time, having residualized school and block fixed effects. In order to leverage untreated observations, we randomly assign the untreated schools in our sample a treatment date. The point estimates presented here are the incremental changes in electricity consumption for treated units, relative to untreated units, in each time period. We find no statistically significant impact of energy efficiency upgrades prior to treatment, but a sharp decline in electricity consumption in the years following treatment. Because our panel is strongly unbalanced, we pool two or more years before treatment into a single coefficient, and four or more years after treatment into a single coefficient. We do not find a strong contemporaneous effect of energy efficiency upgrades, but given that many upgrades take place during the summer, we would not expect to see large effects during year zero.

no upgrades to 100% of a school's upgrades. This parameterization allows us to estimate treatment effects in a setting where schools undergo multiple energy efficiency upgrades. We allow the coefficient on the treatment, β_h, to be different for each block, as we expect upgrades to have heterogeneous effects on energy consumption throughout the day. We also include school fixed effects, α_i, to absorb school-specific average energy consumption, for instance related to its physical characteristics or average enrollment. We include hour-block fixed effects, κ_h, to absorb variation over the daily cycle in energy use, and γ_t, date or month-of-sample fixed effects, to capture common shocks across the sample. ε_{ith} is an idiosyncratic error term. We cluster our standard errors at the school level, allowing for arbitrary dependence between any two observations within the same school. In all estimates that follow, we trim our sample to remove the 1st and 99th percentile observations of the dependent variable, in order to prevent extreme outliers from affecting the results.

We present results for several specifications, each with a different combination of fixed effects, which range from fairly parsimonious specifications, including only school and block fixed effects, to much more flexible ones, allowing for, for instance, school-specific hour-block by month-of-year effects. Under the assumption that, conditional on fixed effects, treatment is as good as randomly assigned, or, that conditional on fixed effects, there are no remaining time-varying differences between treated and untreated schools, we can use this approach to identify the causal effects of energy efficiency upgrades on electricity consumption.

2.3.2 Difference-in-Difference Results

Table 2.3.2 reports the results from a series of difference-in-difference specifications. As we move across columns, we progressively include richer fixed effects and controls. The top panel reports estimates of the aggregate treatment effect of the retrofits across all hours. Below it, we report differential treatment effects for each three-hour block.

Looking at the aggregate effects, we find that the average project implemented at the schools in our sample delivered a reduction in electricity consumption between 2 and 5%. The estimated effect is sensitive to the presence of month-of-sample fixed effects and controls, which are included in columns (4) and (5), respectively. This could reflect common trends across schools. Examining the block patterns, we find that the largest reductions accrue during the hours of school operation, which seems intuitive. Note that consumption of electricity (in levels) is larger in those hours

Table 2.3.2: Difference-in-Difference Results by Hour-Block

	(1)	(2)	(3)	(4)	(5)
Any Intervention: (aggregate)	-0.0454	-0.0450	-0.0488	-0.0191	-0.0174
	(0.0068)	(0.0068)	(0.0068)	(0.0082)	(0.0082)
Midn. to 3 AM x Treat	-0.0560	-0.0505	-0.0538	-0.0242	-0.0222
	(0.0096)	(0.0081)	(0.0082)	(0.0092)	(0.0092)
3 AM to 6 AM x Treat	-0.0571	-0.0487	-0.0519	-0.0223	-0.0204
	(0.0093)	(0.0082)	(0.0083)	(0.0092)	(0.0093)
6 AM to 9 AM x Treat	-0.0229	-0.0264	-0.0306	-0.0007	0.0008
	(0.0079)	(0.0066)	(0.0064)	(0.0084)	(0.0082)
9 AM to Noon x Treat	-0.0416	-0.0490	-0.0518	-0.0237	-0.0205
	(0.0084)	(0.0061)	(0.0062)	(0.0082)	(0.0081)
Noon to 3 PM x Treat	-0.0362	-0.0454	-0.0481	-0.0201	-0.0167
	(0.0088)	(0.0067)	(0.0068)	(0.0085)	(0.0084)
3 PM to 6 PM x Treat	-0.0454	-0.0438	-0.0503	-0.0180	-0.0188
	(0.0090)	(0.0082)	(0.0082)	(0.0092)	(0.0092)
6 PM to 9 PM x Treat	-0.0514	-0.0453	-0.0506	-0.0192	-0.0191
	(0.0089)	(0.0080)	(0.0081)	(0.0091)	(0.0091)
9 PM to Midn. x Treat	-0.0524	-0.0511	-0.0535	-0.0247	-0.0220
	(0.0089)	(0.0078)	(0.0079)	(0.0089)	(0.0089)
Observations	19,253,018	19,252,986	19,251,960	19,252,986	19,251,960
School FE, Block FE	Yes	Yes	Yes	Yes	Yes
School-Block FE	No	Yes	Yes	Yes	Yes
School-Block-Month FE	No	No	Yes	No	Yes
Month of Sample FE	No	No	No	Yes	No
Month of Sample Ctrl.	No	No	No	No	Yes

Notes: Standard errors clustered at the school level. Dependent variable is log hourly electricity consumption, averaged by blocks of three hours.

(often twice as high), so the difference in electricity savings in levels across hours is substantial.[13]

Table 2.3.3 presents the results for two subsample: treated schools that installed HVAC upgrades and treated schools that installed lighting upgrades (both compared to untreated schools that installed no upgrades over our sample period). The results for HVAC and lighting differ substantially. We estimate reductions of between 2 and 5 % for HVAC interventions, but the HVAC estimates are relatively noisy. Lighting interventions appear to drive larger reductions in electricity consumption: between 4 and 7 %. Furthermore, the HVAC effects appear to be relatively consistent throughout the day, while the lighting effects are stronger during the main school hours.

One potential explanation for this difference is that the effects of lighting interventions are relatively homogeneous across schools, while the effects of HVAC interventions are much more sensitive to the local climate, current weather conditions, and other factors. Therefore, it might be harder to control for all confounding factors in a parsimonious way. For example, the inclusion of school-specific month effects appears to affect some of the estimates for the HVAC specifications (e.g., columns (3) and (5)), whereas they are much less important for the lighting interventions.[14]

[13]Boomhower and Davis (2016) measure the benefits of efficiency investments by time of day and show that reductions in the middle of the day are worth significantly more in California.

[14]In ongoing work, we are exploring heterogeneity in more detail.

Table 2.3.3: Difference-in-Difference Results by Type of Intervention

	(1)	(2)	(3)	(4)	(5)
HVAC Interventions: (aggregate)	-0.0489	-0.0484	-0.0523	-0.0238	-0.0213
	(0.0089)	(0.0089)	(0.0089)	(0.0108)	(0.0103)
Midn. to 3 AM x Treat	-0.0654	-0.0538	-0.0582	-0.0304	-0.0293
	(0.0120)	(0.0100)	(0.0101)	(0.0111)	(0.0108)
3 AM to 6 AM x Treat	-0.0671	-0.0521	-0.0560	-0.0286	-0.0271
	(0.0115)	(0.0100)	(0.0101)	(0.0111)	(0.0109)
6 AM to 9 AM x Treat	-0.0193	-0.0195	-0.0271	0.0034	0.0018
	(0.0100)	(0.0082)	(0.0080)	(0.0098)	(0.0093)
9 AM to Noon x Treat	-0.0408	-0.0457	-0.0496	-0.0230	-0.0209
	(0.0103)	(0.0076)	(0.0075)	(0.0094)	(0.0090)
Noon to 3 PM x Treat	-0.0302	-0.0416	-0.0444	-0.0190	-0.0157
	(0.0107)	(0.0083)	(0.0082)	(0.0100)	(0.0095)
3 PM to 6 PM x Treat	-0.0320	-0.0406	-0.0477	-0.0177	-0.0188
	(0.0110)	(0.0103)	(0.0102)	(0.0113)	(0.0108)
6 PM to 9 PM x Treat	-0.0526	-0.0498	-0.0561	-0.0265	-0.0273
	(0.0112)	(0.0101)	(0.0102)	(0.0111)	(0.0108)
9 PM to Midn. x Treat	-0.0526	-0.0557	-0.0591	-0.0323	-0.0302
	(0.0111)	(0.0094)	(0.0094)	(0.0105)	(0.0102)
Observations	14,939,790	14,939,760	14,938,895	14,939,760	14,938,895
Light Interventions: (aggregate)	-0.0674	-0.0669	-0.0705	-0.0421	-0.0427
	(0.0095)	(0.0095)	(0.0097)	(0.0124)	(0.0121)
Midn. to 3 AM x Treat	-0.0575	-0.0486	-0.0510	-0.0223	-0.0196
	(0.0128)	(0.0110)	(0.0112)	(0.0122)	(0.0122)
3 AM to 6 AM x Treat	-0.0593	-0.0489	-0.0537	-0.0227	-0.0223
	(0.0126)	(0.0111)	(0.0113)	(0.0122)	(0.0124)
6 AM to 9 AM x Treat	-0.0274	-0.0370	-0.0443	-0.0112	-0.0131
	(0.0105)	(0.0084)	(0.0083)	(0.0105)	(0.0103)
9 AM to Noon x Treat	-0.0515	-0.0629	-0.0668	-0.0376	-0.0356
	(0.0110)	(0.0080)	(0.0081)	(0.0103)	(0.0105)
Noon to 3 PM x Treat	-0.0416	-0.0606	-0.0622	-0.0354	-0.0311
	(0.0114)	(0.0089)	(0.0089)	(0.0109)	(0.0110)
3 PM to 6 PM x Treat	-0.0582	-0.0559	-0.0608	-0.0303	-0.0295
	(0.0116)	(0.0109)	(0.0110)	(0.0122)	(0.0123)
6 PM to 9 PM x Treat	-0.0647	-0.0531	-0.0581	-0.0271	-0.0268
	(0.0116)	(0.0104)	(0.0106)	(0.0116)	(0.0118)
9 PM to Midn. x Treat	-0.0659	-0.0553	-0.0561	-0.0290	-0.0247
	(0.0119)	(0.0108)	(0.0110)	(0.0120)	(0.0121)
Observations	12,940,064	12,940,036	12,939,239	12,940,036	12,939,239
School FE, Block FE	Yes	Yes	Yes	Yes	Yes
School-Block FE	No	Yes	Yes	Yes	Yes
School-Block-Month FE	No	No	Yes	No	Yes
Month of Sample FE	No	No	No	Yes	No
Month of Sample Ctrl.	No	No	No	No	Yes

Notes: Standard errors clustered at the school level. Dependent variable is log hourly electricity consumption, averaged by blocks of three hours. Only schools that experienced either an HVAC or light upgrade are included in the treated sample, respectively.

2.3.3 Robustness

Despite the promising event study shown above, several features of our data suggest that the treatment effects estimated using the difference-in-difference approach may be biased. For one, the coefficient estimates appear sensitive to the sets of fixed effects we include. Also, the number and composition of schools in the sample (and the treatment group) changes over time, which means it is important to flexibly control for time-varying, confounding factors. Since electricity consumption follows seasonal, weekly and diurnal patterns, controlling for all potential changes can be difficult.

Here, we present two sets of robustness checks on the difference-in-difference approach. First, we show a placebo exercise, in which we randomly assign "treatments" to schools in the period prior to actual treatment, to test the robustness of our fixed effects approaches. We observe patterns in these placebo estimates that suggest that our specifications do not adequately control for time-varying unobservable characteristics that may bias our difference-in-difference results. We next present results from a series of nearest-neighbor matching estimators, in an effort to avoid issues with selection bias that may arise given the differences between treated and untreated schools in our sample. We find that these results are highly unstable, and sensitive to different matching specifications.

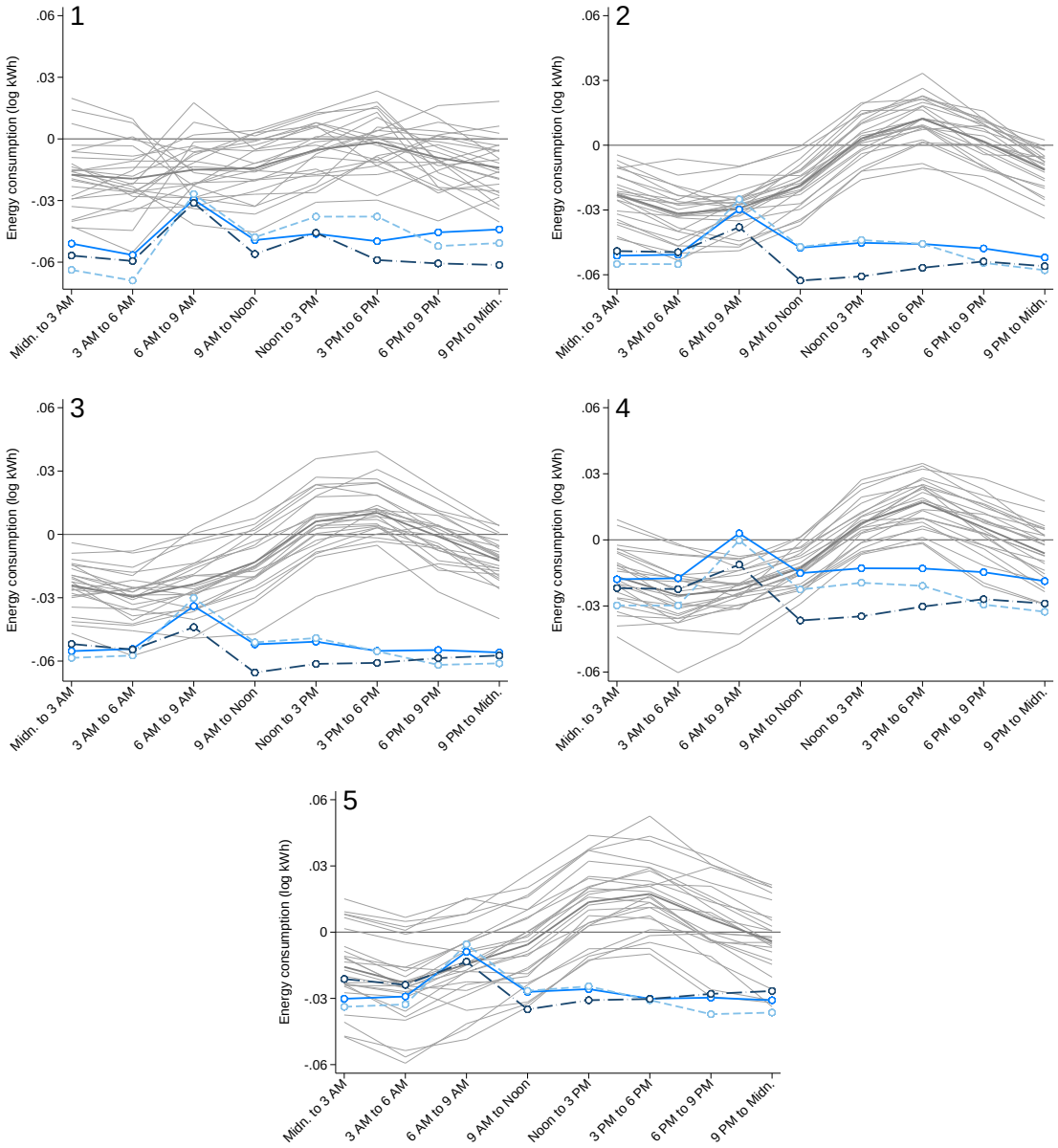
2.3.3.1 Placebo Tests

We first conduct a series of placebo tests using the same specifications summarized in Tables 2.3.2 and 2.3.3 to gauge the extent to which our difference-in-difference approach is appropriately controlling for time-varying unobservable characteristics which may threaten our identification strategy. To do this, we drop all post-treatment observations and randomly assign approximately 50% of schools into a placebo "treated" group, to match the proportion of schools actually in treatment in the real sample. We then randomly assign these "treated" schools a "treatment date" by taking a uniform date draw between their first appearance in the sample and their last appearance in the sample.[15] We then estimate specifications (1) through (5) and store the block-wise coefficients. We repeat this process 25 times.

Figure 2.3.3 reports the results of this exercise. Panels 1 to 5 match the specifications in columns (1) to (5) of Tables 2.3.2 and 2.3.3. We plot treatment effects for each placebo run in gray, and overlay our estimated treatment effects for any upgrades, HVAC upgrades and lighting upgrades in shades of blue. We also plot the average coefficient across all placebo estimates (solid thick gray line). Notably,

[15] We do not allow schools to be "treated" in either their first or last month in the sample.

Figure 2.3.3: Placebo Treatment Effects – Difference-in-Difference



Notes: This figure presents difference-in-difference treatment effects using real and placebo data. Each panel corresponds to the column of the same number from Tables 2.3.2 and 2.3.3. The light gray lines present the placebo effects, with the solid gray line showing the average placebo effect. Placebo effects are generated using pre-treatment data only, and randomly assigning treatment status and timing according to the distribution present in real data. The solid blue line shows treatment effects from any upgrade; the dashed aqua line shows treatment effects for HVAC upgrades; and the dash-dotted navy line shows treatment effects from lighting upgrades, all using real data. Even with the most flexible specifications, which include school-by-hour-block fixed effects, we see marked hourly patterns in the placebos.

the placebo coefficient estimates display a systematic pattern by block across all of the simulations. Specifically, estimates appear negative in the early part of the day and positive after noon. These results suggest changes in the temporal pattern of consumption over time that are not captured by the rich fixed effects we include. This calls into question the validity of these difference-in-difference specifications.

2.3.3.2 Matching

Given the systematic differences between treated and untreated schools summarized in Table 2.2.1, we also explore the sensitivity of the coefficient estimates to the set of untreated schools included in the estimation sample. In particular, we implement nearest neighbor matching based on several candidate choice sets, with the goal of choosing matches that are most closely aligned on observable characteristics, in order to reduce selection bias.

Matching presents a challenge in the school setting. An unrestricted nearest neighbor match will tend to select untreated schools from the same district as the treated schools, since schools within the same district tend to be demographically similar and have similar weather and electricity consumption. However, school infrastructure decisions are often made at the *district* level, meaning that unrestricted matches and matches that are restricted to be within the same school district may be problematic. In particular, if a district selects one of its schools to receive an energy efficiency upgrade and not others for reasons that are unobservable to the econometrician, this type of nearest-neighbor matching strategy may induce selection bias. On the other hand, restricting matches to schools that are in different districts makes it more difficult to find untreated schools that are comparable on observable characteristics.

We present results from three types of matches: unrestricted, restricted to schools in the same district, and restricted to schools in other districts. Our treatment effect estimates exhibit strong sensitivity to the matching criteria that we choose. Table 2.3.4 presents aggregate treatment effect estimates, analogous to the top row of Table 2.3.2.

In each panel, matches are drawn from an unrestricted set in the first row, the set of schools in the same district in the second row and the set of schools in outside districts in the third row.[16] Examining the top panel ("Any interventions"), it is clear that estimates are sensitive to modeling choices. Matching on "Any district" yields larger estimates than when candidate matches are restricted, although the

[16]The matching variables are summarized in the table notes and include both demographic variables and electricity usage patterns. Results using other sets of matching variables are similarly sensitive to the specification and match issues summarized in Table 2.3.4.

Table 2.3.4: Matching Results

	(1)	(2)	(3)	(4)	(5)
Any Intervention:					
Any district	-0.0504	-0.0506	-0.0541	-0.0289	-0.0304
	(0.0124)	(0.0124)	(0.0131)	(0.0123)	(0.0130)
Same district	-0.0149	-0.0147	-0.0096	0.0008	0.0100
	(0.0141)	(0.0141)	(0.0149)	(0.0143)	(0.0147)
Opposite district	-0.0471	-0.0471	-0.0506	-0.0215	-0.0258
	(0.0119)	(0.0120)	(0.0127)	(0.0123)	(0.0130)
Observations	4,828,122	4,828,108	4,826,977	4,828,108	4,826,977
HVAC Interventions:					
Any district	-0.0098	-0.0100	-0.0054	-0.0113	-0.0068
	(0.0263)	(0.0263)	(0.0273)	(0.0265)	(0.0265)
Same district	0.0079	0.0081	0.0132	0.0136	0.0157
	(0.0222)	(0.0222)	(0.0236)	(0.0213)	(0.0215)
Opposite district	-0.0665	-0.0668	-0.0679	-0.0394	-0.0335
	(0.0151)	(0.0151)	(0.0156)	(0.0155)	(0.0166)
Observations	2,379,037	2,379,033	2,378,466	2,379,033	2,378,466
Lighting Interventions:					
Any district	-0.0602	-0.0599	-0.0562	-0.0339	-0.0251
	(0.0212)	(0.0212)	(0.0220)	(0.0205)	(0.0235)
Same district	-0.0459	-0.0455	-0.0451	-0.0270	-0.0287
	(0.0158)	(0.0158)	(0.0168)	(0.0168)	(0.0193)
Opposite district	-0.0461	-0.0462	-0.0492	-0.0040	-0.0051
	(0.0122)	(0.0122)	(0.0128)	(0.0190)	(0.0203)
Observations	1,914,567	1,914,563	1,914,147	1,914,563	1,914,147
School FE, Block FE	Yes	Yes	Yes	Yes	Yes
School-Block FE	No	Yes	Yes	Yes	Yes
School-Block-Month FE	No	No	Yes	No	Yes
Month of Sample FE	No	No	No	Yes	No
Month of Sample Ctrl.	No	No	No	No	Yes

Notes: This table displays regression results where the untreated group is chosen via nearest-neighbor matching. We match one untreated school to each treatment school. Each row in the table employs a different restriction on which schools are allowed to be matched to any given treatment school. "Any district" matches allow any control school to be matched to a treatment school; "same district" matches are restricted to untreated schools in the same school district, and "opposite district" matches are restricted to untreated schools from different districts. In each case, the matching variables are the mean, maximum, and standard deviation of electricity consumption in each three-hour block (e.g. 9 AM-Noon) from the pre-treatment period; demographic indicators at the census block level, including the poverty rate, log of per capita income, school-level indicators (enrollment; age of the school; grades taught; an academic performance index; and climate).

differences are greater when matches are drawn from the same district. Looking across panels, which isolate HVAC and lighting interventions, a similar narrative holds. The sensitivity to matching criteria and specification reflects a tension between quality of candidate matches and selection bias on imperfect matches. These results suggest that it is difficult to use matching methods to construct an appropriate counterfactual in this setting.

2.4 Machine Learning Analysis

Our machine learning approach deploys forecasting methods to predict electricity consumption at each school and uses this prediction as a counterfactual to estimate energy savings. Broadly, these methods use algorithms to construct prediction models that are designed to perform well out of sample. The researcher provides the machine learning tool with a dataset and the candidate variables for model selection. The machine learning algorithm will then subset the data into "training" and "test" subsamples, fit models using the training data only, and test the out-of-sample fit of these models using the test data.[17] These methods are designed to trade off bias and variance, and penalize overfitting, such that the resulting model provides the best out-of-sample fit. There are a variety of machine learning algorithms that can used to do this, including LASSO, trees, random trees, a combination of models (bagging), and others. Ultimately, these methods return a forecasting model, which we use to generate a counterfactual data series.

2.4.1 Machine Learning Approach

The goal of the predictive model in our setting is to provide a counterfactual during the treated period that describes what would have happened to energy consumption at treated schools if they had not implemented an energy efficiency project. Machine learning tools are particularly well-suited to constructing counterfactuals, since the goal of building the counterfactual is not to individually isolate the effect of any particular variable, but rather to provide a good overall prediction fit. With the goal of achieving best predictive power, machine learning techniques give substantial flexibility to the algorithms to build a statistical model that considers many potential regressors. Machine learning models tend to out-perform models that are chosen by the researcher in a more idiosyncratic fashion when it comes to predictive power, since they use an algorithm and penalty rule that trades off fit and variance. This

[17] Often, this subsampling procedure is repeated several times to improve the performance of predictions.

enables the econometrician to select models from a much wider space than would be possible with trial-and-error. Importantly, in our empirical setting, we generate a prediction model separately for each school This allows for a great deal of flexibility in the control variables used to create our counterfactual and makes it feasible to retrieve estimates of heterogeneous treatment effects.

Relative to the difference-in-differences estimators, a key difference in our machine learning approach is that we only use pre-treatment data to generate the counterfactual model. By providing only pre-treatment data to the algorithm, we generates a predictive model that is a function of co-variates and, by construction, is uncontaminated by the treatment effect itself.[18]

2.4.1.1 Methodological Contribution

Until now, the intersection of machine learning and causal inference has focused on two applications: randomized controlled trials and selection-on-observables designs. Our approach combines machine learning with selection on unobservables. We leverage high-frequency data to construct unit-specific counterfactuals in a panel fixed-effects framework.

The first strand of the existing literature to combine machine learning and causal inference focuses on randomized trials. Under random (or as good as random) assignment, there is little need for machine learning to identify average treatment effects. In many randomized settings, however, researchers are often also interested in retrieving heterogeneous effects while minimizing concerns about "cherry-picking". Athey and Imbens 2015 propose a recursive partitioning methodology which uses random forests to estimate heterogeneous treatment effects with proper inference. While these methods are useful when units are randomly assigned to treatment, the non-random selection present in our setting makes this method undesirable, as the partitioning itself may introduce greater selection bias.

Machine learning methods have also been developed for settings without random assignment into treatment. The existing work in this area focuses on identifying average treatment effects in selection-on-observables designs, where, conditional on observable characteristics, the identifying assumption is that the outcome of interest would be equal across groups in the absence of treatment (i.e., no selection on unobservables). Three broad classes of methods have arisen to use machine learning to improve selection on observables methods. The first approach is akin to existing propensity score methods (cross-sectional units are assigned PS weights according to

[18]Note that the pre-treatment data is further subsetted by the algorithm into a "training" and "test" sample, as described above. None of the post-treatment data is used to "test" the predictive model. The method is described in detail below.

their ability to predict treatment status), but uses machine learning tools for model selection (McCaffrey, Ridgeway, and Morral 2004). A closely-related method forces covariate "balance" by directly including a covariate balancing constraints in the machine learning algorithm (e.g., Wyss et al. 2014). Finally, Belloni, Chernozhukov, and Hansen (2014) proposes a "double selection" approach, using machine learning in the form of LASSO, to both predict selection into treatment as well as to predict an outcome, using both the covariates that predict treatment assignment and the outcome in the final step.

These double selection methods improve upon existing selection-on-observables machine learning designs, but all three methods are subject to the same identifying assumptions as traditional selection-on-observables methods. Furthermore, these methods perform best with large N, and when the degree of heterogeneity is small relative to the sample size (i.e., it is possible to find control units that are suitable matches on observable characteristics). In our setting, it is challenging to find good matches across the treated and untreated groups, as we observe a limited number of units and the nature of selection in our setting makes matching difficult. In particular, it seems attractive to match treated schools to untreated schools in the same district, but it is likely that selection into treatment occurs within district. Alternatively, restricting our matches to occur across districts means that our matches will often remain relatively different on observables.

Though our setting features non-random assignment to treatment and small N, we have large T: our electricity consumption data consists of tens- or hundreds-of-thousands of observations per cross-sectional unit. As a result, we propose a new method for combining machine learning and causal inference. Rather than using machine learning to predict selection into treatment, we use untreated time periods to create unit-specific predictions of the outcome in the absence of treatment. For each school, we generate a LASSO-based counterfactual electricity consumption time series for the treated periods. We can then use these data in conjunction with a within estimator to recover the causal effect of treatment. While our method has some particular data requirements that limit its suitability to settings with large-T samples, it also offers many benefits. There is an intuitive appeal to using untreated observations within a subsequently treated unit as a counterfactual. Machine learning allows the researcher to take an agnostic approach to model selection, where outcomes of different cross-sectional units may be optimally predicted using different variables. Furthermore, the machine learning approach protects against oversaturation. Our method also provides benefits when some features of the data-generating process are exposed to measurement error, as we will demonstrate.

Our method proceeds in two steps. In a first step, we build a statistical model that explains electricity consumption at a given school, as a function of a set of observable

variables, only using pre-treatment data. We repeat this step for each school separately, generating an individual prediction model tailored to match past consumption patterns. In a second step, we compare actual energy consumption to the prediction from the model during the treatment period. Exploiting the same structure as in the difference-in-difference specifications, we take advantage of untreated schools by comparing prediction errors both for treated and untreated schools.

2.4.1.2 Step 1

We begin by forecasting electricity consumption for each school using machine learning. The machine learning algorithms pick and/or create explanatory variables and generate a prediction model based on a LASSO regression.

A key feature of our approach is that we only use data from the pre-period to "train" and "test" the model. For treated schools, the pre-period is defined as the period before any intervention occurs. For untreated schools, we select a subset of the data to be the pre-treatment period by randomly assigning a treatment date between the 20th percentile and 80th percentile calendar dates available.[19] Data after the treatment or pseudo-treatment date is set aside and not used for the purposes of estimating the model.[20] Once the pre-treatment period is defined, we use a cross-validation method for each school separately to algorithmically select the best predictive model.

The resulting output of this first step is a school-specific prediction model for electricity consumption that we apply to each observation in our sample. Using the coefficients from the prediction model in this first step, and observations on the covariates during our entire sample, we predict energy consumption for the whole sample period for each school.

2.4.1.3 Step 2

Armed with this prediction, we can apply a variety of estimators. The simplest treatment effect estimate is to take the difference between actual electricity consumption at treatment schools and our prediction from the machine learning model resulting from the first step in the post-treatment period. The resulting estimator of the

[19] We set the threshold to be between the 20th and 80th percentile to have a more balanced number of observations in the pre- and post-sample.

[20] Imagine an untreated school that we observe between 2009 and 2013. We randomly select a cutoff date for this particular school, e.g., March 3, 2011, and only use data prior to this cutoff date when generating our prediction model.

average effect across treated schools becomes

$$\hat{\beta}^T = \mathbb{E}[Y^T_{i,post} - \hat{Y}^T_{i,post}],$$

where $Y^T_{i,post}$ represents actual average consumption of electricity during the treatment period, and $\hat{Y}^T_{i,post}$ is the predicted electricity consumption during the same treatment period, based on the predictive model built using pre-treatment data only. If an energy efficiency project achieved savings, we would expect $\hat{\beta}^T$ to be negative, as the prediction $\hat{Y}^T_{i,post}$ would overstate expected electricity consumption.

To account for any systematic biases in the prediction, we also use pre-treatment data in a differenced framework, and obtain,

$$\hat{\beta}^{TD} = \mathbb{E}[Y^T_{i,post} - \hat{Y}^T_{i,post}] - \mathbb{E}[Y^T_{i,pre} - \hat{Y}^T_{i,pre}],$$

where $Y^T_{i,pre}$ and $\hat{Y}^T_{i,pre}$ are actual and predicted log electricity consumption during the pre-period, respectively. In the context of prediction models, this correction will tend to be very minor on average, as by construction the average prediction in the pre-treatment period should be close to average actual consumption. For example, these two averages cancel each other by construction if the prediction is based on OLS.[21]

These estimators do not leverage the presence of untreated schools. As with any case study, a before-and-after comparison can suffer from substantial pitfalls. Whereas a richer predictive model can help control for nonlinear effects of observable variables, there is still the concern that there might be trends and other shocks that are not properly captured by the model.

To partially address these concerns, we can use untreated schools as in the previous section, as long as trends are common across schools. We can check the performance of the prediction during the treatment period by examining prediction errors at untreated schools, i.e.,

$$\hat{\beta}^U = \mathbb{E}[Y^U_{i,post} - \hat{Y}^U_{i,post}].$$

If the model has good performance out of sample, $\hat{\beta}^U_i$ should be zero in expectation. In practice, however, the predictive model might fail at capturing behavioral changes over time or other trends in unobservables, leading to some systematic differences that would imply $\hat{\beta}^U \neq 0$.

[21] When using more sophisticated prediction methods that trade-off overfitting, or when allowing for heterogeneous treatment effects within a school (e.g., across hours, for different temperature ranges, etc.), the average prediction might not exactly match average actual consumption at a given cell, making this correction potentially more relevant.

We can also test the model's performance in a differenced approach similar to estimating $\hat{\beta}^{TD}$, where we examine the differences between real and predicted consumption in the post and pre period for untreated schools only, and obtain,

$$\hat{\beta}^{UD} = \mathbb{E}[Y^U_{i,post} - \hat{Y}^U_{i,post}] - \mathbb{E}[Y^U_{i,pre} - \hat{Y}^U_{i,pre}],$$

Again, $\hat{\beta}^{UD}$ should be zero in expectation if the model is successful.

To the extent that there are systematic differences between the prediction and the observed outcomes for untreated schools, and to the extent that these differences reflect trends and biases in the predictive model that are common across schools, we can use these differences as a bias correction for the treated schools. We have,

$$\begin{aligned}\hat{\beta}^{PD} &= \mathbb{E}[Y^T_{i,post} - \hat{Y}^T_{i,post}] - \mathbb{E}[Y^U_{i,post} - \hat{Y}^U_{i,post}] \\ &= \hat{\beta}^T - \hat{\beta}^U,\end{aligned}$$

which will provide a corrected treatment effect estimate under the assumption of common trends and shocks between treated and untreated schools. As before, we can also rely on a triple difference that exploits the differences in predictions between treated and untreated schools during the pre- and post-period,

$$\begin{aligned}\hat{\beta}^{DD} &= \Big(\mathbb{E}[Y^T_{i,post} - \hat{Y}^T_{i,post}] - \mathbb{E}[Y^T_{i,pre} - \hat{Y}^T_{i,pre}]\Big) \\ &\quad - \Big(\mathbb{E}[Y^U_{i,post} - \hat{Y}^U_{i,post}] - \mathbb{E}[(Y^U_{i,pre} - \hat{Y}^U_{i,pre}]\Big) \\ &= \hat{\beta}^{TD} - \hat{\beta}^{UD}.\end{aligned}$$

As mentioned above, this triple difference will tend to provide very similar results to those only using post data when looking at average effects. Note that this triple difference relies on the same identifying assumptions as the difference-in-difference estimator described in the regression methodology section above, namely, that conditional on covariates, treatment and control schools are trending similarly.

Additionally, we consider embedding our machine learning predictions into a regression framework to account for potential additional confounding factors, such as the composition of schools over time. In particular, we consider a regression in which we regress the differences between actual energy consumption and the prediction on the treatment dummy, school-hour fixed effects, month-of-sample fixed effects, and other controls. The approach parallels a standard difference-in-differences methodology, with the major difference that the dependent variable is the error from the prediction model, instead of electricity consumption itself. The only difference with the regression approach is that we include an additional control, "post-period", which

is a dummy to account for the fact that there is a break between the predictions in- and out-of-sample. Including this control is an additional way to take into account that predictions are not as accurate during the post-period, which could introduce an artificial structural break in the data.

2.4.2 Methods Comparison: Monte Carlo

The regression and machine learning methodologies use different sources of variation to generate their respective counterfactuals, and as a result they may retrieve different estimates of the average treatment effect (ATE). Without ex ante knowledge of the true ATE and the true underlying data generating process (DGP), it is not possible to evaluate the relative performance of the estimators as they are applied to our observational data. Monte Carlo simulations, on the other hand, provide a way to control the DGP and thereby compare the relative merits of various approaches against a known benchmark. In this subsection, we describe a series of Monte Carlo simulations, which allow us to understand how well the regression and machine learning approaches perform when applied to settings with known potential confounders.

The baseline DGP is chosen to reflect the main features of our observational data on school energy use. Cross-sectional units (call them "schools") arrive in the dataset at different dates and are observed at high frequency for different durations. The outcome variable of interest at these schools ("energy use") varies according to school size and temperature, which in turn varies by geographic location (inland versus coastal), season, and hour-of-day. At some schools, a treatment of known magnitude and timing is superimposed on this DGP; the remaining schools are untreated.

There are many potential confounders that one may want to consider. The ones that we highlight in this Monte Carlo exercise are motivated by actual uncertainty that we encounter when attempting to correctly specify our regression approaches in the context of school energy efficiency investments: 1) the treatment effect may be nonlinear with respect to temperature; and, 2) the treatment date may be observed with error.

There are several intuitive reasons why the machine learning approach described above can offer benefits relative to a regression approach that seeks to retrieve the ATE. In order to recover a consistent estimate of the ATE, the regression model must be correctly specified. In a setting with a complex DGP, properly selecting controls can be extremely difficult, and there are often not clear rules governing variable selection. In contrast, machine learning uses an algorithm, including cross-validation, to select control variables, imposing discipline on variable selection. Furthermore, because we can implement school-specific machine learning predictions, we are able

to much more flexibly specify the counterfactual-generating model than would be computationally feasible in a regression context. Regression models are also subject to over-fitting. While omitting variables from a regression may lead to bias, including variables that are correlated with treatment can also introduce bias, causing controls to soak up some of the treatment effect. Because our machine learning predictions are constructed using pre-treatment data only, the counterfactual model will not suffer from these problems.

Such overfitting would not occur with the machine learning approach. The prediction model is tested and trained on pre-treatment data only, and will capture the true relationship between Y_{ith} and covariates in the absence of the intervention. This allows the researcher to retrieve the correct ATE by comparing the prediction to the true data. There is an important potential benefit, therefore, from using only pre-intervention data to estimate control parameters. This is potentially feasible in a regression approach as well, but it is rarely the way regressions are implemented. In addition, measurement error in the timing of treatment may cause bias in a regression model, as mistakenly attributing post-treatment data to the pre-treatment period and vice versa will attenuate estimated treatment effects. Machine learning will be subject to less bias than regression if the treatment date is perceived to be earlier than it actually is, since the counterfactual will remain uncontaminated. In the case where the treatment date is thought to be later than it actually is, we may have concerns that the testing and training dataset used to build the model includes treated observations, but in cross-validating the model, these observations should be given less weight, as they will appear to be outliers.

Ultimately, the machine learning algorithm results in a parsimonious prediction model that allows the researcher to remain more agnostic about variable selection, prevents model overfitting, and reduces concerns about measurement error. Our Monte Carlo results provide evidence that there are real benefits to implementing such an approach.

Table 2.4.5 presents summary statistics from the Monte Carlo simulations. We test three DGPs, each of which includes a treatment effect that is nonlinearly increasing in temperature. The first DGP ("No Measurement Error") assumes that the effective treatment date used by the researcher is accurate. Two alternate DGPs assume that the researcher is using a mismeasured treatment date that is either before ("early") or after ("late") the true date. We attempt to retrieve the true treatment effect using the regression and machine learning approaches under four different levels of controls. Specification 0 is reported only for the machine learning method, and it controls for nothing in the second stage. Specifications 1-3 deploy controls

that sequentially increase saturation.[22] The table reports the percentage distance between the estimated ATE and the truth, and the standard error.

Several qualitative conclusions emerge. The regression estimates are systematically attenuated (i.e., the statistics in Table 2.4.5 are negative), and appear to be sensitive to the specification. When there is no measurement error in the treatment date, the fully-specified regression retrieves the true ATE and is precise. This is consistent with the efficiency of OLS (best linear unbiased estimator). Regression also performs well when slightly underspecified (Specification 2), although this feature may not be generalizable since correlations between observed and unobserved variables may be stronger in the real world. Machine learning without controls (Specification 0) performs significantly better than regression under Specification 1, when the regression model is poorly specified. Including additional controls with the machine learning model seems to have little effect on its overall performance under this scenario. This suggests, as expected, that machine learning may have substantial benefits relative to a misspecified regression.

Measurement error in the treatment date introduces significant attenuation bias into the regression estimates. While machine learning does not solve these problems entirely, it appears to yield results with superior properties relative to the regression estimates. First, machine learning estimates are less sensitive to the presence of controls in the second stage. This is a direct result of the prediction itself having been generated using LASSO-selected variables at the school level, which already embody the most important sources of variation. This can be seen most clearly in Specification 0, which performs remarkably well in the absence of any second-stage control variables.

Second, while regression estimates are attenuated across the board, machine learning estimates are far less so. This is particularly true in the presence of measurement error. When the treatment date is mismeasured "early", the counterfactual will be uncontaminated but the "treatment period" will include untreated periods, leading to attenuation. One might be concerned when mismeasurement of the treatment date is late, since this may expose the training/testing procedure to some contamination of the treatment effect. However, the LASSO procedure may perceive these as outliers. Recall that it penalizes the inclusion of additional variables, and thus trades off parsimony against goodness of fit. The large reduction in bias that we observe from machine learning is encouraging, and aligns with our intuition about the channels of potential bias.

[22]Specification 1 includes school fixed effects. Specification 2 adds month of year and year fixed effects. Finally, specification 3 include month-of-sample fixed effects, school interacted with time-of-day fixed effects, and school-specific temperature controls.

Table 2.4.5: Monte Carlo Results, Percent Deviations from "True" Effect

DGP	Specification	Regression	Machine Learning
No Measurement Error	0		0.0213
			(0.0200)
	1	-0.1656	-0.0029
		(0.0071)	(0.0107)
	2	-0.0032	0.0057
		(0.0027)	(0.0068)
	3	0.0005	0.0130
		(0.0014)	(0.0018)
Wrong treatment date (Early)	0		-0.1870
			(0.0212)
	1	-0.4259	-0.2301
		(0.0107)	(0.0149)
	2	-0.3002	-0.2216
		(0.0089)	(0.0094)
	3	-0.3102	-0.2113
		(0.0106)	(0.0093)
Wrong treatment date (Late)	0		0.0050
			(0.0197)
	1	-0.3538	-0.1569
		(0.0148)	(0.0111)
	2	-0.2818	-0.1516
		(0.0133)	(0.0090)
	3	-0.2827	-0.1438
		(0.0126)	(0.0083)

Notes: This table reports percentage deviations from the "true" treatment effect using regression and machine learning approaches under different data generating processes. "No Measurement Error" indicates that the reported date of treatment is accurate. "Wrong treatment date (Early)" and "Wrong treatment date (Late)" refer to reported treatment dates (used by the researcher) that are before and after the true treatment date, respectively. Specification 0 has no second-stage control variables, Specification 1 includes school FEs, Specification 2 includes school, month, and year FEs, and Specification 3 includes school-by-block, month-of-sample, and school-specific temperature controls. Standard errors, clustered by school, in parentheses.

There is some question as to the necessity of implementing the second stage at all. A clear tradeoff exists between bias and precision, which can be seen most clearly in "Wrong treatment date (Late)". In that case, the second stage biases the estimates, presumably due to measurement error; but there are clear gains in precision. The extent to which this trade-off favors implementing the second stage or not will depend on the true DGP, and it is impossible to draw a general conclusion from the single DGP that we simulate here.

In summary, machine learning appears to offer some meaningful advantages over the regression approach. While we have tested it on a limited class of potential confounding factors, our intuition is that some of the benefits will extend to a broad class of DGPs. Parsimonious variable selection, robustness to second-stage (researcher) control variable selection, and training/testing on pre-treatment period data all have appealing features in a general sense.

2.4.3 Machine Learning Results

2.4.3.1 Step 1

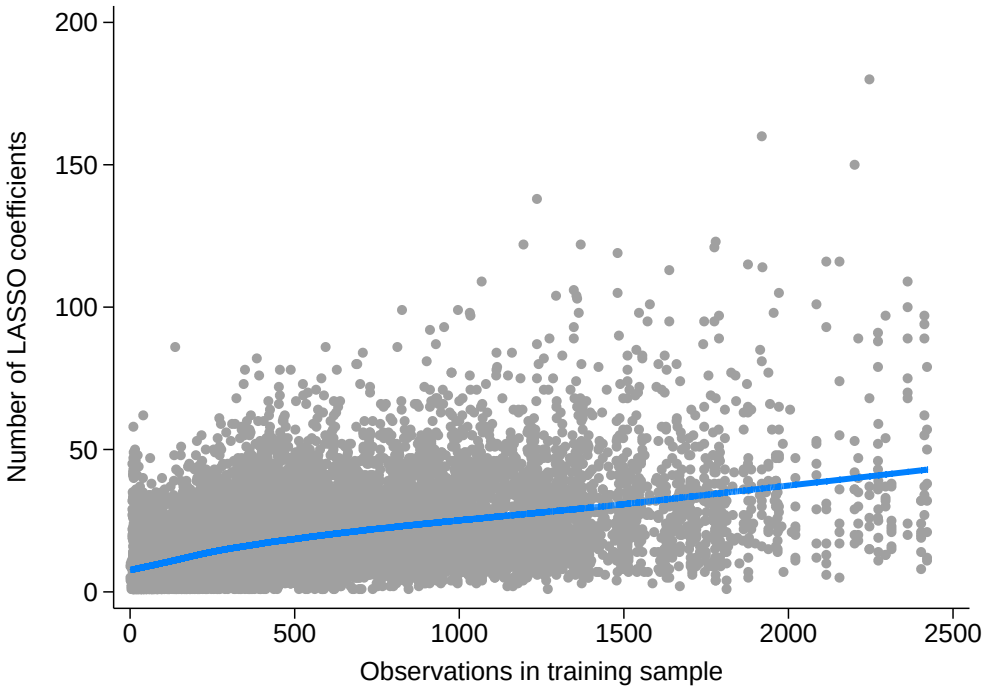
We use a LASSO estimator with pre-treatment data to estimate each school-specific electricity consumption model. We are flexible in the allowed regressors that are considered by the estimator, including: block fixed effects, day of the week, a holiday dummy, a seasonal spline, a temperature spline, and several multiple combinations of these variables (e.g., day of the week interacted with holiday, the seasonal spline interacted with temperature and the blocks, etc). In total, this generates over 3,000 candidate variables.[23]

The machine learning estimator tries to balance the number of explanatory variables and the prediction errors of the model, using cross-validation. Naturally, given that we are estimating these coefficients for each school separately, the optimal models do not include all regressors. In fact, we find that the optimal models usually include between 50 and 100 variables. Importantly, however, the variables are not the same for each school. In fact, we find that the joint set of variables across all schools covers all of the more than 3,000 candidate variables that we consider.

The richness of the selected model depends substantially on the available data. Figure 2.4.4 shows the relationship between the amount of data available to "train" the model and the number of non-zero coefficients in the prediction model. One can see that when data availability is limited, the LASSO coefficient will try to avoid overfitting and limit the number of coefficients that produce the predictive model.

[23]Note that if we were to run all schools together, we would have over 6,000,000 candidate variables, as each variable is allowed to have a school-specific coefficient!

Figure 2.4.4: Number of LASSO coefficients by school



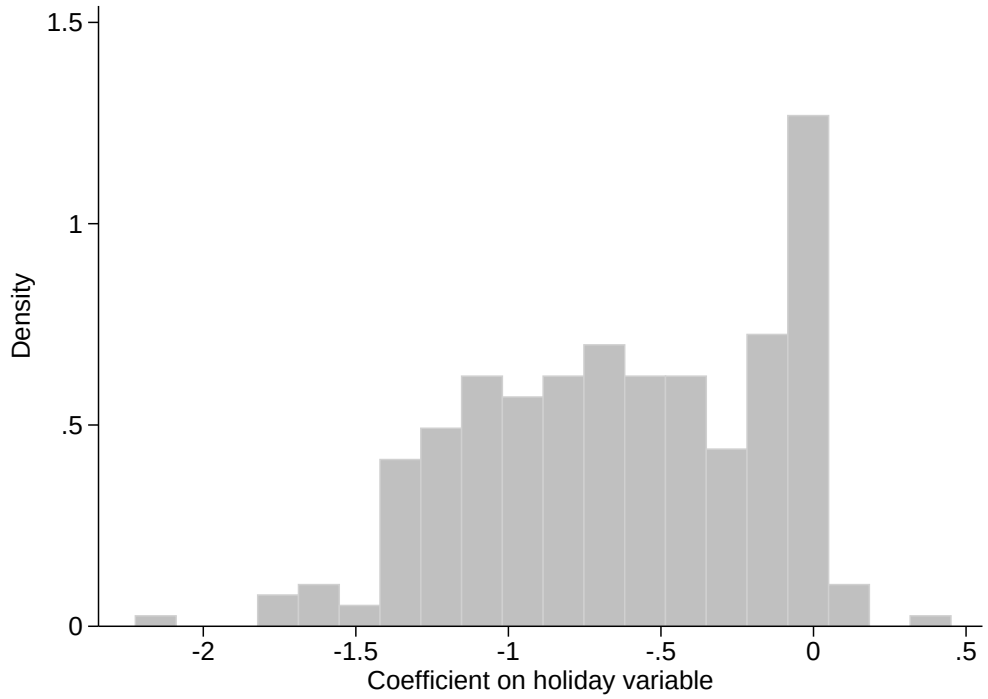
Notes: This figure shows the relationship between the number of observations for a school in the pre-treatment ("training") dataset and the number of variables the LASSO selects to include in the prediction model for that school, across every school in the sample. As expected, the more data available to the LASSO, the more variables it chooses to include. This provides evidence that the LASSO is not overfitting.

One can also see that, even with larger data sets, at some point the inclusion of additional variables is limited.

The variables related to the holiday dummy provide a good example to understand the power of the machine learning approach applied to each school *individually*.[24] Holidays dramatically affect electricity consumption in general, and certainly

[24]We define holiday to include major national holidays, as well as the Thanksgiving and Winter break common to most schools. Unfortunately, we do not have school-level data for the summer break, although the seasonal splines should help account for long spells of inactivity at the schools.

Figure 2.4.5: Holiday Effects in LASSO models



Notes: This figure displays the marginal effect of holidays on predicted energy consumption from the school-specific LASSOs.

at K-12 schools. Indeed, almost all the school-specific models pick up at least one variable containing the holiday dummy, potentially interacted with other regressors (e.g., interacted with the time of day or the day of week). We allow for over fifty such interactions. The median model picks up eight of them, but some school models pick up over twenty holiday variables, depending on data availability and the importance of holidays to that school. The LASSO methodology tries to avoid including too many variables that could lead to collinearity issues and poor performance out of sample. The coefficient estimates on those variables are mostly negative, and we do not observe many large outliers, as seen in Figure 2.4.5.

2.4.3.2 Step 2

Table 2.4.6 presents the baseline estimates comparing prediction errors across schools. In column (1), one can see that the average prediction error for untreated schools during the post-treatment period (i.e., out of sample) is zero. The algorithm appears to predict average consumption well on average, although the prediction errors are not zero at the block level, suggesting that the prediction model performs less well when predicting particular times of the day. This is because the prediction methodology was built to minimize average errors, but was not stratified to also balance errors across hours.[25]

Column (2) also reports average prediction errors, but in this case for treated schools. One can see that the differences for this set of schools are more systematic, as implied by the average effect. At the block level, all of the estimates are negative, implying energy consumption reductions during the treatment period for most blocks of hours. Overall, one can see that the prediction errors for treated and untreated schools are systematically different, with prediction errors for untreated schools precisely estimated around zero, and prediction errors for treated schools significantly negative, i.e., implying energy savings.

Column (3) presents the difference between prediction errors for treated and untreated schools. To the extent that both predictive models have common biases, $\hat{\beta}^{DD}$ should provide a better estimate of the treatment effect. One can see that the estimated effects between column (2) and (3) are not very different on average, although the effects by block appear to be more balanced after performing the correction. Columns (4) and (5) exploit the pre-treatment data to correct for potential prediction biases already present in the baseline. The results are very similar to column (3), suggesting that the estimated treatment effect is not an artifact of the predictive model performing poorly for treated schools in-sample.

The previous results present differences-in-differences across treated and untreated schools, without controlling for any other factors such as school fixed-effects or seasonality and time trends. However, the prediction model has limitations, and it is thus important to further control for confounding factors. Table 2.4.7 presents post-period average effects including additional controls, which parallels Table 2.3.2 in the difference-in-differences results. In column (1), one can see that the effects for

[25] We also consider a predictive model for each school-block, to correct some of these biases. We find that the main findings in this section remain, with most of the biases at the block level that we observe being corrected even without controls.

Table 2.4.6: Prediction Results - Average prediction errors

	(1) $\hat{\beta}^{U}$	(2) $\hat{\beta}^{T}$	(3) $\hat{\beta}^{PD}$	(4) $\hat{\beta}^{UD}$	(5) $\hat{\beta}^{TD}$	(6) $\hat{\beta}^{DD}$
Aggregate	-0.0042	-0.0465	-0.0423	-0.0113	-0.0529	-0.0501
	(0.0050)	(0.0061)	(0.0079)	(0.0050)	(0.0060)	(0.0063)
Midn. to 3 AM	-0.0149	-0.0510	-0.0361	-0.0165	-0.0513	-0.0361
	(0.0073)	(0.0081)	(0.0109)	(0.0074)	(0.0081)	(0.0109)
3 AM to 6 AM	-0.0146	-0.0467	-0.0321	-0.0169	-0.0472	-0.0321
	(0.0074)	(0.0081)	(0.0110)	(0.0074)	(0.0081)	(0.0110)
6 AM to 9 AM	-0.0051	-0.0272	-0.0221	-0.0126	-0.0335	-0.0221
	(0.0055)	(0.0058)	(0.0080)	(0.0054)	(0.0057)	(0.0080)
9 AM to Noon	0.0024	-0.0360	-0.0384	-0.0164	-0.0558	-0.0384
	(0.0050)	(0.0053)	(0.0073)	(0.0051)	(0.0053)	(0.0073)
Noon to 3 PM	0.0087	-0.0402	-0.0489	-0.0092	-0.0596	-0.0489
	(0.0051)	(0.0057)	(0.0076)	(0.0050)	(0.0056)	(0.0076)
3 PM to 6 PM	0.0075	-0.0572	-0.0647	0.0024	-0.0617	-0.0647
	(0.0065)	(0.0073)	(0.0097)	(0.0065)	(0.0073)	(0.0097)
6 PM to 9 PM	-0.0055	-0.0582	-0.0527	-0.0083	-0.0586	-0.0527
	(0.0066)	(0.0075)	(0.0100)	(0.0066)	(0.0075)	(0.0100)
9 PM to Midn.	-0.0111	-0.0547	-0.0437	-0.0126	-0.0554	-0.0437
	(0.0071)	(0.0077)	(0.0105)	(0.0071)	(0.0076)	(0.0105)
Observations	3,434,982	7,341,034	10,776,016	6,916,585	12,331,889	19,248,474

Notes: Standard errors clustered at the school level. Dependent variable is the prediction error in log electricity consumption consumption by blocks of three hours.

treated schools remain negative and significant, albeit smaller, as we include school fixed effects that take away some persistent heterogeneity. We obtain similar results if we also include school-specific block effects, as seen in column (2). Column (3) includes school-specific block times month effects, to control for seasonality. One can see that the results are not very different. As in the regression section, to control for trends, we include month-of-sample fixed effects in column (4), and a month-of sample time trend in column (5). We find that controlling for seasonality, trends, and changes in the composition of the panel is important and reduces the treatment effect, but the main patterns remain.

Table 2.4.7: Prediction Results by Hour-Block

	(1)	(2)	(3)	(4)	(5)
Treatment (aggregate)	-0.0400	-0.0403	-0.0417	-0.0266	-0.0237
	(0.0068)	(0.0068)	(0.0071)	(0.0076)	(0.0076)
Midn. to 3 AM x Treat	-0.0375	-0.0411	-0.0426	-0.0273	-0.0245
	(0.0087)	(0.0082)	(0.0085)	(0.0087)	(0.0088)
3 AM to 6 AM x Treat	-0.0315	-0.0368	-0.0386	-0.0230	-0.0205
	(0.0088)	(0.0083)	(0.0086)	(0.0088)	(0.0089)
6 AM to 9 AM x Treat	-0.0205	-0.0201	-0.0213	-0.0064	-0.0033
	(0.0068)	(0.0064)	(0.0066)	(0.0074)	(0.0074)
9 AM to Noon x Treat	-0.0422	-0.0457	-0.0469	-0.0321	-0.0289
	(0.0064)	(0.0061)	(0.0065)	(0.0074)	(0.0074)
Noon to 3 PM x Treat	-0.0457	-0.0457	-0.0457	-0.0321	-0.0277
	(0.0067)	(0.0064)	(0.0068)	(0.0075)	(0.0076)
3 PM to 6 PM x Treat	-0.0528	-0.0475	-0.0500	-0.0339	-0.0320
	(0.0081)	(0.0078)	(0.0081)	(0.0084)	(0.0085)
6 PM to 9 PM x Treat	-0.0480	-0.0430	-0.0444	-0.0291	-0.0264
	(0.0083)	(0.0081)	(0.0084)	(0.0086)	(0.0087)
9 PM to Midn. x Treat	-0.0421	-0.0430	-0.0445	-0.0292	-0.0264
	(0.0084)	(0.0080)	(0.0083)	(0.0085)	(0.0086)
Observations	19,253,016	19,252,988	19,251,882	19,252,988	19,251,882
School FE, Block FE	Yes	Yes	Yes	Yes	Yes
School-Block FE	No	Yes	Yes	Yes	Yes
School-Block-Month FE	No	No	Yes	No	Yes
Month of Sample FE	No	No	No	Yes	No
Month of Sample Ctrl.	No	No	No	No	Yes

Notes: Standard errors clustered at the school level. Dependent variable is the prediction error in log electricity consumption consumption by blocks of three hours.

As shown in the regression results, our data include both HVAC and lighting interventions. Using the alternative machine learning approach, we also explore whether the treatment effects across interventions appear to be different. Table 2.4.8 presents the results for schools with only HVAC interventions and those with only lighting interventions. We find results that are broadly consistent with the findings in Table 2.3.3, although substantially larger in some specifications.

For HVAC, we find that there is no effect at night, and most of the effects are highest during sunlight hours, between 9am and 3pm, and are statistically significant between 3pm and 6pm. We also find some positive (not significant) results in the early hours of the morning, which can be explained by potential ramping up of new HVAC systems, since these ramps tend to be better timed in newer systems. The results appear to be less sensitive across specifications than in the case of regression. For lighting, we find that there is a treatment effect both during the day and at night, although the effect is largest during the day and only significant during school hours, suggesting that the interventions appear to be most useful during active hours of operation. The results appear to be stable and large, in the order of 3-7% during the day.

Table 2.4.8: Prediction Results by Type of Intervention

	(1)	(2)	(3)	(4)	(5)
HVAC Interventions: (aggregate)	-0.0444	-0.0447	-0.0464	-0.0319	-0.0292
	(0.0081)	(0.0081)	(0.0083)	(0.0092)	(0.0091)
Midn. to 3 AM x Treat	-0.0451	-0.0457	-0.0471	-0.0332	-0.0315
	(0.0102)	(0.0094)	(0.0096)	(0.0099)	(0.0099)
3 AM to 6 AM x Treat	-0.0387	-0.0407	-0.0421	-0.0282	-0.0265
	(0.0102)	(0.0094)	(0.0097)	(0.0099)	(0.0100)
6 AM to 9 AM x Treat	-0.0175	-0.0139	-0.0166	-0.0016	-0.0010
	(0.0075)	(0.0070)	(0.0072)	(0.0080)	(0.0079)
9 AM to Noon x Treat	-0.0355	-0.0401	-0.0430	-0.0278	-0.0273
	(0.0070)	(0.0066)	(0.0068)	(0.0077)	(0.0076)
Noon to 3 PM x Treat	-0.0393	-0.0412	-0.0426	-0.0289	-0.0271
	(0.0073)	(0.0070)	(0.0073)	(0.0080)	(0.0079)
3 PM to 6 PM x Treat	-0.0438	-0.0441	-0.0471	-0.0318	-0.0315
	(0.0094)	(0.0091)	(0.0094)	(0.0096)	(0.0096)
6 PM to 9 PM x Treat	-0.0514	-0.0472	-0.0494	-0.0346	-0.0338
	(0.0098)	(0.0095)	(0.0098)	(0.0100)	(0.0100)
9 PM to Midn. x Treat	-0.0487	-0.0488	-0.0502	-0.0364	-0.0345
	(0.0098)	(0.0090)	(0.0093)	(0.0095)	(0.0095)
Observations	14,939,789	14,939,763	14,938,822	14,939,763	14,938,822
Light Interventions: (aggregate)	-0.0620	-0.0626	-0.0638	-0.0510	-0.0478
	(0.0093)	(0.0092)	(0.0096)	(0.0105)	(0.0105)
Midn. to 3 AM x Treat	-0.0277	-0.0352	-0.0358	-0.0237	-0.0192
	(0.0108)	(0.0106)	(0.0108)	(0.0110)	(0.0112)
3 AM to 6 AM x Treat	-0.0240	-0.0355	-0.0370	-0.0241	-0.0205
	(0.0112)	(0.0107)	(0.0110)	(0.0111)	(0.0113)
6 AM to 9 AM x Treat	-0.0269	-0.0311	-0.0329	-0.0198	-0.0163
	(0.0085)	(0.0077)	(0.0078)	(0.0088)	(0.0089)
9 AM to Noon x Treat	-0.0596	-0.0610	-0.0617	-0.0498	-0.0452
	(0.0078)	(0.0072)	(0.0076)	(0.0087)	(0.0089)
Noon to 3 PM x Treat	-0.0641	-0.0587	-0.0572	-0.0476	-0.0408
	(0.0081)	(0.0076)	(0.0080)	(0.0090)	(0.0092)
3 PM to 6 PM x Treat	-0.0675	-0.0561	-0.0567	-0.0450	-0.0402
	(0.0102)	(0.0097)	(0.0101)	(0.0104)	(0.0107)
6 PM to 9 PM x Treat	-0.0565	-0.0475	-0.0481	-0.0361	-0.0316
	(0.0103)	(0.0100)	(0.0103)	(0.0106)	(0.0107)
9 PM to Midn. x Treat	-0.0425	-0.0436	-0.0439	-0.0321	-0.0273
	(0.0106)	(0.0105)	(0.0107)	(0.0109)	(0.0111)
Observations	12,940,115	12,940,089	12,939,223	12,940,089	12,939,223
School FE, Block FE	Yes	Yes	Yes	Yes	Yes
School-Block FE	No	Yes	Yes	Yes	Yes
School-Block-Month FE	No	No	Yes	No	Yes
Month of Sample FE	No	No	No	Yes	No
Month of Sample Ctrl.	No	No	No	No	Yes

Notes: Standard errors clustered at the school level. Dependent variable is the prediction error in log electricity consumption consumption by blocks of three hours. Only treated schools that experienced either an HVAC or light upgrade are included.

2.4.4 Machine Learning Robustness

In order to demonstrate that these machine learning results are robust, we provide two tests in this section. We first demonstrate that we recover similar estimates to our main results when we "predict backwards in time." That is, in our main results, we use only pre-treatment data to train the machine learning model, and use this model to generate predictions about counterfactual electricity use in the post-treatment period.[26] As demonstrated above, we find that energy efficiency upgrades cause a reduction in consumption of between 2 and 4 percent.

In this exercise, we instead train the model on the *post*-treatment data, and generate predictions about counterfactual electricity consumption in the *pre*-treatment period to create this "trained-on-post" model (as opposed to our original "trained-on-pre" version). We now expect to see an increase in electricity consumption when we move from the training sample to the testing sample, because the training sample now includes the treatment effect, and the testing sample does not. In our analysis, we scale the treatment effects generated by this trained-on-post model by -1, such that the magnitudes of the changes generated by the two models are directly comparable. This exercise provides evidence about the validity of the identifying assumption that the treated and untreated groups are not trending differently; if these groups were exhibiting similar trends, the main analysis and this trained-on-post analysis would differ. Table 2.4.9 presents the results of this exercise.

The first panel of Table 2.4.9 displays results that are analogous to our main regressions, where we use the pre-treatment data to forecast into the post-treatment period. The dependent variable is the log of the prediction error. Note here that the explanatory variable of interest is a post-treatment dummy variable, which will allow us to make a clean comparison between this regression and the model where we forecast from the post-treatment period into the pre-treatment period.[27] When we implement this model, as expected, we find similar results to Table 2.4.7 above. The next panel presents the results of using the trained-on-post model. We find similar results to the trained-on-pre model across all specifications. If anything,

[26] As described above, we assign untreated schools a randomized "treatment" date.

[27] The reason that we implement this regression rather than a regression with treatment fractions on the right-hand side is that this would require that we split the sample into training and test sets at different points across the two models - in the normal model, splitting at the point where any treatment is implemented; in the trained-on-post model at the point where all treatments have been implemented. We intend to do this in the future, but re-implementing the machine learning procedure is computationally intensive.

Table 2.4.9: Prediction Results: Pre- vs. Post-Period Training

	(1)	(2)	(3)	(4)	(5)
Trained on pre	-0.0451	-0.0465	-0.0498	-0.0251	-0.0397
	(0.0077)	(0.0077)	(0.0082)	(0.0082)	(0.0082)
Observations	19,253,016	19,252,988	19,251,882	19,252,988	19,251,882
Trained on post	-0.0461	-0.0485	-0.0531	-0.0309	-0.0486
	(0.0078)	(0.0078)	(0.0083)	(0.0082)	(0.0085)
Observations	19,103,288	19,101,250	19,100,030	19,101,250	19,100,030
Pooled	-0.0393	-0.0412	-0.0420	-0.0413	-0.0420
	(0.0074)	(0.0074)	(0.0075)	(0.0074)	(0.0075)
Observations	38,350,824	38,343,320	38,343,004	38,343,320	38,343,004
School FE, Block FE	Yes	Yes	Yes	Yes	Yes
School-Block FE	No	Yes	Yes	Yes	Yes
School-Block-Month FE	No	No	Yes	No	Yes
Month of Sample FE	No	No	No	Yes	No
Month of Sample Ctrl.	No	No	No	No	Yes

Notes: Standard errors clustered at the school level. Dependent variable is the prediction error in log electricity consumption consumption by blocks of three hours. Independent variable is an indicator, equal to zero in the training period and equal to one in the post-training period. The training period in the "Trained on pre" panel is the pre-treatment period; the training period in the "Trained on post" panel is the post-treatment period. In the "Trained on post" and "pooled" panels, the dependent variable is normalized such that the sign of the effect matches the "Trained on pre" panel.

the estimates from this trained-on-post model are slightly larger. Finally, we show the results from a model where we pool both trained-on-pre and trained-on-post observations. Here, we find again find nearly identical results. These results provide confidence in our machine learning estimates.

For a second robustness check, we implement a variant on the double selection procedure described by Belloni, Chernozhukov, and Hansen (2014). In the standard version of this procedure (typically used in cross-sectional models), the econometrician fits a first LASSO to predict selection into treatment, and stores the variables with non-zero coefficients. Next, she fits a second LASSO to predict the outcome of interest, again storing the non-zero-coefficient variables. Finally, she regresses the outcome variable on the treatment indicator, as well as the non-zero-coefficient re-

gressors from *both* LASSOs. This procedure allows for proper inference after model selection.

In our panel data setting, school fixed effects preclude us from including time invariant predictors of selection into treatment in our regressions. We modify the double selection procedure to fit our context. We first estimate a LASSO to predict the timing of treatment, next estimate a second LASSO to predict electricity consumption, and finally estimate a third LASSO with time as the dependent variable, in order to handle trends. Similar to the standard double selection procedure, we then regress energy consumption on treatment timing and the union of the non-zero-coefficient variables from all three LASSOs.[28] This procedure allows us to control for selection into treatment timing, which is potentially endogenous in our context, and also allows us to do proper inference in a context where we have used LASSO for model selection. Table 2.4.10 displays the results of this exercise for any intervention, HVAC interventions, and lighting interventions.

Note that because we are predicting the timing of treatment, the double selection procedure is designed to work for one specification only. In particular, Table 2.4.10 presents the results of the double selection procedure in a model with school-block fixed effects and a month of sample control. There are two noteworthy features of Table 2.4.10. First, the effects of energy efficiency upgrades on electricity consumption estimated via double-selection are remarkably similar to specifications (4) and (5) of our main estimates, which are the closest comparisons. This suggests that selection into the timing of treatment is not substantially biasing our main effects. Second, the standard errors from the double selection procedure are substantially smaller than those from our main estimates. This is consistent with results from Belloni, Chernozhukov, and Hansen (2014), and, since the double selection procedure is designed in part to enable proper inference, suggests that the standard errors in our main estimates are quite conservative.

[28] To make this computationally tractable, we residualize each dependent variable by the full set of controls, and implement the final step by regressing residualized prediction errors on residualized treatment date error and residualized time error. These procedures are mathematically equivalent, via Frisch-Waugh-Lovell.

Table 2.4.10: Double Selection Prediction Results

	(1) Any Intervention	(2) HVAC Interventions	(3) Lighting Interventions
Treatment (aggregate):	-0.0265	-0.0271	-0.0341
	(0.0065)	(0.0076)	(0.0089)
Midn. to 3 AM x Treat	-0.0211	-0.0264	-0.0171
	(0.0079)	(0.0092)	(0.0108)
3 AM to 6 AM x Treat	-0.0189	-0.0239	-0.0201
	(0.0080)	(0.0094)	(0.0109)
6 AM to 9 AM x Treat	-0.0095	-0.0049	-0.0228
	(0.0062)	(0.0072)	(0.0084)
9 AM to Noon x Treat	-0.0345	-0.0292	-0.0522
	(0.0062)	(0.0068)	(0.0082)
Noon to 3 PM x Treat	-0.0374	-0.0328	-0.0518
	(0.0064)	(0.0071)	(0.0085)
3 PM to 6 PM x Treat	-0.0369	-0.0335	-0.0490
	(0.0073)	(0.0088)	(0.0100)
6 PM to 9 PM x Treat	-0.0283	-0.0335	-0.0350
	(0.0076)	(0.0092)	(0.0103)
9 PM to Midn. x Treat	-0.0258	-0.0322	-0.0266
	(0.0077)	(0.0089)	(0.0107)
Observations	18,955,082	14,739,263	12,676,858
School FE, Block FE	Yes	Yes	Yes
School-Block FE	Yes	Yes	Yes
School-Block-Month FE	No	No	No
Month of Sample FE	No	No	No
Month of Sample Ctrl.	Yes	Yes	Yes

Notes: Standard errors clustered at the school level. Dependent variable is the prediction error in log electricity consumption consumption by blocks of three hours. This table displays the results of estimating a model with school-by-hour-block fixed effects where prediction models are generated via the double selection approach.

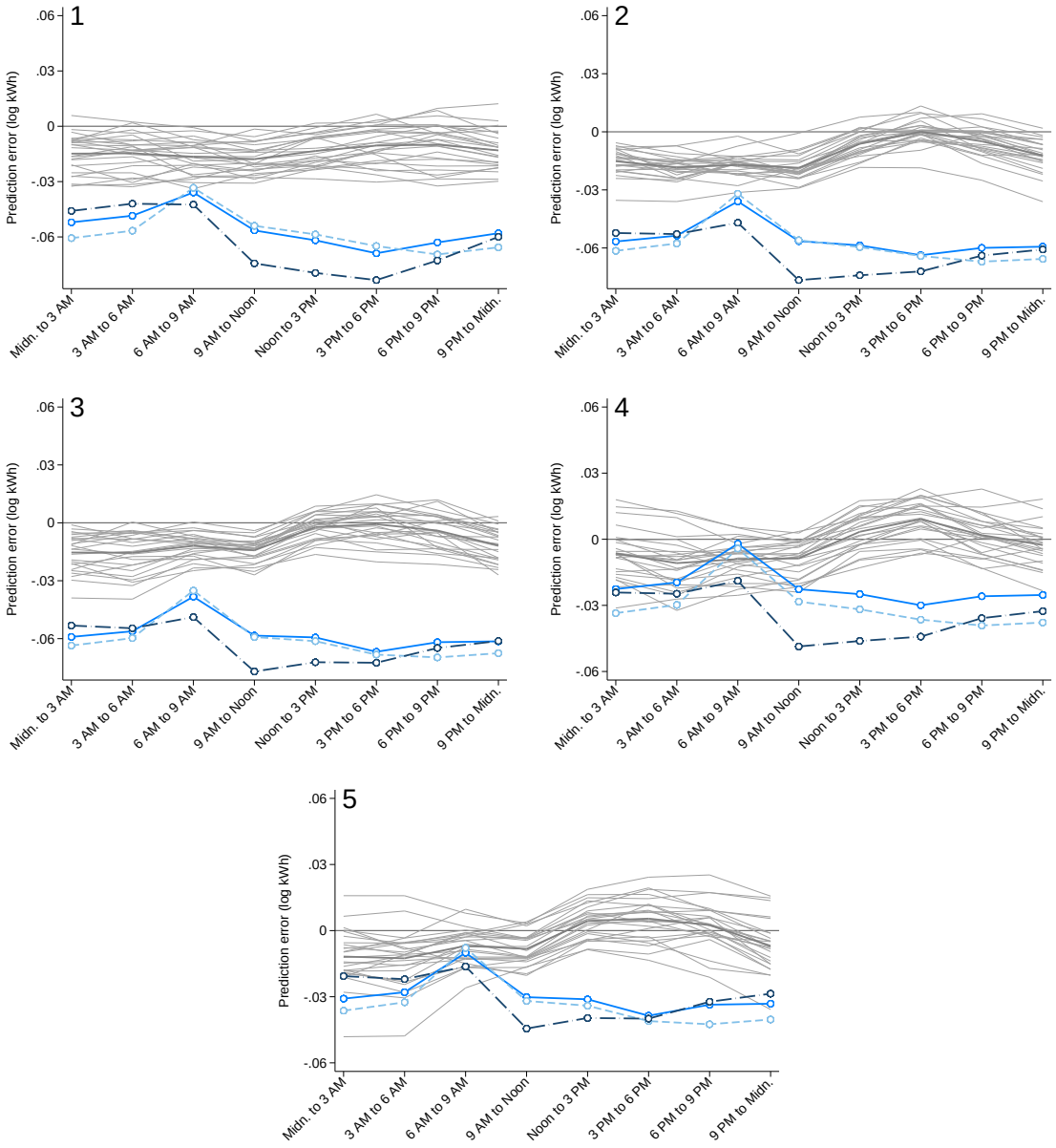
2.4.5 Methods Comparison: Results

The results from the machine learning approach are broadly consistent with our findings using the difference-in-differences approach. However, there are some differences. One major difference is that the traditional regressions are more sensitive to the inclusion of fixed effects and controls. An extreme version of this difference is the fact that regression results without including school fixed effects are severely biased, whereas the school-specific predictions inherent to the machine learning method already removes substantial differences across schools. This means that fixed effects in the prediction model do not need to absorb as many remaining differences across schools. It is clear that that estimates from the machine learning approach appear to be larger, which is consistent with our initial hypothesis and Monte Carlo results, and could be due to the presence of measurement error and/or specification bias.

One way to see the power of the machine learning methodology is by running a series of placebo simulations akin to the ones we implemented for the difference-in-differences methodology. As above, we drop all post-treatment observations, thereby eliminating usage data that are exposed to treatment, and randomly assign approximately 40 percent of schools to be "treated". For the prediction placebos, and for computational reasons, we keep the same random treatment date that we used to generate the prediction.

Figure 2.4.6 presents a series of placebo tests with 25 placebo realizations for the prediction regressions. As was the case for the difference-in-differences placebos, the average placebo effects are centered around zero for all the simulations. Comparing the two sets of placebo results (Figures 2.3.3 and Figures 2.4.6), one can see that the hourly patterns that previously emerged in the difference-in-differences placebos are no longer as systematic for the machine learning estimates. Thus, the machine learning methodology appears to more successfully control for compositional changes over time.

Figure 2.4.6: Placebo Treatment Effects – Machine Learning



Notes: This figure presents machine learning treatment effects using real and placebo data. Each panel corresponds to the column of the same number from Tables 2.4.7 and 2.4.8. The light gray lines present the placebo effects, with the solid gray line showing the average placebo effect. Placebo effects are generated using pre-treatment data only, and randomly assigning treatment status and timing according to the distribution present in real data. The solid blue line shows treatment effects from any upgrade; the dashed aqua line shows treatment effects for HVAC upgrades; and the dash-dotted navy line shows treatment effects from lighting upgrades, all using real data. Unlike when we use a difference-in-difference method (as in Figure 2.3.3), we see little to no hourly patterns in the placebo treatment effects with machine learning.

2.5 Realized versus Ex-Ante Predicted Energy Savings

Both the regression and machine learning approach provide estimates that suggest average energy savings at schools that range between 2 to 5%, depending on the specification. While these effects provide an average sense of the treatment effect, the interventions across schools can be quite different, ranging from replacing a few fixtures, to upgrading a whole HVAC system. Therefore, it might be difficult to interpret such measures in terms of success or failure of these interventions.

In this section, we take advantage of the fact that each intervention has an engineering estimate of expected savings as a first step to understand the effectiveness of these heterogeneous measures. We estimate a treatment effect that is proportional to these measures of expected savings, so that interventions that are quite different in scope can be easily compared.

2.5.1 Assessing realized vs. expected savings

To estimate the effectiveness of the interventions when compared to ex-ante expected savings, we consider the following set of regressions (and variants):

$$Y_{ith} = -\beta S_{it} + \alpha_i + \kappa_h + \gamma_t + \epsilon_{ith},$$

where Y represents electricity consumption for the difference-in-differences approach, and prediction error for the machine learning methodology. The variable S_i represents expected ex-ante energy savings for a given intervention, which are estimated at the annual level. In particular, we define the expected energy savings at any point in time to only include the energy savings that belong to efficiency projects that have already taken place.

The main coefficient of interest is β. Note that these regressions are in levels, as opposed to logs, so that a coefficient of one can be interpreted as ex-post expected savings matching on average ex-ante energy savings. A coefficient larger than one would suggest that the observed energy savings are larger than those estimated ex-ante. On the contrary, an estimate smaller than one would suggest that the ex-post realized savings are not as large as anticipated.

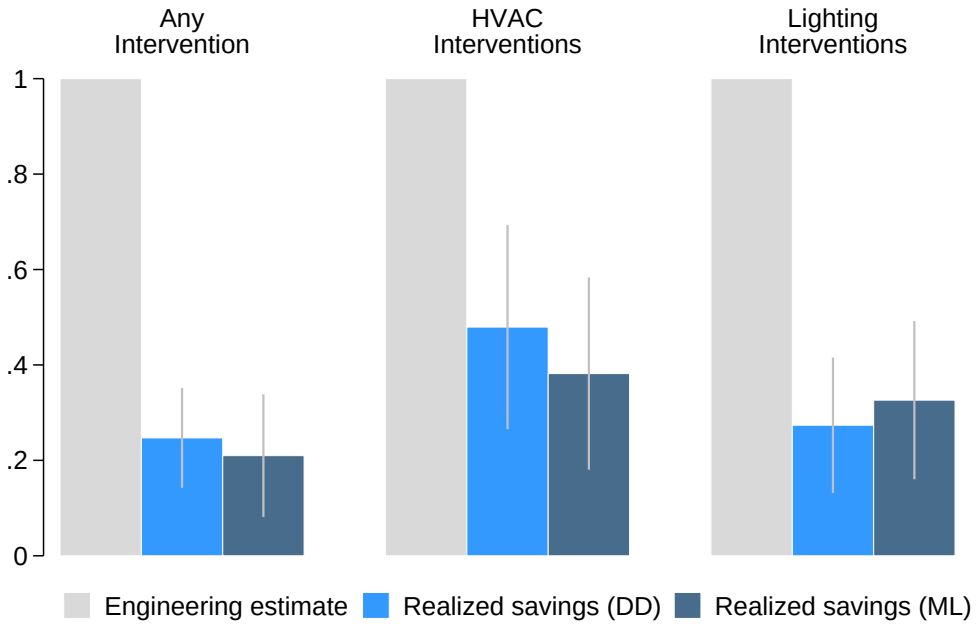
Table 2.5.11 presents the results for all interventions, HVAC interventions, and lighting interventions across specifications. In each panel, we present the estimates

Table 2.5.11: Ex-Post vs. Ex-Ante Savings

	(1)	(2)	(3)	(4)	(5)
Any Intervention:					
Energy consumption (kWh)	0.3008	0.3031	0.3167	0.2411	0.2470
	(0.0531)	(0.0536)	(0.0561)	(0.0531)	(0.0533)
Prediction error (kWh)	0.2517	0.2522	0.2534	0.2043	0.2098
	(0.0657)	(0.0663)	(0.0683)	(0.0650)	(0.0656)
Observations	19,249,960	19,249,932	19,248,920	19,249,932	19,248,920
HVAC Interventions:					
Energy consumption (kWh)	0.5834	0.5877	0.6230	0.5179	0.4793
	(0.1030)	(0.1040)	(0.1112)	(0.0999)	(0.1092)
Prediction error (kWh)	0.4637	0.4658	0.4668	0.4075	0.3819
	(0.1004)	(0.1014)	(0.1065)	(0.1012)	(0.1027)
Observations	14,939,791	14,939,764	14,938,902	14,939,764	14,938,902
Light Interventions:					
Energy consumption (kWh)	0.4001	0.4023	0.4189	0.2771	0.2736
	(0.0643)	(0.0643)	(0.0673)	(0.0693)	(0.0724)
Prediction error (kWh)	0.4153	0.4151	0.4168	0.3343	0.3262
	(0.0762)	(0.0769)	(0.0776)	(0.0851)	(0.0846)
Observations	12,940,115	12,940,088	12,939,300	12,940,088	12,939,300
School FE, Block FE	Yes	Yes	Yes	Yes	Yes
School-Block FE	No	Yes	Yes	Yes	Yes
School-Block-Month FE	No	No	Yes	No	Yes
Month of Sample FE	No	No	No	Yes	No
Month of Sample Ctrl.	No	No	No	No	Yes

Notes: Standard errors clustered at the school level. Estimates report the coefficient on expected savings, scaled over time as projects get implemented. A coefficient of one implies that ex-post savings were realized as expected on average.

Figure 2.5.7: Estimated realization rates



Notes: This figure displays the realization rates from various energy efficiency upgrades, estimated from a specification with school-by-block-by-month fixed effects and a month of sample control (Column (5) in Table 2.5.11). Difference in difference (DD) estimates use the log of electricity consumption as the dependent variable; machine learning (ML) estimates use the log of the prediction error as the dependent variable. Standard errors are clustered at the school level.

for the regression approach, in which the dependent variable is electricity consumption followed by estimates based on the machine learning approach, in which the dependent variable is prediction error from the machine learning model. Figure 2.5.7 displays these same results in graphical form, using the regression specification in Column (5) of Table 2.5.11.

The results suggest that realized savings are consistently lower than predicted ex ante. None of the estimated coefficients, which are akin to realization rates, are above 50 percent and some are closer to 20 percent. Across interventions, the results suggest that expected energy savings from HVAC and lighting interventions match ex ante expected savings more closely than other interventions. These highlights

some of the difficulties in assessing energy efficiency investment in the presence of very heterogeneous measures.

Energy efficiency policy discussions sometimes distinguish between "net" and "gross" savings, where the former subtracts energy efficiency investments customers would make absent a utility program. Because the machine learning approach provides school-specific counterfactuals, it allows us to disentangle the extent to which untreated schools (i.e., schools who are not receiving rebates from the utility) are also reducing their consumption over the time period. For example, comparing estimates of $\hat{\beta}^U$ and $\hat{\beta}^T$ in Table 2.4.6 suggests that untreated schools are not reducing consumption over time. This suggests that the low realized savings are not driven by unmeasured efficiency upgrades at untreated schools, and are more likely driven by overly optimistic ex ante predictions or rebound.

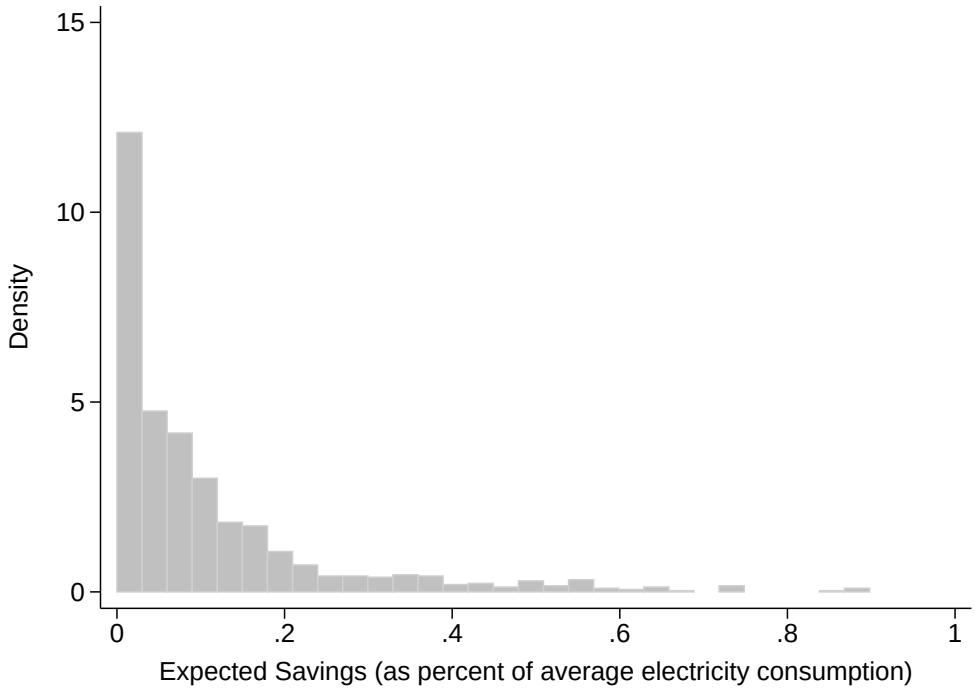
2.5.2 Sensitivity to outliers

When evaluating the effectiveness of energy efficiency upgrades, measurement error is a substantial concern. There are many ways in which energy efficiency upgrade data can fail to reflect on-the-ground realities. One plausible dimension of mismeasurement is in upgrade dates: schools may differ in whether they report the date an upgrade was initiated vs. completed vs. paid for; or installations may occur over multiple days; or there may simply be enumeration error. As shown in the Monte Carlo exercise in Section 2.4.2, mismeasured treatment dates can drive large wedges between real and estimated treatment effects. We demonstrate that our machine learning procedure is less biased than a standard regression model in the presence of this type of measurement error.

A second important dimension of measurement error is in expected savings. Our own estimates, as well as other recent work (e.g. Fowlie, Greenstone, and Wolfram 2015a) suggest that ex ante engineering estimates do not accurately reflect real-world savings. One potential reason for this discrepancy in our context is expected savings estimates that are implausibly large. Figure 2.5.8 shows the distribution of expected savings in the population of schools that implement energy efficiency upgrades. Most interventions represents savings of less than 10% of a school's average electricity consumption. However, there is a long tail of interventions that represent much larger savings.

There is vast heterogeneity across measures, which highlights challenges in estimating the effectiveness of these energy efficiency interventions. Futhermore, Figure

Figure 2.5.8: Ex-ante energy savings

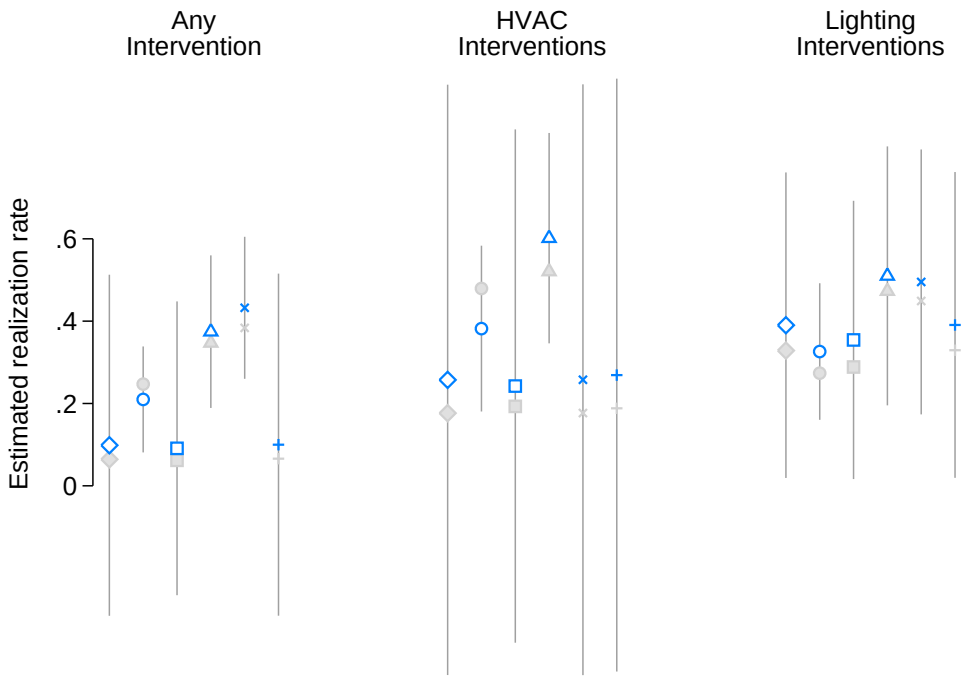


Notes: This figure displays the distributions of ex-ante expected savings as a percent over average electricity consumption. Only schools with interventions (i.e., non-zero expected savings are reported). The figure excludes schools for which expected savings are above 100%.

2.5.8 highlights measurement issues in the data. For example, we find that some schools have reported expected energy savings that exceed their average electricity consumption in our data, which cannot be possibly true. The measurement issues could be due to some measurement error in expected savings, or due to a mismatch between schools and interventions. The clear mismeasurement in expected savings displayed in Figure 2.5.8 has real consequences for policy, as ex ante expected savings estimates are being used to determine whether measures pass a cost-benefit test.

Furthermore, this type of measurement error will lead our ex post analysis to estimate small realization rates: if a school installs an upgrade that, ex ante, is projected to save 100% of its energy consumption, but in fact this upgrade only saves 5% of consumption, we will estimate a 5% realization rate. If this same upgrade were ex ante estimated to reduce consumption by 10%, we would instead estimate a 50%

Figure 2.5.9: Realization rates: sensitivity to outliers



Notes: This figure displays estimated realization rates for any intervention, HVAC interventions, and lighting interventions across a range of samples, estimated estimated from a specification with school-by-block-by-month fixed effects and a month of sample control. Blue markers indicate point estimates and gray lines indicate 95% confidence intervals from the machine learning method; gray markers indicate point estimates from the regression method. Each symbol shape represents a different sample. Diamonds display results using the full sample. Circles display results trimming 1st and 99th percentile outliers in prediction errors across the full sample – this is the main trim performed throughout the paper. Squares trim 1st and 99th percentile outliers in prediction errors within each school. Triangles trim 1st and 99th percentile schools by average electricity consumption. Xs trim 1st and 99th percentile schools by total expected savings. +s trim 1st and 99th percentile schools by total expected savings divided by consumption. There is a large range of estimates, but the point estimates never exceed 0.65.

realization rate. In order to check the sensitivity of our estimates to outliers, we re-estimate our main results under a variety of sample trimming exercises.

Figure 2.5.9 displays the results of this exercise. Using a specification with school-by-block-by-month fixed effects and a month of sample control, the machine learning method estimates realization rates for any intervention ranging from 0.091 (trimming the 1st and 99th percentile of prediction errors within school) to 0.432 (trimming at the 1st and 99th percentile of expected savings). In general, removing the smallest and largest schools from the sample yields the largest realization rates - but across upgrade types, these rates never exceed 0.601 (estimated using HVAC, trimming schools by average electricity consumption). This range of results highlights the challenges inherent in measuring the returns to energy efficiency upgrades when a central parameter, the expected savings, is measured with substantial error. Despite these difficulties, however, nearly all of the estimates fall within the confidence interval of our main estimate (trimming the 1st and 99th percentile of prediction errors within school, displayed with circles). We can reject realization rates of 1 in all cases, but cannot always reject realization rates of 0. This suggests that while our realization rate estimates do vary with our choice of sample, even our most favorable estimates imply that energy efficiency investments undertaken in schools fall short of delivering the expected energy savings.

2.6 Conclusions

We study the incidence and impacts of energy efficiency investments at public K-12 schools in California. Using high-frequency hourly electricity data, we estimate the treatment effect of energy efficiency projects at treated schools, leveraging the presence of control schools that did not participate in these programs. We explore two complementary methodologies: a regression approach and a machine learning approach. We find that results are comparable across methodologies and a battery of controls, although machine learning estimates are systematically larger. We discuss in the methodological section how these differences can be explained by potential measurement error and specification bias.

Focusing on HVAC and lighting interventions, we find that these energy efficiency investments delivered about 2 to 4% electricity consumption savings on average, when compared to control schools. These energy savings appear to be a substantial share of ex-ante expected savings for the lighting and HVAC interventions, with actual savings predicted to be around 70-90% of projected savings. However, realized savings appear to be noisily measured and small when we consider a wider battery of measures, with estimates of at most 15% of ex-ante expected savings. Heterogeneity in how expected savings are defined, as well as other sources of measurement error (e.g., in treatment date), could be driving some of these results.

Chapter 3

Panel Data and Experimental Design[1]

3.1 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.[2] 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.[3]

[1] The material in this chapter is from Energy Institute at Haas Working Paper #277, coauthored with Louis Preonas and Matt Woerman. The original version can be found online at `https://ei.haas.berkeley.edu/research/papers/WP277.pdf`.

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

[3] 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

In a widely cited paper based on results from Frison and Pocock (1992), McKenzie (2012) recommends considering 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 data.[4] As data collection becomes cheaper and easier, 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".[5]

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. 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 er-

McKenzie (2012)). Baird et al. (2014) extends the classic cross-sectional setup to randomized saturation designs, capable of measuring spillover and general equilibrium effects. Athey and Imbens (2016a) discusses statistical power using a randomization inference approach.

[4]These include Bloom et al. (2013), Blattman, Fiala, and Martinez (2014), Jessoe and Rapson (2014), Bloom et al. (2015), Fowlie, Greenstone, and Wolfram (2015b), and Fowlie et al. (2016).

[5]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).

rors, 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 3.2 provides background on power calculations. Section 3.3 presents analytical power calculations expressions for panel data, and demonstrates their importance in Monte Carlo simulations with serially correlated errors. Section 3.4 applies these results to real experimental data. Section 3.5 discusses a few alternative estimation strategies and extensions. Section 3.6 discusses issues in designing panel RCTs and conducting power calculations in practice. Section 3.7 concludes.

3.2 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 under treatment and control conditions 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:

$$MDE = \left(t^d_{1-\kappa} + t^d_{\alpha/2}\right) \sqrt{\text{Var}\left(\hat{\tau} \mid \mathbf{X}\right)} \tag{3.1}$$

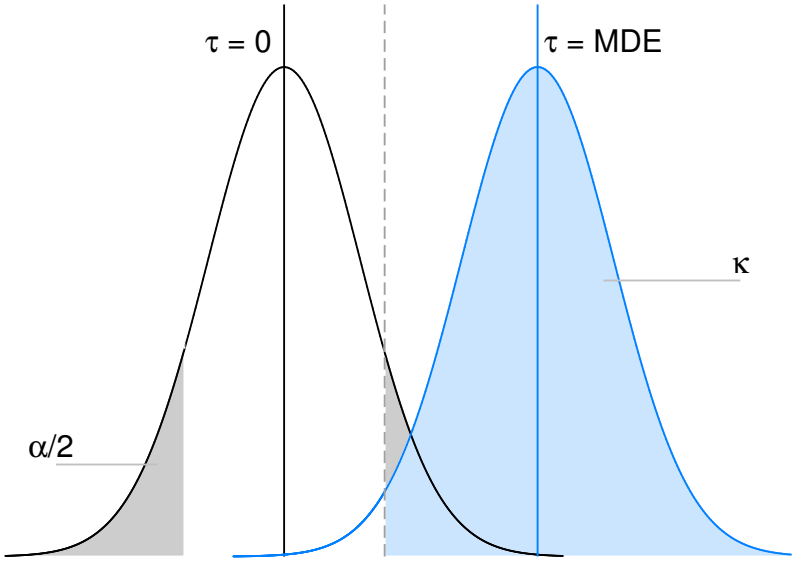
where $\text{Var}(\hat{\tau} \mid \mathbf{X})$ is the exact finite sample variance of the treatment effect estimator, conditional on independent variables $\mathbf{X}$; $t^d_{\alpha/2}$ is the critical value of a t distribution with d degrees of freedom associated with the probability of a Type I error, 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 3.2.1 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 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 3.2.1 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:

[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.

Figure 3.2.1: 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.

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* $\mathbf{X} = [\beta \; D]$*. In practice, this follows from random assignment of D_i.*

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

$$\mathrm{E}[\hat{\tau} - \tau \mid \mathbf{X}] = 0$$

$$\mathrm{Var}(\hat{\tau} \mid \mathbf{X}) = \frac{\sigma_\varepsilon^2}{P(1-P)J}$$

$$MDE = \left(t_{1-\kappa}^{J-2} + t_{\alpha/2}^{J-2}\right)\sqrt{\frac{\sigma_\varepsilon^2}{P(1-P)J}} \tag{3.2}$$

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 D, $J - 1$ unit fixed effects, and $(m + r) - 1$ time fixed effects. 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.*

[7] See Appendix B.1.1.1 for a full derivation of the variance of $\hat{\tau}$ in this model.

[8] This is the standard model used in panel RCTs. We assume here that these models are properly specified.

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

$$\mathrm{E}[\hat{\tau} - \tau \mid \mathbf{X}] = 0$$
$$\mathrm{Var}(\hat{\tau} \mid \mathbf{X}) = \left(\frac{\sigma_\omega^2}{P(1-P)J}\right)\left(\frac{m+r}{mr}\right)$$
$$MDE = \left(t_{1-\kappa}^{J} + t_{\alpha/2}^{J}\right)\sqrt{\left(\frac{\sigma_\omega^2}{P(1-P)J}\right)\left(\frac{m+r}{mr}\right)} \tag{3.3}$$

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, $\sigma_\omega^2 \leq \sigma_\varepsilon^2$ 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 (3.3) in terms of σ_ε^2. Let ρ_υ and ρ_δ denote the proportion of the composite variance σ_ε^2 contributed by σ_υ^2 and σ_δ^2, respectively:

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

Then, (3.3) can be rewritten as:

$$MDE = \left(t_{1-\kappa}^{J} + t_{\alpha/2}^{J}\right)\sqrt{\left(\frac{\sigma_\varepsilon^2}{P(1-P)J}\right)\left(\frac{m+r}{mr}\right)(1-\rho_\upsilon-\rho_\delta)} \tag{3.4}$$

[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 B.1.2.1 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.

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 (3.4) 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.

3.3 Theory

3.3.1 Correlated errors in panel models

Power calculation formulas such as the standard cross-sectional model (Equation (3.2)) and the FP model (Equation (3.3)) 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 cases where treatment is randomly assigned across cross-sectional units, however, the OLS variance estimator is an unbiased estimator of the true variance, even in the presence of cross-sectional error correlations:

Lemma 1. *In a cross-sectional model with random assignment to treatment,* $\frac{\sigma_\varepsilon^2}{P(1-P)J}$ *is an unbiased estimator of* $\text{Var}(\hat{\tau} \mid \mathbf{X})$ *even if* $\text{E}[\varepsilon_i \varepsilon_j \mid \mathbf{X}] \neq 0$ *for some* $i \neq j$. *See Appendix B.1.3 for a proof.*[12]

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

[11] Replacing $\rho_\delta = 0$, $P = 0.5$, and $J = 2n$, Equation (3.4) is equivalent to the power calculation formula for difference-in-differences derived by FP and discussed in McKenzie (2012). See Appendix B.1.2.1 for complete derivations. Equation (3.3) 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 $\sigma_\delta^2 > 0$, in keeping with assumptions economists typically make about data generating processes. Hence, Equation (3.3) 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 B.1.3, we provide a proof which allows for arbitrary cross-sectional error dependence. Athey and Imbens (2016a, 2016b) still recommend using Eicker-Huber-White standard errors in this case, to allow for heteroskedasticity. We do not know of a paper that discusses power calculations in the presence of heteroskedastic disturbances.

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.

3.3.2 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, including arbitrary correlations within cross-sectional units over time, which are present in real-world data.[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-

[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 (3.3)) 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.

[15]We do not consider cross-sectional correlation, 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 B.1.2.3.

treatment periods, and $D_{it} = 1$ in post-treatment periods; for control units, $D_{it} = 0$ in all periods.

Assumption 6 (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 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 D, $J - 1$ unit fixed effects, and $(m + r) - 1$ time fixed effects. 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.*

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:*

$$\psi_T^B \equiv \frac{2}{PJm(m-1)} \sum_{i=1}^{PJ} \sum_{t=-m+1}^{-1} \sum_{s=t+1}^{0} \mathrm{Cov}\left(\omega_{it}, \omega_{is} \mid \mathbf{X}\right)$$

$$\psi_T^A \equiv \frac{2}{PJr(r-1)} \sum_{i=1}^{PJ} \sum_{t=1}^{r-1} \sum_{s=t+1}^{r} \mathrm{Cov}\left(\omega_{it}, \omega_{is} \mid \mathbf{X}\right)$$

$$\psi_T^X \equiv \frac{1}{PJmr} \sum_{i=1}^{PJ} \sum_{t=-m+1}^{0} \sum_{s=1}^{r} \mathrm{Cov}\left(\omega_{it}, \omega_{is} \mid \mathbf{X}\right)$$

to be the average pre-treatment, post-treatment, and across-period covariance between different error terms of the same treated unit, respectively. Define ψ_C^B, ψ_C^A, and ψ_C^X analogously, where we consider the $(1 - P)J$ control units instead of the PJ treated units. Using these definitions, assume that $\psi^B = \psi_T^B = \psi_C^B$; $\psi^A = \psi_T^A = \psi_C^A$; and $\psi^X = \psi_T^X = \psi_C^X$.[16]

[16] 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

The OLS estimator with unit and time fixed effects remains $\widehat{\tau} = (\ddot{D}'\ddot{D})^{-1}\ddot{D}'\ddot{Y}$ and again, $\mathrm{E}[\hat{\tau} - \tau \mid \mathbf{X}] = 0$. However, Assumptions 21–10 extend FP to a more general power calculation formula that incorporates arbitrary within-unit correlations:[17]

$$\mathrm{Var}(\hat{\tau} \mid \mathbf{X}) = \left(\frac{1}{P(1-P)J}\right)\left[\left(\frac{m+r}{mr}\right)\sigma_\omega^2 + \left(\frac{m-1}{m}\right)\psi^B + \left(\frac{r-1}{r}\right)\psi^A - 2\psi^X\right] \tag{3.5}$$

$$MDE = (t^J_{1-\kappa} + t^J_{\alpha/2})\sqrt{\left(\frac{1}{P(1-P)J}\right)\left[\left(\frac{m+r}{mr}\right)\sigma_\omega^2 + \left(\frac{m-1}{m}\right)\psi^B + \left(\frac{r-1}{r}\right)\psi^A - 2\psi^X\right]} \tag{3.6}$$

Throughout the remainder of the paper, we refer to Equation (3.6) 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 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* $\mathrm{Var}(\hat{\tau} \mid \mathbf{X})$*, even in the presence of arbitrary within-period cross-sectional correlations. See Appendix B.1.3 for a proof, and see Appendix B.1.2.3 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, $\vec{\omega}$, and the corresponding variance-covariance matrix, Ω, can be

periods $\{1, \ldots, r\}$. In a randomized setting, $\mathrm{E}\left[\psi_T^B\right] = \mathrm{E}\left[\psi_C^B\right]$, $\mathrm{E}\left[\psi_T^A\right] = \mathrm{E}\left[\psi_C^A\right]$, and $\mathrm{E}\left[\psi_T^X\right] = \mathrm{E}\left[\psi_C^X\right]$, making this a reasonable assumption *ex ante*. However, it is possible for treatment to alter the covariance structure of treated units only.

[17] We present the formal derivation of this formula in Appendix B.1.2.2. Note that if $m = 1$ (or $r = 1$), ψ^B (or ψ^A) is not defined and is multiplied by 0 in Equation (3.6).

represented as follows:[18]

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

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 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

[18] 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.

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:

$$\left(\frac{m-1}{m}\right)\psi^B + \left(\frac{r-1}{r}\right)\psi^A > 2\psi^X \tag{3.7}$$

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.

3.3.3 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:

$$Y_{it} = \beta + \upsilon_i + \delta_t + \omega_{it} \tag{3.8}$$

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

$$\omega_{it} = \gamma\omega_{i(t-1)} + \xi_{it} \tag{3.9}$$

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, 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.[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^{FP} varies only with panel length.

Figure 3.3.2 displays the results of this exercise. The left column shows rejection rates under the FP formula using OLS standard errors, which assumes zero serial

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

[20] We provide these formal derivations in Appendix B.2.1, along with further details on these Monte Carlo simulations. We also provide additional simulation results that separately vary m and r in Appendix B.3.

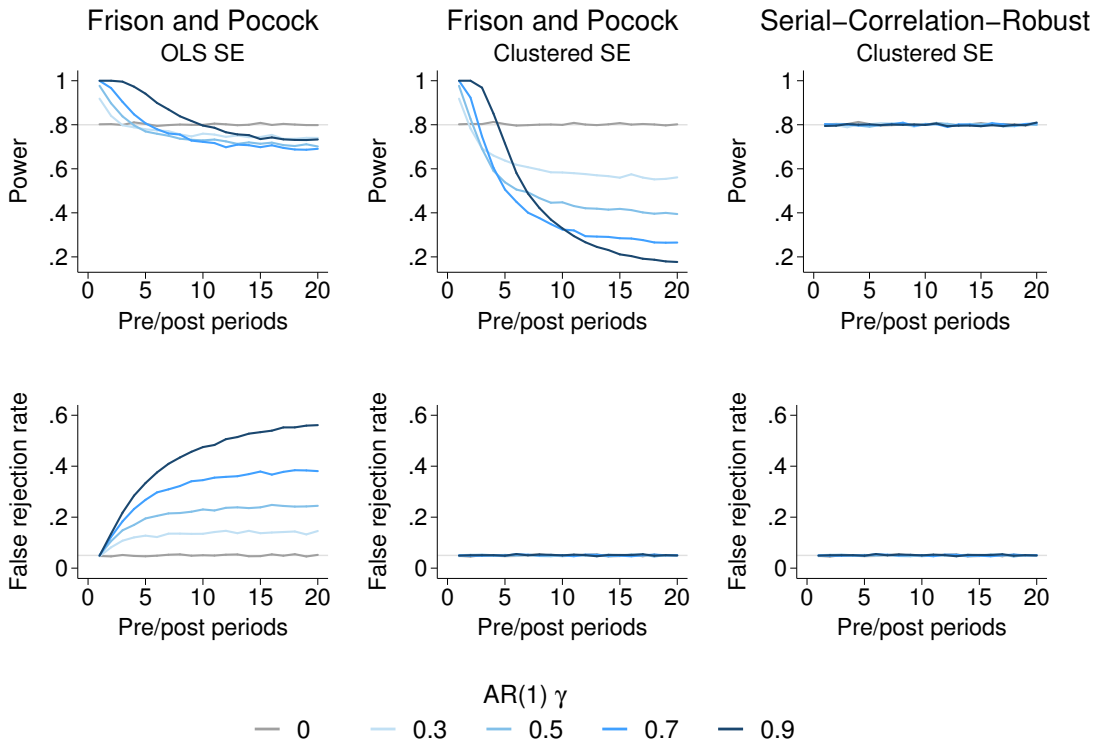
correlation both *ex ante* and *ex post*. The middle column shows rejection rates under the FP formula using CRVE standard errors, which accounts for serial correlation *ex post* only. The right column show rejection rates under our SCR formula using CRVE standard errors, which allows for serial correlation both *ex ante* and *ex post*. The top row plots realized power as a function 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 (3.3) 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 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 (3.7). For shorter panels, using the FP formula instead of our SCR formula yields dramatically overpowered experiments. While this may seem counterintuitive, (3.7) 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

Figure 3.3.2: 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 (3.3)), 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.

γ yields more dramatically underpowered experiments for longer panels, higher γ yields more dramatically *over*powered experiments for shorter panels.[21]

[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,

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. This exercise suggests that, in theory, researchers must account for serial correlation in experimental design. 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*.

3.4 Applications to real-world data

3.4.1 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. The authors estimate the following equation to derive the central result, which is found in Table 2 in the original paper:

$$Performance_{it} = \alpha Treat_i \times Experiment_t + \beta_t + \gamma_i + \varepsilon_{it} \tag{3.10}$$

This is a standard DD estimating equation with fixed effects for individual i and week t. The paper's dataset consists of weekly performance measures for the 249 workers enrolled in the experiment between January 2010 and August 2011. We perform simulations using data from the pre-treatment period only, to avoid contamination from the Bloom et al. (2015) intervention. 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.[22] Table 3.4.1 provides summary statistics, including the AR(1) $\hat{\gamma}$ parameter recovered by estimating

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 B.3 for additional short-panel results.

[22]This leaves us with a very different sample from Bloom et al. (2015). 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.

Equation (3.9) on residuals from this dataset. Our estimate of $\hat{\gamma} = 0.233$ is highly statistically significant, which indicates that these worker performance data exhibit weak serial correlation.

Table 3.4.1: Summary statistics – Bloom et al. (2015)

Mean	Std. Dev.	Min	Max	AR(1) $\hat{\gamma}$	Individuals	Periods	Observations
0.153	0.943	−2.766	3.665	0.233	79	48	3,792

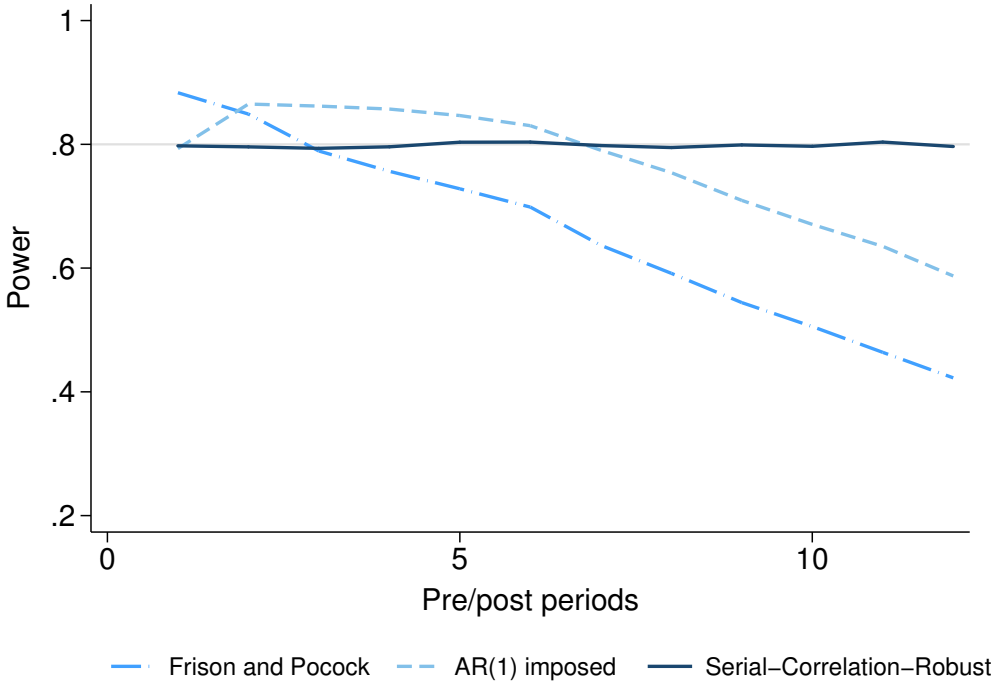
Notes: This table shows summary statistics for worker productivity in the Bloom et al. (2015) data. The data are weekly job performance z-scores, constructed by taking the average of normalized performance measures, where each measure is standardized to have a mean of 0 and standard deviation of 1 across the sample. Our sample consists only of individuals that had no missing observations throughout the entire pre-treatment period, January 1, 2010 through November 28, 2010. We compute $\hat{\gamma}$ by estimating Equation (3.9) on residuals from this dataset. In doing so, we cluster standard errors at the individual level. The 95% confidence interval is [0.165, 0.300]. For more details on the standardized job performance measures and the actual experimental design, see Bloom et al. (2015).

We perform Monte Carlo simulations on this dataset which 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 panel length, we randomly assign three treatment effect sizes, τ^{FP}, $\tau^{AR(1)}$, and τ^{SCR}, at the individual level and estimate Equation (3.10) separately for each treatment effect size. We calibrate τ^{FP} using the FP formula, assuming zero serial correlation and estimating $\sigma^2_{\hat{\omega}}$ using residuals from these data. We calibrate $\tau^{AR(1)}$ using the SCR formula, deriving the ψ parameters analytically by assuming an AR(1) error structure with $\gamma = 0.233$. We calibrate τ^{SCR} using the SCR formula, where we non-parametrically estimate $\psi^B_{\hat{\omega}}$, $\psi^A_{\hat{\omega}}$, and $\psi^X_{\hat{\omega}}$ from residuals.[23]

Figure 3.4.3 reports the results of this exercise, demonstrating that only the SCR formula achieves the desired statistical power in the Bloom et al. (2015) data. In contrast, failing to account for serial correlation leads to experiments that deviate dramatically from the desired 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.

[23]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. Appendix B.4.1 outlines how to estimate these parameters, and Appendix B.2.2 provides further details on this simulations.

Figure 3.4.3: 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 (3.9). The solid line applies the serial-correlation-robust formula, where we non-parametrically generate estimates of $\psi_{\tilde{\omega}}^{B}$, $\psi_{\tilde{\omega}}^{A}$, and $\psi_{\tilde{\omega}}^{X}$ 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.

This confirms our previous finding from simulated data: in a panel data setting with serial correlation, the FP equation does not yield the desired power. By contrast, simulations calibrated with the SCR formula yield 80 percent power, even using real experimental data. These results show that it is possible to design a panel RCT to achieve 80 percent power, so long as the *ex ante* formula properly accounts for the within-unit correlation structure of the data.

3.4.2 Pecan Street data

Having demonstrated the importance of properly accounting for serial correlation using data from an actual RCT in a developing country setting, we now turn to a much higher-frequency dataset on residential electricity consumption in the United States. Electricity data are particularly well-suited for studying power calculations for panel RCTs. Hourly electricity data exhibit high within-household autocorrelation, so we should expect to have to adjust our standard errors, and also our power calculations, accordingly. Furthermore, we can aggregate the hourly data by averaging hourly electricity consumption over different time spans, each of which exhibits a distinct correlation structure.

We use publicly available household electricity data from Pecan Street (2016), a research organization based at the University of Texas at Austin, which makes high-resolution energy usage data available to academic researchers. We downloaded a dataset of hourly electricity consumption, measured in kilowatt hours (kWh), from 699 households over 26,888 hours.[24] We restrict the sample to households that appear in every hour between January 1, 2013 and December 31, 2014 in order to construct a balanced panel, and we drop any household that ever reports zero consumption.[25] We are left with a final sample of 97 households over 17,520 hours. We create three additional datasets by collapsing these hourly data into daily, weekly, and monthly averages for each household.

Table 3.4.2 presents basic summary statistics, and Figure 3.4.4 displays the time series of data for one randomly selected household in our sample at varying levels of aggregation. Unsurprisingly, while mean electricity use is consistent across different collapses of the data, the standard deviation decreases as we move from higher- to lower-frequency datasets.

As observation frequency decreases, the number of observations declines from 1,699,440 total observations over 17,520 periods in the hourly data to 2,328 observations over 24 periods in the monthly dataset. Taken together, this suggests that different frequencies will shed light on the performance of different power calculation approaches, since the hourly data likely has a different correlation structure than the monthly data. Finally, RCTs using energy consumption data are becoming increas-

[24]The raw data are available with a login from `https://dataport.pecanstreet.org/data/interactive`.

[25]These observations are likely to be errors.

ingly common in economics, making our Pecan Street application relevant to this growing literature.[26]

Table 3.4.2: Summary statistics – Pecan Street

Dataset	Mean	Std. Dev.	Min	Max	AR(1) $\hat{\gamma}$	Households	Periods	Observations
Hourly	1.200	1.164	0.019	13.501	0.628	97	17,520	1,699,440
Daily	1.200	0.789	0.082	6.013	0.651	97	730	70,810
Weekly	1.198	0.739	0.105	5.175	0.713	97	106	10,282
Monthly	1.197	0.712	0.169	4.296	0.654	97	24	2,328

Notes: This table shows summary statistics for electricity consumption in the Pecan Street data. All values are in average kW of electricity consumed. The raw data are at the hourly level, in kWh. To construct the daily, weekly, and monthly dataset, we average hourly kWh consumption data across the relevant time period. We compute $\hat{\gamma}$ by estimating Equation (3.9) on residuals from this dataset. In doing so, we cluster standard errors at the household level. All are statistically significant at less than the 1% level.

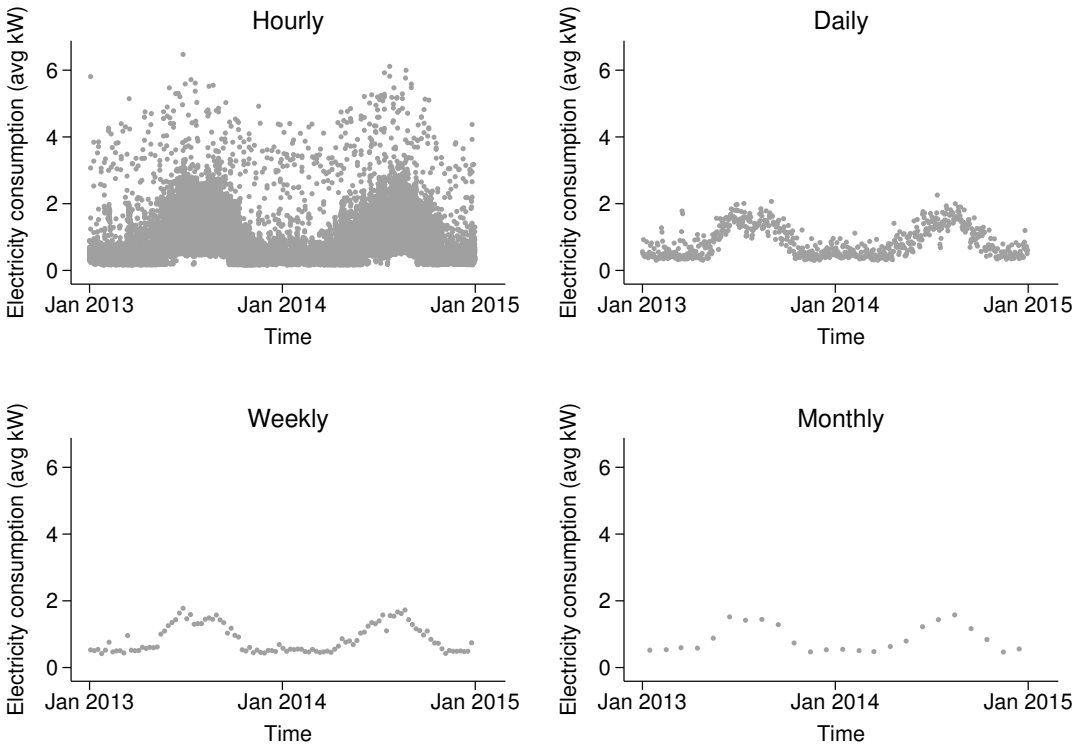
We conduct Monte Carlo simulations on all four versions of the Pecan Street data to assess the performance of alternative power calculation assumptions. These follow the exact same procedure as the Bloom et al. (2015) simulations, performed separately for the hourly, daily, weekly, and monthly datasets. We subset consecutive periods to create panels of differing lengths, and we calibrate τ^{FP} using the FP formula, $\tau^{AR(1)}$ using the SCR formula assuming AR(1) errors, and τ^{SCR} using the SCR formula, non-parametrically estimating the $\psi_{\hat{\omega}}$ terms to leverage the full correlation structure of each subsetted dataset.[27] Again, the realized power for each set of simulations is the rejection rate of the null hypothesis, $\tau = 0$, across 10,000 iterations.

Figure 3.4.5 shows that in all four versions of the Pecan Street data, improperly accounting for serial correlation *ex ante* yields realized power that sharply deviates from the desired 80 percent. Neither the FP assumption of i.i.d. errors nor an assumed AR(1) structure *ex ante* manages to account for the serial correlation present in electricity consumption data. As with the Bloom et al. (2015) simulations, we achieve correctly powered experimental designs only by applying the SCR method and accounting for the full covariance structure of the Pecan Street data. This result holds at both high and low temporal frequencies.

[26]For example, see Allcott (2011), Jessoe and Rapson (2014), Ito, Ida, and Tanaka (2015), Fowlie, Greenstone, and Wolfram (2015b), and Fowlie et al. (2016). There is also a large quasi-experimental literature that uses energy consumption data.

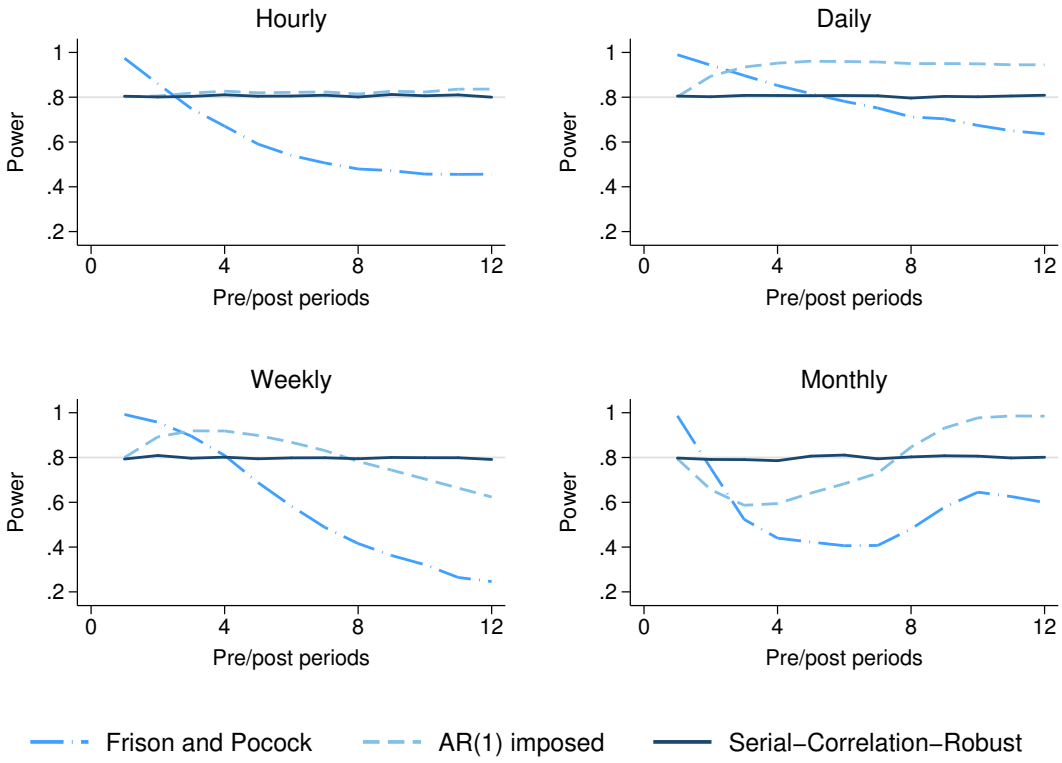
[27]See Appendix B.2.3 for details on the Pecan Street simulations.

Figure 3.4.4: Pecan Street data – Varying levels of aggregation



Notes: This figure shows the time series of Pecan Street electricity consumption data for one randomly selected household in our sample. Each panel displays the data at a different level of aggregation. The data are in units of average kW. These data are highly serially correlated: when we estimate an AR(1) model, Equation (3.9), on residuals from these datasets, we recover the AR(1) parameters of 0.628, 0.651, 0.713, and 0.654 for the hourly, daily, weekly, and monthly data, respectively.

Figure 3.4.5: 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 (3.3)). The short-dashed lines use Equation (3.6) assuming an AR(1) error structure, where we estimate the AR(1) parameters by estimating Equation (3.9) separately for each dataset. The solid lines apply Equation (3.6), by estimating separate covariance components of $\psi_{\hat{\omega}}^{B}$, $\psi_{\hat{\omega}}^{A}$, and $\psi_{\hat{\omega}}^{X}$ 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.

3.4.3 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, 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.[28] 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.[29]

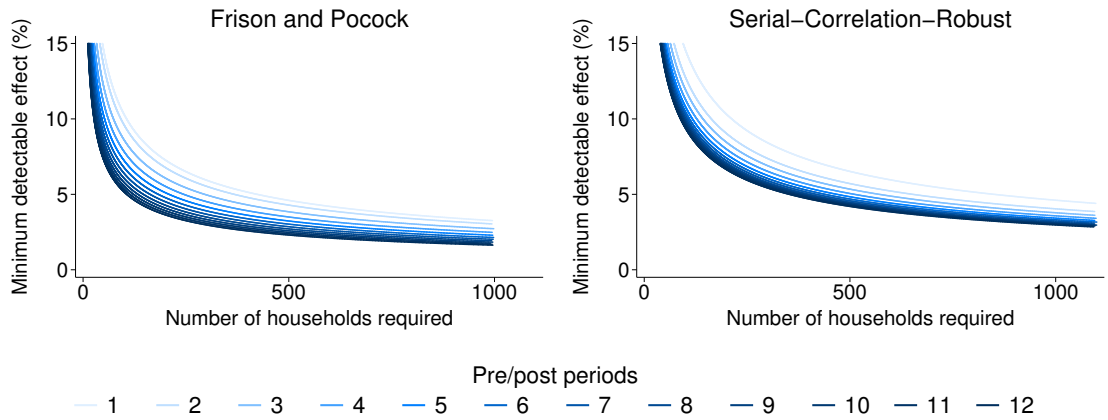
We plot the results of this exercise in Figure 3.4.6. The left panel depicts power calculations using the FP formula, which assumes an i.i.d. error structure. The right panel applies the SCR formula to compute the number of units required for the same set of MDEs, using our non-parametric estimates of $\psi^B_{\hat{\omega}}$, $\psi^A_{\hat{\omega}}$, 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 smaller number of households than the SCR power calculation. Hence, if an researcher in this setting applies

[28] Appendix B.4.1 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 B.5 proves that power calculations using estimated parameters recover the same MDE in expectation as those using true parameters. Whether estimating these parameters is plausible 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 Section 3.6 for more details on how to overcome a lack of pre-experimental data.

[29] See Appendix B.2.4 for details.

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.

Figure 3.4.6: 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.

3.5 Discussion

3.5.1 Collapsing data

Above, we demonstrate the importance of considering serial correlation when performing power calculations. Applying the SCR power calculation formula will appropriately account for within-unit correlation, yielding correctly powered experiments. This is analogous to using CRVE standard errors to recover the correct false rejection rate. BDM suggest an alternative way of achieving the correct false rejection rates in the presence of serial correlation: ignore the time-series structure of the data by averaging the pre-treatment and post-treatment data for each unit, then estimate a

two-period panel DD regression on this collapsed dataset and apply the OLS variance estimator.

While this approach does recover the desired false rejection rate, simply applying the FP formula to a collapsed dataset will *not* yield correctly powered experiments. The FP formula for collapsed data is:

$$MDE = \left(t_{1-\kappa}^{J-2} + t_{\alpha/2}^{J-2}\right) \sqrt{\left(\frac{2}{P(1-P)J}\right) \sigma_{\omega^C}^2} \tag{3.11}$$

where $\sigma_{\omega^C}^2$ is the variance of ω_{it}^C, the error terms in the *collapsed* data. In contrast, the SCR formula with data collapsed to one pre- and one post-treatment period is:

$$MDE = \left(t_{1-\kappa}^{J-2} + t_{\alpha/2}^{J-2}\right) \sqrt{\left(\frac{2}{P(1-P)J}\right) \left(\sigma_{\omega^C}^2 - \psi^X\right)} \tag{3.12}$$

where ψ^X is the same average covariance term for the uncollapsed data.[30] Even when using a collapsed dataset, the FP formula fails to account for the cross-period covariance represented by ψ^X, and will therefore yield an incorrect MDE. While this may seem to contradict the BDM result, the two results are in fact consistent.

Intuitively, to achieve the desired false rejection rate using the OLS variance estimator on collapsed data, it need only be the case that this estimator produce an unbiased estimate of the true variance. Because the collapsed model is a two-period panel, it is isomorphic to a (cross-sectional) first-difference model. Since this model has only one observation per unit, it cannot exhibit within-unit correlation over time. This means that the OLS variance estimator is unbiased, and so too is the OLS variance estimator from the equivalent two-period panel model. A researcher aiming to achieve the correct false rejection rate can be agnostic to the actual value of $\text{Var}(\hat{\tau})$; she simply needs to know that the estimator of $\text{Var}(\hat{\tau})$ will be unbiased.

In contrast, for a power calculation, the researcher must know (or have an estimate of) $\text{Var}\left(\hat{\tau} \mid \mathbf{X}\right)$. Applying the first-difference model to the collapsed two-period

[30] The collapsed model begins with Assumptions 21–10, and then averages all m pre-treatment observations and r post-treatment observations for each unit, in order to estimate a two-period pre/post panel with $2J$ observations. We can express the idiosyncratic error variance of this collapsed model, $\sigma_{\omega^C}^2$, in terms of uncollapsed parameters, where $\sigma_{\omega^C}^2 = \frac{1}{2}\Big[\left(\frac{m+r}{mr}\right)\sigma_\omega^2 + \left(\frac{m-1}{m}\right)\psi^B + \left(\frac{r-1}{r}\right)\psi^A\Big]$. This shows that Equation (3.12) is identical to the uncollapsed SCR formula in Equation (3.6). By contrast, Equation (3.11) differs from the uncollapsed FP formula in Equation (3.3), because $\sigma_{\omega^C}^2$ captures a portion of the serial correlation (represented by ψ^B and ψ^A). We provide this proof in Appendix B.1.2.4.

panel, the resulting expression for the variance of the errors of the collapsed-and-first-differenced data is $\sigma^2_{\Delta\omega^C} = 2\left(\sigma^2_{\omega^C} - \psi^X\right)$. That is, we can represent the variance of the collapsed data with a single term, which incorporates both the idiosyncratic variance and serial correlation.[31] To accurately perform power calculations for a collapsed DD model, a researcher may either plug $\sigma^2_{\Delta\omega^C}$ into the *cross-sectional* power calculation formula (Equation (3.2)), or apply the SCR formula which directly incorporates both $\sigma^2_{\omega^C}$ and ψ^X. In both cases, generating appropriately powered experiments requires accounting for ψ^X, and therefore accounting for serial correlation — simply plugging $\sigma^2_{\omega^C}$ into FP will not yield the desired result.

3.5.2 ANCOVA

The vast majority of experiments that have been published in top economics journals, and of ongoing experiments that have been pre-registered at the American Economics Association's RCT registry, estimate treatment effects using analysis of covariance (ANCOVA) methods. These regression specifications include only post-treatment observations, and control for the average pre-treatment level of the outcome variable $\overline{Y}_{i,pre} = \frac{1}{m}\sum_{t\in m} Y_{it}$, and post-period fixed effects δ_t:[32]

$$Y_{it} = \beta + \tau D_i + \theta \overline{Y}_{i,pre} + \delta_t + \varpi_{it} \tag{3.13}$$

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) comments that, 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.

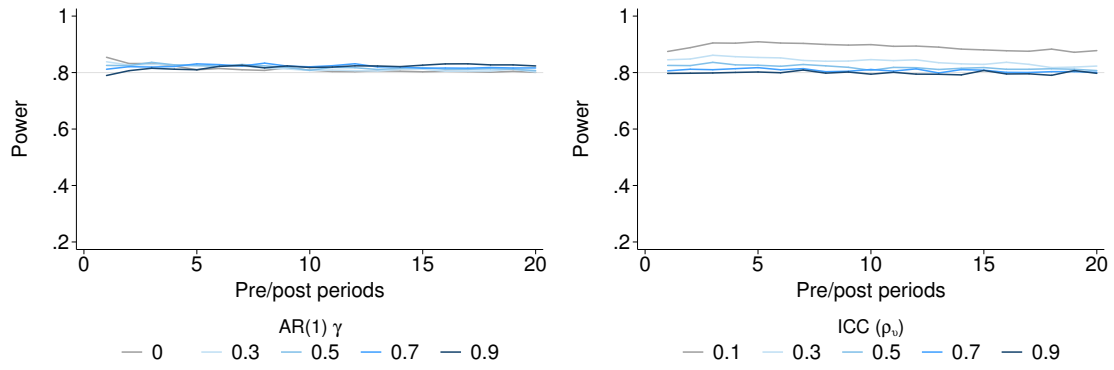
We examine the relative benefits of ANCOVA in settings with serial correlation, by conducting an analogous simulation exercise to that in Section 3.3. As above, we generate data with an AR(1) parameter governing the error structure. We use the SCR power calculation formula to calibrate τ^{MDE}, such that τ^{MDE} equals the

[31] For example, when compared to data with no serial correlation, a collapsed dataset with positive serial correlation will exhibit smaller first-difference error terms, and hence a smaller error variance.

[32] This model again follows Assumptions 21–10, and then averages all m pre-treatment observations for each unit, in order to estimate a model with Jr post-treatment observations only. Note that in most cases, researchers collect one pre-treatment ("baseline") wave of data, and one post-treatment ("follow-up") wave of data. This yields the cross-sectional ANCOVA specification: $Y_{i,post} = \beta + \tau D_i + \theta Y_{i,pre} + \varpi_i$.

minimum detectable effect for a panel fixed effects model, and randomly assign treatment at the unit level. Whereas above we estimate $\hat{\tau}$ using a panel DD model, here we estimate $\hat{\tau}$ using Equation (3.13). We again record whether we reject the null of $\tau = 0$, and repeat 10,000 times. We do this twice: once, as above, for varying levels of the AR(1) parameter γ; and also for varying levels of intracluster correlation coefficient ρ_υ.[33] Figure 3.5.7 presents the results of this exercise, on the same scale as Figure 3.3.2.

Figure 3.5.7: The benefits of ANCOVA are limited with serial correlation



Notes: This figure displays realized power from performing power calculations on data generated with AR(1) processes, where we calibrate the treatment effect size using the serial-correlation-robust power calculation formula from a panel DD model, but estimate the treatment effect using the ANCOVA specification of Equation (3.13). A realized power of 0.80 indicates that the ANCOVA estimator is exactly as efficient as the DD estimator; power above 0.80 indicates that ANCOVA is more efficient; power below 0.80 indicates that ANCOVA is less efficient. In left panel, we display realized rejections rates for different panel lengths, while varying the level of serial correlation γ (setting $m = r$, with $\rho_\upsilon = 0.5$). In the right panel, we alternatively vary the intracluster correlation coefficient ρ_υ for different panel lengths (keeping $m = r$, with $\gamma = 0.5$). We apply the cluster-robust variance estimator in all cases. While the ANCOVA estimator is generally more efficient than the DD estimator, these gains are relatively small in the presence of serial correlation. As expected, the gains decay as ρ_υ increases.

Across the range of AR(1) γ values, intracluster correlations, and panel lengths, estimation using ANCOVA provides realized power greater than 80 percent. This means that ANCOVA is indeed more efficient than a panel DD estimator, even in the presence of serial correlation. However, the magnitude of these power gains

[33]We adjust ρ_υ by varying σ^2_υ. See Appendix B.2.1 for further detail on these ANCOVA simulations.

from ANCOVA is quite modest, especially for moderate levels of γ and moderate-to-large ρ_υ. It is intuitive that introducing serial correlation and proper inference will erode the relative efficiency gains of ANCOVA. In a world with i.i.d. errors, much of ANCOVA's efficiency advantage relative to DD results from its use of just one degree of freedom in predicting the post-treatment outcome with pre-treatment data (i.e. in estimating θ), as opposed to J degrees of freedom for J unit fixed effects in a DD model. In settings with serial correlation, however, both ANCOVA and DD require the CRVE with J degrees of freedom, thereby eroding the degrees-of-freedom advantage of ANCOVA. Our results suggest that ANCOVA only marginally improves the efficiency of DD in the presence of serial correlation.

3.5.3 Trading off J for m and r

Recruiting participants, administering treatment, and collecting data are all costly. Implementation costs are often the limiting factor in study size. Power calculations enable researchers to calibrate experimental design parameters to a particular MDE. 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:

$$\min_{P,J,m,r} MDE(P,J,m,r) \quad \text{s.t.} \quad p_T PJ + p_C(1-P)J + p_m Jm + p_r Jr \leq B \tag{3.14}$$

Here, p_T and p_C are the costs of including each treated and control unit in the experiment, respectively; p_m and p_r are the costs of collecting one pre- and post-treatment observation, respectively; 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. This tradeoff arises not only from differences in these costs, but also from differences in the marginal effects of additional units and time periods on the MDE. These marginal effects can be calculated by taking the partial derivative of the MDE formula with respect to each of these design parameters. Using the SCR power calculation formula, the "elasticities" of the MDE with respect to number of units and number of time periods are:

$$\frac{\partial MDE/MDE}{\partial J/J} = -\frac{1}{2}$$

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

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

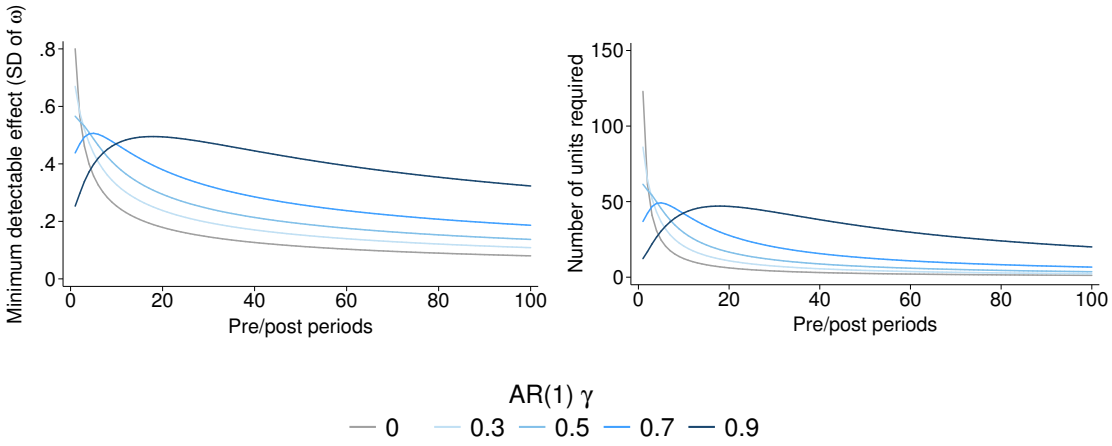
There is a constant elasticity of MDE with respect to J of -0.5, meaning that a 1 percent increase in the number of units always yields a 0.5 percent reduction in the MDE. However, the elasticity of MDE with respect to m and r varies as a function of the error structure and the number of time periods.[34] 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.[35]

Figure 3.5.8 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. We normalize the MDE to be in units of σ_ω. The right panel depicts the tradeoff between additional units and additional time periods, by plotting the combinations of J and $m = r$ that yield an MDE equal to σ_ω. To analytically

[34]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 Equations (B.39)–(B.41) of Appendix B.1.1.1).

[35]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.

Figure 3.5.8: 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.

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. In fact, for $\gamma \geq 0.6$, marginally increasing m or r in a relatively short panel actually increases the MDE for a given J, and likewise increases the J required to achieve a given MDE. While such non-monotonicities only occur in settings with a high degree of serial correlation, we recover $\hat{\gamma} > 0.6$ for all four versions of the Pecan Street dataset. This suggests that for experiments using highly correlated data, additional periods of data might decrease statistical power if the panel is not sufficiently long.

3.6 Power calculations in practice

3.6.1 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 DGPs 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 — 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.[36]

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).[37] 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. 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.

[36] See Appendix B.1.2.3 for the full derivation and resulting power calculation equation.

[37] 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 B.4.2.

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.

3.6.2 Practical challenges

All power calculations are a combination of art and science (Duflo, Glennerster, and Kremer (2007)). Even in cases where researchers are able to simulate power calculations using a large pre-existing dataset, they must assume that these data are representative of the underlying DGP governing their future experiment.[38] Imperfect pre-existing datasets can be used to inform *ex ante* experimental design, even if they are less representative than desired, and especially if they are smaller than the expected size of the experiment. For example, if a pre-existing dataset has too few cross-sectional units, researchers can bootstrap these units to construct a larger dataset for the purposes of simulating power calculations. Likewise, if a pre-existing dataset has too few time periods, it may be possible to estimate the error structure in this short panel in order to bound estimates of the ψ^B, ψ^A, and ψ^X covariance parameters and apply the analytical formula.[39]

In some settings, researchers may have no pre-existing data whatsoever, meaning that power calculations will require substantial guesswork. In this scenario, we recommend that researchers apply our SCR formula analytically, assuming a range of reasonable parameters. It may be possible to benchmark these "reasonable" parameter values to previous research, coupled with appropriate sensitivity analyses.

While accounting for non-i.i.d. errors increases the number of parameters needed to perform power calculations, ignoring the potential for serially correlated errors can significantly undermine an experiment's chances of success. Hence, it is important to consider the error structure *ex ante*, even in the absence of pre-existing data. In

[38]It is always possible that experimental data will differ systematically from a seemingly representative pre-existing dataset, due to either cross-sectional sampling differences or exogenous time trends affecting the outcome of interest.

[39]See Appendix B.4.3 for further detail on strategies to overcome shortcomings in pre-existing data.

any case, we recommend that researchers perform power calculations under a range of assumed parameter values, in order to test the sensitivity of their experiment's MDE to different assumptions about the data generating process and to experimental design parameters.

3.7 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 demonstrating that serial correlation matters, even when collapsing data or implementing ANCOVA methods; providing researchers with a framework for trading off units for effect sizes that takes cost into account; and discussing implementation of

power calculations in practice.[40] Ultimately, we provide researchers with new tools for experimental design that properly account for serial correlation in panel data settings, which enable the development of well-powered experiments with multiple waves of data.

[40]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`.

References

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

Allcott, Hunt, Allan Collard-Wexler, and Stephen D. O'Connell. 2016. "How Do Electricity Shortages Affect Industry? Evidence from India." *American Economic Review* 106 (3): 587–624.

Allcott, Hunt, and Michael Greenstone. 2012. "Is there an energy efficiency gap?" *The Journal of Economic Perspectives* 6 (1): 3–28.

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.

Asher, Sam, and Paul Novosad. 2016. "Market Access and Structural Transformation: Evidence from Rural Roads in India." `http://www.dartmouth.edu/~novosad/asher-novosad-roads.pdf`.

Athey, Susan, and Guido Imbens. 2015. "Recursive Partitioning for Heterogeneous Causal Effects." *Proceedings of the National Academy of Science* 113 (27): 7353–7360.

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

———. 2016b. "The State of Applied Econometrics – Causality and Policy Evaluation." Working Paper.

Baird, Sarah, J. Aislinn Bohren, Craig McIntosh, and Berk Ozler. 2014. "Designing Experiments to Measure Spillover Effects." World Bank Policy Research Working Paper WPS 6824.

Banerjee, Abhijit, Esther Duflo, and Nancy Qian. 2012. "On the Road: Access to Transportation Infrastructure and Economic Growth in China." NBER Working Paper 17897.

Barbose, Galen L., Charles A. Goldman, Ian M. Hoffman, and Megan A. Billingsley. 2013. "The future of utility customer-funded energy efficiency programs in the United States: projected spending and savings to 2025." *Energy Efficiency Journal* 6, no. 3 (January): 475–493.

Barron, Manuel, and Maximo Torero. 2016. "Household Electrification and Indoor Air Pollution." `https://ideas.repec.org/p/pra/mprapa/61424.html`.

Belloni, Alexandre, Victor Chernozhukov, and Christian Hansen. 2014. "Inference on Treatment Effects After Selection Amongst High-Dimensional Controls." *The Review of Economic Studies* 81 (2): 608–650.

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.

Boomhower, Judson, and Lucas Davis. 2016. "Do Energy Efficiency Investments Deliver at the Right Time?" *Energy Institute at Haas Working Paper,* no. WP 271.

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.

CEC. 2017. *Proposition 39: California Clean Energy Jobs Act, K-12 Program and Energy Conservation Assistance Act 2015-2016 Progress Report.* Technical report. California Energy Commission.

Central Electricity Authority. 2011. *Load Generation Balance Report: 2011-12.*

Chakravorty, Ujjayant, Kyle Emerick, and Majah-Leah Ravago. 2016. "Lighting up the last mile: The benefits and costs of extending electricity to the rural poor." `https://works.bepress.com/chakravorty/45/download/`.

Chand, T. R. Kiran, K. V. S. Badarinath, Christopher D. Elvidge, and B. T. Tuttle. 2009. "Spatial characterization of electrical power consumption patterns over India using temporal DMSP–OLS night–time satellite data." *International Journal of Remote Sensing* 30 (3): 647–661.

Chen, Xi, and William D. Nordhaus. 2011. "Using luminosity data as a proxy for economic statistics." *Proceedings of the National Academy of Sciences* 108 (21): 8589–8594.

Cicala, Steve. 2017. "Imperfect Markets versus Imperfect Regulation in U.S. Electricity Generation." *NBER Working Paper.*

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

Davis, Lucas, Alan Fuchs, and Paul Gertler. 2014. "Cash for coolers: evaluating a large-scale appliance replacement program in Mexico." *American Economic Journal: Economic Policy* 6 (4): 207–238.

Dinkelman, Taryn. 2011. "The Effects of Rural Electrification on Employment: New Evidence from South Africa." *American Economic Review* 101 (7): 3078–3108.

Donaldson, Dave. Forthcoming. "Railroads of the Raj: Estimating the Impact of Transportation Infrastructure." *American Economic Review.*

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.

EIA. 2015. *Electric Power Monthly.* Technical report. Energy Information Administration, November.

Eichholtz, Piet, Nils Kok, and John M. Quigley. 2013. "The Economics of Green Building." *Review of Economics and Statistics* 95 (1): 50–63.

Faber, Benjamin. 2014. "Trade Integration, Market Size, and Industrialization: Evidence from China's National Trunk Highway System." *The Review of Economic Studies* 83 (2): 1046–1070.

Fisher, Anthony, Michael Haneman, Michael Roberts, and Wolfram Schlenker. 2012. "The Economic Impacts of Climate Change: Evidence from Agricultural Output and Random Fluctuations in Weather: Comment." *American Economic Review* 102 (7): 1749–1760.

Fowlie, Meredith, Michael Greenstone, and Catherine Wolfram. 2015a. "Do Energy Efficiency Investments Deliver? Evidence from the Weatherization Assistance Program." *Mimeograph.*

———. 2015b. "Do Energy Efficiency Investments Deliver? Evidence from the Weatherization Assistance Program." NBER Working Paper 21331.

Fowlie, Meredith, Catherine Wolfram, C. Anna Spurlock, Annika Todd, Patrick Baylis, and Peter Cappers. 2016. "Default Effects, Follow-on Behavior and Welfare in Residential Electricity Pricing Programs." Working Paper.

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.

Gerarden, Todd D, Richard G Newell, and Robert N Stavins. 2015. *Assessing the Energy-Efficiency Gap.* Technical report. Harvard Environmental Economics Program, January.

Gertler, Paul, Orie Shelef, Catherine Wolfram, and Alan Fuchs. 2016. "The Demand for Energy-Using Assets among the World's Rising Middle Classes." *American Economic Review* 106 (6): 1366–1401.

Gillingham, Kenneth, and Karen Palmer. 2014. "Bridging the energy efficiency gap: policy insights from economic theory and empirical evidence." *Review of Environmental Economics and Policy* 8, no. 1 (January): 18–38.

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

Henderson, J. Vernon, Adam Storeygard, and David N. Weil. 2012. "Measuring Economic Growth from Outer Space." *American Economic Review* 102 (2): 994–1028.

IEA. 2015. *World Energy Outlook.* Technical report. International Energy Agency, June.

Imbens, Guido W., and Thomas Lemieux. 2008. "Regression discontinuity designs: A guide to practice." *Journal of Econometrics* 142 (2): 615–635.

Imbens, Guido, and Karthik Kalyanaraman. 2012. "Optimal Bandwidth Choice for the Regression Discontinuity Estimator." *Review of Economic Studies* 79 (3): 933–959.

International Energy Agency. 2011. *World Energy Outlook 2011: Energy for All.*

Ito, Koichiro, Takanori Ida, and Makoto Tanaka. 2015. "The Persistence of Moral Suasion and Economic Incentives: Field Experimental Evidence from Energy Demand." NBER Working Paper 20910.

Jayachandran, Seema, Joost de Laat, Eric Lambin, and Charlotte Stanton. 2016. "Cash for Carbon: A Randomized Controlled Trial of Payments for Ecosystem Services to Reduce Deforestation." `http://faculty.wcas.northwestern.edu/~sjv340/jayachandran_etal_pes_rct.pdf`.

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

Joskow, Paul L, and Donald B Marron. 1992. "What does a negawatt really cost? Evidence from utility conservation programs." *The Energy Journal* 13 (4): 41–74.

Kahn, Matthew, Nils Kok, and John Quigley. 2014. "Carbon emissions from the commercial building sector: The role of climate, quality, and incentives." *Journal of Public Economics* 113:1–12.

Kok, Nils, and Maarten Jennen. 2012. "The impact of energy labels and accessibility on office rents." *Energy Policy* 46 (C): 489–497.

Kushler, Martin. 2015. "Residential energy efficiency works: Don't make a mountain out of the E2e molehill." *American Council for an Energy-Efficient Economy Blog* (June).

Lee, David S., and Thomas Lemieux. 2010. "Regression Discontinuity Designs in Economics." *Journal of Economic Literature* 48 (2): 281–355.

Lee, Kenneth, Edward Miguel, and Catherine Wolfram. 2016. "Experimental Evidence on the Demands for and Costs of Rural Electrification." NBER Working Paper 22292.

Levinson, Arik. 2016. "How Much Energy Do Building Energy Codes Save? Evidence from California Houses." *American Economic Review* 106, no. 10 (October): 2867–2894.

Lipscomb, Molly, A. Mushfiq Mobarak, and Tania Barham. 2013. "Development Effects of Electrification: Evidence from the Topographic Placement of Hydropower Plants in Brazil." *American Economic Journal: Applied Economics* 5 (2): 200–231.

McCaffrey, Daniel, Greg Ridgeway, and Andrew Morral. 2004. "Propensity Score Estimation with Boosted Regression for Evaluating Causal Effects in Observational Studies." *RAND Journal of Economics* 9 (4): 403–425.

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

McKinsey & Company. 2009. *Unlocking energy efficiency in the U.S. economy.* Technical report. McKinsey Global Energy and Materials.

McRae, Shaun. 2015. "Infrastructure Quality and the Subsidy Trap." *American Economic Review* 105 (1): 35–66.

Min, Brian. 2011. "Electrifying the Poor: Distributing Power in India." `https://courses.cit.cornell.edu/patel/psac/Min_PSAC_4_6.pdf`.

Min, Brian, and Kwawu Mensan Gaba. 2014. "Tracking Electrification in Vietnam Using Nighttime Lights." *Remote Sensing* 6 (10): 9511–9529.

Min, Brian, Kwawu Mensan Gaba, Ousmane Fall Sarr, and Alassane Agalassou. 2013. "Detection of rural electrification in Africa using DMSP-OLS night lights imagery." *International Journal of Remote Sensing* 34 (22): 8118–8141.

Ministry of Power. 2005. *Rajiv Gandhi Grameen Vidyutikaran Yojna Scheme of Rural Electricity Infrastructure and Household Electrification.*

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.

Myers, Erica. 2014. "Asymmetric information in residential rental markets: implications for the energy efficiency gap." *Job Market Paper.*

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.

Rud, Juan Pablo. 2012. "Electricity provision and industrial development: Evidence from India." *Journal of Development Economics* 97 (2): 352–367.

Tsujita, Yuko, ed. 2014. *Inclusive Growth and Development in India: Challenges for Underdeveloped Regions and the Underclass.* United Kingdom: Palgrave Macmillan.

UNDP. 2015. "Energy access." `http://www.undp.org/content/undp/en/home/ourwork/climate-and-disaster-resilience/sustainable-energy/energy-access.html`.

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

World Bank. 2015. "Energy Overview." `http://www.worldbank.org/en/topic/energy/overview`.

Wyss, Richard, Alan Ellis, Alan Brookhart, Cynthia Girman, Michele Funk, Robert LoCasale, and Til Sturmer. 2014. "The Role of Prediction Modeling in Propensity Score Estimation: An Evaluation of Logistic Regression, bCART, and the Covariate-Balancing Propensity Score." *American Journal of Epidemiology* 180 (6): 645–655.

Appendix A

Out of the Darkness and Into the Light? Development Effects of Rural Electrification – Appendix

A.1 Data

A.1.1 RGGVY program data

The Rural Electrification Corporation maintains an online database of RGGVY implementation status, and also hosts two separate portals for "Villages Covered" and "Villages Completed" under the 10th and 11th Plans of RGGVY.[1] Each village-specific page within these portals reported that village's pre-program electrification status, proposed RGGVY implementation details (e.g. number of households to be electrified, new transformer capacity to be installed), actual progress of RGGVY implementation (e.g. number of household connections completed, new transformer capacity installed), and implementation status (e.g. whether work has been completed in this village).[2] Unfortunately, these village-level data are of very poor quality, with rampant missing information, internal inconsistencies, and program outcomes that

[1] We downloaded these data from the RGGVY home page, `http://rggvy.gov.in`, which has since been deactivated as RGGVY has ben subsumed into DDUGJY. The new program website is `http://www.ddugjy.in/`.

[2] Notably, neither portal recorded the date on which a village was sanctioned for electrification, nor the date on which works were begun or completed. The only timing recorded in either set of webpages describes, to the best of our knowledge, was the latest upload of data to the internet. Each of these datasets had a separate tab on the RGGVY homepage, respectively: "Villages Covered (X & XI Plan DPR)" and "Villages Completed (X & XI Plan)". We scraped these datasets between August and October 2014.

conflict with village-level Census data.[3] As detailed in Table A.1.1, RGGVY programs outcomes are missing for 65 percent of villages that were eligible under the 10th Plan (our analysis focuses on this earlier wave of the RGGVY program). These data also report a greater number of covered habitations than exist for 32 percent of villages.[4]

Table A.1.1: RGGVY Microdata Irregularities

Type of Irregularity	Percent of Villages	Percent of 10th-Plan Villages
RGGVY outcomes disagree across Covered and Completed datasets	26.8	32.7
Outcomes missing from either Covered or Completed dataset	77.9	65.3
Outcomes missing from both Covered and Completed datasets	74.4	59.9
All outcomes missing from both Covered and Completed datasets	33.4	22.3
Completed dataset reports status not energised	24.4	14.1
RGGVY covers more habitations than exist in village	32.2	31.7

Note. — This table shows data irregularities across the RGGVY Covered and Completed village datasets, which we do not use in our analysis. The right two columns shows the percent of all villages and the percent of 10th-Plan villages that satisfy each irregularity criterion, where the denominator excludes missing and unmatched villages. Program outcomes considered in the first four rows include the count of household connections, aggregate transformer capacity installed, and aggregate transmission capacity installed. (The first three rows count villages where *any* outcome disagrees or is missing; the fourth row counts only villages for which *all* of these outcomes are missing.)

Even if these village-level program data were of better quality, we might worry about using them for our analysis. We would expect the RGGVY microdata to identify precisely which villages were treated under the program. However, we would still need to construct a control group from the subset of villages *not* represented in the RGGVY microdata. Because the RGGVY datasets do not include information on villages left out of the program, we would need to merge the RGGVY microdata into village-level Census data and denote villages included in the RGGVY dataset as "treated" and all other villages as "control." Any imperfect merge or missing RGGVY microdata would cause us to incorrectly categorize a village. We might also worry about manipulation of RGGVY village-level outcomes, if implementing

[3]The Census data seems to be of relatively high quality, with no evidence of population manipulation. Asher and Novosad (2016) shows that PCA data has a high correlation with another Indian dataset, the Socioeconomic and Caste Census; we also find high correlations between Census data and National Rural Drinking Water Programme's census of habitations (described further in Section A.1.5). For these reasons, we are inclined to trust the Census data over the RGGVY data.

[4]We base this total number of habitations on our merge of the Habitation Census to the village-level Census panel, considering only those RGGVY villages that match to a *matched* village from the habitation census merge (described below).

agencies (or data tabulators) had an incentive to overstate the extent of electrification attributable to RGGVY.[5]

For these reasons, we have chosen to exclude the RGGVY village-level data from our analysis entirely. Instead, we rely on district-level RGGVY summary reports to determine the Five-Year Plan under which each electrification project was sanctioned, the implementing agency responsible for implementation, and the approximate timing of electrification. Since the program was implemented based on Detailed Project Reports (DPR) at the district level, we can use these aggregate data to link villages to DPRs, and, importantly, to the 10th or 11th Plan.[6] Our analysis only relies on these district-level reports to assign districts to Five-Year Plans. As this is a matter of public record and involves large transfers of public funds from the Rural Electrification Corporation to decentralized implementing agencies, we are much more confident in the accuracy of this aggregate information.

Table A.1.2 summarizes these district-level progress reports.[7] We see that under both the 10th and 11th Plans, the majority of district-level DPRs were implemented by local electricity distribution companies. However, there were also many districts whose implementing agencies included large public sector undertakings, state departments of power, and state electricity boards. This table also shows that the majority of villages covered under both Plans were categorized as "electrified" villages, meaning that there was a minimum of power access prior to RGGVY.[8] For these villages, RGGVY sought to provide "more intensive electrification," including bringing energy access to below-poverty-line households that still lacked connections. In early 2015, the program reported that over 97 percent of 10th-Plan projects had been completed, and that over 90 percent of 11th-Plan projects had been completed.

[5]We are not the only researchers to find inconsistencies in microdata from Indian programs. Asher and Novosad (2016) document striking irregularities in PMGSY population data.

[6]These reports were available at `http://rggvy.gov.in` under the "Progress Reports" menu, and can now be found at `http://ddugjy.gov.in`. Most districts were appear in exactly one DPR. For districts with multiple DPRs, we have aggregated DPRs up to the district-level, in order to create a one-to-one mapping between districts and DPRs. Of the 530 RGGVY districts, only 30 districts had DPRs aggregated across both the 10th and 11th Plans.

[7]The state of Goa was excluded from RGGVY along with all 7 union territories, because 100 percent of their villages were electrified prior to 2005 (Ministry of Power (2012)). We treat Telangana as part of Andhra Pradesh, since its 2014 split from Andhra Pradesh occurred after our period of analysis.

[8]The definition of "electrification" has changed over time in India, but as long as "one bulb was burning" anywhere in the village, a village was considered electrified.

Table A.1.2: Summary Statistics – RGGVY Implementation and Scope

Type of Implementing Agency	States	Districts	Award Dates	Unelectrified Villages	Electrified Villages	BPL Connections
A. 10th Plan						
Public Sector	11	57	2005–2007	32,638	20,126	2,386,042
State Departments	5	10	2007–2010	542	1,088	45,921
State Electricity Boards	3	6	2006–2007	4,482	2,604	441,639
Distribution Companies	14	156	2005–2008	26,429	76,495	4,649,733
Total	25	229		64,091	100,313	7,523,335
B. 11th Plan						
Public Sector	9	78	2008–2011	30,298	62,705	6,768,765
State Departments	5	33	2008–2010	2,710	3,850	212,887
State Electricity Boards	3	34	2008–2011	80	18,166	149,882
Distribution Companies	15	183	2008–2010	13,118	135,038	7,363,814
Rural Electricity Coops	2	5	2008–2011	0	755	79,220
Total	25	331		46,206	220,514	14,574,568

Note. — Data summarize district-level progress reports, previously available at `http://rggvy.gov.in`. Public sector undertakings include government-owned generating companies, such as Power Grid Corporation of India and National Hydroelectric Power Corporation. The right three columns show the number of previously unelectrified and previously electrified villages covered by the program, as well as the the number of below poverty line households to receive electric connections. Villages classified as electrified had basic electricity infrastructure with at least 10 percent of households electrified prior to RGGVY implementation. 23 (of 27) states contain both 10th and 11th Plan districts, while 30 (of 530) individual districts were targeted under both Plans. For a few districts, we correct financial award dates reported to have occurred *before* their respective project sanction dates or before the official announcement of the program.

A.1.2 Geospatial data

Our main source of geospatial information is ML Infomap's VillageMap.[9] This dataset includes village boundary shapefiles for nearly every village in India, as defined by the 2001 village-level Census. We take the 2001 boundaries as fixed, which is consistent with our decision to use the 2001 village as our unit of analysis.[10] Every square meter in India belongs to a village or city/town; the only "blank spaces" in the village maps are forests, bodies of water, and "towns" (urban areas). Village boundaries are set by the Census Organization of India (Census of India (2011)). Figure A.1.1 shows an example of the level of detail included in the village boundary dataset.

We were unable to acquire village shapefiles for Arunachal Pradesh, Meghalaya, Mizoram, Nagaland, and Sikkim. This forces us to drop these 5 states from our

[9]These data are also used by Min (2011) and Asher and Novosad (2015), among others, to map villages in India.

[10]In the 2 percent of cases where a 2001 village matches to multiple 2011 villages in the 2001/2011 Census concordance, we aggregate 2011 data up to the 2001 village definition.

geospatial dataset (and, subsequently, from our analysis of nighttime lights), though we are still able to include them for the remainder of our analysis. Fortunately, these small states comprise only 2 percent of all Indian villages and less than 1 percent of India's rural population. They also represent less 3 percent of villages covered under RGGVY, and only 1 percent of villages covered by the 10th Plan.

We also exclude Assam, Himachal Pradesh, Jammu and Kashmir, Uttar Pradesh, and Uttarakhand from our geospatial dataset, because we believe that the shapefiles for these states are extremely poor quality. For these 5 states, the correlation between the village area measurement implied by the shapefiles and the village area reported in the Indian Census (the official body in charge of defining village boundaries) is below 0.35. Table A.1.3 shows these correlations by state, demonstrating a clear gap between the 12 correlated and 5 uncorrelated states. These uncorrelated states represent 39 percent of villages in RGGVY 10th-Plan districts, most of which are in Uttar Pradesh. While we restrict our analysis of nighttime brightness to these 12 correlated states, our analysis of Census outcomes includes villages in all 22 states in Table A.1.3.

Our full geospatial dataset includes village boundaries for 172,013 villages across the 12 RGGVY states remaining after the sample restrictions detailed above.[11] Each boundary shapefile is identified by its 2001 Census code, as well as village attributes from the 2001 Primary Census Abstract. We use this identifying information to link geospatial information into our administrative datasets.

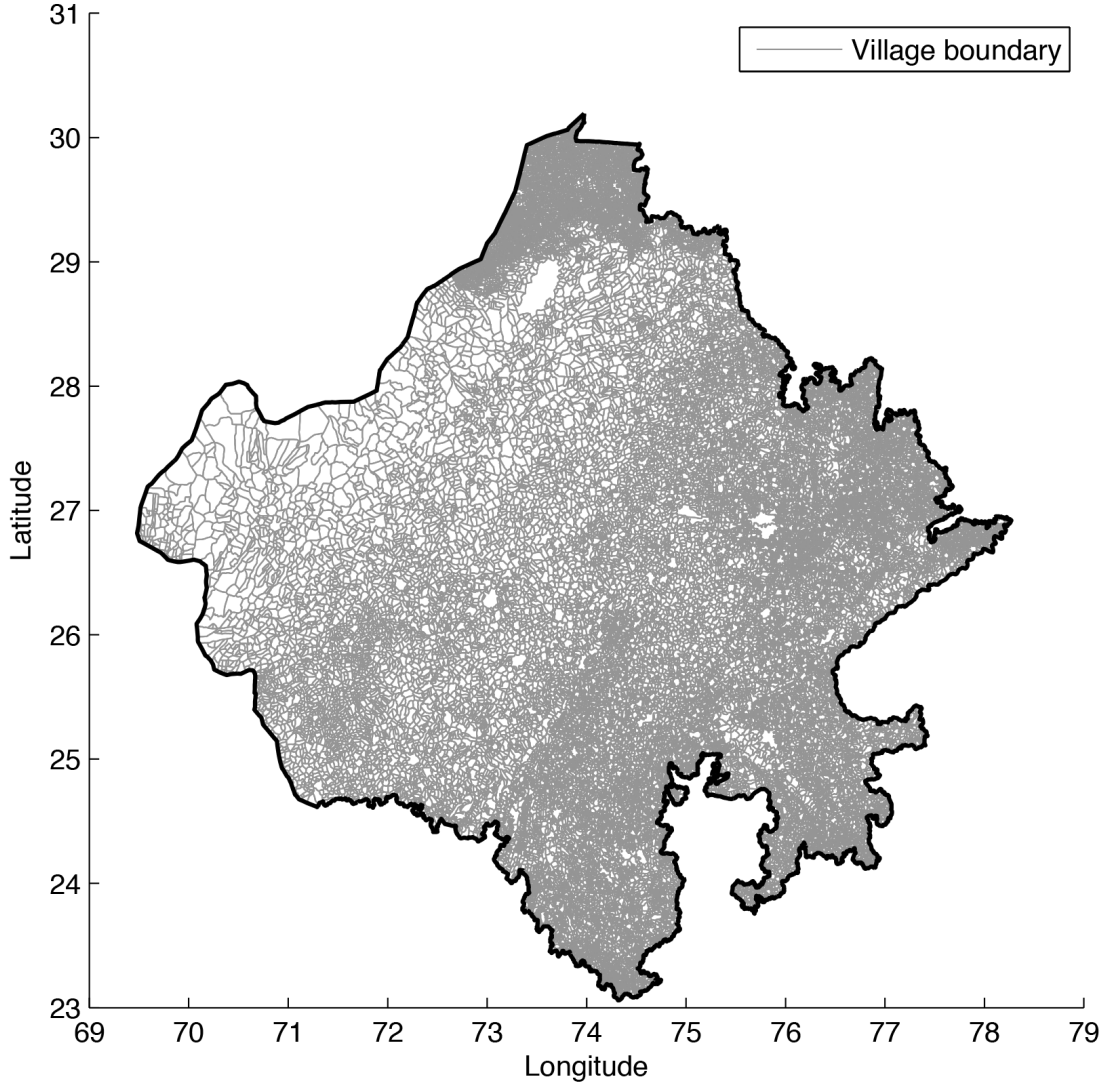
[11]This number excludes village with null or missing populations, which appear to have been miscoded.

Table A.1.3: Correlation of Shapefiles with Village Areas

State	Area Correlation	Percent of Total Villages	Percent of 10th-Plan Villages
Jharkhand	0.978	5.0	5.9
West Bengal	0.954	6.4	10.8
Bihar	0.932	6.6	9.8
Gujarat	0.896	3.1	0.8
Haryana	0.873	1.1	0.4
Karnataka	0.806	4.6	6.5
Maharashtra	0.781	7.0	1.5
Andhra Pradesh	0.772	4.5	7.2
Rajasthan	0.714	6.8	9.9
Orissa	0.680	8.1	2.5
Madhya Pradesh	0.638	8.9	3.4
Chhattisgarh	0.605	3.3	1.2
Uttarakhand	0.326	2.7	5.3
Uttar Pradesh	0.281	16.6	31.8
Himachal Pradesh	0.138	3.0	0.4
Assam	0.106	4.1	1.0
Jammu and Kashmir	0.002	1.1	0.5
Arunachal Pradesh	missing	0.6	0.3
Meghalaya	missing	0.9	0.3
Mizoram	missing	0.1	0.1
Nagaland	missing	0.2	0.1
Sikkim	missing	0.1	0.1
States with correlation > 0.35		68.3	59.9
States with correlation < 0.35		28.8	39.0
States with missing shapefiles		2.9	1.0

Note. — This table shows the correlation between areas calculated from village shapefiles and village areas reported in the Census's 2001 Village Directory. Our spatial dataset includes only the 12 states for which this correlation is at least 0.35. We omit the 5 states with shapefile areas that are uncorrelated with reported village areas, as we take this as a sign of low quality shapefiles. The middle column reports the percent of Indian villages contained in each state, while the right column shows the percent of villages in districts eligible for RGGVY under the 10th Plan. Omitted from this table are 3 states which were not eligible under RGGVY's 10th Plan (Goa, Punjab, and Tamil Nadu), as well as 2 states which were eligible under RGGVY's 10th Plan but contain no single-habitation 10th-Plan villages in our RD bandwidth (Kerala and Tripura).

Figure A.1.1: Rajasthani Village Boundaries



Note. — This figure shows the approximately 41,575 villages in Rajasthan. The 1st and 99th percentiles of Rajasthani village area are 0.2 km^2 and 29.5 km^2, respectively.

A.1.3 Nighttime lights

In order to credibly measure electrification, we use the National Oceanic and Atmospheric Administration (NOAA)'s Defense Meteorological Satellite Program - Operational Line Scan (DMSP-OLS) Nighttime Lights data[12]. Descriptions of these data can be found in Elvidge et al. (1997) and Doll (2008).[13] Data are publicly available from 1992 to 2013; we use images from 1999 to 2013 in this paper. These images are compiled from nightly satellite photographs taken between 8:30 and 10:00 PM local time, and they are extremely high resolution: pixels are 30 arc-second squares.[14] The Indian subcontinent alone contains 417,876 pixels. Each pixel is assigned a digital number (DN) indicating brightness, ranging from 0 to 63. This DN is approximately proportional to average luminosity.[15] While the images often top-code very bright locations such as urban centers (Henderson, Storeygard, and Weil (2012)), we focus our attention on rural areas with very low risk of top-coding.

We might instead worry that the DMSP-OLS sensors are not sensitive enough to detect the subtle changes in brightness associated with rural electrification. However, a substantial body of evidence suggests otherwise. Elvidge et al. (1997) and Elvidge et al. (2001) use DMSP-OLS data to estimate electrification rates around the world at the national level. A variety of papers have also mapped nighttime lights to electrification rates at the sub-national level, including Ebener et al. (2005), Doll, Muller, and Morley (2006), Chand et al. (2009), and Townsend and Bruce (2010). Three studies are particularly relevant to our work: Min et al. (2013) use original survey data from Senegal and Mali to show that electrified rural villages are significantly brighter in the DMSP-OLS data than their unelectrified counterparts. Min and Gaba (2014) find a strong correlation between village-level ground-based electricity usage survey data and DMSP-OLS nighttime lights imagery in rural Vietnam, showing that both streetlights and electrified homes are correlated with higher DMSP-OLS DN readings. Finally, Min (2011) shows a strong correlation between power distribution and nighttime lights and administrative data on electrification

[12]The data are available for download here: `http://ngdc.noaa.gov/eog/dmsp/downloadV4composites.html`

[13]Researchers have also used these nighttime lights data as proxies for economic activity, including Noor et al. (2008), Bleakley and Lin (2012), Henderson, Storeygard, and Weil (2012), Li, Ge, and Chen (2013), Michalopoulos and Papaioannou (2013), and Michalopoulos and Papaioannou (2014). Given that we are studying electrification directly, we refrain from using nighttime lighting as a proxy for GDP.

[14]These pixels are squares with approximately 500 meter sides at the equator.

[15]See Chen and Nordhaus (2011) for details.

and nighttime lights in the Indian state of Uttar Pradesh.[16] Based on these findings, we are confident that the DMSP-OLS data are capable of accurately measuring rural electrification — any activity in rural India bright enough to be visible from space is likely to require electricity.[17] Furthermore, any bottom coding in our data will lead us to underestimate the effects of RGGVY on nighttime brightness.[18]

NOAA releases three different DMSP-OLS lights products: average visible lights; stable lights; and average visible × percent lights. The average visible lights dataset contains the average DN over the year's observations for each pixel. The stable lights dataset is a more heavily processed version of the average visible lights: it contains the lighting from persistently lit places, and does not include the light from fires and other sporadic events. Finally, the average visible × percent lights takes the average visible DN in a pixel and multiplies it by the frequency with which it is observed.[19] Our analysis focuses on the average visible lights data, which are best equipped to detect the low levels of lighting associated with electrification of rural villages (Min et al. (2013)). Section A.2.1 performs a robustness check using the stable lights dataset.

We construct village-level nighttime lights values by combining village shapefiles with the nighttime brightness data in ArcGIS, overlaying the 2001 village boundaries on top of lights images for each year. Figure A.1.2 shows an example of this overlay from a region of Andhra Pradesh. For each village, we calculate the maximum DN value over all pixels whose centroids are contained within a village boundary. In other words, a village's brightness for a given year is equal to the brightness of its brightest

[16]Note that this analysis was done at the district level, obviating the need for village-level shapefiles.

[17]There are two obvious exceptions: the first is agricultural fires; the second is car headlamps. Because of the yearly nature of the DMSP-OLS data, ephemeral fires do not appear in the final stable lights averaged datasets (details can be found in NOAA's DMSP-OLS data description). Roads are sparse in the villages we are looking at; it is extremely unlikely that there is enough road traffic to appear in the DMSP-OLS dataset. Even if there were enough road traffic to be detected in any given flyover of the satellite, this road traffic is likely to be erratic enough, and the satellite images are taken at variable enough times of night, that it is unlikely that we mistake vehicle traffic for consistent village electrification (Min and Gaba (2014)).

[18]Suppose, for example, that the satellite can detect brightnesses of λ or greater only, and that at baseline, villages A and B both have brightness of $\lambda - 5$. These are both coded as λ. Village B is electrified under RGGVY, and now has brightness of $\lambda + 5$. We observe a difference of 5 between A and B, when in fact, the true difference was 10.

[19]These lights have been used by Alam (2013) to examine power quality in India, but they are much less frequently used in the economics literature.

pixel.[20] For villages too small to contain a pixel's centroid, we assign the value of the pixel at the village centroid as the maximum lights value.[21] Between 1999 and 2007, NOAA had two satellites operating DMSP-OLS equipment.[22] For these years, we calculate the mean and maximum lights values for each satellite separately, and then we take an unweighted average across satellites to obtain village-year DN statistics.

One concern about the DMSP-OLS dataset is that the sensors used to calculate the DN tend to degrade and become dirty over time. The satellites do not contain on-board calibration equipment, and the sensors are only adjusted before being loaded onto the satellite platform. NOAA's Earth Observation Group, which manages DMSP-OLS data, has done some ex-post calibration in order to bring different satellite-years in line with one another. However, this algorithm must assume that brightness in one region (usually the island of Sicily is used) remains fixed over time, thereby calibrating the sensors to that region's DN values. This process is imperfect and not fully transparent. Most economists who use these data in a panel setting include satellite or year fixed effects to control for inconsistencies in sensors over time (e.g. Henderson, Storeygard, and Weil (2012)). In our empirical analysis of nighttime brightness, we use an estimator that relies primarily on cross-sectional variation in brightness, which should assuage these concerns.

We undertake an additional procedure to remove measurement error from the lights data. After constructing a village-year panel of the maximum nighttime brightness, we linearly project lights values from 2001 and 2011 on the values of the two years before and after. We run the following OLS regressions, weighting by village area[23]:

$$\begin{aligned} L_v^{2001} &= \alpha_0 + \alpha_1 L_v^{1999} + \alpha_2 L_v^{2000} + \alpha_3 L_v^{2002} + \alpha_4 L_v^{2003} + \varepsilon_v \\ L_v^{2011} &= \kappa_0 + \kappa_1 L_v^{2009} + \kappa_2 L_v^{2010} + \kappa_3 L_v^{2012} + \kappa_4 L_v^{2013} + \upsilon_v \ , \end{aligned} \tag{A.1}$$

where L_v^y is the maximum brightness of village v in year y. We use the estimated $\hat{\alpha}_i$s and $\hat{\kappa}_i$s (for $i = \{0, 1, 2, 3, 4\}$) to construct the nighttime lights outcome variables we use in our regressions, $\hat{L}_v^{2001}$ and $\hat{L}_v^{2011}$. This removes random year-to-year noise in

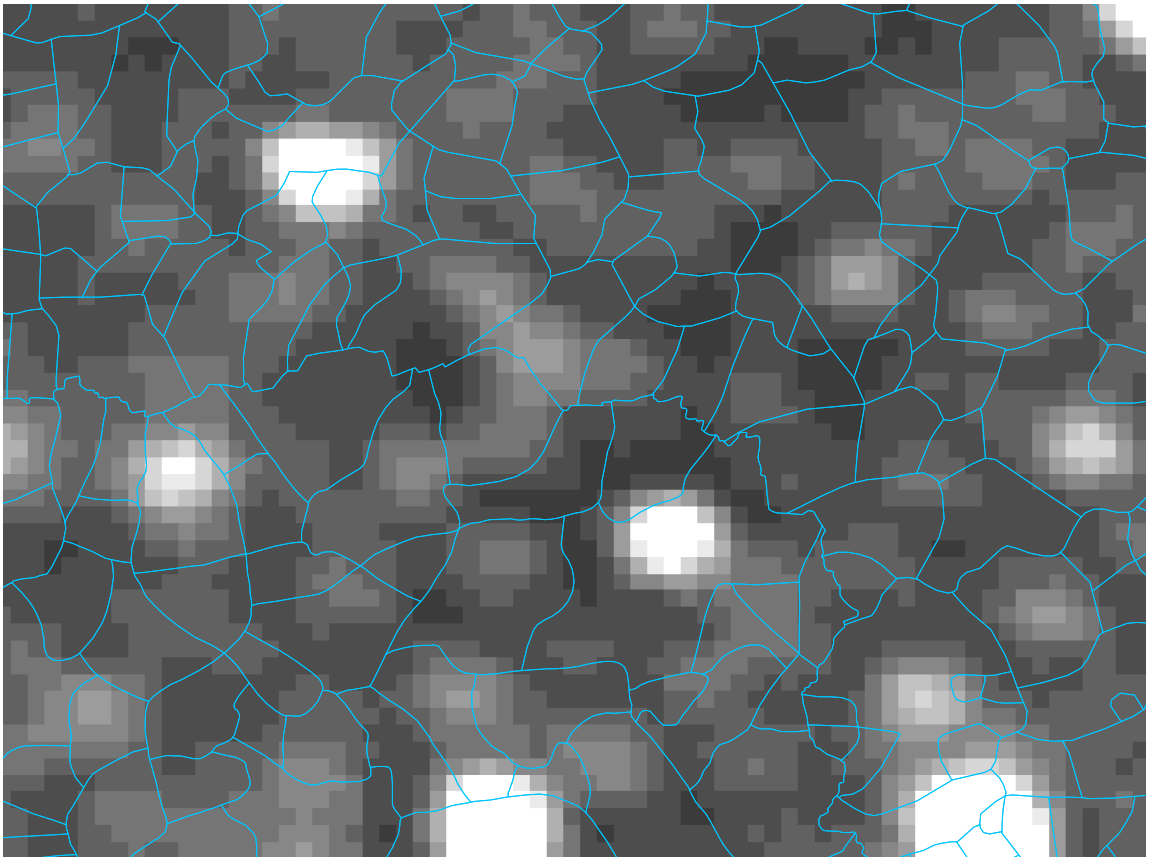
[20] We also conduct sensitivity analysis on assigning each village the average brightness across all its pixels. We use the the standard `Zonal Statistics as Table` operation in ArcGIS to calculate both the maximum and mean brightness.

[21] All of the villages that did not contain a pixel centroid only overlapped with one pixel, so this is the correct operation for these very small villages.

[22] In 1999, F12 and F14 were active; from 2000 to 2003, F14 and F15 were active; from 2004 to 2007, F15 and F16 were active; from 2008 to 2009, only F16 was active; and from 2010 to 2013, only F18 was active.

[23] The results are not sensitive to the decision to weight. Unweighted regressions produce nearly identical results.

Figure A.1.2: Example of Nighttime Lights with Village Boundaries



Note. — This image shows a close-up of averge visible nighttime brightness overlaid with village boundaries, for an area in Rajasthan. The $\approx$1km^2 pixels in this image range in brightness values from 3 to 38.

the 2001 and 2011 lights data that cannot be explained by the brightness observed in adjacent years. It also isolates the more stable year-to-year changes in brightness that we would associate with new electricity infrastructure. In Section A.2.1, we include a sensitivity analysis that varies the number of adjacent years in these linear projections, while also considering simple unweighted averages of adjacent years.

A.1.4 Census data

We construct a village-level panel dataset using three datasets from the Census of India's 2001 and 2011 decennial Censuses. These datasets are all available for download

from the Census of India's website.[24] Below, we describe each dataset separately, along with the process we used to construct our 2001–2011 village panel.

A.1.4.1 Primary Census Abstract

The Primary Census Abstract (PCA) reports village population and employment information for all geographic units across India's 28 states and 8 union territories.[25] This dataset includes the total number of households in each village, along with village population broken down by gender, the 0–6 age cohort, scheduled caste, and scheduled tribe.[26] Village-level literacy rates are also included. The 2001 PCA dataset contains 593,643 villages, while the 2011 dataset contains 597,483 villages.

The PCA reports employment counts by gender, for three disjoint subsets of the village population: "main workers" who participate in any economically productive activity (with or without compensation) for at least 6 months of the year; "marginal workers" who do so for less than 6 months of the year; and "non-workers" who do not participate in economically productive activity. Within each of these groups, workers of each gender are separately categorized as either cultivators, agricultural laborers, household industry workers, or other. The distinction between cultivators and agricultural laborers is that agricultural laborers work for wages, while taking on no risk in cultivation and owning no right to cultivate land. "Other" workers include all workers not covered by the other three categories, such as government workers, teachers, doctors, and factory workers, and includes all non-farm, non-household employment.

A.1.4.2 Houselisting Primary Census Abstract

The Houselisting Primary Census Abstract (HPCA) reports on a variety of household attributes and amenities at the village level. These include physical housing stock characteristics such as type of floor/wall/roof and number of rooms; drinking water source; type of latrine; primary cooking fuel; and main source of in-home lighting (e.g., electricity vs. kerosene). This dataset also includes information on household inhabitants, including the number of members; number of married couples; and whether houses are owned or rented. Finally, the HPCA includes data on

[24] `http://censusindia.gov.in`. We downloaded these data between September 2014 and August 2015.

[25] India currently has 29 states, but the Andhra Pradesh-Telangana split occurred in 2014, after the 2011 data were collected and published.

[26] Scheduled castes (SC) and scheduled tribes (ST) are official designations for castes and ethnic groups that have been historically disadvantaged. Since its independence, India has targeted SC and ST communities for affirmative action in social programs and political representation.

household asset ownership, including whether houses own mobile phones, televisions, motorcycles, radios, and other durable goods.

For each of the above categories, the HPCA reports the proportion of households within each administrative unit that satisfy each respective criterion. This allows us to treat each variable as continuous, with considerable variation across villages. The 2011 HPCA is publicly available at the village level, and it reports on 597,502 villages. The 2001 HPCA is only available at the Census block level (i.e. the administrative unit between district and village), and it reports block-specific values for most variables across all 5,415 blocks. However, a few variables are only reported at the district level, including indicators of physical upkeep, ownership status, number of rooms, and number of married couples per household.

A.1.4.3 Village Directory

The Village Directory (VD) provides detailed village-level information on a variety of amenities.[27] These data are analogous to a community survey that is often included with household survey data. Unlike the amenities featured in the HPCA, the VD reports public, community-level characteristics that are not specific to individual households. These include the number of primary/middle/secondary schools and other educational facilities; the number of hospitals, community health centers, and other health facilities; community drinking water sources; phone, post office, and other communication services; bus/rail service and road quality; and the presence of banking facilities and credit societies. The dataset also includes 1/0 indicators for the availability of electric power services, broken out by agricultural, domestic, and commercial end-use sectors; the 2011 VD additionally reports average hours per day of electric power received by each sector. For most villages, the VD lists the most important manufacturing and agricultural commodities (the latter in 2011 only). Finally, the VD contains several geographic variables, including village area, area of cropland irrigated (by water source), distance to the nearest road and navigable waterway, and distance to the nearest town. After removing villages with populations that are either zero or missing, the 2001 and 2011 VD datasets contain 593,643 and 596,615 villages, respectively.

A.1.4.4 Census panel dataset

We are able to match villages across the above six datasets using their official census codes. Within each Census year (2001 and 2011), state, district, and village codes are

[27] The 2001 VD was a standalone product, while the 2011 VD was distributed as part of the District Census Handbook (DCHB).

coded consistently across PCA, HPCA, and VD datasets, so merging these data is straightforward.[28] In order to link villages across Census years, we take advantage of the Census's 2001–2011 concordance. We treat the 2001 PCA village as our master cross-sectional unit, re-aggregating any 2001 villages that split into multiple villages by 2011.[29] Our final panel includes only those villages that match to all 5 village level datasets – 2001/2011 PCA, 2011 HPCA, and 2001/2011 VD. Since village-level data do not exist for the 2001 HPCA, we instead assign each village the values from its parent block (or district, when block-level data is unavailable).[30]

In many cases, the variables reported in the 2011 Census datasets differ from those published in the corresponding 2001 datasets. This is especially true for the VD and certain sections of the HPCA. Because our preferred specification includes a control for the 2001 level of the outcome variable wherever possible, we combine and redefine variables such that our final panel contains only consistent variables.[31] The full panel dataset contains 580,643 villages, with over 200 matching pre/post variables. Tables A.1.4, A.1.5, and A.1.6 provide summary statistics for our Census panel dataset, for three subsets of variables originating from PCA, HPCA, and VD, respectively.

[28]Subdistrict, tehsil, and block codes are not consistently coded across datasets, which reflects different administrative conventions across states. For example, while the PCA and HPCA assigns a single tehsil code to each village, the VD assigns separate tehsil and block codes. In our merges, we match on only state, district, and village code, ignoring block code. Because village code is *virtually* unique within each district, this does not affect the accuracy of this merge.

[29]There are many 2001 villages that match to more than one 2011 village, based on the 2001–2011 concordance. We interpret these as administrative splits, and re-aggregation of all 2011 variables affords us a consistent comparison across years. We drop the *very* few cases (i.e. < 0.01 percent) where multiple 2001 villages appear to have merged into a single 2011 village.

[30]While this is imperfect, the alternative would be to ignore any block-level information on the 2001 levels of 2011 HPCA variables. Our RD analysis of 2011 village-level outcomes greatly benefits from the use of 2001 controls to increase precision.

[31]For example, the 2001 VD lists multiple types of tubewells, while the 2011 VD lists only a single tubewell indicator. We construct a single tubewell indicator from the 2001 VD, such that this indicator is coded identically across VD years.

Table A.1.4: Summary Statistics – Primary Census Abstract

Village Characteristics	All Districts		10th-Plan Districts		10th-Plan Districts 150–450 Population	
	2001	2011	2001	2011	2001	2011
Village population	1222.19	1416.42	1234.40	1442.17	297.60	359.46
	(1713.62)	(1963.73)	(1607.80)	(1879.55)	(86.60)	(154.96)
Number of HH	226.02	286.50	218.46	277.65	54.08	70.55
	(337.69)	(418.29)	(303.64)	(382.86)	(18.50)	(31.10)
Share of pop. SC or ST	0.36	0.37	0.33	0.33	0.39	0.40
	(0.32)	(0.32)	(0.30)	(0.30)	(0.36)	(0.36)
Literacy rate	0.46	0.57	0.44	0.56	0.44	0.55
	(0.17)	(0.15)	(0.16)	(0.13)	(0.17)	(0.15)
Employment rate	0.44	0.45	0.42	0.42	0.44	0.45
	(0.13)	(0.14)	(0.13)	(0.14)	(0.14)	(0.15)
Male employment rate	0.53	0.54	0.52	0.52	0.52	0.53
	(0.09)	(0.10)	(0.10)	(0.10)	(0.10)	(0.11)
Female employment rate	0.35	0.35	0.32	0.32	0.36	0.36
	(0.21)	(0.22)	(0.21)	(0.22)	(0.23)	(0.23)
% Male ag workers	0.41	0.40	0.40	0.40	0.42	0.42
	(0.14)	(0.15)	(0.14)	(0.15)	(0.15)	(0.16)
% Male household workers	0.01	0.01	0.01	0.01	0.01	0.01
	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)
% Male other workers	0.11	0.12	0.10	0.11	0.09	0.10
	(0.11)	(0.12)	(0.10)	(0.11)	(0.10)	(0.12)
% Female ag workers	0.30	0.28	0.27	0.25	0.32	0.30
	(0.22)	(0.22)	(0.21)	(0.21)	(0.23)	(0.23)
% Female household workers	0.01	0.01	0.01	0.01	0.01	0.01
	(0.05)	(0.05)	(0.05)	(0.05)	(0.06)	(0.05)
% Female other workers	0.04	0.05	0.03	0.05	0.03	0.05
	(0.08)	(0.08)	(0.07)	(0.08)	(0.07)	(0.09)
% Male main workers	0.44	0.41	0.43	0.38	0.42	0.37
	(0.13)	(0.16)	(0.13)	(0.17)	(0.15)	(0.19)
% Female main workers	0.18	0.18	0.16	0.16	0.17	0.17
	(0.19)	(0.19)	(0.18)	(0.18)	(0.20)	(0.20)
% Male marginal workers	0.09	0.13	0.09	0.14	0.10	0.16
	(0.11)	(0.14)	(0.10)	(0.14)	(0.12)	(0.17)
% Female marginal workers	0.17	0.17	0.16	0.16	0.19	0.19
	(0.18)	(0.19)	(0.17)	(0.18)	(0.20)	(0.21)
Number of villages	580,643	580,643	290,067	290,067	62,647	62,647

Note. — This table reports means and standard deviations from the 2001 and 2011 Primary Census Abstract. The left two columns include all villages that match across Census datasets, for all 27 RGGVY states. The middle two columns include the subset of those villages located in districts eligible for RGGVY under the 10th Plan. The right two columns include only 10th-Plan villages with 2001 populations between 150 and 450 (i.e. our RD bandwidth). Worker by sector is presented as the fraction of total workers in the village (main + marginal for each gender, respectively). Agricultural workers include both cultivators and agricultural laborers; other workers are classified as non-agricultural, non-household workers. By definition, main workers work at least 6 months per year, while marginal worker work less than 6 months per year. The employment rate (by gender) divides the sum of main and marginal workers by the village population (of that gender). SC and ST refer to Schedule Caste and Scheduled Tribe designations.

Table A.1.5: Summary Statistics – Houselisting Primary Census Abstract

Household Characteristics	All Districts		10th-Plan Districts		10th-Plan Districts 150–450 Population	
	2001	2011	2001	2011	2001	2011
Average household size	5.42	5.03	5.60	5.24	5.53	5.16
	(0.61)‡	(0.82)	(0.64)‡	(0.86)	(0.61)‡	(0.86)
Average number of rooms	2.11	2.03	2.18	2.04	2.20	2.06
	(0.52)‡	(0.73)	(0.54)‡	(0.73)	(0.54)‡	(0.81)
% Good condition	0.43	0.43	0.42	0.43	0.43	0.43
	(0.14)‡	(0.31)	(0.11)‡	(0.30)	(0.12)‡	(0.35)
% Livable condition	0.51	0.50	0.51	0.51	0.50	0.51
	(0.12)‡	(0.29)	(0.10)‡	(0.28)	(0.11)‡	(0.33)
% Dilapidated condition	0.06	0.06	0.06	0.07	0.06	0.07
	(0.04)‡	(0.10)	(0.03)‡	(0.11)	(0.03)‡	(0.12)
% Owns phone	0.03	0.51	0.02	0.54	0.02	0.51
	(0.04)†	(0.27)	(0.02)†	(0.26)	(0.02)†	(0.29)
% Owns TV	0.17	0.28	0.15	0.24	0.14	0.23
	(0.14)†	(0.25)	(0.10)†	(0.21)	(0.10)†	(0.22)
% Owns bicycle	0.43	0.47	0.48	0.51	0.45	0.50
	(0.23)†	(0.29)	(0.25)†	(0.30)	(0.26)†	(0.33)
% Owns motorcycle/scooter	0.06	0.13	0.05	0.12	0.05	0.11
	(0.05)†	(0.13)	(0.03)†	(0.12)	(0.03)†	(0.13)
% Owns car	0.01	0.02	0.01	0.02	0.01	0.02
	(0.01)†	(0.04)	(0.01)†	(0.04)	(0.01)†	(0.04)
% Electric/gas cooking	0.05	0.08	0.04	0.06	0.05	0.06
	(0.07)†	(0.16)	(0.06)†	(0.13)	(0.07)†	(0.13)
% Non-elec/gas cooking	0.94	0.92	0.96	0.93	0.95	0.94
	(0.08)†	(0.16)	(0.06)†	(0.13)	(0.07)†	(0.14)
% Thatched roof	0.27	0.21	0.30	0.23	0.27	0.22
	(0.26)†	(0.26)	(0.24)†	(0.25)	(0.24)†	(0.27)
% Mud floor	0.76	0.70	0.78	0.73	0.79	0.76
	(0.18)†	(0.29)	(0.17)†	(0.28)	(0.16)†	(0.28)
% Electric/solar lighting	0.40	0.52	0.31	0.45	0.32	0.49
	(0.29)†	(0.37)	(0.25)†	(0.36)	(0.26)†	(0.38)
% Non-electric/solar lighting	0.59	0.48	0.68	0.54	0.68	0.51
	(0.30)†	(0.37)	(0.25)†	(0.36)	(0.26)†	(0.38)
% Indoor water	0.25	0.29	0.25	0.29	0.21	0.23
	(0.19)†	(0.30)	(0.19)†	(0.29)	(0.17)†	(0.28)
Number of villages		580,643		290,067		62,647

Note. — This table reports means and SDs from the 2001 and 2011 Houselisting Primary Census Abstract. The left two columns include all villages that match across Census datasets, for all 27 RGGVY states. The middle two columns include the subset of those villages located in districts eligible for RGGVY under the 10th Plan. The right two columns include only 10th-Plan villages with 2001 populations between 150 and 450 (our RD bandwidth). The 2001 HPCA reports at either the block or district level, so 2001 columns report these means. A ‡ indicates district SDs, while a † indicates block SDs. Share variables are reported in the HPCA as the share of households satisfying each condition. Households are categorized as good, livable, or dilapidated based on their physical structure. Phone ownership includes both landline and mobile phones; non-electric/gas cooking includes households that cook with kerosene, charcoal, crop residue, cowdung, and firewood; non-electric/solar lighting sources include kerosene and other oil.

Table A.1.6: Summary Statistics – Village Directory

Village Characteristics	All Districts		10th-Plan Districts		10th-Plan Districts 150–450 Population	
	2001	2011	2001	2011	2001	2011
Village area (ha)	432.89	413.25	402.66	378.37	178.09	185.39
	(1369.13)	(972.28)	(1721.84)	(1163.87)	(562.11)	(1664.20)
% of area irrigated	0.43	0.40	0.63	0.56	0.51	0.45
	(10.28)	(14.46)	(14.48)	(20.28)	(6.77)	(9.32)
Drinking water facilities	0.99	0.99	0.99	0.99	0.99	0.99
	(0.07)	(0.08)	(0.08)	(0.08)	(0.08)	(0.09)
Educational facilities	0.80	0.84	0.76	0.80	0.58	0.64
	(0.40)	(0.36)	(0.43)	(0.40)	(0.49)	(0.48)
# of primary schools	1.12	1.41	1.05	1.31	0.59	0.71
	(1.19)	(1.50)	(1.18)	(1.51)	(0.56)	(0.63)
# of secondary schools	0.12	0.24	0.09	0.22	0.01	0.05
	(0.37)	(0.60)	(0.33)	(0.60)	(0.13)	(0.24)
Medical facilities	0.32	0.43	0.29	0.45	0.12	0.28
	(0.47)	(0.50)	(0.45)	(0.50)	(0.32)	(0.45)
# of health ctrs	0.03	0.10	0.03	0.16	0.00	0.01
	(0.18)	(39.55)	(0.17)	(55.95)	(0.07)	(0.11)
Number of dispensaries	0.08	0.07	0.07	0.04	0.01	0.01
	(0.50)	(8.48)	(0.48)	(1.24)	(0.14)	(0.14)
Agricultural credit societies	0.14	0.14	0.11	0.10	0.03	0.03
	(0.34)	(0.34)	(0.31)	(0.30)	(0.16)	(0.16)
Banking facilities	0.06	0.07	0.06	0.07	0.01	0.02
	(0.24)	(0.26)	(0.23)	(0.25)	(0.11)	(0.13)
Electric power	0.77	0.90	0.70	0.91	0.62	0.88
	(0.42)	(0.31)	(0.46)	(0.29)	(0.49)	(0.33)
Elec. power for ag.	0.60	0.68	0.52	0.67	0.41	0.59
	(0.49)	(0.47)	(0.50)	(0.47)	(0.49)	(0.49)
Elec. power for dom. use	0.75	0.89	0.68	0.90	0.59	0.88
	(0.43)	(0.31)	(0.47)	(0.30)	(0.49)	(0.33)
Elec. power for all uses	0.49	0.55	0.38	0.52	0.29	0.43
	(0.50)	(0.50)	(0.49)	(0.50)	(0.46)	(0.49)
Post office	0.23	0.11	0.20	0.10	0.04	0.03
	(0.42)	(0.32)	(0.40)	(0.30)	(0.20)	(0.18)
Bus service	0.75	0.45	0.63	0.35	0.43	0.22
	(0.43)	(0.50)	(0.48)	(0.48)	(0.49)	(0.41)
Rail service	0.05	0.02	0.04	0.03	0.01	0.01
	(0.21)	(0.15)	(0.19)	(0.16)	(0.11)	(0.11)
# of villages	580,643	580,643	290,067	290,067	62,647	62,647

Note. — This table reports means and standard deviations from the 2001 and 2011 Village Directory. The left two columns include all villages that match across Census datasets, for all 27 RGGVY states. The middle two columns include the subset of those villages located in districts eligible for RGGVY under the 10th Plan. The right two columns include only 10th-Plan villages with 2001 populations between 150 and 450 (i.e. our RD bandwidth). Educational facilities include schools and adult training centers. Medical facilities include hospitals, family or maternity welfare centers, clincs, and dispensaries (i.e. small outpatient facilities). Electric power end uses include agriculture, domestic use, and commercial use (the latter is not reported separately in the 2001 Village Directory). All variables are 1/0 when not otherwise noted.

A.1.5 Habitation merge

Because the RGGVY program determines eligibility based on habitation population (only villages with constituent habitations of at least 300 people were eligible for electrification under the 10th Plan), the 2001 village population as reported in the PCA provides an imperfect indicator of eligibility status.[32] Any village with a population below 300 cannot contain a habitation that is eligible under the 10th Plan. However, a village with a population above 300 may or may not be eligible, depending on the number and size of its constituent habitations. In order to accurately assign eligibility status, we supplement our Census panel dataset with information about the habitations within each village.

To the best our knowledge, there exists only one comprehensive nationwide habitation-level data source: the National Rural Drinking Water Programme conducted a census of habitations in 2003 and 2009 with the purpose of assessing the drinking water availability for all rural habitations in India.[33] The two waves of the census list the habitation names and populations for 483,510 and 567,406 villages, respectively, and include over 1.3 and 1.6 million individual habitations, respectively. Together, these datasets cover over 95 percent of India's villages. Unlike the Census products described above, the habitation census datasets do not include village census codes, meaning that the only village identifier present is the village name. Linking these datasets to our master Census village panel is therefore not straightforward.

We apply a multi-step string matching algorithm in order to merge the 2003 and 2009 habitation census data into our Census panel. First, for each village census code, we construct a list of the various transliterations of that village's name that appear in each of our datasets.[34] Second, we search for exact string matches between villages with census codes and village names that appear in the Water Habitation Census. We repeat this procedure for each version of the village name. Because transliterations of names from Hindi, Bengali, and other Indian languages into English are not standardized from Hindi, Bengali, and other Indian languages, merging on multiple spellings increases the likelihood of an exact string match. We save the

[32]Recall that a habitation is a sub-village administrative unit, similar to a neighborhood. Habitations are not official census administrative units, but are frequently used in making policy that affects rural village in India.

[33]The data, along with more information on the National Rural Drinking Water Programme are available from `http://indiawater.gov.in/imisreports/nrdwpmain.aspx`. In fact, the RGGVY program documentation lists this habitation census as a reference to be used by implementing agencies (Ministry of Power (2014b)).

[34]These include the 2001 and 2011 PCA, the 2001 VD, the Census 2001–2011 village concordance, and the RGGVY microdata. The latter simply provide another set of variant spellings, in order to increase the chances of an exact string match.

matches in a separate file, and then remove them from subsequent steps. Third, we repeat this process using the reclink fuzzy string match function in Stata. Finally, for the remaining unmatched villages, we apply the `Masala merge` fuzzy match routine developed by Asher and Novosad (2016). This algorithm computes the Levensthein distance between strings, while accounting for letter substitutions and interpolations common to Hindi transliterations, and produces a set of village name matches.[35]

After completing this matching procedure separately for the 2003 and 2009 habitation datasets, we combine the results into a single village-level dataset to merge with the Census panel. This dataset includes indicators for villages having matched to the 2003 and/or 2009 habitation census; the 2003 and 2009 counts of the number of habitations per village; the 2003 and 2009 populations of the largest habitation in each village; and the 2003 and 2009 populations summed over all constituent habitations. Overall, 86.1 percent of villages match to either the 2003 or 2009 habitation census, and 50.6 percent of matched villages have exactly 1 habitation.[36]

Table A.1.7 reports match rates after each step in this fuzzy merge process. For 91.3 percent of matches (and 88.6 percent of matches with 2001 village populations between 150 and 450), the village population implied by the habitation census (i.e. the sum of all habitation populations in a village) deviates from the village population reported by the Census by less than 20 percent.[37] We suspect large population disparities to indicate erroneous matches, hence we exclude the 11.4 percent of matched villages with disparities greater than 20 percent from all RD specifications.[38] Figure A.1.3 shows habitation matches across the support of 2001 village populations, with lower match rates for smaller villages closer to our RD bandwidth.

[35] `Masala merge` is more accurate and flexible than standard fuzzy merging routines, such as reclink. However we use reclink to remove close matches before applying Masala merge, as this significantly reduces the time required to execute the computationally intensive Masala merge program. We thank the authors for sharing this algorithm with us. A longer description of Masala merge, as well as the code, can be found here: `http://www.dartmouth.edu/~novosad/code.html`.

[36] These match rates are very close to those achieved by Asher and Novosad (2016). For villages that match to both the 2003 and 2009 habitation census, there is a correlation of 0.98 between 2003 and 2009 habitation counts.

[37] Since the habitation censuses were conducted two years after the 2001 Census and two years before the 2011 Census, we should not expect 2003/2009 habitation population to correspond exactly with 2001/2011 village populations — even with 100 percent match accuracy.

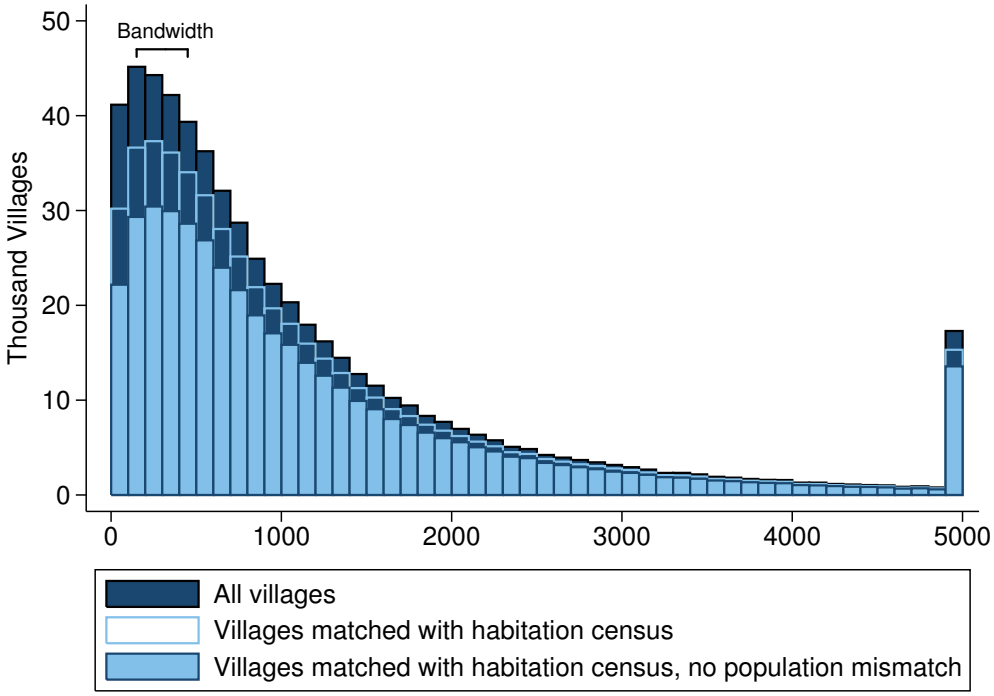
[38] Mismatched villages would lead to measurement error in our RD threshold, which relies on correctly identifying villages with a single habitation. Tables A.2.3 and A.2.12 include these villages as sensitivity analysis.

Table A.1.7: Summary of Habitation Census Merge Results

Habitation census match	2003 and 2009	2003 only	2009 only	Unmatched
A. Match rates (all villages)				
Exact matches	0.486	0.065	0.132	0.317
+ `reclink`	0.639	0.050	0.126	0.185
+ `Masala` merge	0.651	0.085	0.125	0.139
B. Match rates (150–450 population)				
Exact matches	0.471	0.046	0.140	0.343
+ `reclink`	0.627	0.048	0.139	0.186
+ `Masala` merge	0.641	0.069	0.137	0.153
C. Summary statistics (all villages)				
Average habitations per village	2.672	2.397	3.352	
Share single-habitation villages	0.571	0.402	0.517	
Share with population mismatch $> 20\%$	0.087	0.478	0.095	
D. Summary statistics (150–450 population)				
Average habitations per village	1.875	1.776	2.057	
Share single-habitation villages	0.654	0.517	0.607	
Share with population mismatch $> 20\%$	0.114	0.545	0.106	

Note. — This table shows results from the habitation merge algorithm described above. Panels A and B report the share of villages that have merged after each step of the algorithm. Panels C and D calculate summary statistics on the subset of Census panel villages that successfully merge to the habitation dataset. Panels A and C report match counts and summary statistics for all 580,643 villages, while Panels B and D report only the 129,453 villages with 2001 populations between 150 and 450. We define a population mismatch over 20 percent to be a matched village for which the sum of constituent habitation populations deviates from both 2001 and 2011 Census populations by at least 20 percent.

Figure A.1.3: Habitation Merge Results, by 2001 Village Population



Note. — This figure shows a histogram of Indian villages by 2001 population (solid navy), and the subset of villages that we successfully matched with the habitation census (hollow blue). The solid light blue bars show the subset of matched villages with population disparities of less than 20 percent, which we include in our RD analysis. Match rates are lower for smaller villages, yet we still match 84.7 percent of villages with 2001 populations between 150 and 450. Exclude villages with population disparities, this leaves us with 69.6 percent of villages with 2001 populations between 150 and 450.

A.1.6 Socio-Economic Caste Census dataset

We use microdata from the Socioeconomic and Caste Census (SECC) to examine the effects of RGGVY on income. These data were collected between 2011 and 2012, with the intention of recording data on the socioeconomic status of every single Indian.[39] The SECC is a follow-up to the 2002 Below Poverty Line Census, which only included households that were likely to be below the poverty line.[40] The 2011 SECC was expanded to the entire population, and while it used the Enumeration Blocks from the 2011 Census, it was collected separately, using an electronic tablet-based data collection platform. We obtained a subset of these data from the Ministry of Petroleum and Natural Gas, whose liquid petroleum gas subsidy program, Pradhan Mantri Ujjwala Yojana, uses SECC data to determine eligibility.[41] As a result, we observe the universe of rural individuals that are eligible for this fuel subsidy program. This includes "households having one of the Deprivations [in the SECC]"[42], where a "deprivation" is a household poverty indicator. In the SECC, to be considered for poverty status, a household must not be automatically deemed either too wealthy or too destitute. After removing these affluent and destitute households, the remaining households are considered for poverty status. That is, only the subset of households without one (or more) affluence indicator(s) and without one (or more) destitution indicator(s) were tested for poverty indicators. Among the tested households, the households found to display at least one poverty indicator were eligible for the program, and therefore included in our dataset.

Specifically, our sample of fuel-subsidy-eligible households was constructed by first removing all households that satisfied at least one affluence indicator, or "exclusion."[43] In particular, households with motorized 2/3/4 wheelers or fishing boats; with mechanized 3-4 wheeler agricultural equipment; with a Kisan credit card (issued by the government to assist farmers) with a credit limit over Rs. 50,000; with a government employee member; operating a non-agricultural enterprise registered with the government; with a member earning more than Rs. 10,000 per month; paying income tax; paying professional tax; with 3 or more rooms with "pucca" (essentially permanent) walls and roof; with a refrigerator; with a landline phone; with more

[39] See `http://www.secc.gov.in/aboutusReport` for further details.

[40] We do not use the 2002 data in our analysis. It does not comprehensively survey the entire population, and we have been unable to gain access to the data.

[41] The Ministry of Rural Development, who collected the SECC, are in the process of making the full dataset publicly available. As of now, only district-level aggregates are posted at `http://secc.gov.in/welcome`. We downloaded our data in Excel format from `http://lpgdedupe.nic.in/secc/secc_data.html`.

[42] `http://www.pmujjwalayojana.com/faq.html`

[43] See `http://secc.gov.in/reportlistContent`.

than 2.5 acres of irrigated land and irrigation equipment; with 5 or more acres of irrigated land for two or more crop seasons; **or** owning at least 7.5 acres of land and irrigation equipment were all excluded from our SECC dataset. Next, destitute households were "automatically" included if they were without shelter; were destitute or living on alms; earned income from manually scavenging; belonged to a primitive tribal group; or were engaged in legally released bonded labor.

Our SECC dataset includes all remaining households that satisfy **at least one** poverty (or "deprivation") critertion. These criteria are: households with one or fewer rooms, "kuccha" (non-pucca) walls and roof; households with no adult members between the age of 18 and 59; female-headed households with no adult male member between 16 and 59; households with a "differently-able" member with no other able-bodied member; scheduled caste and scheduled tribe households; households with no literate adults above age 25; or landless households deriving a majority of their income from manual labor. These households yield a dataset the 332 million individuals from 81 million households with no affluence indicators (auto-exclusions), no destitution indicators (auto-inclusions), and at least one poverty indicator (deprivation).[44] This represents roughly half of all households in rural India.

This SECC datset contains individual-level data on age, gender, employment, caste, and marital status; and household-level data on the housing stock, land ownership, asset ownership, income sources, and the household head. Importantly for our analysis, the SECC includes data on income, in the form of an indicator for whether the main income earner in each household receives less than 5,000 rupees per month, or between 5,000 and 10,000 rupees per month.[45] Ideally, we would like to have a continuous measure of income, yet we believe this to be the best measurement of income available in any Indian Census product. We cannot use income data included in other surveys, such as the more comprehensive NSS or ICRISAT's VDSA, because these datasets do not contain a large enough sample of villages near the 300-person RGGVY cutoff to be useful for our identification strategy. The SECC also contains information on the main source of household income, and whether at least one household member has a salaried job. We use these variables to test for

[44] We were unable to download SECC data from several districts. While most of these are urban districts with virtually no rural villages, six of these districts contain a nontrivial number of rural villages. These missing districts are: Chamoli, Uttarakhand (1,246 villages); Jalor, Rajasthan (802 villages); Jalpaiguri, West Bengal (768 villages); Dhanbad, Jharkhand (1,760 vilages); Dindigul, Tamil Nadu (481 villages); and Thanjavur, Tamil Nadu (516 villages). Together, these districts contain 5,573 villages, representing less than 1 percent of the total number of villages in our Census dataset.

[45] As described above, households whose primary income earner earns above 10,000 rupees per month were automatically excluded from our sample.

the effects of RGGVY on incomes. The SECC also contains information on land and asset ownership and household head characteristics. Because the SECC contains records at the individual level, we can also use it to corroborate overall employment results from the PCA, and to test for heterogeneity by age group.

In particular, we use data each individual supplies on occupation to test for the effects of RGGVY on sectoral changes. Unlike the other data in the SECC, the occupation field in the SECC was not implemented via a drop-down menu in the survey program, but rather was left open-ended. This results in a large number of employment categories, rampant misspellings, and transliterations from Indian languages to English. We attempt to consolidate these to sectors that match the PCA: "agriculture," "household," and "other" labor. To do this, for each district, we use the `strgroup` command in `Stata` to group similar words. We then replace each group of words with the word in the group that appears in the largest number of entries, under the assumption that it is most likely the correct spelling. From this cleaned subset of words, we use regular expressions to categorize agriculture, students, dependents, shopkeepers, retirees, drivers, household laborers, children, and generic workers. We deploy this grouping algorithm a second time, again keeping the most-represented word in each group. Finally, we sort the remaining occupations into three mutually-exclusive worker categories: "agriculture," "household," and "other" (which are as consistent as possible with the PCA coding of these variables). We also create three mutually exclusive non-labor categories: "students," "dependents," and "none" (the latter includes both unemployed individuals and individuals that do not report an occupation).

We merge these SECC data with our village-level Census dataset using a fuzzy match algorithm similar to that described in Section A.1.5. While the SECC data include Census codes for state, district, and block, they do not include the same village codes that are used by the Census. However, upon inspection we discovered that in nearly all districts, SECC village codes simply reindex Census village codes and largely preserve the relative order of village codes within a block. Hence, the first step of our algorithm searches for exact matches on (reindexed) villages codes, allowing for discrepancies between village names of up to 2 characters. Second, for the remaining unmatched villages, we search for exact matches on villages name. Third, for the few remaining unmatched villages, we apply the `Masala` merge algorithm discussed above. Ultimately, we are able to match 94.2 percent of SECC villages to a village in our Census dataset, with very few (i.e., less than 5 percent) of matches relying on the fuzzy `Masala` merge algorithm.[46] However, because our subset of SECC

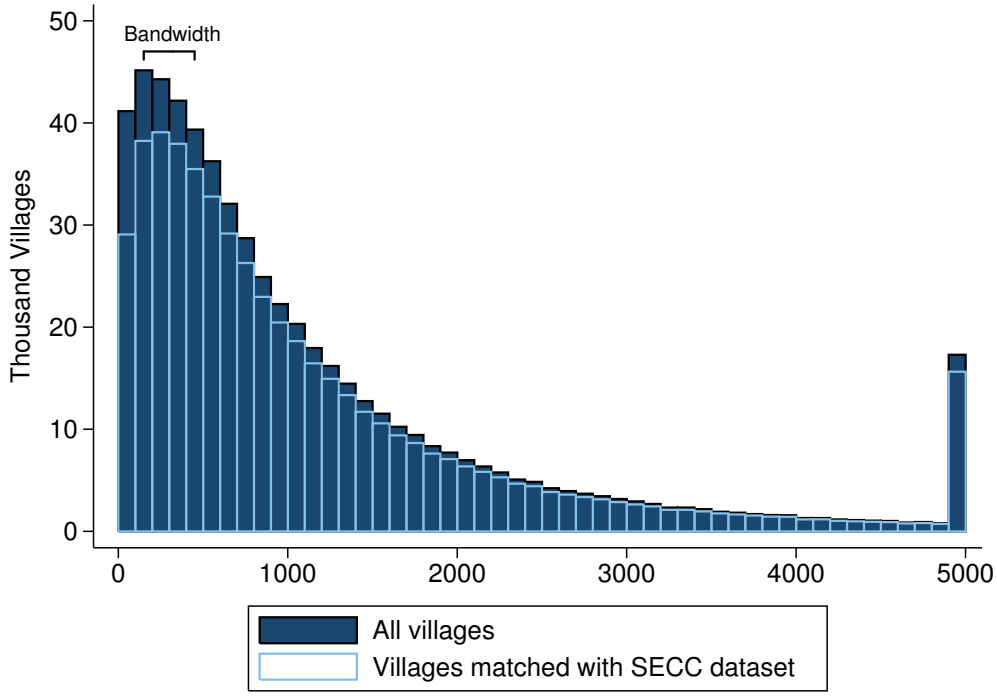
[46]This excludes the 2.7 percent of SECC-Census matches for which the SECC data include either a village population or a village household count over 10 percent larger than those reported by the

data does not include the full population of Indian villages, only 89.7 percent of villages in our Census dataset match to an SECC village. Figure A.1.4 summarizes the proportion of total villages that match to our SECC dataset in a histogram by population.

Table A.1.8 displays summary statistics from our SECC dataset. The first column includes all villages in our raw subset of the SECC data, while the remaining columns report on SECC villages that match to the Census. They reveal no systematic differences induced by our matching algorithm. The bottom row reveals that proportion of village households included in our SECC dataset (i.e. the proportion of households with at least one poverty indicator) is very similar after restricting the matched sample to only villages in RGGVY 10th-Plan districts and within our 150–450 RD bandwidth. Consistent with Table 1.4.1, villages in our main RD analysis sample are slightly more agricultural, and households in these villages are more likely to own land, more likely to rely on cultivation as a source of income, and less likely to earn income from manual labor. Panel B reports employment rates separately by gender and age, and we see that both male and female youth unemployment rates are quite low across all subsets of the SECC data.

2011 Census. We also omit this 2.7 of villages from our analysis, as these population and household count discrepancies make these SECC-Census matches suspect.

Figure A.1.4: SECC Merge Results, by 2001 Village Population



Note. — This figure shows a histogram of Indian villages by 2001 population, and the subset of villages that we successfully matched with a village in the SECC dataset (hollow blue). Overall, we match 88.9 percent of Census villages to the SECC dataset, and for 94.1 percent of these matches, at least 10 percent of total households are included in our SECC dataset (because they have at least one poverty indicator). Within our 150–450 population bandwidth, we match 88.7 percent of Census villages to the SECC dataset, and for 91.9 percent of these matches, our SECC dataset includes at least 10 percent of total village households.

Table A.1.8: Summary Statistics – SECC Village-Level Dataset

2011 Village Characteristics	All Districts	All Districts	10th-Plan Districts	10th-Plan Districts 150–450 Population
	Raw Data	Matched	Matched	Matched
A. Share of households				
Monthly income >Rs 5,000	0.08	0.08	0.09	0.08
	(0.17)	(0.16)	(0.17)	(0.19)
Main income: cultivation	0.27	0.27	0.28	0.34
	(0.31)	(0.30)	(0.30)	(0.35)
Main income: non-farm enterprise	0.01	0.01	0.01	0.01
	(0.05)	(0.05)	(0.04)	(0.05)
Main income: manual/casual labor	0.64	0.65	0.63	0.57
	(0.33)	(0.33)	(0.33)	(0.37)
Main income: domestic service	0.02	0.02	0.02	0.02
	(0.07)	(0.07)	(0.07)	(0.08)
Main income: other sources	0.06	0.06	0.06	0.06
	(0.16)	(0.16)	(0.16)	(0.18)
> 1 HH member w/ salaried job	0.02	0.02	0.02	0.02
	(0.09)	(0.09)	(0.09)	(0.10)
Owning any land	0.36	0.37	0.39	0.47
	(0.32)	(0.31)	(0.31)	(0.35)
Owning irrigation equipment	0.03	0.03	0.04	0.04
	(0.09)	(0.09)	(0.10)	(0.11)
HH head is female	0.16	0.16	0.16	0.18
	(0.14)	(0.14)	(0.15)	(0.18)
HH head is literate	0.45	0.45	0.43	0.43
	(0.24)	(0.23)	(0.23)	(0.26)
HH head w/ middle school educ.	0.19	0.19	0.18	0.18
	(0.16)	(0.16)	(0.16)	(0.18)

Note. — This table reports means and standard deviations from the SECC village-level dataset, which includes all individuals residing in households with at least one poverty indicator in 2011. Variables in Panel A are coded as the share of households in each village, while variables in Panel B are calculated by aggregating up from the individual level. For all summary statistics, the denominator is the total number of households (individuals of a given subpopulation) included in this SECC dataset for each village. The left column reports raw means and standard deviations for all villages in the SECC dataset, while the remaining columns include only SECC villages that match to villages in our Census dataset. We define "adult" to include all individuals at least 16 years old. The last row reports the average fraction of each village's total number of households (per the 2011 Census) that is reported in the SECC dataset, with standard deviations in parentheses.

Table A.1.9: Summary Statistics – SECC Village-Level Dataset (Cont'd)

2011 Village Characteristics	All Districts	All Districts	10th-Plan Districts	10th-Plan Districts 150–450 Population
	Raw Data	Matched	Matched	Matched
B. Share of each subpopulation				
Male adult employment rate	0.73	0.73	0.71	0.71
	(0.20)	(0.19)	(0.20)	(0.23)
Female adult employment rate	0.75	0.76	0.73	0.74
	(0.20)	(0.19)	(0.20)	(0.22)
Male youth employment rate	0.09	0.09	0.08	0.09
	(0.15)	(0.15)	(0.16)	(0.18)
Female youth employment rate	0.09	0.09	0.09	0.09
	(0.16)	(0.15)	(0.16)	(0.18)
Adult literacy rate	0.50	0.50	0.48	0.49
	(0.22)	(0.21)	(0.21)	(0.23)
Number of villages	548,489	516,456	255,989	54,481
Number of households	80.6M	74.6M	36.6M	2M
Number of individuals	331.8M	308.5M	158.7M	8.5M
Share of village households included		0.51	0.49	0.50
		(0.26)	(0.25)	(0.28)

Note. — This table reports means and standard deviations from the SECC village-level dataset, which includes all individuals residing in households with at least one poverty indicator in 2011. Variables in Panel A are coded as the share of households in each village, while variables in Panel B are calculated by aggregating up from the individual level. For all summary statistics, the denominator is the total number of households (individuals of a given subpopulation) included in this SECC dataset for each village. The left column reports raw means and standard deviations for all villages in the SECC dataset, while the remaining columns include only SECC villages that match to villages in our Census dataset. We define "adult" to include all individuals at least 16 years old. The last row reports the average fraction of each village's total number of households (per the 2011 Census) that is reported in the SECC dataset, with standard deviations in parentheses.

A.1.7 DISE schools dataset

In order to include educational outcomes in our analysis, we construct a panel dataset containing the universe of primary (grades 1–5) and upper primary (grades 6–8) schools in India for the 2005–2006 through 2014–2015 school years. These data are publicly available online through the District Information System for Education (DISE)'s School Report Cards website.[47] In total, these data cover 1.68 million schools across over 600,000 villages. However, this yearly school panel is quite unbalanced, with the average school appearing in only 7 out of 10 total years.[48]

The DISE data were originally intended to inform policymakers about the effectiveness of the District Primary Education Programme (DPEP), and data collection for DPEP districts began in 1995. In the early 2000s, DISE was extended to cover the rest of the country in the early 2000s. The information in this dataset is collected by school headmasters or head teachers, and submitted to district- and subsequently state-level authorities before entering the official national system. Quality control is performed by the cluster resource coordinator, and again at the district level.

This dataset includes a large number of variables, although all variables do not appear in every year of the data. We restrict our analysis to student enrollment counts, which appear consistently throughout the full panel. DISE records the number of students, broken down by grade and gender, for each school-year pair.[49] Table A.1.10 summarizes enrollment counts from this dataset at the village level. Across India, there are approximately 200 children in primary school (grades 1–5) and approximately 70 children in upper primary school (grades 6–8) on average per village. There are slightly more boys on average than girls in primary school, and this gap grows among upper primary students. On average, we observe between 1.9 and 2.3 schools per village in the DISE data; we observe between 536,330 and and 559,442 unique villages in each year, reflecting the unbalanced panel nature of these data.

DISE also reports information on various school facility characteristics. The DISE dataset includes, for example, what building materials each school is made out of, whether a school has toilets, whether a school is private or public, the funding received by the school, and the number of teachers employed by each school. Figure A.1.5 provides an example of the data included in the 2012–2013 school year's DISE

[47] These data can be found here: `http://schoolreportcards.in/SRC-New/`.

[48] While this partly reflects new school construction between 2005 and 2014, we encountered a number of errors when downloading data from the DISE website that contributed to the unbalanced nature of this panel. This means that 7 percent of schools in our DISE panel are missing data for at least 1 year between their first and last reporting years. We have no reason to believe that these errors are anything other than random.

[49] In some years, these variables are also broken down by students who belong to Scheduled Castes and Scheduled Tribes, as well as students with special needs.

dataset for a randomly selected school, Government Private School Gujro Ka Jhopra in Rajasthan.

In constructing our DISE panel dataset, we link schools across years based on their unique numerical school codes. DISE also reports the state, district, block, and village name of each school. We use this information to match schools to villages in our main analysis dataset. DISE does not report village census codes, so we are forced to undertake a fuzzy match procedure similar to that used to merge habitations into our village dataset (see Section A.1.5). We begin by building a concordance between school codes and village names as reported in the DISE dataset. In many cases, one school will be associated with multiple village names in different years — this can occur both when villages are split or combined, and when single villages are spelled differently across datasets.[50] We conservatively allow for multiple village spellings for a single school, in order to maximize our chances of matching to a village name in the census.

We search for exact string matches between villages with census codes and village names that appear in the DISE data, following the exact string match, `reclink`, and `Masala merge` algorithm described in Section A.1.5. After establishing a list of census villages matched to DISE villages, we remove duplicates by reintroducing the school code data. In cases where one school code matches to multiple census villages, we select the match which occurs most frequently.[51] We use the results of this matching procedure to construct two datasets: first, a dataset of schools matched to census villages, which we use to generate our school-level RD results in the main text and in Appendix A.2.8; and second, a village-level dataset that sums enrollment across schools in a village to a single village-level observation, which we use for sensitivity analysis below.

Table A.1.11 reports the proportion of census villages that we successfully match to the DISE dataset, after each stage of the matching algorithm. We achieve an overall match rate of 67 percent across all villages, and 58 percent for villages with populations within our main RD bandwidth of 150–450 people. Note that unlike the habitation match described above, we should not expect to be able to match 100 percent of villages *ex ante*: not every village in India is home to a school. According to the Village Directory, only 84 percent of villages actually had schools in 2011. Of these villages, we are able to successfully match 74 percent of villages to schools, and 66 percent of villages with populations between 150 and 450.

[50] We only observe each school once per school year. These different linkages occur when we observe schools across different years.

[51] For example: if school A matches to village 1 in 2005, village 1 in 2006, and village 2 in 2007, we keep the match that links school A with village 1.

Figure A.1.6 displays our match results graphically, demonstrating both that small villages are less likely to have schools than large villages, and that our matching algorithm is more successful in larger villages than in smaller villages. Our final match rates for villages *containing schools* (according to the Village Directory) within our preferred RD bandwidth are comparable to those achieved with habitation merges. Our RD analysis on school enrollment ultimately includes only 41 percent of the villages in our RD regressions for Census outcomes, because only 51 percent of single-habitation villages in our main RD sample match to the DISE data. We also must exclude the 10 percent of school-village matches with missing enrollment data in either 2011–2012 or 2005–2006, as our main RD specification uses 2011–2012 enrollment as an outcome and 2005–2006 enrollment as a control variable.[52]

[52]In many cases, these data appears "missing" because schools likely did not exist during the 2005–2006 school year. We perform sensitivities on the choice of school years in Section A.2.8.

Figure A.1.5: Sample DISE data, 2012–2013

Change Language - Hindi, Marathi, Kannada, Malayalam, Tamil, Telugu, Gujarati, Punjabi

Click to print this page Previous year report card - 2011-12, 2010-11, 2009-10, 2008-09 Click Here to see **Detailed Report** **Click here to see RTE Report Card**

SCHOOL REPORT CARD:2012-13*

State RAJASTHAN	**District Name** BUNDI	Grade* 6/10
School Code 08230302902	**School Name** GOVT. PS GUJRO KA JHOPRA	
Block Name NAINWA	**Cluster Name** BAMANGAON, GGUPS	
Village Name BAMANGAON	**Name of Head Master** KAMLESH (Post Graduate)	

General Information (PINCODE:323801)

Rural / Urban	Rural	Distance from BRC (Km.)	17	Distance From CRC (Km.)	0		
Type of Residential School	NA	Residential School	No	Approachable by all weather roads	Yes		
School Category	Primary only	Lowest Class in school	1	Highest class in school	5		
Pre-primary Section	No	Total Students (Pre-primary)	0	Total Teachers (Pre-primary)	0		
Year of Establishment	2001	Year of recognition	2001	Year of Upgradation from Pri. to U.Pri.			
Management	Local Body	Academic Inspections	13	**School Funds (In Rs.)**	**Recd.**	**Expd.**	
Type of School	Co-Educational	Shift School	No	Teaching Learning Material fund	500	500	
Special School for CWSN	No	No.of visits by Resource teacher for CWSN	0	School Development Fund	5000	5000	
No. of Visits by BRC Coordinator	10	No. of Visits by CRC Coordinator	3	Collection from Students	0	0	

Staff Category (Primary & Upper Primary only)

Teaching Staff	**Pri.**	**U.Pri.**			
Sanctioned	2	0	Part-time instructor (Upper Primary only)	0	Non-teaching Staff	0
In Position	1	0	Teachers Involved in Non-teaching assignments	0	Head Master/Head Teacher	Yes
Contract Teachers	0	0	Avg. working days spent on Non-tch assignments	0	Teachers Received in service Training	1
Graduate & Above	1		Teachers with Professional Qualification	1	Teachers Aged above 55	

School Building, Equipment & Facilities

Number of Building Blocks	**Pucca**	1	**Partially Pucca**	0	**Kuccha**	0	**Tent**	0
Classrooms Require Major Repairs	0	# of Classrooms for Teaching		2	Number of Other Rooms			1
Classrooms Require Minor Repairs	0	Status of School Building		Government	Separate Room for Head Master			No
# of Classrooms in Good Condition	2	Playground		No	Land available for playground			No
Ramp for Disabled Children Needed	Yes	Ramps for Disabled Children Available		Yes	Hand rails for Ramp			No
Medical check-up of Students	Yes	Electricity		No	Computer Aided Learning Lab			No
Furniture for Students	No	# of Computers Available		0	# of Computers Functional			0
Toilets	**Boys**	**Girls**	Library	Yes	# of Books in School Library			158
Total	0	1	Drinking Water Facility	Handpump	Drinking Water Functional			No
Functional	0	1	Measured campus plan prepared	Yes	Boundary Wall			No

Enrolment & Repeaters

Enrolment	2011-12	Total			SC		ST		OBC		Repeaters		CWSN	Muslim	
Grade	All	All	Boys	Girls	Boys	Girls	Boys	Girls	Boys	Girls	Boys	Girls		Boys	Girls
I	5	10	5	5	0	0	0	0	3	4	0	0	0	0	0
II	8	7	3	4	0	0	0	0	1	4	0	0	0	0	0
III	5	6	2	4	0	0	0	0	1	4	0	0	0	0	0
IV	1	5	1	4	0	0	0	0	1	4	0	0	0	0	0
V	2	2	1	1	0	0	0	0	1	0	0	0	0	0	0
VI	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
VII	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
VIII	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Total	21	30	12	18	0	0	0	0	7	16	0	0	0	0	0

Incentives (Previous Academic Year)

	Primary only										**Upper Primary only**									
	General		SC Students		ST Students		OBC Students		Muslim Minority		General		SC Students		ST Students		OBC Students		Muslim Minority	
	Boys	Girls	Boys	Girls	Boys	Girls	Boys	Girls	Boys	Girls	Boys	Girls	Boys	Girls	Boys	Girls	Boys	Girls	Boys	Girls
Textbooks	1	0	0	1	0	0	5	14	0	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Stationary	0	0	0	0	0	0	0	0	0	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Uniform	0	0	0	0	0	0	0	0	0	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Scholarship	0	0	0	0	0	0	0	0	0	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Transport Facility	0	0	0	0	0	0	0	0	0	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Residential Facility	0	0	0	0	0	0	0	0	0	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA

Key Indicators

Pupil : Teacher Ratio	30	Student : Classroom Ratio	15	% Change in Enr. Over Prev. Year	42.86
% SC Students	0.00	%SC Girls to SC Enrolment	0.00	% Girls Enrolment	60.00
% ST Students	0.00	%ST Girls to ST Enrolment	0.00	% CWSN Enrolment	0.00
%Muslim Students	0.00	% Muslim Girls to Muslim Enrolment	0.00	%Classrooms Require Major Repair	0.00
% OBC Enrolment	76.67	% Repeaters to Total Enrolment	0.00	Teachers with Prof. Qualification	100.00

** Based on availability of Ramp, Playground, Boundary Wall, Drinking Water, Boys Toilet, Girls Toilet, Library, PTR≤30 at Primary Schools, PTR ≤35 at Upper Primary Level, SCR≤30 at Primary Schools, SCR ≤35 at Upper Primary Level and Classroom-Teacher Ratio≥1.*

CWSN : Children with Special Needs, BRC : Block Resource Center, CRC : Cluster Resource Center , NA : Not Applicable, na : Not Available

*As on 30th September 2012

Note. — This figure displays the variables available in the 2012–2013 DISE dataset, from the Government Private School Gujro Ka Jhopra in Rajasthan.

Table A.1.10: Summary Statistics – DISE Schools Dataset

	School year beginning in									
	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014
A. Students enrolled per village										
Grades 1–5, male	100.3	104.8	104.2	103.3	100.8	101.0	100.9	97.5	69.0	72.8
	(144.6)	(152.2)	(154.5)	(153.8)	(157.7)	(158.6)	(162.9)	(163.7)	(149.3)	(142.4)
Grades 1–5, female	91.8	97.3	97.3	97.3	95.4	95.6	95.3	92.2	64.4	68.4
	(129.1)	(136.7)	(139.6)	(137.8)	(142.5)	(142.4)	(145.9)	(146.6)	(131.2)	(127.7)
Grades 6–8, male	33.7	36.3	37.5	38.2	39.0	41.0	43.2	44.0	33.7	37.5
	(85.7)	(88.6)	(90.9)	(89.9)	(93.4)	(94.7)	(97.9)	(101.8)	(94.2)	(93.7)
Grades 6–8, female	27.8	31.1	32.9	34.5	36.3	38.8	41.4	42.6	31.9	35.8
	(74.0)	(78.5)	(81.8)	(82.5)	(87.2)	(89.4)	(92.1)	(96.7)	(87.1)	(89.2)
Total	253.6	269.4	271.9	273.4	271.5	276.4	280.9	276.3	198.9	214.6
	(391.3)	(412.4)	(424.5)	(421.8)	(440.2)	(445.9)	(460.4)	(469.8)	(431.6)	(425.1)
B. Observations per year										
Number of villages	536,330	542,135	552,028	526,765	558,708	562,035	564,904	567,565	568,382	559,442
Number of schools	0.99M	1.06M	1.11M	1.09M	1.16M	1.21M	1.24M	1.26M	1.32M	1.28M
Schools per village	1.9	1.9	2.0	2.1	2.1	2.1	2.2	2.2	2.3	2.3
	(1.8)	(2.0)	(2.1)	(2.1)	(2.1)	(2.2)	(2.3)	(2.3)	(2.5)	(2.4)

	Unique villages	Years per village	Unique schools	Years per school
C. Observation counts	667,514	8.30	1,678,545	6.98
		(2.86)		(3.32)

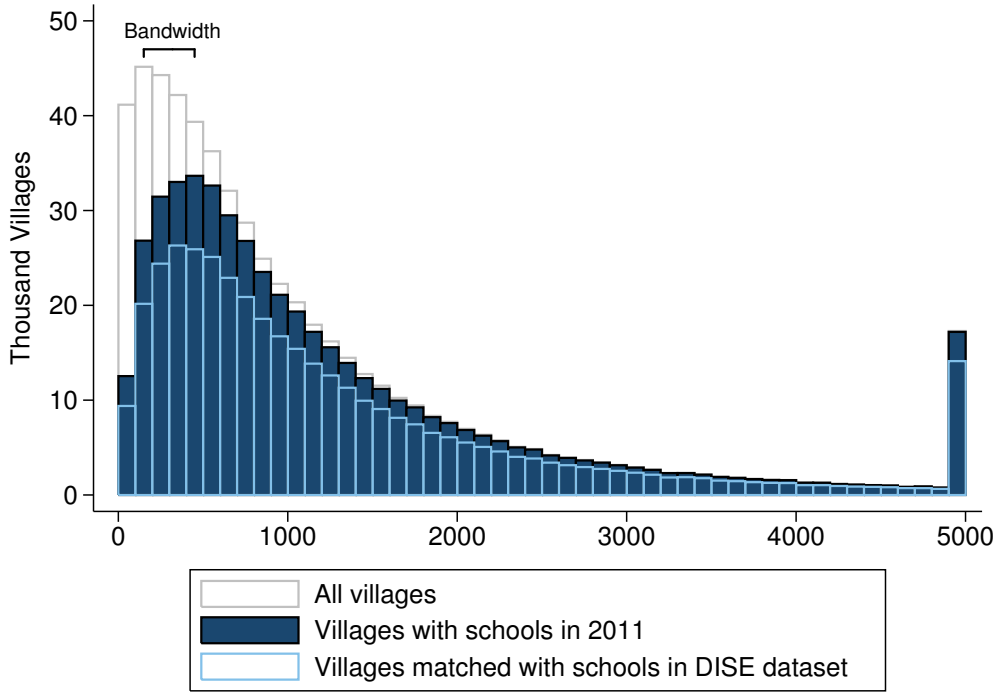
Note. — This table reports means and standard deviations for school enrollment, by village by year. Panel A sums enrollment across all schools within a village, for boys/girls and for primary/upper primary levels. Panel B reports the number of unique villages and schools in each annual cross-section of this unbalanced panel. Panel C counts the number of unique villages and schools present across all ten years, as well as the average number of observations per cross-sectional unit. Note that a "village" in these data is defined by DISE school code identifiers, and this does not exactly correspond to the Census definition of a village.

Table A.1.11: Summary of School Merge Results

	All Villages	150–450 Population
A. Match rates		
Exact matches	0.371	0.324
+ `reclink`	0.624	0.541
+ `Masala merge`	0.667	0.578
B. Summary statistics		
Average number of schools per matched village	3.594	1.972
Share of villages with school in 2011	0.844	0.741
Match rates for villages with school in 2011	0.734	0.654

Note. — This table shows results from the school merge algorithm described above, and it is analogous to Table A.1.7. Panel A reports the share of villages that have merged after each step of the algorithm. Panel B calculates summary statistics on the subset of Census panel villages that successfully merge to the DISE schools dataset. The first column reports match counts and summary statistics for all 580,643 villages, while the second column considers only the 129,453 villages with 2001 populations between 150 and 450. Our merge algorithm does not restrict matches to only the subset of villages reported to have schools in the 2011 Census.

Figure A.1.6: School Merge Results, by 2001 Village Population



Note. — This figure shows a histogram of Indian villages by 2001 population (hollow white), the subset of villages with schools 2011 (as reported by the 2011 Village Directory; solid navy), and the subset of villages that we successfully matched with a school in the the DISE dataset (hollow blue). After adjusting for the relatively share of small villages without schools, we achieve match rates closer to those shown in Figure A.1.3.

A.1.8 Village counts by dataset

Our master dataset includes RGGVY district-level implementation details, village characteristics from the 2001 and 2011 Census, and a count of habitations per village as determined by the fuzzy merge to the habitation census. Each merge between different data sources is imperfect, and Table A.1.12 shows the number of villages present after each step in this merging process. We focus on single-habitation villages in RGGVY 10th-Plan districts (in the middle and right columns), in order to ensure the internal validity of our RD design. Our analysis of nighttime brightness requires an additional merge between this panel dataset and village-specific brightness for the restricted 12-state geospatial sample (as indicated by the fourth row of Table A.1.12).[53] Likewise, our analyses of SECC outcomes and school enrollment each require an additional (fuzzy) merge with the SECC and DISE datasets, respectively (as indicated by rows 4 and 6 of Table A.1.12). We do not enforce the lights match when running regressions on non-spatial outcomes, nor we do not enforce SECC or schools matches when running regressions on non-SECC/non-enrolllment outcomes.

[53]Village shapefiles include attribute tables with 2001 Census codes, allowing a straightforward merge between Census datasets and shapefiles.

Table A.1.12: Count of Villages by Merged Dataset

Number of Villages	Total	RGGVY 10th-Plan	RGGVY 10th-Plan Single-Hab.	RGGVY 10th-Plan Single-Hab. 150–450
Raw Census datasets (village-level)	>593,000			
2001–2011 Census panel	580,643	290,067		
2001–2011 Census panel + habitations	499,799	218,841	115,444	29,765
2001–2011 Census panel + habitations + SECC	380,528	193,493	100,827	25,942
2001–2011 Census panel + habitations + lights	309,105	137,519	70,790	18,686
2001–2011 Census panel + habitations + schools	297,330	140,196	68,872	15,215†

Note. — All village counts exclude Goa and the 7 Union Territories, which were not covered under RGGVY. Rows labeled "habitations" count only villages that we can successfully match to the 2003 or 2009 census of habitations, with population disparities of less than 20 percent. The row labeled "SECC" includes only villages that we successfully link to our SECC dataset. The row labeled "lights" counts only villages that we successfully match village shapefiles. This excludes 5 states for which we do not have shapefiles (Arunachal Pradesh, Meghalaya, Mizoram, Nagaland, Sikkim) and 5 states for which the shapefiles are not correlated with reported village areas (Assam, Himachal Pradesh, Jammu and Kashmir, Uttar Pradesh, Uttarakhand). The row labeled "schools" includes only villages that we successfully link to a school in the DISE dataset. We conduct our RD analysis on villages in the right-most column, which restricts the middle column to villages with 2001 populations between 150 and 450.

† Only 81 percent of village-school matches report nonmissing enrollment data for both 2011 and 2005, further reducing our sample size for regressions on enrollment outcomes.

A.2 Empirics

In this appendix, we provide a variety of robustness and falsification exercises to complement the results in the main text. We first discuss sensitivity of our nighttime brightness results, and then move to the economic outcomes of interest.

A.2.1 Nighttime brightness: RD robustness

A.2.1.1 Sample and outcome variable definition

As discussed above in Section A.1.2, because of missing or low-quality shapefiles, we are forced to drop ten states from our nighttime lighting analysis sample. This leaves us with twelve states with 10th-Plan RGGVY districts, which contain over 18,000 single-habitation villages within our RD bandwidth. Our main specification assigns village brightness based on each village's brightest pixel. We focus on the brightest pixel due to the typical organizational structure of South Asian villages, which have concentrated inhabited regions surrounded by agricultural lands. Since RGGVY is targeted at electrifying public places and homes, the brightest pixel will best reflect brightness that is attributable to RGGVY infrastructure upgrades.

We also linearly project village brightness on the values in adjacent years in order to remove year-to-year measurement error and focus on more permanent year-to-year changes in electricity use (see Equation (A.1) in Appendix A.1.3). Table A.2.1 demonstrates the degree to which this cross-year calibration refines our estimates, while Table A.2.2 compares these weight-averaged linear projections to unweighted 3- and 5-year averages. These unweighted averages yield very similar point estimates, however we see that the linear projections provide greater precision. In addition, Table A.2.2 shows that our results are almost identical when we average village brightness across all pixels (i.e., Columns (2), (4), and (6)), as opposed to using the maximum brightness.

We have chosen to use NOAA's average lights product, as opposed to the more processed stable nighttime lights. While the latter images exclude fires and other sporadic lights, they are also less likely to detect the low levels of lighting we might expect from small recently electrified villages. Table A.2.3 compares RD estimates using each data product, and we see that the stable lights actually yield larger RD coefficients than the (preferred) average visible lights. Columns (2) and (4) of Table A.2.3 display the results of an additional sensitivity test, by including 2,373 single-

Table A.2.1: RD Sensitivity – Raw vs. Projected Lights

2011 village brightness	Raw Lights	Projected (2010–2012)	Projected (2009–2013)
	(1)	(2)	(3)
1[2001 pop $\geq$ 300]	0.0788	0.1408**	0.1493**
	(0.0698)	(0.0680)	(0.0603)
2001 population	−0.0002	−0.0006	−0.0008
	(0.0009)	(0.0007)	(0.0007)
1[2001 pop $\geq$ 300] $\times$ 2001 pop	0.0012	0.0007	0.0008
	(0.0011)	(0.0009)	(0.0008)
2001 Control	Yes	Yes	Yes
State FEs	Yes	Yes	Yes
RD bandwidth	150	150	150
Number of observations	18,686	18,686	18,686
Number of districts	130	130	130
Mean of dependent variable	6.244	6.368	6.370
R^2	0.673	0.746	0.766

Note. — This table shows results from estimating Equation (1.1), using raw and projected 2011 brightness. Column (1) uses raw 2011 lights as the dependent variable. Column (2) uses a linear projection of 2011 lights on 2010 and 2012 values. Column (3) reproduces our preferred specificaton from Table 1.5.2, using a linear projection of 2011 lights on 2009, 2010, 2012, and 2013 values. For each column, the 2001 lights control uses the analogous projection, for consistency within each regression. Each regression includes all single-habitation villages in 10th-Plan districts with 2001 populations in the RD bandwidth (a 150-person bandwidth includes villages with 2001 populations between 150 and 450), for the 12 states with available village shapefiles that match to Census village areas with a correlation above 0.35. Standard errors are clustered at the district level. Significance: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

habitation villages whose official village Census populations differ from their matched habitation populations by over 20 percent. Such substantial population disparities suggest that these Census villages may be wrongly matched to single-habitation villages, hence we have excluded them from all other specifications. As expected, including these potentially erroneous matches attenuates our RD estimates, by having introduced measurement error in the RD indicator variable.

Table A.2.2: RD Sensitivity – Alternative Lights Variables

2011 village brightness	Projected		3-Year Average		5-Year Average	
	Max	Mean	Max	Mean	Max	Mean
	(1)	(2)	(3)	(4)	(5)	(6)
1[2001 pop $\geq$ 300]	0.1493**	0.1386**	0.1163*	0.1146*	0.1219*	0.1188**
	(0.0603)	(0.0556)	(0.0671)	(0.0622)	(0.0631)	(0.0592)
2001 population	−0.0008	−0.0008	−0.0004	−0.0004	−0.0004	−0.0004
	(0.0007)	(0.0007)	(0.0008)	(0.0008)	(0.0008)	(0.0007)
1[2001 pop $\geq$ 300] × 2001 pop	0.0008	0.0006	0.0007	0.0004	0.0006	0.0004
	(0.0008)	(0.0007)	(0.0009)	(0.0009)	(0.0009)	(0.0008)
2001 Control	Yes	Yes	Yes	Yes	Yes	Yes
State FEs	Yes	Yes	Yes	Yes	Yes	Yes
RD bandwidth	150	150	150	150	150	150
Number of observations	18,686	18,686	18,686	18,686	18,686	18,686
Number of districts	130	130	130	130	130	130
Mean of dependent variable	6.370	6.077	7.047	6.711	6.735	6.411
R^2	0.766	0.762	0.753	0.751	0.761	0.760

Note. — This table shows results from estimating Equation (1.1), using alternative definitions of 2011 brightness as the outcome variable. Columns (1)–(2) use a linear projection of 2011 lights on 2009, 2010, 2012, and 2013 values, with Column (1) reproduces our preferred specification from Table 1.5.2. Columns (3)–(4) use unweighted averages of 2010–2012 values, while Columns (5)–(6) use unweighted averages of 2009–2013 values. Columns (1), (3), and (5) assign village brightness based on the brightest pixel, whereas Columns (2), (4), and (6) average village brightness across all pixels contained in the village boundary. We construct 2001 lights controls to be analogous to their respective outcome variables, for consistency within each regression. Each regression includes all single-habitation villages in 10th-Plan districts with 2001 populations in the RD bandwidth (a 150-person bandwidth includes villages with 2001 populations between 150 and 450), for the 12 states with available village shapefiles that match to Census village areas with a correlation above 0.35. Standard errors are clustered at the district level. Significance: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

Table A.2.3: RD Sensitivity – NOAA DMSP–OLS Datasets

2011 village brightness	Average Visible Lights		Stable Lights	
	(1)	(2)	(3)	(4)
1[2001 pop $\geq$ 300]	0.1493**	0.1170**	0.1810**	0.1419**
	(0.0603)	(0.0530)	(0.0726)	(0.0680)
2001 population	−0.0008	−0.0004	−0.0012	−0.0008
	(0.0007)	(0.0006)	(0.0008)	(0.0007)
1[2001 pop $\geq$ 300] $\times$ 2001 pop	0.0008	0.0004	0.0008	0.0005
	(0.0008)	(0.0007)	(0.0010)	(0.0009)
Forced population match	Yes	No	Yes	No
2001 Control	Yes	Yes	Yes	Yes
State FEs	Yes	Yes	Yes	Yes
RD bandwidth	150	150	150	150
Number of observations	18,686	21,059	18,686	21,059
Number of districts	130	130	130	130
Mean of dependent variable	6.370	6.333	4.873	4.825
R^2	0.766	0.775	0.782	0.789

Note. — This table shows results from estimating Equation (1.1), using alternative NOAA DMSP–OLS lights data. Columns (1)–(2) show the average visible lights data, which is our preferred measure of nighttime brightness (Column (1) is reproduced from Table 1.5.2). Columns (3)–(4) show results for NOAA's more processed stable lights product. Columns (2) and (4) include villages that match to the 2003 and/or 2009 habitation census datasets, but have population disparities of greater than 20 percent (indicating potentially erroneous matches). For each specification, 2011 and 2001 brightness values are constructed using a linear projection on the brightness values of adjacent years, using their respective NOAA data products. Each regression includes all single-habitation villages in 10th-Plan districts with 2001 populations in the RD bandwidth (a 150-person bandwidth includes villages with 2001 populations between 150 and 450), for the 12 states with available village shapefiles that match to Census village areas with a correlation above 0.35. Standard errors are clustered at the district level. Significance: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

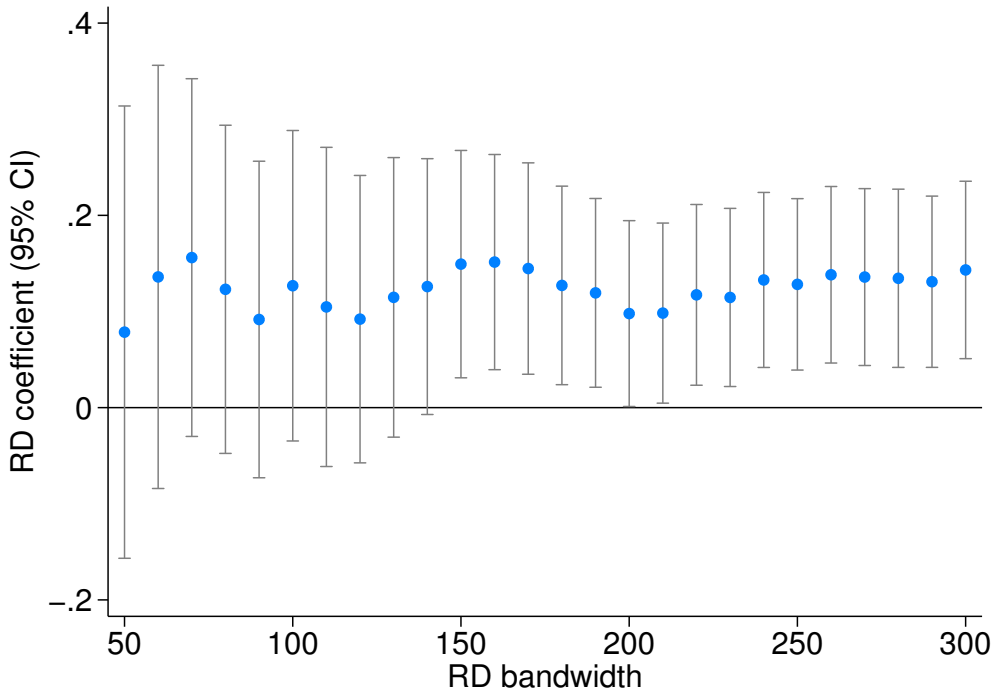
A.2.1.2 Bandwidths

Because RD designs rely on data close to the threshold to estimate a local average treatment effect (LATE) *at* the threshold, any sensitivity of RD estimates to bandwidth selection could undermine their internal validity. Our preferred specification uses a 150-person bandwidth on either side of the 300-person cutoff, including villages of between 150 and 450 people. Figure A.2.1 estimates Equation (1.1) using bandwidths ranging from 50 to 300 people. This demonstrates that our RD point estimate is not sensitive to bandwidth selection, as our point estimates are quite stable. At the same time, they lose significance at narrower bandwidths, which reduce the RD sample size.

We choose a 150-person bandwidth for two primary reasons. First, we want to ensure that our estimation window does not overlap with two other program eligibility cutoffs — the 100-person cutoff implemented under the 11th Plan of RGGVY, and the 500-person cutoff used in the PMGSY road-building program. Both thresholds may have led to increases in nighttime brightness with the potential to confound our estimates at the 300-person cutoff, either directly through electrification or indirectly due to the economic benefits of road infrastructure. Second, we want to exclude very small villages (i.e. population less than 50) in our RD sample. These villages are likely quite different than villages of 200–400 inhabitants.

As an alternative strategy, we implement on the optimal RD bandwidth procedure formalized by Imbens and Kalyanaraman (2012). Using this technique, we derive an optimal bandwidth of 137 using a uniform kernel, 162 using an Epanechnikov kernel, and 174 using a triangular kernel. Figure A.2.1 shows that had we chosen any of these three bandwidths, our results would have be nearly identical.

Figure A.2.1: RD Sensitivity – Nighttime Brightness, Bandwidth



Note. — This figure presents our bandwidth sensitivity analysis for Equation (1.1), estimated separately on bandwidths ranging from 50 (i.e., 250–350 people) to 300 (i.e., 0–600 people). Each dot represents the point estimate on the RD discontinuity at a given bandwidth around the 300-person cutoff, with 95 percent confidence intervals clustered at the district level. Our chosen bandwidth of 150 includes villages with populations between 150 and 450. The optimal bandwidth, calculated using the algorithm proposed by Imbens and Kalyanaraman (2012), is 137 using a uniform kernel, 162 using an Epanechnikov kernel, and 174 using a triangular kernel.

A.2.1.3 Functional form

Our preferred RD specification excludes higher-order polynomials, following Gelman and Imbens (2014). We control for only a linear function of the running variable, allowing the slope to differ on either side of the RD threshold. While this has become standard practice for implementing RD designs, we also test for sensitivity of our estimates to higher-order polynomials. Table A.2.4 compares our preferred speci-

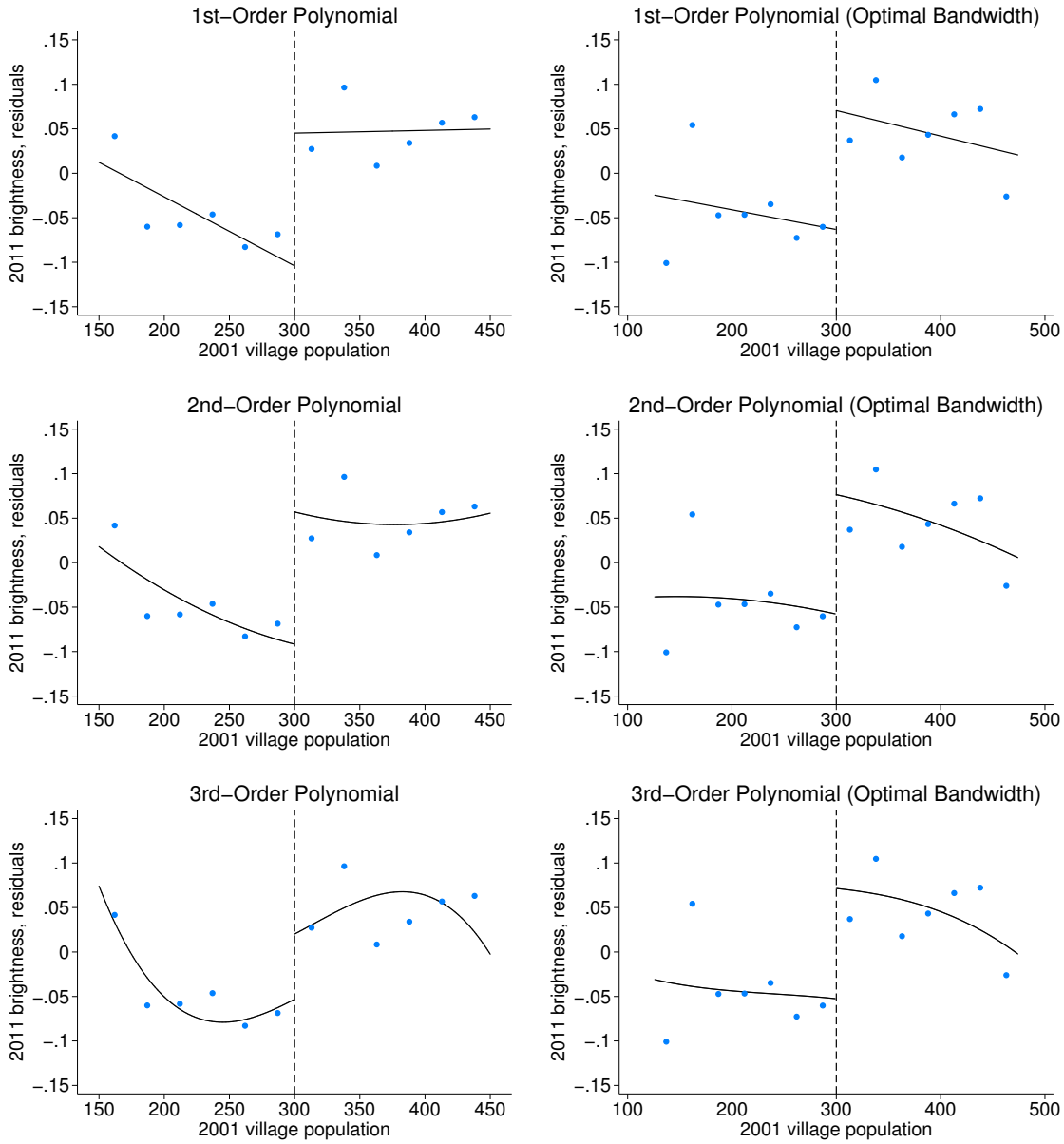
fication (in Column (1)) to specifications with 2nd- and 3rd-order terms. Our RD estimate is robust to the inclusion of a quadratic function of the running variable, but we lose precision when we include higher-order terms. Figure A.2.2 presents these results graphically, and we see that the 3rd-order polynomial appears to be affected by observations far from the RD threshold (as Gelman and Imbens (2014) warn might occur with higher-order polynomials). The right-hand panels of Figure A.2.2 present the same three specifications using a 174-person bandwidth, also reported in Columns (2), (4), and (6) of Table A.2.4. This is our largest Imbens-Kalyanaraman optimal bandwidth (which we calculated using uniform, Epanechnikov, and triangular kernels). We see that extending the bandwidth affects the curvature of the 3rd-order polynomial estimates, while the fitted linear and quadratic curves appear mostly unchanged.

Table A.2.4: RD Sensitivity – Higher Order Polynomials

2011 village brightness	Linear		Quadratic		Cubic	
	(1)	(2)	(3)	(4)	(5)	(6)
1[2001 pop $\geq$ 300]	0.1493**	0.1337**	0.1497**	0.1372**	0.0742	0.1261
	(0.0603)	(0.0549)	(0.0607)	(0.0550)	(0.0987)	(0.0864)
2001 population	−0.0008	−0.0002	−0.0011	−0.0021	0.0001	−0.0019
	(0.0007)	(0.0005)	(0.0016)	(0.0013)	(0.0016)	(0.0015)
1[2001 pop $\geq$ 300] × 2001 pop	0.0008	−0.0001	0.0014	0.0036	0.0014	0.0036
	(0.0008)	(0.0006)	(0.0029)	(0.0023)	(0.0029)	(0.0024)
(2001 population)^2			−0.0000	−0.0000	−0.0000	−0.0000
			(0.0000)	(0.0000)	(0.0000)	(0.0000)
(2001 population)^3					−0.0000	−0.0000
					(0.0000)	(0.0000)
2001 Control	Yes	Yes	Yes	Yes	Yes	Yes
State FEs	Yes	Yes	Yes	Yes	Yes	Yes
RD bandwidth	150	174	150	174	150	174
Number of observations	18,686	21,551	18,686	21,551	18,686	21,551
Number of districts	130	130	130	130	130	130
Mean of dependent variable	6.370	6.344	6.370	6.344	6.370	6.344
R^2	0.766	0.775	0.766	0.775	0.766	0.775

Note. — This table compares our main RD specification to two specificaions with higher-order polynomials, as presented graphically in Figure A.2.2. Columns (1)–(2) estimate our main specification using a linear function of the running variable, 2001 population. Columns (3)–(4) use a quadratic function of population, while Columns (5)–(6) use a cubic function of population. Each regression includes all villages meeting the above sample criteria with 2001 populations in the RD bandwidth (either our preferred bandwidth of 150 or our largest optimal bandwidth of 174), for the 12 states with available village shapefiles that match to Census village areas with a correlation above 0.35. Standard errors are clustered at the district level. Significance: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

Figure A.2.2: RD Sensitivity – Nighttime Brightness, Higher Order Polynomials



Note. — This figure presents our RD for 2011 nighttime brightness, estimated using 1st-, 2nd-, and 3rd-order polynomials. The three figures in the left-hand column correspond to the regressions in Table A.2.4, and the top-left panel reproduces Figure 1.5.4. The three figures in the right-hand column use the Imbens-Kalyanaraman optimal bandwidth of 174 (which contains 21,551 villages). Blue dots show average residuals from regressing 2011 nighttime brightness on 2001 nighttime brightness and state fixed effects. Each dot contains approximately 1,600 villages, averaged in 25-person population bins. Linear terms are estimated separately on each side of the threshold, and higher-order terms are restricted to be the same on each side of the threshold. Plots include all within-bandwidth, single-habitation, 10th-Plan villages, for the 12 states with available village shapefiles that correspond to Census village areas (with a correlation above 0.35).

A.2.1.4 Fixed effects and 2001 controls

RD designs do not rely on fixed effects or controls for identification, but their inclusion can greatly improve efficiency (Lee and Lemieux (2010)). Our main specification controls for both state fixed effects and the 2001 level of nighttime brightness. Table A.2.5 estimates this specification with and without this 2001 control, as well as with no fixed effects, state fixed effects, and district fixed effects. We see that the RD estimates without the 2001 control are noisy and imprecise, meaning that our RD requires this baseline control to find a statistically detectable effect. Stated differently, we find that RGGVY eligibility has led to a statistically significant increase in brightness *conditional* on 2001 levels of brightness, but we are unable to statistically detect the *unconditional* effect of RGGVY eligibility on 2011 brightness. This makes sense: nighttime brightness prior to RGGVY is very heterogeneous; conditioning on the pre-period level dramatically improves the signal-to-noise ratio. This is not an uncommon practice in papers using remotely sensed data, as Jayachandran et al. (2016) also rely on pre-period controls in order to detect treatment effects. Table A.2.5 also reveals that our estimates are not sensitive to the inclusion of fixed effects. We have chosen to include state fixed effects in our main specification, in order to control for large differences in nighttime brightness across our 12 sample states (see Table A.2.6).

Table A.2.7 introduces additional 2001 village controls to our main RD specification. Our RD point estimates are not affected by the inclusion of these pre-RGGVY controls, even though two (literacy rate and presence of road) have statistically significant associations with 2011 nighttime brightness. Interestingly, 2001 electric power indicator variables are poor predictors of 2011 brightness. This is not surprising, considering that these Census variables do not capture the intensity of electrification within the village.

Table A.2.5: RD Sensitivity – Fixed Effects and 2001 Control

	2011 village brightness					
	(1)	(2)	(3)	(4)	(5)	(6)
1[2001 pop $\geq$ 300]	0.1002	0.0836	0.0745	0.1451**	0.1493**	0.1292**
	(0.1201)	(0.1149)	(0.1144)	(0.0635)	(0.0603)	(0.0568)
2001 population	0.0006	0.0007	0.0009	−0.0012*	−0.0008	−0.0005
	(0.0013)	(0.0013)	(0.0012)	(0.0007)	(0.0007)	(0.0006)
1[2001 pop $\geq$ 300] $\times$ 2001 pop	0.0008	0.0005	−0.0001	0.0012	0.0008	0.0005
	(0.0017)	(0.0016)	(0.0015)	(0.0008)	(0.0008)	(0.0008)
2001 Control	No	No	No	Yes	Yes	Yes
State FEs	No	Yes	No	No	Yes	No
District FEs	No	No	Yes	No	No	Yes
RD bandwidth	150	150	150	150	150	150
Number of observations	18,686	18,686	18,686	18,686	18,686	18,686
Number of districts	130	130	130	130	130	130
Mean of dependent variable	6.370	6.370	6.370	6.370	6.370	6.370
R^2	0.001	0.062	0.169	0.748	0.766	0.791

Note. — This table shows results from estimating Equation (1.1), using with and without fixed effects and the 2001 control. For each regression, the dependent variable is the maximum village brightness for 2011. Column (5) reproduces our preferred specification from Table 1.5.2. Each regression includes all single-habitation villages in 10th-Plan districts with 2001 populations in the RD bandwidth (a 150-person bandwidth includes villages with 2001 populations between 150 and 450), for the 12 states with available village shapefiles that match to Census village areas with a correlation above 0.35. Standard errors are clustered at the district level. Significance: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

Table A.2.6: Nighttime Brightness by State

RD Sample State	2001 Average Village Brightness	2011 Average Village Brightness	Villages in RD Sample
Andhra Pradesh	5.021	7.049	1,020
Bihar	3.347	4.559	2,555
Chhattisgarh	3.999	6.807	364
Gujarat	5.111	6.007	333
Haryana	12.329	17.625	31
Jharkhand	4.211	6.537	549
Karnataka	5.495	8.002	2,747
Madhya Pradesh	4.774	5.191	1,568
Maharashtra	4.946	5.563	315
Orissa	4.258	5.681	471
Rajasthan	4.757	7.039	4,138
West Bengal	4.413	6.568	4,595
Total	4.583	6.492	18,686

Note. — This table shows the average 2001 and 2011 brightness by state, for villages in our main RD sample. The 2001 and 2011 brightness variables used in this table are the same linear projections used in our main specification.

Table A.2.7: RD Sensitivity – 2001 Village Controls

	2011 village brightness					
	(1)	(2)	(3)	(4)	(5)	(6)
1[2001 pop $\geq$ 300]	0.1489**	0.1484**	0.1510**	0.1431**	0.1468**	0.1808***
	(0.0603)	(0.0605)	(0.0606)	(0.0615)	(0.0604)	(0.0573)
2001 population	−0.0009	−0.0008	−0.0008	−0.0008	−0.0008	−0.0013**
	(0.0007)	(0.0007)	(0.0007)	(0.0007)	(0.0007)	(0.0006)
1[2001 pop $\geq$ 300] × 2001 pop	0.0009	0.0008	0.0008	0.0009	0.0009	0.0016**
	(0.0008)	(0.0008)	(0.0008)	(0.0008)	(0.0008)	(0.0007)
2001 literacy rate	0.9232***					
	(0.2937)					
2001 share SC/ST		−0.1558				
		(0.1181)				
2001 share of area irrigated			0.2959			
			(0.2699)			
2001 presence of road (0/1)				0.2690***		
				(0.0630)		
2001 electricity, any use (0/1)					0.0072	
					(0.1343)	
2001 electricity, all uses (0/1)						0.1593
						(0.1333)
2001 Brightness	Yes	Yes	Yes	Yes	Yes	Yes
State FEs	Yes	Yes	Yes	Yes	Yes	Yes
RD bandwidth	150	150	150	150	150	150
Number of observations	18,686	18,686	18,669	18,358	18,647	16,907
Number of districts	130	130	130	130	130	130
Mean of dependent variable	6.370	6.370	6.370	6.398	6.365	6.451
R^2	0.767	0.766	0.766	0.766	0.766	0.773

Note. — This table introduces 2001 village controls to our main RD specification. SC/ST refer to official Scheduled Caste and Scheduled Tribe designations, while the 2001 presence of a road indicator includes both paved and mud roads. 2001 electric power indicators consider three distinct end uses: domestic, agricultural, and commercial. Electricity for any use (all uses) indicates whether *any one* (*all*) of these three end-use sectors had electric power in 2001. Each regression includes all villages meeting the above sample criteria with 2001 populations in the RD bandwidth (a 150-person bandwidth includes villages with 2001 populations between 150 and 450), for the 12 states with available village shapefiles that match to Census village areas with a correlation above 0.35. Standard errors are clustered at the district level. Significance: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

A.2.1.5 Pre-RGGVY lights

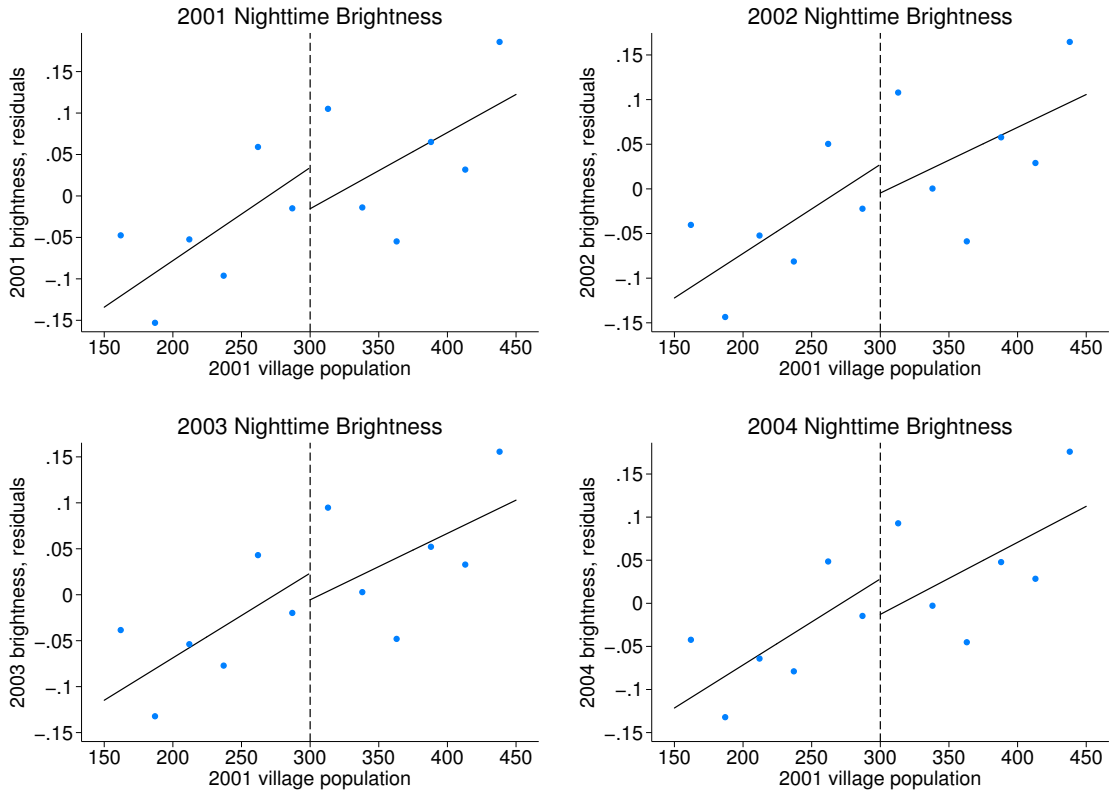
Covariate smoothness across the threshold is a key identifying assumption in RD designs. If pre-existing factors were to jump discontinuously at the threshold, this would have the potential to confound our RD estimates. Figure A.2.3 and Table A.2.8 show that lights from 2001–2004 (i.e., *before* the announcement of RGGVY) do not exhibit any significant breaks at the 300-person cutoff. This supports our assumption that selection into eligibility around the 300-person cutoff was as-good-as-random.

Table A.2.8: RD Sensitivity – Pre-RGGVY Brightness

Village brightness	2001	2002	2003	2004
	(1)	(2)	(3)	(4)
$\mathbf{1}[\text{2001 pop} \geq 300]$	−0.0496	−0.0319	−0.0290	−0.0411
	(0.0736)	(0.0721)	(0.0649)	(0.0648)
2001 population	0.0011	0.0010	0.0009	0.0010
	(0.0007)	(0.0007)	(0.0007)	(0.0007)
$\mathbf{1}[\text{2001 pop} \geq 300] \times$ 2001 pop	−0.0002	−0.0003	−0.0002	−0.0002
	(0.0011)	(0.0011)	(0.0010)	(0.0010)
State FEs	Yes	Yes	Yes	Yes
RD bandwidth	150	150	150	150
Number of observations	18,686	18,686	18,686	18,686
Number of districts	130	130	130	130
Mean of dependent variable	4.512	4.448	3.603	3.918
R^2	0.053	0.055	0.058	0.062

Note. — This table shows results from estimating our RD specification using brightness outcomes for years prior to the announcement of RGGVY. For each regression, the dependent variable is the maximum village brightness for a given year, after applying a linear projection onto brightness in adjacent years. Unlike Equation (1.1), these specifications do not control for pre-RGGVY brightness. Each regression includes all single-habitation villages in 10th-Plan districts with 2001 populations in the RD bandwidth (a 150-person bandwidth includes villages with 2001 populations between 150 and 450), for the 12 states with available village shapefiles that match to Census village areas with a correlation above 0.35. Standard errors are clustered at the district level. Significance: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

Figure A.2.3: RD Sensitivity – Pre-RGGVY Brightness



Note. — These four figures estimate our preferred RD specification using nighttime brightness for 2001–2004, corresponding to the regressions in Table A.2.8. Each graph uses a different year's nighttime brightness as the dependent variable, after having linearly projected these values on adjacent years. Blue dots show average residuals from regressing that year's nighttime brightness on state fixed effects (without any other controls). Each dot contains approximately 1,600 villages, averaged in 25-person population bins. Lines are estimated separately on each side of the 300-person threshold, for all 18,686 single-habitation villages between 150–450 people, in 10th-Plan districts, for the 12 states with available village shapefiles that correspond to Census village areas (with a correlation above 0.35).

A.2.1.6 Standard errors

Finally, our main specification clusters standard errors at the district level. This allows for arbitrary dependence in the error structure between any two villages in the same district. Because RGGVY projects were approved based on district-specific implementation plans (or DPRs), and funds were awarded to district implementing agencies, we allow for within-district dependence to control for any unobserved factors affecting RGGVY implementation. Since districts are geographically contiguous area, this also accounts for spatial correlations between nearby villages.

Table A.2.9 reports standard errors on our RD point estimate for alternative assumptions about the error structure. We see that clustering by census block yields slightly larger standard errors than clustering by district.[54] We also calculate Conley "spatial HAC" standard errors, which are robust to spatial dependencies between villages within a given geographic radius, as well as heteroscedasticity and autocorrelation.[55] Table A.2.9 shows that spatial standard errors estimated with a 50-km and 250-km bandwidth are smaller than our preferred standard error estimates.

Table A.2.9: RD Sensitivity – Alternative Standard Errors

	Clustered by Block	Clustered by District	Spatial HAC (50 km)	Spatial HAC (250 km)
Standard Error on RD coefficient	(0.0680)**	(0.0603)**	(0.0602)**	(0.0542)***

Note. — This table shows robustness of our RD point estimate ($\hat{\beta}_1$ from Table 1.5.2) to alternative standard error assumptions. Column (1) clusters standard errors by census block, which is the administrative unit between district and village. Column (2) clusters by district, which is our preferred method. Columns (3)–(4) apply standard errors that are robust to heteroscedasticity and spatial correlation, with bandwidths of 50 and 250 km. Significance: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

[54] Census block is the administrative unit that is smaller than district but larger than village. The only administrative unit larger than district is state. Because our nightlights regressions only include 12 states, we do not cluster at the state level for fear of bias resulting from having too few clusters (Cameron and Miller (2015)).

[55] We use code from Fetzer (2014) to implement this Conley HAC procedure. This code can be found online: `http://www.trfetzer.com/conley-spatial-hac-errors-with-fixed-effects/`. It is in turn based on code from Hsiang (2010), itself based on theory from Conley (1999) and Conley (2008). Because our RD regression is cross-sectional and does not include multiple observations for each village, the autocorrelation ("AC") component of the spatial HAC estimator is not relevant for our purposes. Our spatial standard errors apply a uniform kernel, which yields very similar estimates to those generated using a linear Bartlett kernel.

A.2.2 Nighttime brightness: validity tests

In addition to the robustness checks above, we also conduct placebo, randomization, and falsification tests to corroborate our results for nighttime brightness. Section 1.5 in the main text discuss our placebo test and randomization inference procedures, and Figure 1.5.5 reveals that our observed increase in nighttime brightness is very unlikely to result from spurious correlations in the relationship between brightness and village populations.

Section 1.5 describes an additional falsification exercise, where we demonstrate that nighttime brightness only increased for the subset of villages within 10th-Plan eligible districts containing a single habitation. We find no increase in brightness at the 300-person threshold for villages in 11th-Plan districts (for which the eligibility cutoff was reduced from 300 to 100 people) or for villages with more than one habitation (for which the total village population does not correspond to the population used to determine eligibility). Figure 1.5.6 graphically compares our main RD results to three "falsification" samples: multi-habitation villages in 10th-Plan districts, single-habitation villages in 11th-Plan districts, and multi-habitation villages in 11th-Plan districts. Table A.2.10 presents the same results in regression format. As expected, none of these alternative samples exhibits evidence of a positive discontinuity at the 300-person cutoff. This provides further evidence that the RGGVY program is what is driving the increase in nighttime brightness for 10th-Plan, single-habitation villages.

Table A.2.10: RD Sensitivity – Falsification Tests

2011 village brightness	10th-Plan Single-Hab.	10th-Plan Multi-Hab.	11th-Plan Single-Hab.	11th-Plan Multi-Hab.
	(1)	(2)	(3)	(4)
1[2001 pop $\geq$ 300]	0.1493**	−0.0326	−0.0109	−0.0781
	(0.0603)	(0.0751)	(0.0595)	(0.0986)
2001 population	−0.0008	−0.0001	−0.0001	0.0007
	(0.0007)	(0.0007)	(0.0005)	(0.0009)
1[2001 pop $\geq$ 300] $\times$ 2001 pop	0.0008	0.0011	0.0005	−0.0005
	(0.0008)	(0.0008)	(0.0008)	(0.0008)
2001 Control	Yes	Yes	Yes	Yes
State FEs	Yes	Yes	Yes	Yes
RD bandwidth	150	150	150	150
Number of observations	18,686	10,304	17,385	11,382
Number of districts	130	116	174	149
Mean of dependent variable	6.370	5.398	5.587	5.010
R^2	0.766	0.793	0.756	0.611

Note. — This table compares our main RD specification in Column (1) to three separate samples for which RGGVY's 300-person eligibility threshold should not be relevant. Columns (2) and (4) estimate Equation (1.1) using villages with multiple habitations, for which the running variable (village population) does not correspond to the habitation populations that determined village eligibility. Columns (3) and (4) estimate Equation (1.1) using villages that were eligibile for RGGVY under the 11th Plan, which moved the eligibility cutoff from 300 to 100 people. Figure 1.5.6 presents these three falsification tests graphically. Each regression includes all villages meeting the above sample criteria with 2001 populations in the RD bandwidth (a 150-person bandwidth includes villages with 2001 populations between 150 and 450), for the 12 states with available village shapefiles that match to Census village areas with a correlation above 0.35. Standard errors are clustered at the district level. Significance: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

A.2.3 Nighttime brightness: timing

As a final validity test, we look at changes in nighttime brightness over time. Since we are attributing the increase in nighttime brightness to RGGVY eligibility, year-on-year changes in differential brightness at the RD cutoff should be consistent with the rollout of the RGGVY program. We should expect our RD estimates to increase incrementally over time, for two reasons. First, RGGVY project funds under the 10th Plan were disbursed gradually between 2005 and 2010. Because we estimate

the average effect on brightness across 128 eligible districts, our RD estimate should increase in magnitude as more districts received RGGVY funding. (There is an additional lag between the date that a district received RGGVY funding and the rollout of project implementation across its constituent villages). Second, the effects of infrastructure improvements on observed brightness are likely not immediate. For a village that received transformer upgrades and additional household electric connections in 2009, we might expect observed brightness to increase incrementally between 2010–2011, as villages invest in appliances that use electricity and emit light.

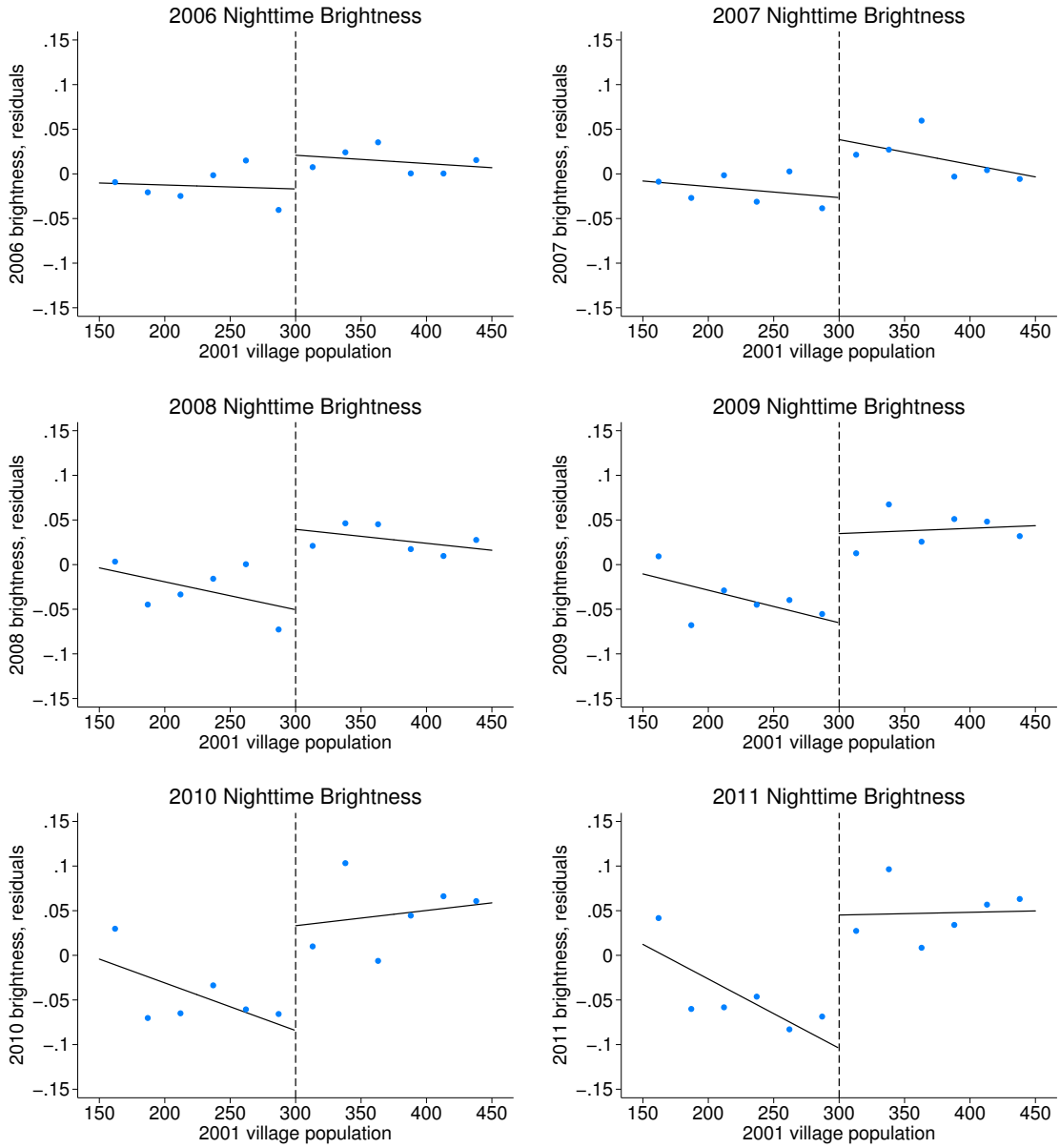
We test for these gradual rollout and investment effects in Figure A.2.4 and Table A.2.11. As expected, we see that the RD point estimate increases monotonically from 2006 to 2011. Hence, the RD effect that we detect from NOAA's satellite images is consistent with the incremental rollout and takeup of electricity use due to the RGGVY program.

Table A.2.11: RD Sensitivity – Brightness by Year

Village brightness	2006	2007	2008	2009	2010	2011
	(1)	(2)	(3)	(4)	(5)	(6)
$\mathbf{1}$[2001 pop $\geq$ 300]	0.0377	0.0649**	0.0900**	0.0999**	0.1175*	0.1493**
	(0.0251)	(0.0321)	(0.0352)	(0.0427)	(0.0632)	(0.0603)
2001 population	−0.0000	−0.0001	−0.0003	−0.0004	−0.0005	−0.0008
	(0.0002)	(0.0003)	(0.0003)	(0.0005)	(0.0008)	(0.0007)
$\mathbf{1}$[2001 pop $\geq$ 300] × 2001 pop	−0.0000	−0.0002	0.0002	0.0004	0.0007	0.0008
	(0.0003)	(0.0003)	(0.0004)	(0.0006)	(0.0009)	(0.0008)
2001 Control	Yes	Yes	Yes	Yes	Yes	Yes
State FEs	Yes	Yes	Yes	Yes	Yes	Yes
RD bandwidth	150	150	150	150	150	150
Number of observations	18,686	18,686	18,686	18,686	18,686	18,686
Number of districts	130	130	130	130	130	130
Mean of dependent variable	3.844	4.344	5.063	5.123	7.542	6.370
R^2	0.914	0.867	0.854	0.801	0.758	0.766

Note. — This table shows results from estimating Equation (1.1), using brightness outcomes from varying years. For each regression, the dependent variable is the maximum village brightness for a given year, after applying a linear projection onto brightness in adjacent years. Each regression includes all single-habitation villages in 10th-Plan districts with 2001 populations in the RD bandwidth (a 150-person bandwidth includes villages with 2001 populations between 150 and 450), for the 12 states with available village shapefiles that match to Census village areas with a correlation above 0.35. Standard errors are clustered at the district level. Significance: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

Figure A.2.4: RD on Nighttime Brightness Over Time



Note. — These six figures estimate our preferred RD specification using nighttime brightness for 2006–2011, corresponding to the regressions in Table A.2.11. Each graph uses a different year's nighttime brightness as the dependent variable, after having linearly projected these values on adjacent years. The lower-right graph reproduces Figure 1.5.4. Blue dots show average residuals from regressing that year's nighttime brightness on 2001 nighttime brightness and state fixed effects. Each dot contains approximately 1,600 villages, averaged in 25-person population bins. Lines are estimated separately on each side of the 300-person threshold, for all 18,686 single-habitation villages between 150–450 people, in 10th-Plan districts, for the 12 states with available village shapefiles that correspond to Census village areas (with a correlation above 0.35).

A.2.4 Census outcome results: RD robustness

Next, we run a series of analogous RD robustness tests for the range of Census outcomes results that we report in the main text.

A.2.4.1 Sample and outcome variables

We normalize our main village-level results in Table 1.5.3 by either village population (by gender) or by the share of village households. This is because our RD specification is cross-sectional, and most village-level outcomes will vary mechanically with the running variable (e.g., number of 2011 female agricultural workers will increase mechanically in 2001 village population). However, this normalization relies on the assumption that village-level demographics were unaffected by RGGVY program eligibility. Figure A.2.5 shows RD plots for 2011 village population and the size of the 0–6 age cohort, as reported in Panel A of Table 1.5.3. We see no evidence that RGGVY eligibility changed either the total village population or the fertility rate.

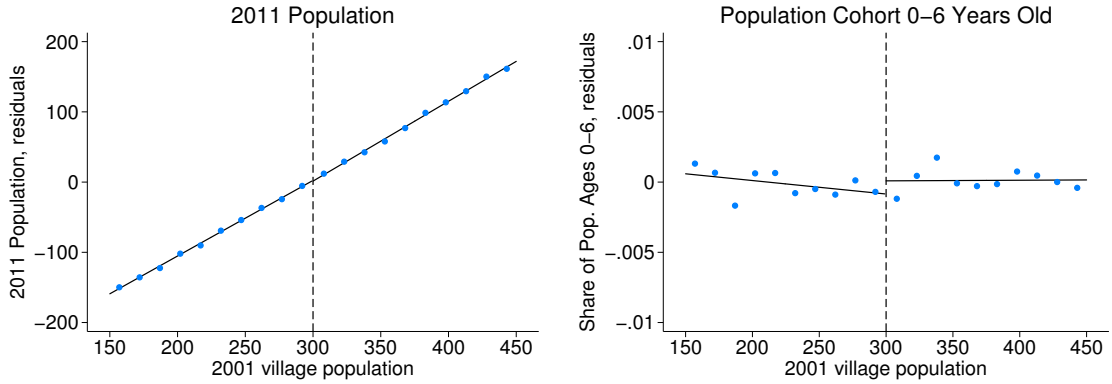
As an additional sensitivity, Table A.2.12 reproduces our main RD results from Table 1.5.3, including villages with population disparities in the habitation census of over 20 percent. After adding nearly 3,962 villages that were potentially miscategorized as having a single habitation, our RD estimates are nearly identical, although slightly more precise due to a greater sample size. In contrast, including villages with population disparities attenuates our nighttime brightness results (see Table A.2.3). As we see no corresponding attenuation in Table A.2.12, this underscores the lack of evidence of any economically meaningful impacts of RGGVY electrification for villages close to the cutoff.

Table A.2.12: RD Sensitivity – Census Outcomes, No Forced Population Match

2011 Outcome Variable	RD Coefficient	Standard Error	95 Percent Confidence	Mean of Outcome
A. Demographic outcomes				
Total population	−0.2473	(2.335)	[−4.824, 4.329]	273.01
0–6 cohort / total population	0.0010	(0.001)	[−0.001, 0.003]	0.15
Average household size	−0.0063	(0.012)	[−0.029, 0.016]	5.12
Literacy rate	−0.0025	(0.002)	[−0.006, 0.001]	0.56
B. Labor outcomes				
Male agricultural workers / male pop	−0.0069***	(0.002)	[−0.012, −0.002]	0.42
Female agricultural workers / female pop	−0.0055	(0.004)	[−0.013, 0.002]	0.29
Male household workers / male pop	−0.0008	(0.001)	[−0.002, 0.000]	0.01
Female household workers / female pop	−0.0015	(0.001)	[−0.003, 0.001]	0.01
Male other workers / male pop	0.0035*	(0.002)	[−0.000, 0.007]	0.10
Female other workers / female pop	−0.0016	(0.002)	[−0.005, 0.002]	0.05
C. Asset ownership				
Share of households with telephone	0.0025	(0.005)	[−0.007, 0.012]	0.53
Share of households with TV	0.0015	(0.004)	[−0.006, 0.009]	0.25
Share of households with bicycle	0.0007	(0.004)	[−0.007, 0.008]	0.49
Share of households with motorcycle	−0.0008	(0.003)	[−0.006, 0.004]	0.13
Share of households without assets	0.0016	(0.004)	[−0.006, 0.009]	0.23
D. Housing stock				
Share of households with elec/gas cooking	0.0008	(0.002)	[−0.004, 0.005]	0.06
Share of households with kerosene lighting	0.0023	(0.006)	[−0.009, 0.014]	0.48
Share of households with mud floors	0.0056	(0.004)	[−0.002, 0.013]	0.74
Share of households with thatched roof	−0.0037	(0.005)	[−0.013, 0.005]	0.23
Share of households dilapidated	−0.0024	(0.003)	[−0.008, 0.003]	0.07
E. Village-wide outcomes				
1/0 Mobile phone coverage in village	0.0051	(0.010)	[−0.015, 0.025]	0.74
1/0 Post office in village	0.0017	(0.003)	[−0.005, 0.008]	0.03
1/0 Ag credit societies in village	0.0005	(0.003)	[−0.006, 0.007]	0.02
1/0 Water from tubewell in village	−0.0033	(0.011)	[−0.024, 0.017]	0.45
Share of village area irrigated	−0.0075*	(0.005)	[−0.016, 0.001]	0.35
Share of village area planted	0.0018	(0.006)	[−0.009, 0.013]	0.58

Note. — This table includes the identical set of regressions in Table 1.5.3, including villages with large population disparities with the Habitation Census (i.e., villages whose matched habitation populations disagree with official Census populations by over 20 percent). The RD bandwidth now includes 33,727 villages with 2001 populations between 150 and 450, across 225 districts. The second column shows the RD point estimate ($\hat{\beta}_1$) for each regression. All specifications control for the 2001 level of the outcome variable, except for share of village area planted (where 2001 values are not available) and 1/0 indicator variables. All specifications also include state fixed effects. Standard errors are clustered at the district level, which we use to calculate 95 percent confidence intervals in the fourth column. The fifth column reports the mean of the dependent variable for each RD regression. Significance: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

Figure A.2.5: RD Reduced Form – 2011 Village Population



Note. — This figure presents results for RD regressions of 2011 village population (left) and the share of the 2011 population less than 7 years old (right). These correspond to the first two rows of Table 1.5.3. Blue dots show average residuals from regressing the 2011 outcome on and state fixed effects and the 2001 level of the outcome (except for the left panel, where 2001 population is the running variable and hence not an additional control). Each dot contains approximately 1,500 villages, averaged in 15-person population bins. Lines are estimated separately on each side of the 300-person threshold, for all 29,765 single-habitation villages between 150 and 450 people, in 10th-Plan districts.

A.2.4.2 Bandwidths

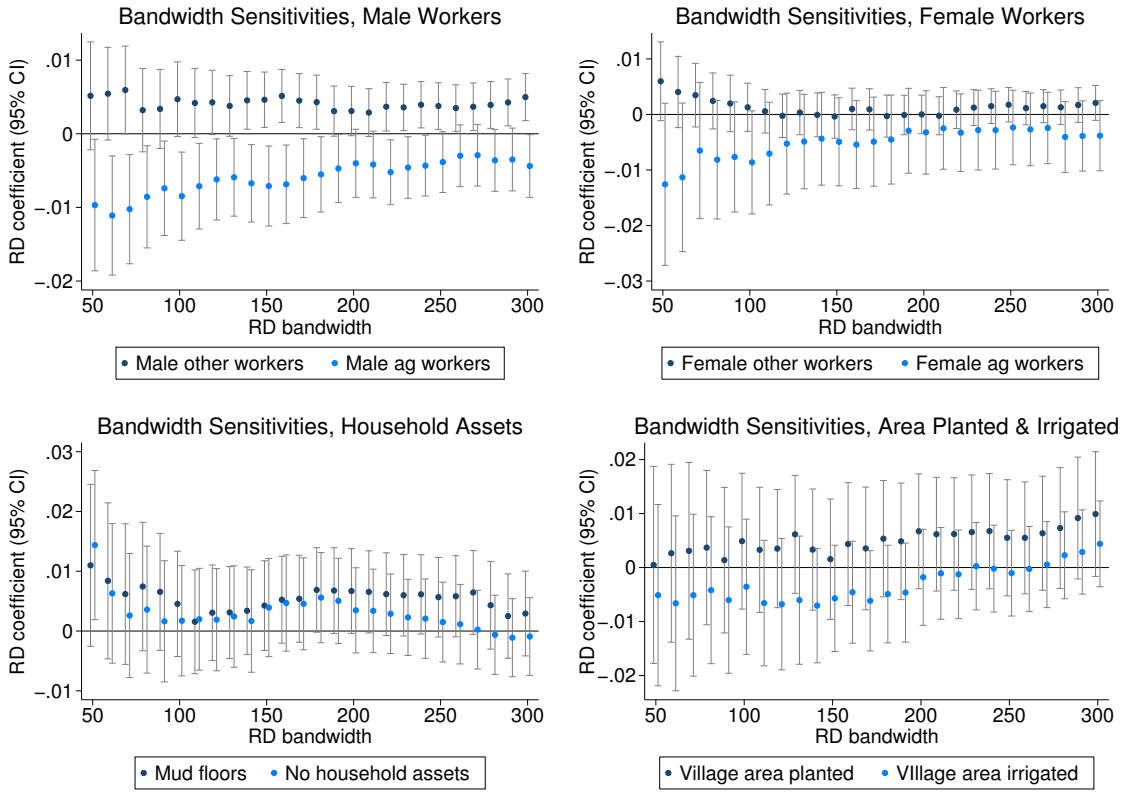
Section A.2.1.2 discusses our choice of a 150-person RD bandwidth around RGGVY's 300-person 10th-Plan cutoff, which avoids overlapping with the 100-person 11th-Plan cutoff and the 500-person PMGSY eligibility cutoff. However, we also demonstrate that our RD results for nighttime brightness are not sensitive to our choice of bandwidth (see Figure A.2.1). We present analogous bandwidth sensitivities in Figure A.2.6, for eight census outcomes reported in Table 1.5.3. This demonstrates that our RD results for these Census outcomes (male and female labor shares; mud floors; asset ownership; share of village area planted and irrigated) are not overly sensitive to our choice of bandwidth, for bandwidths between 100 and 250 (above which includes very small villages). Figure A.2.6 shows the range of feasible symmetric RD bandwidths, which stops at 300 due to our 300-person RD threshold. The optimal bandwidth calculations (following Imbens and Kalyanaraman (2012)) vary across each outcome, ranging from 130 to 353.

A.2.4.3 Functional form

Section A.2.1.3 discusses our choice to exclude higher-order polynomials from our preferred RD specification (following Gelman and Imbens (2014)). Table A.2.13 tests for sensitivity of our RD regressions on village-level census outcomes to the inclusion of a quadratic function of the running variable. This yields point estimates and confidence intervals that are very close to the linear RD specification in Table 1.5.3. Figure A.2.7 presents a subset of these results graphically, where the share of male agricultural workers remains the census outcome that is most obviously discontinuous at the 300-person threshold.

Table A.2.14 re-estimates Equation (1.1) using weighted least squares, weighting observations by their distance from the RD threshold. We define village weights as $w_v \equiv 1 - \frac{|P_v - 300|}{150}$, where P_v is the 2001 village population. These results confirm that upweighting villages close to the RD threshold does not significantly change our results.

Figure A.2.6: RD Sensitivity – Census Outcomes, Bandwidths



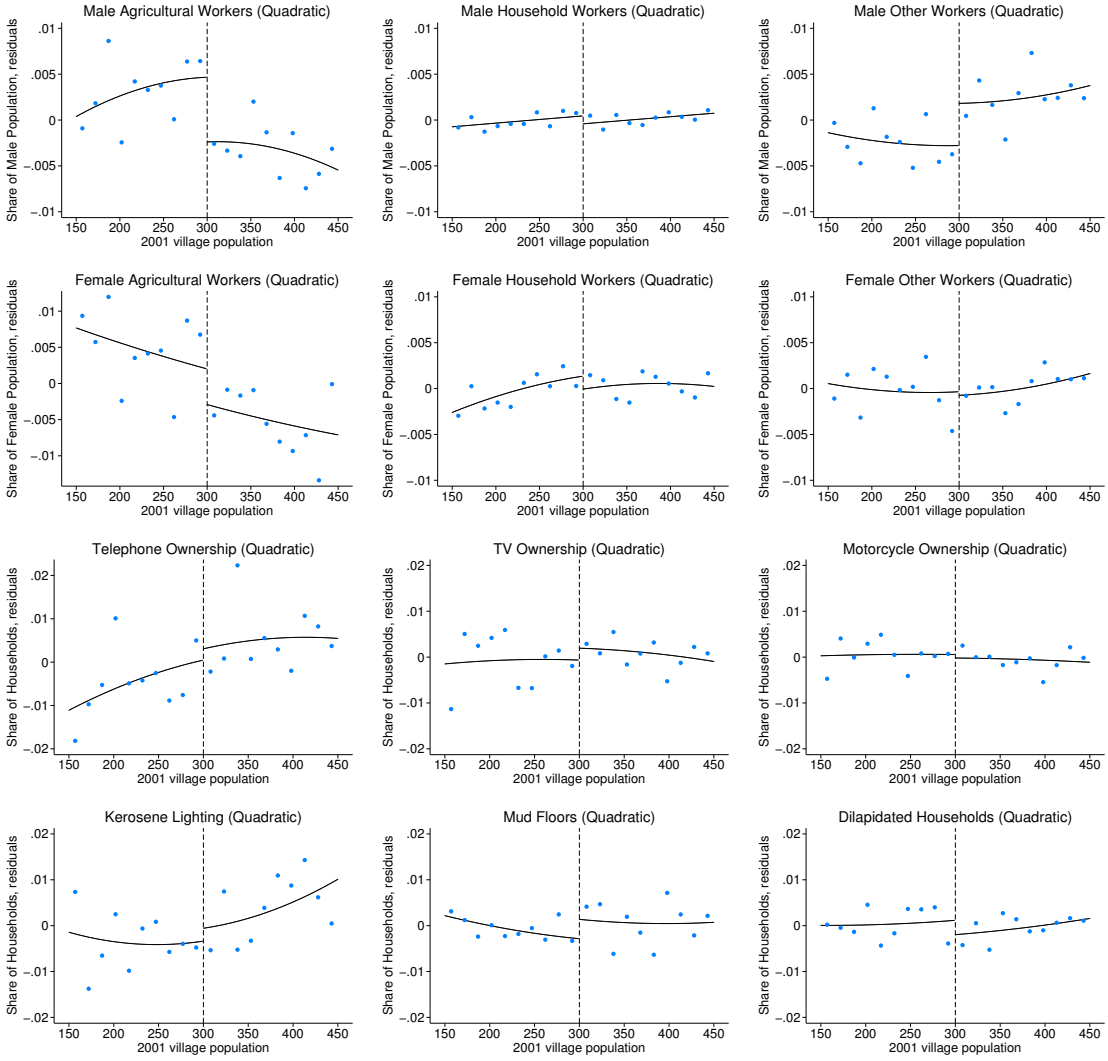
Note. — This figure presents our bandwidth sensitivity analysis for eight separate outcomes in Table 1.5.3. For each outcome, we estimate Equation (1.1) separately on bandwidths ranging from 50 (i.e., 250–350 people) to 300 (i.e., 0–600 people). Each dot represents the point estimate on the RD discontinuity at a given bandwidth around the 300-person cutoff, with 95 percent confidence intervals clustered at the district level. Our chosen bandwidth of 150 includes villages with populations between 150 and 450. The optimal RD bandwidth varies for each outcome, ranging from 130 to 354 for the eight outcomes shown here (calculated using the algorithm proposed by Imbens and Kalyanaraman (2012), using uniform, Epanechnikov, and triangular kernels).

Table A.2.13: RD Sensitivity – Census Outcomes, Quadratic in Population

2011 Outcome Variable	RD Coefficient	Standard Error	95 Percent Confidence	Mean of Outcome
A. Demographic outcomes				
Total population	−0.8933	(2.510)	[−5.813, 4.026]	271.09
0–6 cohort / total population	0.0009	(0.001)	[−0.001, 0.002]	0.14
Average household size	−0.0054	(0.013)	[−0.030, 0.019]	5.13
Literacy rate	−0.0025	(0.002)	[−0.006, 0.002]	0.57
B. Labor outcomes				
Male agricultural workers / male pop	−0.0071**	(0.003)	[−0.012, −0.002]	0.42
Female agricultural workers / female pop	−0.0051	(0.004)	[−0.013, 0.003]	0.29
Male household workers / male pop	−0.0009	(0.001)	[−0.002, 0.000]	0.01
Female household workers / female pop	−0.0015	(0.001)	[−0.004, 0.001]	0.01
Male other workers / male pop	0.0047**	(0.002)	[0.001, 0.008]	0.10
Female other workers / female pop	−0.0003	(0.002)	[−0.004, 0.004]	0.05
C. Asset ownership				
Share of households with telephone	0.0028	(0.005)	[−0.008, 0.014]	0.54
Share of households with TV	0.0026	(0.004)	[−0.005, 0.010]	0.26
Share of households with bicycle	−0.0015	(0.004)	[−0.010, 0.007]	0.50
Share of households with motorcycle	−0.0008	(0.003)	[−0.006, 0.004]	0.13
Share of households without assets	0.0037	(0.004)	[−0.004, 0.012]	0.22
D. Housing stock				
Share of households with elec/gas cooking	0.0005	(0.003)	[−0.005, 0.006]	0.07
Share of households with kerosene lighting	0.0029	(0.006)	[−0.009, 0.015]	0.48
Share of households with mud floors	0.0041	(0.004)	[−0.003, 0.012]	0.73
Share of households with thatched roof	−0.0034	(0.005)	[−0.013, 0.006]	0.23
Share of households dilapidated	−0.0031	(0.003)	[−0.009, 0.003]	0.07
E. Village-wide outcomes				
1/0 Mobile phone coverage in village	−0.0002	(0.011)	[−0.022, 0.022]	0.75
1/0 Post office in village	0.0015	(0.004)	[−0.006, 0.009]	0.03
1/0 Ag credit societies in village	0.0015	(0.004)	[−0.006, 0.009]	0.02
1/0 Water from tubewell in village	−0.0025	(0.011)	[−0.024, 0.019]	0.44
Share of village area irrigated	−0.0056	(0.005)	[−0.015, 0.004]	0.34
Share of village area planted	0.0019	(0.006)	[−0.009, 0.013]	0.58

Note. — This table includes the identical set of regressions in Table 1.5.3, except controlling for a quadratic function of 2001 village population. The RD bandwidth includes 29,765 villages with 2001 populations between 150 and 450, across 225 districts. The second column shows the RD point estimate ($\hat{\beta}_1$) for each regression. All specifications control for a 2nd-order polynomial in the running variable and state fixed effects. All specifications also control for the 2001 level of the outcome variable, except for share of village area planted (where 2001 values are not available) and 1/0 indicator variables. Standard errors are clustered at the district level, which we use to calculate 95 percent confidence intervals in the fourth column. The fifth column reports the mean of the dependent variable for each RD regression. Significance: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

Figure A.2.7: RD Sensitivity – Census Outcomes, Second-Order Polynomials



Note. — This figure shows reduced form results estimating Equation (1.1) including a second-order polynomial in population, as reported in Table A.2.13. Blue dots show average residuals form regressing the 2011 outcomes on the 2001 control and state fixed effects. Each dot contains approximately 1,500 villages, averaged in 15-person population bins, including all 29,765 single-habitation villages between 150 and 450 people, in 10th-Plan districts.

Table A.2.14: RD Sensitivity – Census Outcomes, Weighting Inverse Distance from Cutoff

2011 Outcome Variable	RD Coefficient	Standard Error	95 Percent Confidence	Mean of Outcome
A. Demographic outcomes				
Total population	−0.5747	(2.536)	[−5.545, 4.396]	301.22
0–6 cohort / total population	0.0005	(0.001)	[−0.001, 0.002]	0.14
Average household size	−0.0003	(0.014)	[−0.027, 0.027]	5.14
Literacy rate	−0.0009	(0.002)	[−0.005, 0.003]	0.57
B. Labor outcomes				
Male agricultural workers / male pop	−0.0075***	(0.003)	[−0.013, −0.002]	0.43
Female agricultural workers / female pop	−0.0063	(0.004)	[−0.015, 0.002]	0.28
Male household workers / male pop	−0.0008	(0.001)	[−0.002, 0.000]	0.01
Female household workers / female pop	−0.0010	(0.001)	[−0.003, 0.001]	0.01
Male other workers / male pop	0.0044**	(0.002)	[0.000, 0.009]	0.10
Female other workers / female pop	0.0010	(0.002)	[−0.003, 0.005]	0.05
C. Asset ownership				
Share of households with telephone	0.0037	(0.006)	[−0.008, 0.016]	0.54
Share of households with TV	0.0050	(0.004)	[−0.003, 0.013]	0.25
Share of households with bicycle	0.0007	(0.004)	[−0.008, 0.009]	0.51
Share of households with motorcycle	0.0016	(0.003)	[−0.004, 0.007]	0.13
Share of households without assets	0.0030	(0.004)	[−0.005, 0.011]	0.22
D. Housing stock				
Share of households with elec/gas cooking	0.0015	(0.003)	[−0.004, 0.007]	0.06
Share of households with kerosene lighting	0.0000	(0.007)	[−0.014, 0.014]	0.48
Share of households with mud floors	0.0045	(0.004)	[−0.004, 0.012]	0.74
Share of households with thatched roof	−0.0042	(0.005)	[−0.015, 0.006]	0.22
Share of households dilapidated	−0.0032	(0.003)	[−0.010, 0.003]	0.07
E. Village-wide outcomes				
1/0 Mobile phone coverage in village	0.0088	(0.013)	[−0.017, 0.035]	0.75
1/0 Post office in village	0.0056	(0.004)	[−0.002, 0.013]	0.03
1/0 Ag credit societies in village	0.0053	(0.004)	[−0.002, 0.013]	0.02
1/0 Water from tubewell in village	0.0040	(0.012)	[−0.019, 0.028]	0.46
Share of village area irrigated	−0.0063	(0.006)	[−0.018, 0.005]	0.36
Share of village area planted	0.0034	(0.006)	[−0.008, 0.015]	0.59

Note. — This table includes the same regressions as in Table 1.5.3, but running weighted least squares. We weight villages by their absolute distance from the 300-person cutoff, such that a $w_v \equiv 1 - \frac{|P_v - 300|}{150}$. The RD bandwidth includes 29,573 villages with 2001 populations between 150 and 450, across 225 districts. The second column shows the RD point estimate ($\hat{\beta}_1$) for each regression. All specifications control for the 2001 level of the outcome variable, except for share of village area planted (where 2001 values are not available) and 1/0 indicator variables. All specifications also include state fixed effects. Standard errors are clustered at the district level, which we use to calculate 95 percent confidence intervals in the fourth column. The fifth column reports the mean of the dependent variable for each RD regression. Significance: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

A.2.4.4 Fixed effects and controls

As mentioned above, while RD designs do not require fixed effects or controls for identification, they can greatly improve precision of RD point estimates. Section A.2.1.4 demonstrates that the RD results for nighttime brightness are robust to different fixed effects specifications, and our results for census outcomes are also not sensitive to cross-sectional fixed effects. Table A.2.15 and Figure A.2.8 present RD results estimating Equation (1.1) without fixed effects, while Table A.2.16 and Figure A.2.9 include more granular fixed effects at the district level. These results are quite similar to our preferred specification (with state fixed effects), with slight differences in precision.

Section A.2.1.4 also shows that our RD results for nighttime brightness depend on the inclusion of 2001 brightness as a village-level control. This reveals that there is enough cross-sectional variation in brightness levels (even within districts) to obscure within-village changes in brightness due to RGGVY electrification. In contrast, Table A.2.17 and Figure A.2.10 show that our RD results for census outcomes are much less sensitive to the exclusion of 2001 controls. The decrease in male agricultural employment at the 300-person cutoff remains small but precisely estimated, and the confidence intervals on other labor outcomes in Panel B are very similar to those in Table 1.5.3.

Comparing Tables 1.5.3 and A.2.17 highlights two key features of our census outcome data. First, Panels C and D report on outcomes from the Houselisting Primary Census Abstract (HPCA), which is not available at the village level for 2001 (see Section A.1.4.2). This means that for Table 1.5.3 regressions in Panels C and D, 2001 controls are at the block level.[56] Because these regressions cannot control for within-block baseline heterogeneity, it is unsurprising that removing these 2001 block-level controls does not affect their precision.

Second, our regressions in Panel E of Table 3 already do not include 2001 controls for 1/0 indicator variables. Since these outcomes are not continuous, including baseline controls would not greatly increase the precision of our RD point estimates. Moreover, including 2001 controls for 1/0 indicator variables would effectively remove observations for villages that did not change status (i.e. from 0 to 1) between 2001 and 2011. This illustrates why we strongly prefer nighttime brightness as a measure of electricity access: nighttime brightness is a more continuous measure of electricity access (a 0–63 scale, compared to a binary indicator); nighttime brightness measures luminosity, a better proxy for electricity *consumption* than binary indicators of electricity *access* (the latter does not account for variation in reliability or usage); the

[56]The only exception is share of households dilapidated, which was only available in 2001 at the district level.

majority of RGGVY villages were targeted for "more intensive electrification", while already meeting the government's official definition of being "electrified" (we would not expect binary indicators to reflect more intensive electrification); and the official definition of "electrified" changed in 2004, meaning that 2001 and 2011 VD variables might not be apples-to-apples comparisons (see Section A.3 for further detail). Table A.2.17 reports RD results for end-use-specific 1/0 electric power indicators, and we see the RGGVY led to statistically significant increases in electricity access for the commercial sector. We do not see corresponding increases for electricity access in the domestic or agricultural sectors, which reflects the fact that most treated villages already had low levels of electrification prior to RGGVY — meaning that the 1/0 indicators were coded as 1 in the baseline, and RGGVY did not cause them to increase.

A.2.4.5 2001 covariate smoothness

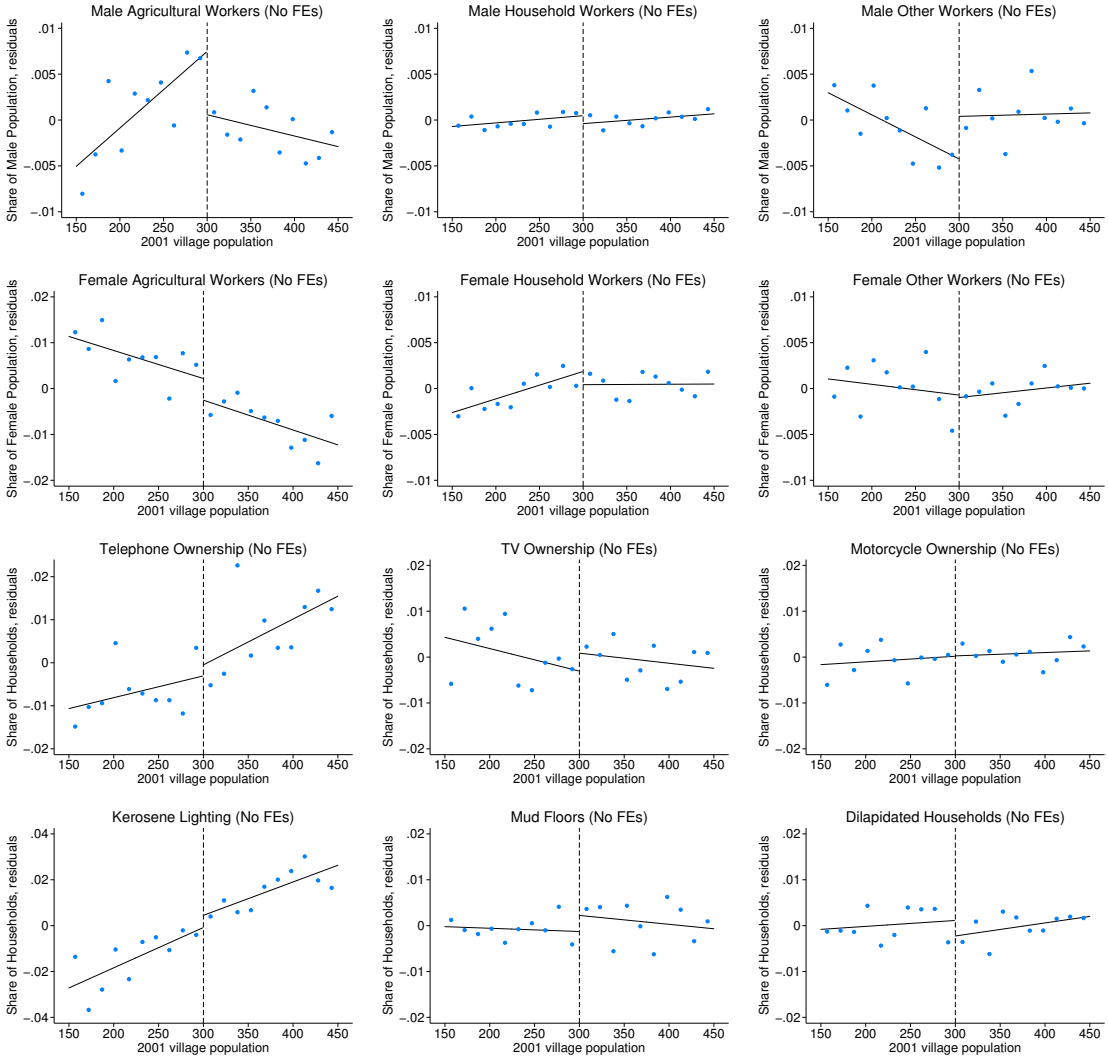
A key RD identifying assumption is smoothness of covariates across the RD cutoff. Section A.2.1.5 demonstrates that nighttime brightness is not discontinuous at the 300-person threshold prior to the 2005 announcement of RGGVY. Table A.2.18 shows that 2001 village-level covariates are also smooth across the 300-person cutoff in 2001 population. This includes all 2001 covariates corresponding to 2011 census outcomes reported in Table 1.5.3, which are available in 2001 at the village level. Figure A.2.11 presents a subset of these results graphically.

Table A.2.15: RD Sensitivity – Census Outcomes, No Fixed Effects

2011 Outcome Variable	RD Coefficient	Standard Error	95 Percent Confidence	Mean of Outcome
A. Demographic outcomes				
Total population	−1.4365	(2.576)	[−6.486, 3.613]	271.09
0–6 cohort / total population	0.0011	(0.001)	[−0.001, 0.003]	0.14
Average household size	−0.0014	(0.013)	[−0.027, 0.024]	5.13
Literacy rate	−0.0025	(0.002)	[−0.006, 0.002]	0.57
B. Labor outcomes				
Male agricultural workers / male pop	−0.0069**	(0.003)	[−0.013, −0.001]	0.42
Female agricultural workers / female pop	−0.0047	(0.004)	[−0.014, 0.004]	0.29
Male household workers / male pop	−0.0009	(0.001)	[−0.002, 0.000]	0.01
Female household workers / female pop	−0.0014	(0.001)	[−0.004, 0.001]	0.01
Male other workers / male pop	0.0046**	(0.002)	[0.001, 0.008]	0.10
Female other workers / female pop	−0.0003	(0.002)	[−0.004, 0.004]	0.05
C. Asset ownership				
Share of households with telephone	0.0026	(0.006)	[−0.009, 0.015]	0.54
Share of households with TV	0.0039	(0.004)	[−0.004, 0.012]	0.26
Share of households with bicycle	−0.0007	(0.004)	[−0.009, 0.008]	0.50
Share of households with motorcycle	0.0001	(0.003)	[−0.005, 0.005]	0.13
Share of households without assets	0.0034	(0.004)	[−0.005, 0.011]	0.22
D. Housing stock				
Share of households with elec/gas cooking	0.0009	(0.003)	[−0.004, 0.006]	0.07
Share of households with kerosene lighting	0.0053	(0.006)	[−0.007, 0.018]	0.48
Share of households with mud floors	0.0035	(0.004)	[−0.004, 0.011]	0.73
Share of households with thatched roof	−0.0032	(0.005)	[−0.013, 0.007]	0.23
Share of households dilapidated	−0.0034	(0.003)	[−0.009, 0.002]	0.07
E. Village-wide outcomes				
1/0 Mobile phone coverage in village	−0.0035	(0.012)	[−0.027, 0.020]	0.75
1/0 Post office in village	0.0020	(0.004)	[−0.005, 0.009]	0.03
1/0 Ag credit societies in village	0.0010	(0.004)	[−0.006, 0.008]	0.02
1/0 Water from tubewell in village	0.0006	(0.011)	[−0.022, 0.023]	0.44
Share of village area irrigated	−0.0068	(0.006)	[−0.019, 0.005]	0.34
Share of village area planted	0.0059	(0.006)	[−0.006, 0.018]	0.58

Note. — This table includes the identical set of regressions in Table 1.5.3, without controlling for state fixed effects. The RD bandwidth includes 29,765 villages with 2001 populations between 150 and 450, across 225 districts. The second column shows the RD point estimate ($\hat{\beta}_1$) for each regression. All specifications control for the 2001 level of the outcome variable, except for share of village area planted (where 2001 values are not available) and 1/0 indicator variables. Standard errors are clustered at the district level, which we use to calculate 95 percent confidence intervals in the fourth column. The fifth column reports the mean of the dependent variable for each RD regression. Significance: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

Figure A.2.8: RD Sensitivity – Census Outcomes, No Fixed Effects



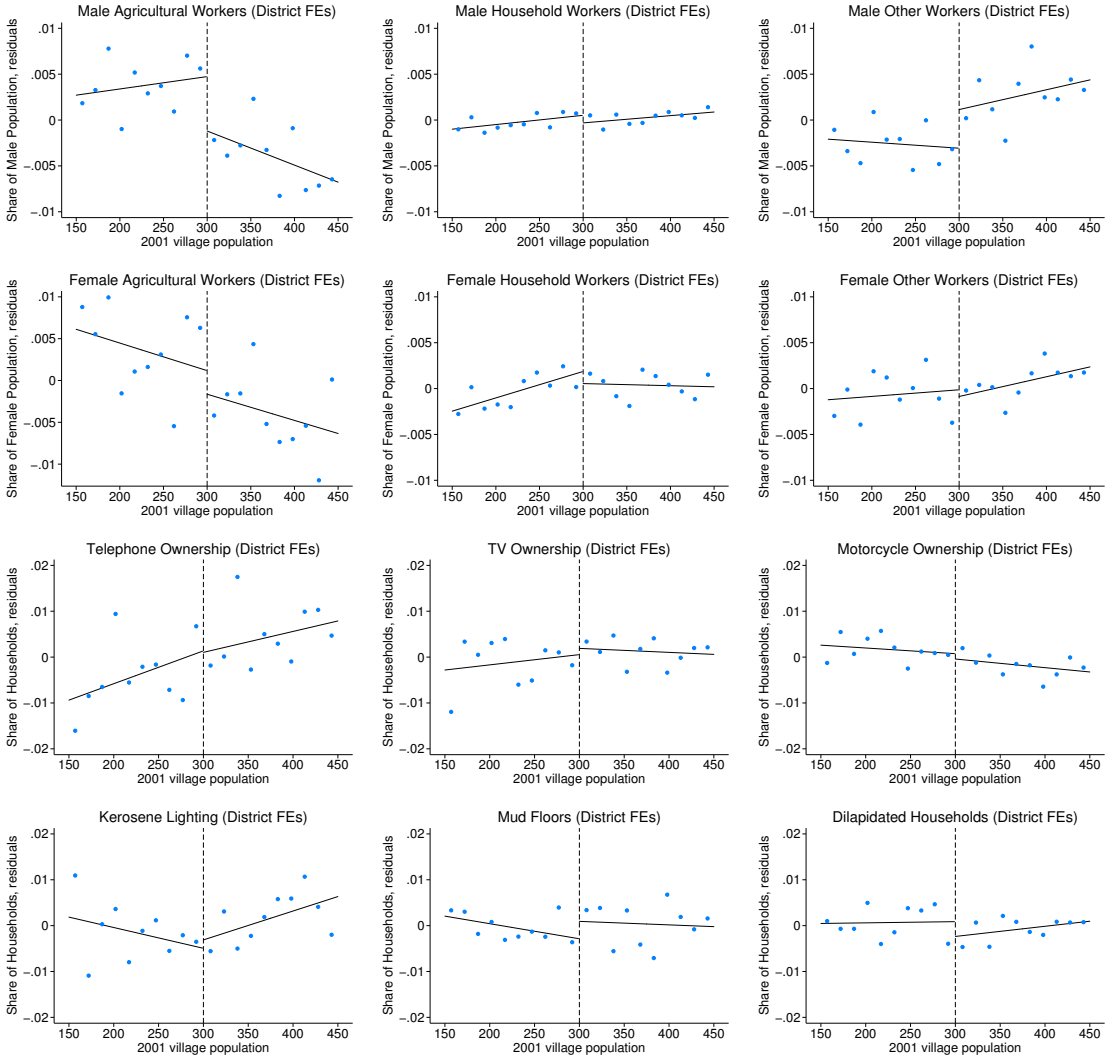
Note. — This figure shows reduced form results estimating Equation (1.1) without fixed effects, as reported in Table A.2.15. Blue dots show average residuals form regressing the 2011 outcomes on the 2001 control. Each dot contains approximately 1,500 villages, averaged in 15-person population bins, including all 29,765 single-habitation villages between 150 and 450 people, in 10th-Plan districts.

Table A.2.16: RD Sensitivity – Census Outcomes, District Fixed Effects

2011 Outcome Variable	RD Coefficient	Standard Error	95 Percent Confidence	Mean of Outcome
A. Demographic outcomes				
Total population	−0.6014	(2.371)	[−5.248, 4.045]	271.09
0–6 cohort / total population	0.0010	(0.001)	[−0.001, 0.003]	0.14
Average household size	−0.0060	(0.012)	[−0.030, 0.018]	5.13
Literacy rate	−0.0031	(0.002)	[−0.007, 0.001]	0.57
B. Labor outcomes				
Male agricultural workers / male pop	−0.0060**	(0.003)	[−0.011, −0.001]	0.42
Female agricultural workers / female pop	−0.0029	(0.004)	[−0.011, 0.005]	0.29
Male household workers / male pop	−0.0008	(0.001)	[−0.002, 0.000]	0.01
Female household workers / female pop	−0.0013	(0.001)	[−0.003, 0.001]	0.01
Male other workers / male pop	0.0043**	(0.002)	[0.001, 0.008]	0.10
Female other workers / female pop	−0.0007	(0.002)	[−0.004, 0.003]	0.05
C. Asset ownership				
Share of households with telephone	−0.0003	(0.005)	[−0.011, 0.010]	0.54
Share of households with TV	0.0014	(0.004)	[−0.006, 0.009]	0.26
Share of households with bicycle	−0.0016	(0.004)	[−0.010, 0.006]	0.50
Share of households with motorcycle	−0.0012	(0.002)	[−0.006, 0.004]	0.13
Share of households without assets	0.0041	(0.004)	[−0.004, 0.012]	0.22
D. Housing stock				
Share of households with elec/gas cooking	−0.0005	(0.003)	[−0.006, 0.005]	0.07
Share of households with kerosene lighting	0.0018	(0.006)	[−0.010, 0.014]	0.48
Share of households with mud floors	0.0038	(0.004)	[−0.003, 0.011]	0.73
Share of households with thatched roof	−0.0018	(0.005)	[−0.011, 0.008]	0.23
Share of households dilapidated	−0.0033	(0.003)	[−0.009, 0.002]	0.07
E. Village-wide outcomes				
1/0 Mobile phone coverage in village	0.0022	(0.011)	[−0.019, 0.024]	0.75
1/0 Post office in village	0.0010	(0.004)	[−0.006, 0.008]	0.03
1/0 Ag credit societies in village	0.0013	(0.004)	[−0.006, 0.008]	0.02
1/0 Water from tubewell in village	−0.0013	(0.010)	[−0.021, 0.018]	0.44
Share of village area irrigated	−0.0059	(0.005)	[−0.015, 0.004]	0.34
Share of village area planted	0.0016	(0.005)	[−0.007, 0.011]	0.58

Note. — This table includes the identical set of regressions in Table 1.5.3, except controlling for district fixed effects (instead of state fixed effects). The RD bandwidth includes 29,765 villages with 2001 populations between 150 and 450, across 225 districts. The second column shows the RD point estimate ($\hat{\beta}_1$) for each regression. All specifications control for the 2001 level of the outcome variable, except for share of village area planted (where 2001 values are not available) and 1/0 indicator variables. Standard errors are clustered at the district level, which we use to calculate 95 percent confidence intervals in the fourth column. The fifth column reports the mean of the dependent variable for each RD regression. Significance: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

Figure A.2.9: RD Sensitivity – Census Outcomes, District Fixed Effects



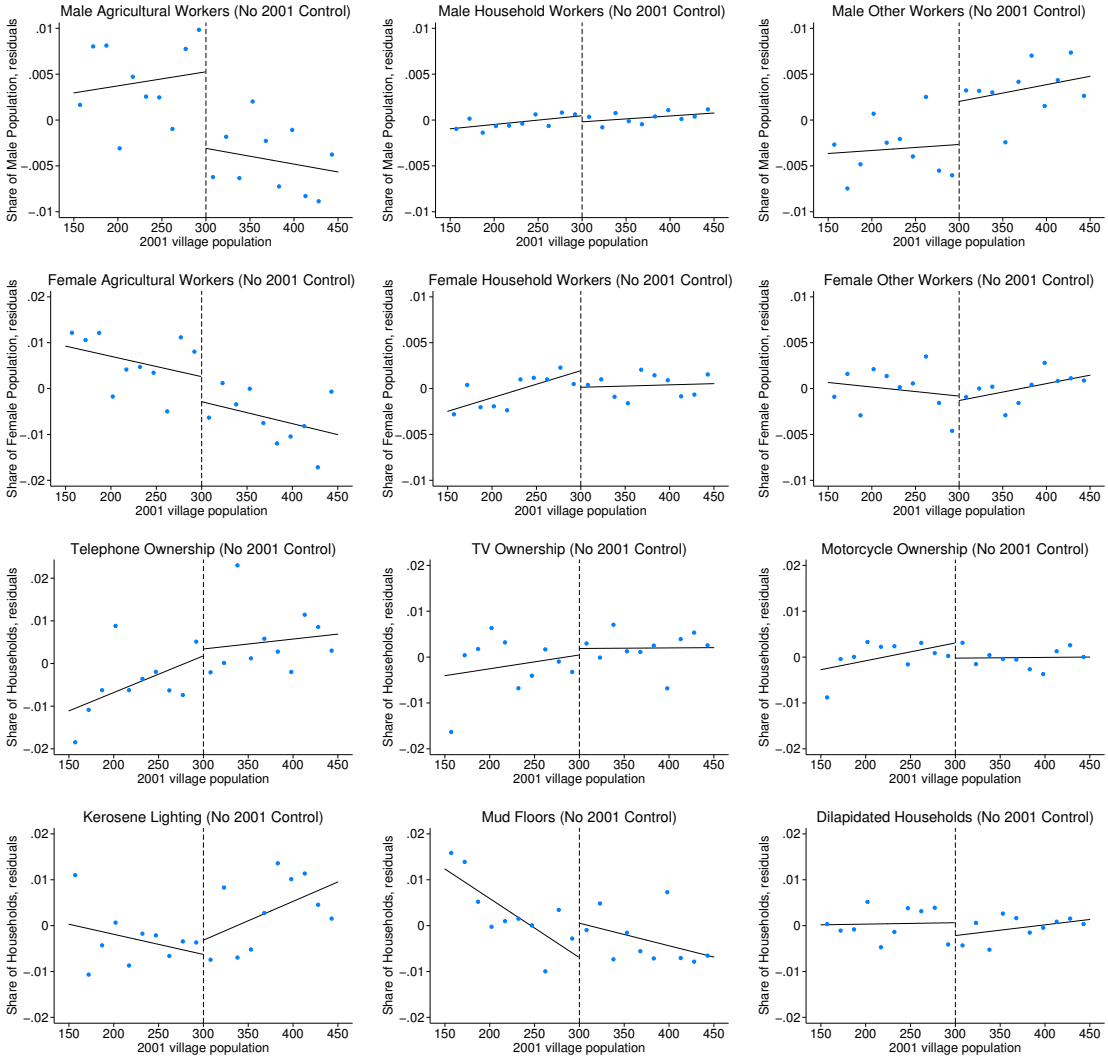
Note. — This figure shows reduced form results estimating Equation (1.1) including district fixed effects, as reported in Table A.2.16. Blue dots show average residuals form regressing the 2011 outcomes on the 2001 control and district fixed effects. Each dot contains approximately 1,500 villages, averaged in 15-person population bins, including all 29,765 single-habitation villages between 150 and 450 people, in 10th-Plan districts.

Table A.2.17: RD Sensitivity – Census Outcomes, No 2001 Controls

2011 Outcome Variable	RD Coefficient	Standard Error	95 Percent Confidence	Mean of Outcome
A. Demographic outcomes				
Total population	−0.8647	(2.528)	[−5.820, 4.091]	271.09
0–6 cohort / total population	0.0007	(0.001)	[−0.001, 0.002]	0.14
Average household size	0.0022	(0.014)	[−0.026, 0.031]	5.13
Literacy rate	−0.0016	(0.003)	[−0.007, 0.004]	0.57
B. Labor outcomes				
Male agricultural workers / male pop	−0.0084***	(0.003)	[−0.014, −0.003]	0.42
Female agricultural workers / female pop	−0.0055	(0.004)	[−0.013, 0.003]	0.29
Male household workers / male pop	−0.0007	(0.001)	[−0.002, 0.001]	0.01
Female household workers / female pop	−0.0018	(0.001)	[−0.004, 0.000]	0.01
Male other workers / male pop	0.0047**	(0.002)	[0.000, 0.009]	0.10
Female other workers / female pop	−0.0005	(0.002)	[−0.004, 0.003]	0.05
C. Asset ownership				
Share of households with telephone	0.0016	(0.006)	[−0.009, 0.013]	0.54
Share of households with TV	0.0014	(0.004)	[−0.007, 0.009]	0.26
Share of households with bicycle	−0.0034	(0.006)	[−0.015, 0.008]	0.50
Share of households with motorcycle	−0.0033	(0.003)	[−0.010, 0.003]	0.13
Share of households without assets	0.0032	(0.005)	[−0.006, 0.012]	0.22
D. Housing stock				
Share of households with elec/gas cooking	−0.0006	(0.003)	[−0.006, 0.005]	0.07
Share of households with kerosene lighting	0.0031	(0.006)	[−0.009, 0.015]	0.48
Share of households with mud floors	0.0075	(0.005)	[−0.003, 0.018]	0.73
Share of households with thatched roof	−0.0022	(0.006)	[−0.013, 0.009]	0.23
Share of households dilapidated	−0.0028	(0.003)	[−0.009, 0.003]	0.07
E. Village-wide outcomes				
1/0 Electricity (domestic use)	−0.0086	(0.006)	[−0.021, 0.004]	0.90
1/0 Electricity (agricultural use)	−0.0098	(0.010)	[−0.030, 0.010]	0.61
1/0 Electricity (commercial use)	0.0243**	(0.011)	[0.003, 0.046]	0.44
1/0 Electricity (all end uses)	0.0240**	(0.011)	[0.002, 0.046]	0.43
Share of village area irrigated	−0.0021	(0.006)	[−0.013, 0.009]	0.35

Note. — Panels A–D of this table included the same sets of regressions in Table 1.5.3. Panel E includes sector-specific 1/0 indicator variables for electricity availability at the village level, while omiting 4 regressions from Table 1.5.3 that already did not include 2001 controls. All RD regressions in this table control only for state fixed effects, not for the 2001 level of the outcome variable. The RD bandwidth includes 29,765 villages with 2001 populations between 150 and 450, across 225 districts. The second column shows the RD point estimate ($\hat{\beta}_1$) for each regression. Standard errors are clustered at the district level, which we use to calculate 95 percent confidence intervals in the fourth column. The fifth column reports the mean of the dependent variable for each RD regression. Significance: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

Figure A.2.10: RD Sensitivity – Census Outcomes, No 2001 Controls



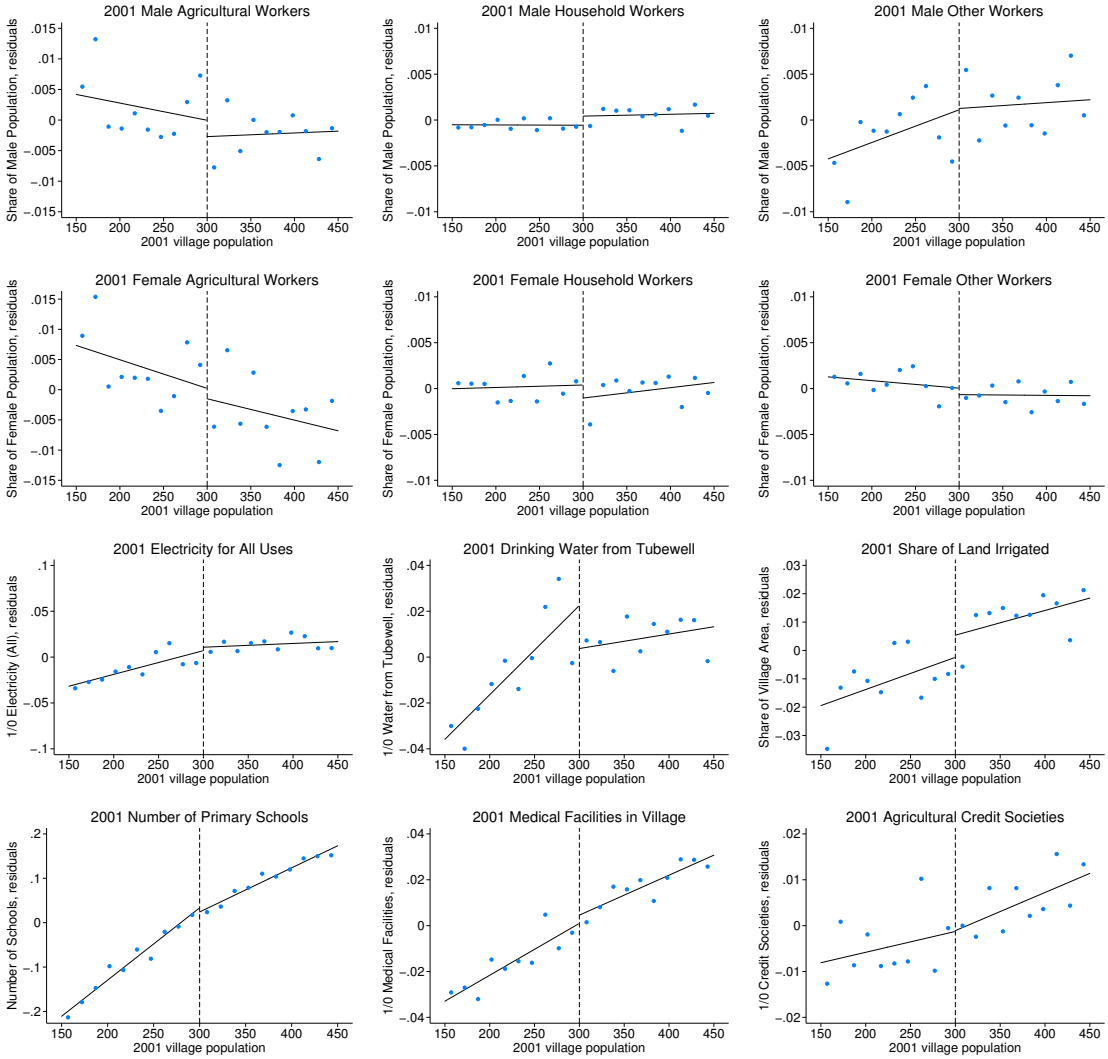
Note. — This figure shows reduced form results estimating Equation (1.1) without 2001 controls, as reported in Table A.2.17. Blue dots show average residuals form regressing the 2011 outcomes on state fixed effects. Each dot contains approximately 1,500 villages, averaged in 15-person population bins, including all 29,765 single-habitation villages between 150 and 450 people, in 10th-Plan districts.

Table A.2.18: RD Sensitivity – Census Outcomes, 2001 Covariate Smoothness

2011 Outcome Variable	RD Coefficient	Standard Error	95 Percent Confidence	Mean of Outcome
A. Demographic indicators				
0–6 cohort / total population	−0.0009	(0.001)	[−0.003, 0.001]	0.18
Share or population SC or ST	−0.0080	(0.007)	[−0.021, 0.005]	0.33
Literacy rate	0.0015	(0.003)	[−0.005, 0.008]	0.45
B. Labor indicators				
Male agricultural workers / male pop	−0.0027	(0.003)	[−0.008, 0.003]	0.43
Female agricultural workers / female pop	−0.0017	(0.005)	[−0.011, 0.007]	0.31
Male household workers / male pop	0.0010	(0.001)	[−0.000, 0.002]	0.01
Female household workers / female pop	−0.0014	(0.001)	[−0.004, 0.001]	0.01
Male other workers / male pop	0.0001	(0.002)	[−0.004, 0.004]	0.09
Female other workers / female pop	−0.0007	(0.002)	[−0.004, 0.003]	0.03
Male main workers / male pop	−0.0017	(0.004)	[−0.009, 0.005]	0.43
Female main workers / female pop	−0.0047	(0.004)	[−0.013, 0.004]	0.17
Male marginal workers / male pop	0.0002	(0.003)	[−0.005, 0.006]	0.09
Female marginal workers / female pop	0.0008	(0.004)	[−0.007, 0.009]	0.18
C. Village-wide indicators				
1/0 Electricity (all end uses)	0.0038	(0.007)	[−0.009, 0.017]	0.29
1/0 Electricity (domestic use)	0.0018	(0.010)	[−0.018, 0.021]	0.59
1/0 Electricity (agricultural use)	−0.0014	(0.009)	[−0.020, 0.017]	0.41
1/0 Water from tubewell in village	−0.0188*	(0.010)	[−0.039, 0.001]	0.51
Share of village area irrigated	0.0079	(0.007)	[−0.006, 0.022]	0.33
1/0 Educational facilities	−0.0081	(0.010)	[−0.028, 0.012]	0.50
Number of primary schools	−0.0093	(0.012)	[−0.032, 0.014]	0.49
Number of secondary schools	0.0039	(0.003)	[−0.001, 0.009]	0.01
1/0 Medical facilities	0.0036	(0.007)	[−0.010, 0.018]	0.10
1/0 Banking facilities	0.0012	(0.002)	[−0.003, 0.006]	0.01
1/0 Agricultural credit societies	0.0002	(0.004)	[−0.008, 0.009]	0.02
1/0 Post office	0.0006	(0.004)	[−0.007, 0.008]	0.03
1/0 Bus service	0.0058	(0.014)	[−0.022, 0.033]	0.41

Note. — This table regresses 2001 village covariates on the RD variables in Equation 1.1 and state fixed effects. Each row represents a separate regression, with the second column reporting the RD point estimate ($\hat{\beta}_1$). This table includes all available 2001 village-level covariates that correspond to a 2011 outcome reported in Table 1.5.3. SC and ST refer to official scheduled caste and scheduled tribe designations. "Main" workers work at least six months of the year, while "marginal" workers work less than six months. The RD bandwidth includes 29,765 villages with 2001 populations between 150 and 450, across 225 districts. Standard errors are clustered at the district level, which we use to calculate 95 percent confidence intervals in the fourth column. The fifth column reports the mean of the dependent variable for each RD regression. Significance: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

Figure A.2.11: RD Sensitivity – Census Outcomes, 2001 Covariate Smoothness



Note. — This figure shows reduced form results estimating Equation (1.1) without 2001 controls, as reported in Table A.2.17. Blue dots show average residuals form regressing the 2011 outcomes on state fixed effects. Each dot contains approximately 1,500 villages, averaged in 15-person population bins, including all 29,765 single-habitation villages between 150 and 450 people, in 10th-Plan districts.

A.2.5 Census outcome results: intensive margin of labor

Our main results examine the effects of electrification on extensive-margin employment. Thus far, we have presented results on the fraction of men and women working in agriculture, in the home, and in other sectors (see Table 1.5.3). We find evidence that eligibility for RGGVY shifts a small number of men out of agriculture and into the more formal sector, and we find a small but not statistically significant impact of RGGVY eligibility on female employment. However, it is also possible that electrification changes employment on the intensive margin, by causing previously employed workers to either increase of decrease the amount they work.

We test for effects on this intensive margin using data from the PCA, which reports separately employment counts for workers are "main", working 6 or more months out of the year, or "marginal", working fewer than 6 months.[57] We conduct an analogous employment analysis using the number of main workers divided by the number of total (main plus marginal) workers for each gender and sector as the dependent variable. In doing this, we drop villages with no workers in a particular gender-sector group, where this fraction is undefined. (This essentially amounts to dropping cases where this intensive margin does not exist.) Our results should therefore be interpreted as the effects of eligibility for electrification on the share of workers working at least 6 months per year, conditional on each village employing people of a given category.

Table A.2.19 and Figure A.2.12 display the results of this exercise. We find no evidence of a discontinuity in the intensive margin of labor at the 300-person RD threshold across any gender-sector category. Our point estimates across all categories are negative, with the exception of female agricultural workers. This demonstrates that RGGVY eligibility did not increase labor on the intensive margin. None of our results are statistically significant, but our confidence intervals are relatively tight, and we can reject small effects.

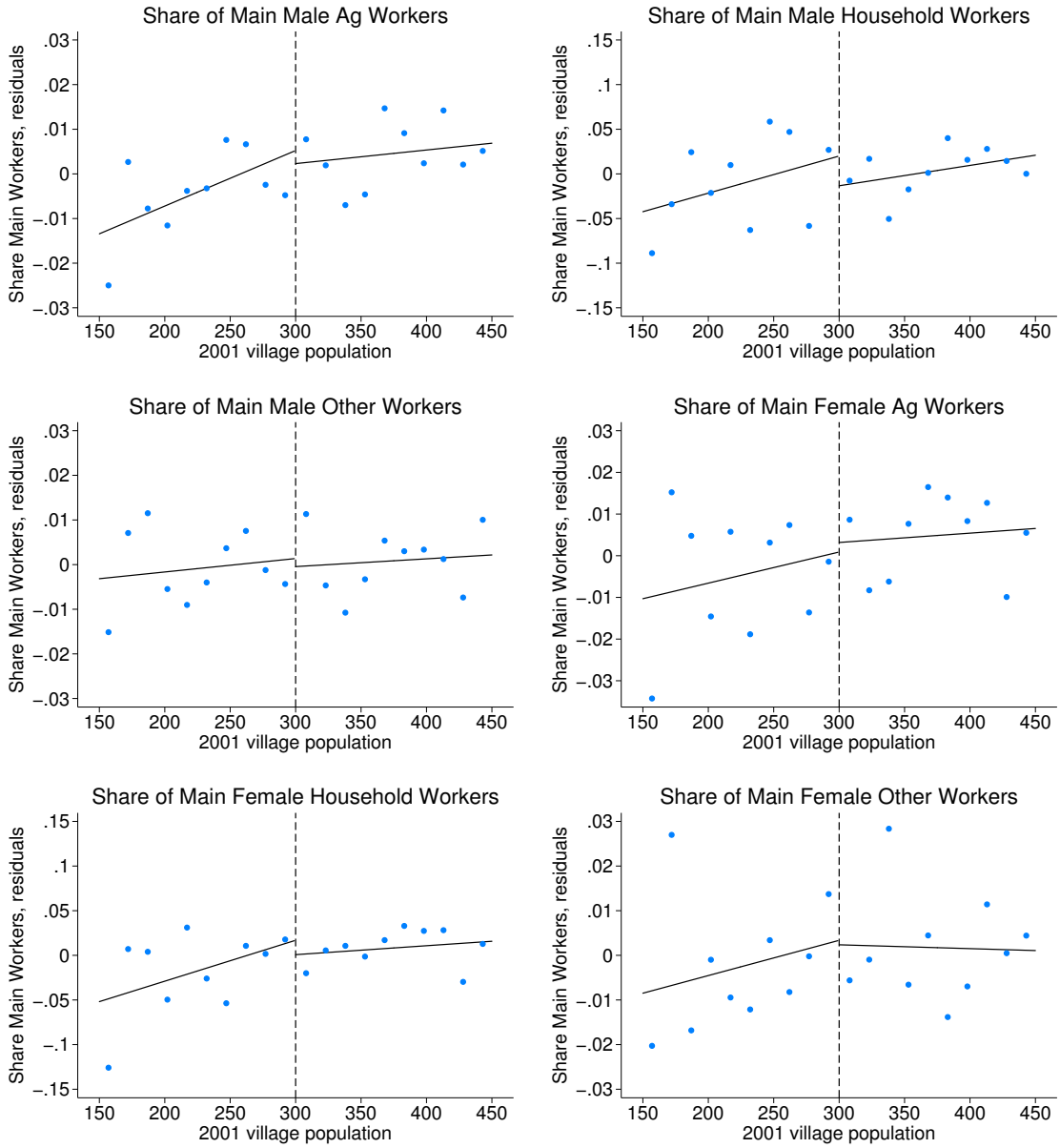
[57]For all employment results in Table 1.5.3 and Appendix A.2.4, we sum main and marginal workers into a single pooled employment metric for each subcategory.

Table A.2.19: RD Results – Share of "Main" Workers by Sector

2011 Outcome Variable	RD Coefficient	Standard Error	95 Percent Confidence	Mean of Outcome	Total Villages
Share of main male agricultural workers	−0.0029	(0.007)	[−0.016, 0.011]	0.72	29,572
Share of main female agricultural workers	0.0023	(0.010)	[−0.017, 0.022]	0.51	26,487
Share of main male household workers	−0.0333	(0.023)	[−0.079, 0.013]	0.65	5,324
Share of main female household workers	−0.0161	(0.035)	[−0.084, 0.052]	0.42	2,807
Share of main male other workers	−0.0018	(0.008)	[−0.017, 0.013]	0.79	24,855
Share of main female other workers	−0.0010	(0.013)	[−0.027, 0.025]	0.69	14,389

Note. — This table estimates Equation (1.1) on the share of "main" workers for each sector/gender. By definition, "main" workers work for at least 6 months of the year, while "marginal" workers work for less than 6 months of the year. Our main results in Table 1.5.3 pooled main and marginal workers, thereby testing for effects on the extensive margin of labor. In this table, we test the intensive margin using outcome variables that divide the number of main workers by the number of (main + marginal) workers in each category. Each row represents a separate regression, with the second column reporting the RD point estimate ($\hat{\beta}_1$). All specifications control for state fixed effects and the 2001 level of the outcome variable. Standard errors are clustered at the district level, which we use to calculate 95 percent confidence intervals in the fourth column. The fifth column reports the mean of the dependent variable for each RD regression. The sixth column shows the number of villages in each regression, which varies since the 2011 outcome and the 2001 control divide by total workers and many villages report zero total workers for a given category. Significance: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

Figure A.2.12: RD Results – Share of "Main" Workers by Sector



Note. — This figure shows RD results estimating Equation (1.1) on the share of workers in each category that work at least 6 months of the year, as reported in Table A.2.19. Blue dots show average residuals form regressing the 2011 outcomes on the 2001 control and state fixed effects. Each dot contains between 150–1,500 villages, averaged in 15-person population bins, including all 29,765 single-habitation villages between 150 and 450 people, in 10th-Plan districts, with nonzero total workers for both 2001 and 2011.

A.2.6 Census outcome results: validity tests

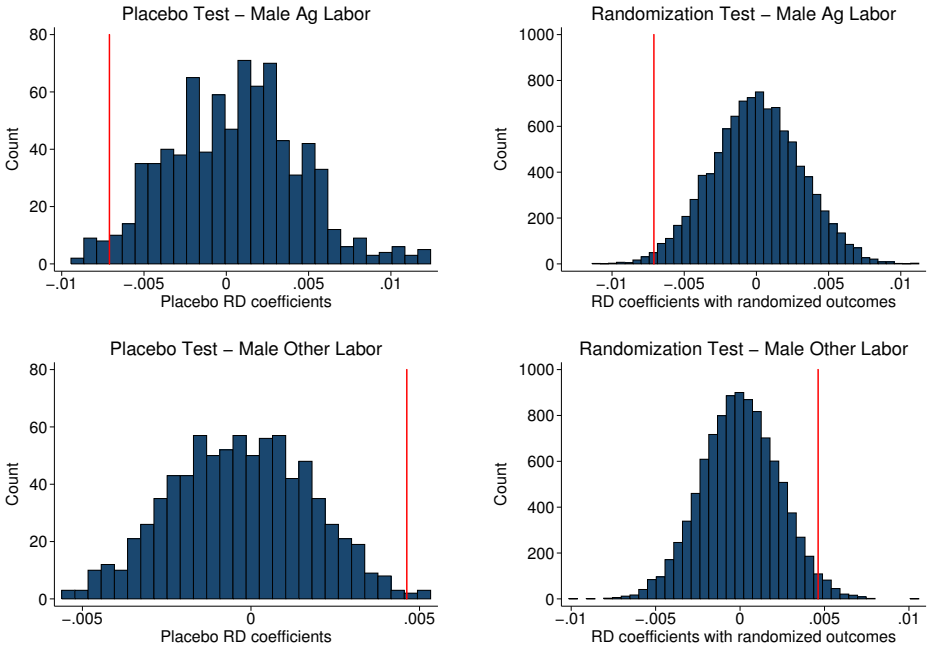
Section 1.5 in the main text supports the validity of the nighttime brightness RD results using a placebo test, a randomization test, and three falsification tests. Below, we conduct the analogous validity tests on the two census outcomes with the strongest non-zero results — the share of male agricultural workers, and the share of male "other" workers. While even these RD results are small in magnitude, they are statistically significant and have RD plots that display a level shift at the 300-person eligibility cutoff.[58] This implies that these validity tests can support our use of RD inference to test the effects of RGGVY program eligibility on these labor outcomes.

Figure A.2.13 reports the results of placebo and randomization tests for 2011 male labor outcomes. We implement these tests separately on each outcome, exactly as described in Section 1.5 in the main text. The decrease in the share of men working in agriculture falls below the 3rd percentile of the placebo distribution and below the 1st percentile of the randomization distribution. This indicates that that this RD result is very unlikely to reflect spurious volatility in the relationship between male agricultural labor and village population data. The increase in the share of male "other" labor also passes the placebo and randomization tests (above the 99th and 97th percentiles, respectively).

Table A.2.20, Figure A.2.14, and Figure A.2.15 conduct three falsification tests, by estimating Equation (1.1) on subsets of villages for which the 300-person cutoff in village population was not the relevant criterion determining RGGVY eligibility (because these villages contain multiple habitations and/or faced a 100-person eligibility cutoff). As expected, none of the three alternative samples shows statistically significant discontinuities at the 300-person cutoff. This provides further evidence that the RGGVY program has driven the small-but-statistically-significant shift of male labor out of the agricultural sector.

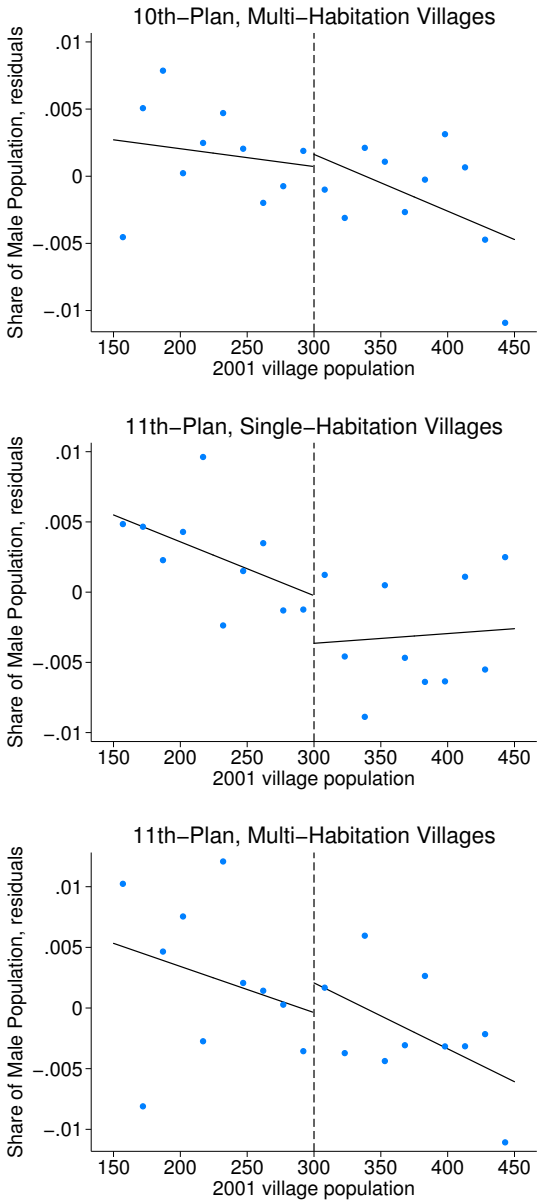
[58] Male agricultural labor is statistically significant at the 5 percent level in every set of sensitivity results we present. Male other labor is mostly significant at the 5 percent level, and quite robust to different RD specifications.

Figure A.2.13: Male Labor Shares – Placebo and Randomization Tests



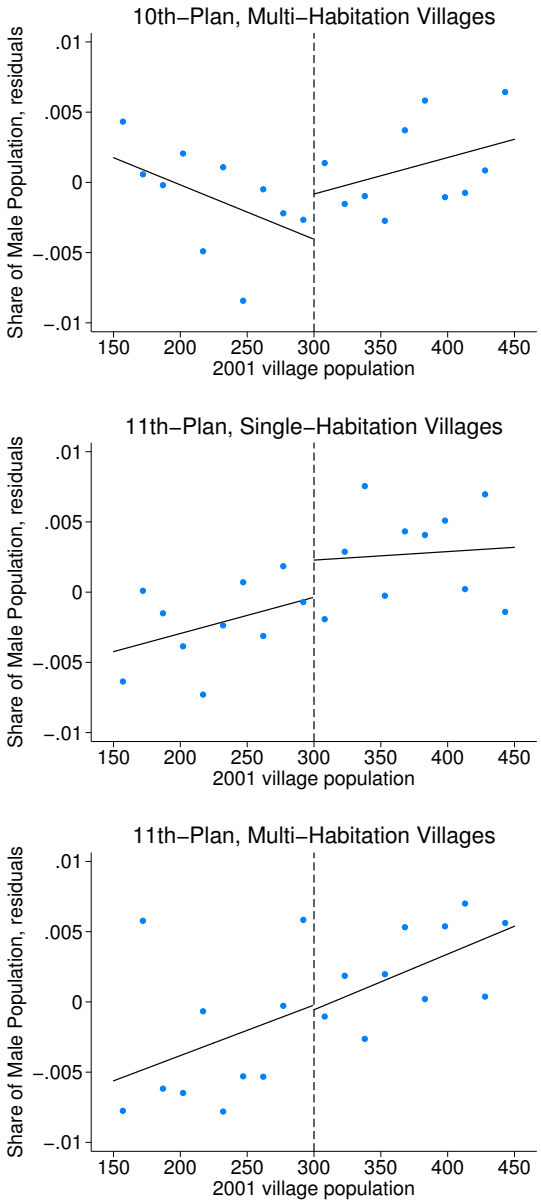
Note. — This figure presents the distributions of placebo RD coefficients and randomized RD coefficients. The left panels were generated by estimating Equation (1.1) using 801 placebo RD thresholds, representing all integer values in $[151, 275] \cup [325, 1000]$. We omit placebo thresholds within 25 people of the true 300-person threshold to ensure that placebo RDs do not detect the true effects of RGGVY eligibility, and we exclude thresholds below 151 due to our 150-person bandwidth. The right panel was generated by scrambling village brightness 10,000 times and re-estimating Equation (1.1) each time. The red lines represent our estimates of the RD coefficients from Table 1.5.3, using the correct 300-person threshold with unscrambled lights data. The RD point estimate for the share of male agricultural workers falls below the 3nd percentile of the placebo distribution and below the 1st percentile of the randomization distribution. The RD point estimate for the share of male other workers falls above the 99th percentile of the placebo distribution and above the 97th percentile of the randomization distribution.

Figure A.2.14: Male Agricultural Labor – Falsification Tests



Note. — This figure presents three falsification tests for our RD on male agricultural labor, corresponding to the first row of Table A.2.20. The top and bottom panels include only villages with multiple habitations, for which the running variable of village population did not determine RGGVY eligibility. The middle and bottom panels include only villages in districts that became eligible for RGGVY under the 11th Plan, for which the appropriate eligibility cutoff was lowered from 300 to 100 people. Blue dots show average residuals form regressing the 2011 outcomes on 2001 male agricultural employment and state fixed effects. Each dot contains approximately 800–1,500 villages, averaged in 15-person population bins. Lines are estimated separately on each side of the 300-person threshold, for villages within the 150–450 population RD bandwidth.

Figure A.2.15: Male Other Labor – Falsification Tests



Note. — This figure presents three falsification tests for our RD on male other labor, corresponding to the second row of Table A.2.20. The top and bottom panels include only villages with multiple habitations, for which the running variable of village population did not determine RGGVY eligibility. The middle and bottom panels include only villages in districts that became eligible for RGGVY under the 11th Plan, for which the appropriate eligibility cutoff was lowered from 300 to 100 people. Blue dots show average residuals form regressing the 2011 outcomes on 2001 male other employment and state fixed effects. Each dot contains approximately 800–1,500 villages, averaged in 15-person population bins. Lines are estimated separately on each side of the 300-person threshold, for villages within the 150–450 population RD bandwidth.

Table A.2.20: RD Sensitivity – Falsification Tests

2011 Outcome Variable	10th Plan Single-Hab.	10th Plan Multi-Hab.	11th Plan Single-Hab.	11th Plan Multi-Hab.
	(1)	(2)	(3)	(4)
Male ag workers / male pop.	−0.0071**	0.0009	−0.0034	0.0024
	(0.0028)	(0.0039)	(0.0033)	(0.0042)
Male other workers / male pop.	0.0046**	0.0032	0.0027	−0.0003
	(0.0019)	(0.0033)	(0.0027)	(0.0033)
RD bandwidth	150	150	150	150
Number of observations	29,765	16,481	24,104	16,164
Number of districts	225	202	261	219

Note. — This table reports results from 8 separate RD regressions, estimating Equation (1.1) on four disjoint subsets of Indian villages. Column (1) reproduces results from Table 1.5.3, using our RD sample of single-habitation villages in 10th-Plan RGGVY districts. Columns (2) and (4) include villages with multiple habitations, for which the the running variable (village population) does not correspond to the habitation populations that determined village eligibility. Columns (3) and (4) includes villages that were eligibile for RGGVY under the 11th Plan, after the cutoff had moved from 300 to 100 people. Figures A.2.14 and A.2.15 present these falsification tests graphically. All specifications control for the 2001 level of the outcome variable and state fixed effects. Standard errors are clustered at the district level. Significance: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

A.2.7 Socio-Economic Caste Census results: RD robustness

Table 1.5.4 presents in the main text results from ten SECC outcomes relating to household wealth and adult employment. Figure 1.5.9 displays RD plots for four of these ten outcomes, and we report the remaining six RD plots below in Figure A.2.16. The two household figures show no visual evidence of discontinuities at the 300-person threshold, confirming our results from Table 1.5.4 that reject economically significant effect as a result of RGGVY electrification. The four employment figures are qualitatively similar to those in Figure 1.5.7, except for suggestive evidence of an increase in the share of adult women working on other jobs.

Figure A.2.17 conducts RD bandwidth sensitivities on two SECC outcomes reported in the main text: the share of total households with at least one poverty indicator, and the share of these households earning a monthly income greater than Rs 5,000. We see that for bandwidths above 80 people, our RD estimates are not sensitive to the choice of bandwidth.[59]

We might be concerned that when we average SECC outcomes to the village level for villages containing very few households with a poverty indicator in 2011, these averages are noisy and sensitive to outliers. Table A.2.21 tests for this possibility, by dropping villages with fewer than 10 percent of households in our SECC dataset in Panel A, and fewer than 20 percent of households in Panel B. In Panel C, we introduce 2001 controls to the SECC RD specifications, by selecting the 2001 Census variables that most closely align with each 2011 village level outcome. These sensitivities have very little effect on our results from Table 1.5.4.

Table A.2.22 presents results for additional village-level employment outcomes. In Panel A, we report results for the main source of household income earned by the household's main income earner. These income categories ("cultivation", "manual/casual labor", "non-farm enterprise") were recorded directly by SECC enumerators, and they do not map to the employment sectors used in other SECC outcomes. By contrast, the string-parsed employment categories for heads of household produce results very close to the adult male employment regressions in Panel B of Table 1.5.4. This is not surprising, as 83 precent household heads in our RD sample are adult men. Panel B of Table A.2.22 reports analogous results for youth (i.e. ages 0–16) employment, for the subset of households with at least 1 poverty indicator. We see that non-farm, non-household youth employment may increase slightly as a result of RGGVY eligibility, but we can reject effects larger than 1.6 percentage points.

We construct the employment categories "agricultural", "household", and "other" by string parsing occupations reported at the individual level and aggregating up to

[59]The Imbens and Kalyanaraman (2012) optimal bandwidths range from 102 to 141 for these outcomes).

the household/village. While we try to recreate the three labor sectors reported in the Primary Census Abstract as closely as possible (in order to best facilitate apples-to-apples comparisons between Census and SECC outcomes), the "other" category contains an extremely wide range of occupations. In Panel C of Table A.2.22, we restrict this "other" category to exclude occupations containing strings closely associated with manual labor: "labor" and "worker". This yields larger point estimates with larger t-stats, across all four gender/age combinations. Hence, this suggests that the (small) increases in "other" employment caused by RGGVY may represent a (small) shift towards relatively higher paying jobs outside of agriculture.

Table A.2.21: RD Sensitivity – SECC Village-Level Outcomes

2011 Outcome	RD Coefficient	Standard Error	95 Percent Confidence	Mean of Outcome
A. Minimum 10 percent households with poverty indicator				
Monthly income greater than Rs 5,000	0.0017	(0.004)	$[-0.006, 0.010]$	0.08
One member holding salaried job	0.0030	(0.002)	$[-0.001, 0.007]$	0.02
Owning any land	0.0044	(0.009)	$[-0.012, 0.021]$	0.44
Male agricultural workers / adult men	-0.0108^{**}	(0.005)	$[-0.021, -0.000]$	0.29
Female agricultural workers / adult women	-0.0021	(0.005)	$[-0.011, 0.007]$	0.08
Male household workers / adult men	0.0013	(0.001)	$[-0.001, 0.004]$	0.01
Female household workers / adult women	-0.0035	(0.007)	$[-0.018, 0.010]$	0.51
Male other workers / adult men	0.0056	(0.006)	$[-0.007, 0.018]$	0.42
Female other workers / adult women	0.0063	(0.005)	$[-0.004, 0.016]$	0.16
B. Minimum 20 percent households with poverty indicator				
Monthly income greater than Rs 5,000	0.0029	(0.004)	$[-0.006, 0.011]$	0.07
One member holding salaried job	0.0018	(0.002)	$[-0.003, 0.006]$	0.02
Owning any land	0.0049	(0.008)	$[-0.011, 0.020]$	0.43
Male agricultural workers / adult men	-0.0114^{**}	(0.006)	$[-0.022, -0.001]$	0.29
Female agricultural workers / adult women	-0.0015	(0.005)	$[-0.011, 0.008]$	0.08
Male household workers / adult men	0.0011	(0.001)	$[-0.002, 0.004]$	0.01
Female household workers / adult women	-0.0044	(0.007)	$[-0.018, 0.009]$	0.51
Male other workers / adult men	0.0056	(0.007)	$[-0.008, 0.019]$	0.43
Female other workers / adult women	0.0064	(0.006)	$[-0.004, 0.017]$	0.17
C. Adding 2001 village controls				
At least one poverty indicator	0.0001	(0.006)	$[-0.012, 0.013]$	0.48
Monthly income greater than Rs 5,000	0.0053	(0.004)	$[-0.003, 0.014]$	0.08
One member holding salaried job	0.0030	(0.002)	$[-0.002, 0.008]$	0.02
Owning any land	-0.0005	(0.008)	$[-0.017, 0.016]$	0.44
Male agricultural workers / adult men	-0.0072	(0.005)	$[-0.017, 0.003]$	0.29
Female agricultural workers / adult women	-0.0036	(0.005)	$[-0.013, 0.006]$	0.08
Male household workers / adult men	0.0007	(0.001)	$[-0.002, 0.004]$	0.01
Female household workers / adult women	-0.0016	(0.008)	$[-0.016, 0.013]$	0.51
Male other workers / adult men	0.0044	(0.006)	$[-0.008, 0.017]$	0.42
Female other workers / adult women	0.0054	(0.005)	$[-0.005, 0.016]$	0.16

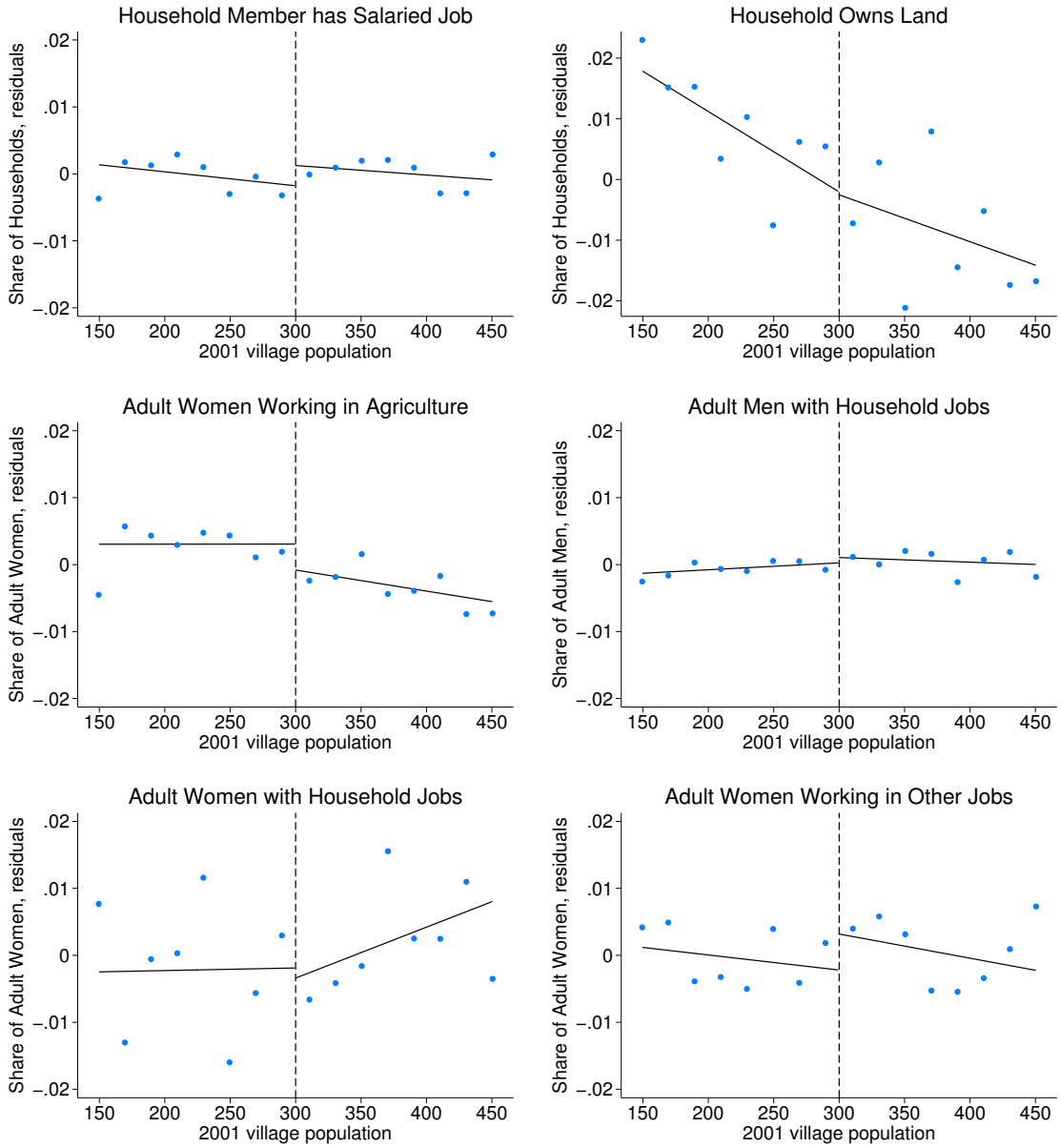
Note. — This table reports results of three sensitivity analyses on regressions reported in Table 1.5.4. Household-level outcomes are coded as the proportion of households with at least one poverty indicator, while adult employment outcomes are coded as the share of men (women) over 16 in households with a poverty indicator with an occupation in each sector. Each row represents a separate RD regression. Panel A includes villages where at least 10 percent of total households are included in our subset of the SECC data, resulting in 23,711 village observations. Panel B includes only villages where at least 20 percent of total households are included in our subset of the SECC data, restricting the analysis to only 21,072 village observations. Panel C includes all 25,942 SECC villages within our RD bandwidth, and also includes 2001 controls from the Census dataset. All regressions include a corresponding 2001 control. The second column shows the RD point estimate ($\hat{\beta}_1$) for each regression. All specifications include state fixed effects. Standard errors are clustered at the district level, which we use to calculate 95 percent confidence intervals in the fourth column. The fifth column reports the mean of the dependent variable for each RD regression. Significance: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

Table A.2.22: RD Results – Additional SECC Village-Level Employment Outcomes

2011 Outcome	RD Coefficient	Standard Error	95 Percent Confidence	Mean of Outcome
A. Share of households				
Main source of income: cultivation	0.0046	(0.007)	$[-0.010, 0.019]$	0.33
Main source of income: manual/casual labor	−0.0081	(0.007)	$[-0.022, 0.006]$	0.58
Main source of income: non-farm enterprise	−0.0006	(0.001)	$[-0.002, 0.001]$	0.01
Head of household occupation: agriculture	-0.0119^{*}	(0.007)	$[-0.025, 0.001]$	0.34
Head of household occupation: household work	0.0026	(0.003)	$[-0.003, 0.008]$	0.06
Head of household occupation: other work	0.0060	(0.007)	$[-0.008, 0.020]$	0.46
B. Youth employment				
Male youth ag workers / male youth	0.0006	(0.001)	$[-0.001, 0.003]$	0.01
Female youth ag workers / female youth	−0.0009	(0.001)	$[-0.003, 0.001]$	0.01
Male youth household workers / male youth	−0.0006	(0.001)	$[-0.003, 0.002]$	0.01
Female youth household workers / female youth	−0.0007	(0.001)	$[-0.004, 0.002]$	0.02
Male youth other workers / male youth	0.0070	(0.005)	$[-0.002, 0.016]$	0.07
Female youth other workers / female youth	0.0067	(0.004)	$[-0.001, 0.015]$	0.07
C. Excluding manual labor				
Adult male other workers / adult men	0.0107^{***}	(0.004)	$[0.003, 0.019]$	0.15
Adult female other workers / adult women	0.0104^{***}	(0.004)	$[0.003, 0.018]$	0.08
Male youth other workers / male youth	0.0086^{*}	(0.005)	$[-0.000, 0.017]$	0.06
Female youth other workers / female youth	0.0080^{*}	(0.004)	$[-0.000, 0.016]$	0.06

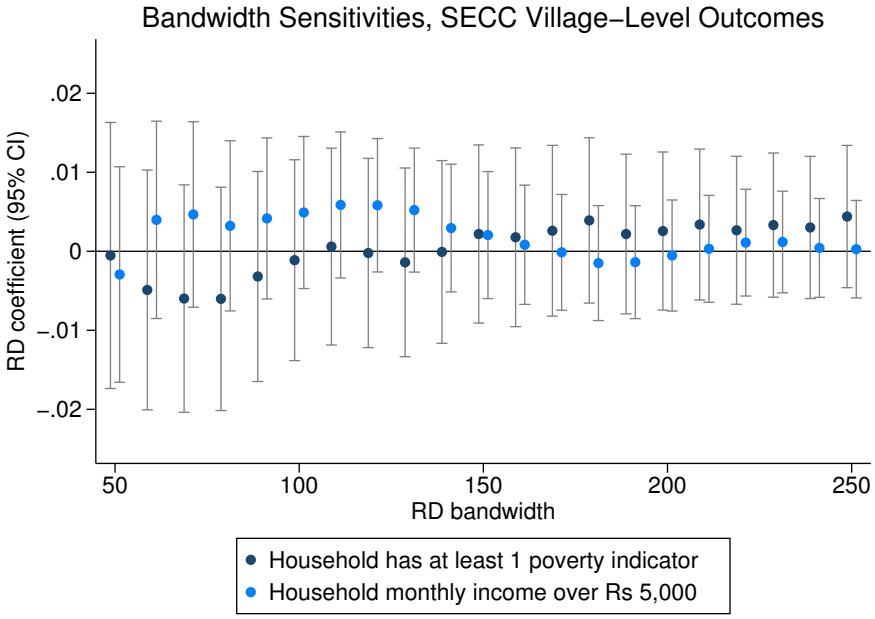
Note. — Each row represents a separate regression estimating Equation (1.1) on a different SECC village-level outcome. Outcomes are coded as the proportion of households (individuals) in each village, conditional on the (individual belonging to a) household having at least one poverty indicator in 2011. In Panel A, we report household income by source, based on categories coded in the SECC data; household head occupations are coded via string parsing each individual's reported occupation. Panel B presents youth employment results analogous to the adults labor results presented in Panel B of Table 1.5.4, defining all individuals less than 16 years old as "youth". Panel C reports adult and youth employment results where we narrow the definition of "other" to exclude occupations with "labor" or "worker" in their description. The second column shows the RD point estimate ($\hat{\beta}_1$) for each regression. All specifications include state fixed effects, but they do not include any additional baseline control variables. The RD bandwidth includes 25,942 villages with 2001 populations between 150 and 450. These regressions contain fewer villages than regressions in Table 1.5.3 because only 87 percent of 10th-Plan, single-habitation, 150–450 villages match to the SECC dataset. Standard errors are clustered at the district level with 222 clusters, which we use to calculate 95 percent confidence intervals in the fourth column. The fifth column reports the mean of the dependent variable for each RD regression. Significance: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

Figure A.2.16: RD Results – SECC Village-Level Outcomes



Note. — This figure presents RD results for SECC outcomes for our preferred specification, as a complement to Figure 1.5.9. They correspond to results reported in Table 1.5.4. Blue dots show average residuals from regressing the 2011 SECC village-level outcome (coded as the share of households in the village with at least one poverty indicator) on state fixed effects. Each dot contains approximately 1,600 villages, averaged in 20-person population bins. Lines are estimated separately on each side of the 300-person threshold, for all 10th-Plan single-habitation villages within our 150–450 population RD bandwidth, which match to the SECC dataset.

Figure A.2.17: RD Sensitivity – SECC, Bandwidths



Note. — This figure presents our bandwidth sensitivity analysis for two SECC outcomes (represented in the first two rows of Table 1.5.4). For each outcome, we estimate Equation (1.1) separately on bandwidths ranging from 50 (i.e., 250–350 people) to 250 (i.e., 50–550 people). Each dot represents the point estimate on the RD discontinuity at a given bandwidth around the 300-person cutoff, with 95 percent confidence intervals clustered at the district level. Our chosen bandwidth of 150 includes villages with populations between 150 and 450. The optimal RD bandwidth for these RD specifications ranges from 102 to 141 (calculated using the algorithm proposed by Imbens and Kalyanaraman (2012), using uniform, Epanechnikov, and triangular kernels).

A.2.8 School enrollment results: RD robustness

Table 1.5.5 reports results for five measures of village-level school enrollment, using our preferred RD specification. Figure 1.5.10 includes RD plots for two of these five outcomes, and we report the remaining three RD plots below in Figure A.2.18. We see no visual evidence of a discontinuity at the RD threshold, confirming our results in Table 1.5.5 that reject economically significant increases in school enrollment as a result of RGGVY electrification.

As an alternative RD specification, we re-estimate these enrollment regressions at the school level. Instead of aggregating enrollment counts across all schools in a village up to a single village-level observation, these regressions treat each school as a separate observation.[60] Table A.2.23 reports results for these school-level RD regressions, while Figure A.2.19 shows the analogous RD graphs. These estimates are very similar to the village-level results, with confidence intervals that reject 10 percent changes in enrollment as a result of RGGVY eligibility. The school-level RD plots in Figure A.2.19 likewise show no evidence of a discontinuity at the 300-person threshold.

Figure A.2.20 conducts RD bandwidth sensitivities using the total enrollment outcome, for both the village-level and school-level specifications. This demonstrates that these RD results are not sensitive to bandwidths above 100 people, which is below the smallest optimal bandwidth calculation for these outcomes (Imbens and Kalyanaraman (2012) optimal bandwidths range from 138 to 179). This provides further evidence in support of the assumptions underpinning our RD design. Table A.2.24 conducts additional specification sensitivities for our village-level enrollment RD. This shows that our RD point estimates are not sensitive to our choice of outcome year (i.e. total enrollment for the school year beginning in 2010, 2011, or 2012), control year (i.e. pre-RGGVY enrollment for the school year beginning in 2005 or 2006), or sample restriction (i.e. removing village-school matches most likely to be inaccurate). Across all sensitivities, we can reject even moderate increases to school enrollment around our 300-person RD threshold.

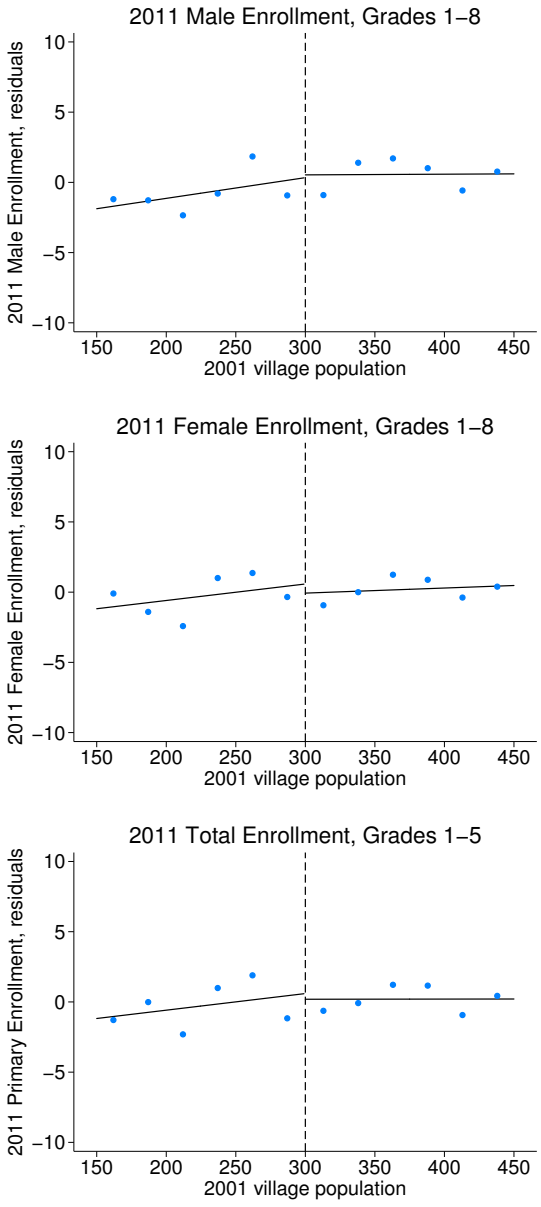
[60] 25 percent of the 15,215 RD bandwidth villages that matched to the DISE schools dataset (via the fuzzy matching algorithm detailed in Section A.1.7) matched to multiple schools reporting enrollment counts in 2005 and 2011.

Table A.2.23: RD Sensitivity – School Enrollment, School-Level Regressions

2011 Outcome Variable	RD Coefficient	Standard Error	95 Percent Confidence	Mean of Outcome
Total enrollment, grades 1–8	−0.742	(2.30)	[−5.25, 3.77]	61.19
Male enrollment, grades 1–8	−0.189	(1.13)	[−2.39, 2.02]	31.07
Female enrollment, grades 1–8	−0.556	(1.22)	[−2.95, 1.84]	30.12
Total enrollment, grades 1–5	−1.205	(1.98)	[−5.08, 2.67]	50.06
Total enrollment, grades 6–8	0.499	(0.75)	[−0.97, 1.97]	11.13

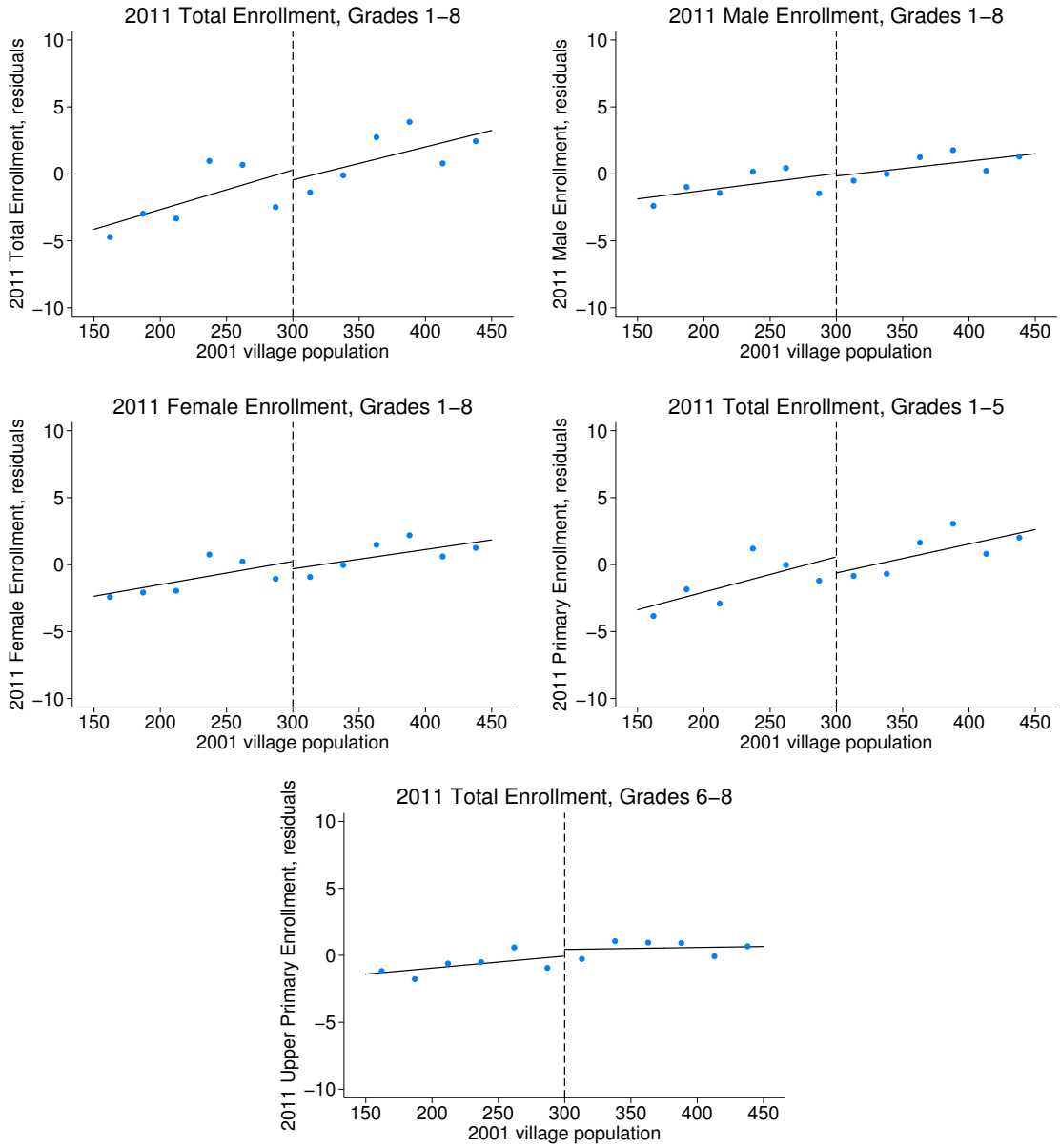
Note. — Each row represents a separate regression estimating Equation (1.1) at the school level, on a different enrollment count. The second column shows the RD point estimate ($\hat{\beta}_1$) for each regression. All specifications control for the 2005 level of the outcome variable and state fixed effects. The RD bandwidth includes 13,150 school-level observations, across 11,578 villages with 2001 populations between 150 and 450. These regressions contain fewer villages than regressions in Table 1.5.5 because some villages have nonmissing enrollment counts for for 2011 and 2005 only after summing across multiple schools (even though no single school in these villages has nonmissing data for both years). Standard errors are clustered at the district level, with 215 clusters, which we use to calculate 95 percent confidence intervals in the fourth column. The fifth column reports the mean of the dependent variable for each RD regression. Significance: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

Figure A.2.18: RD Results – School Enrollment



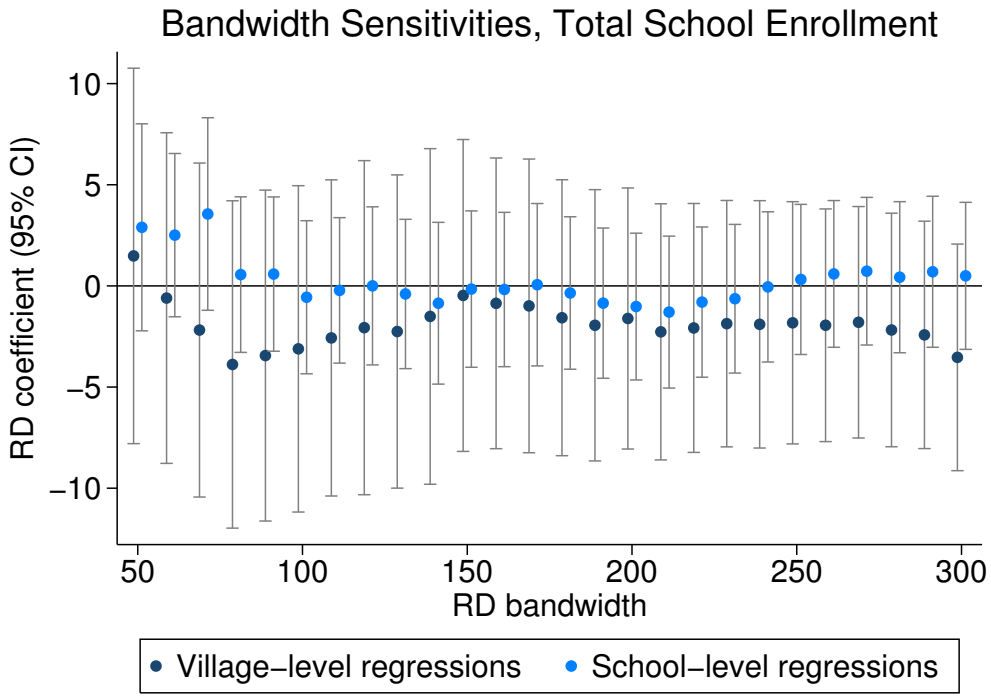
Note. — This figure presents RD results for school enrollment counts for our preferred specification, as a complement to Figure 1.5.10. They correspond to results reported in Table 1.5.5. Blue dots show average residuals form regressing the 2011 village-level enrollment on 2005 village-level enrollment and state fixed effects. Each dot contains approximately 1,000 villages, averaged in 25-person population bins. Lines are estimated separately on each side of the 300-person threshold, for all 10th-Plan single-habitation villages within our 150–450 population RD bandwidth, with school-village matches and nonmissing enrollment counts for both 2005 and 2011.

Figure A.2.19: RD Sensitivity – School-Level Enrollment Regressions



Note. — This figure presents RD results for school-level enrollment, corresponding to results reported in Table A.2.23. Blue dots show average residuals form regressing the 2011 school enrollment on 2005 enrollment and state fixed effects. Each dot contains approximately 1,000 schools, averaged in 25-person population bins. Lines are estimated separately on each side of the 300-person threshold, for all 10th-Plan single-habitation villages within our 150–450 population RD bandwidth, with schools reporting nonmissing enrollment data for both 2011 and 2005.

Figure A.2.20: RD Sensitivity – School Enrollment, Bandwidths



Note. — This figure presents our bandwidth sensitivity analysis for total school enrollment (grades 1–8), at the village level (represented in the first row of Table 1.5.5) and at the school level (represented in the first row of Table A.2.23). For each outcome, we estimate Equation (1.1) separately on bandwidths ranging from 50 (i.e., 250–350 people) to 300 (i.e., 0–600 people). Each dot represents the point estimate on the RD discontinuity at a given bandwidth around the 300-person cutoff, with 95 percent confidence intervals clustered at the district level. Our chosen bandwidth of 150 includes villages with populations between 150 and 450. The optimal RD bandwidth for these RD specifications ranges from 138 to 179 (calculated using the algorithm proposed by Imbens and Kalyanaraman (2012), using uniform, Epanechnikov, and triangular kernels).

Table A.2.24: RD Sensitivity – Total Grade 1–8 Enrollment, Village-Level Regressions

Outcome Year	Control Year	Subsample	RD Coeff	Standard Error	95 Percent Confidence	Mean of Outcome	Village Observations
2011	2005		−0.472	(3.93)	[−8.18, 7.24]	74.05	12,251
2011	2006		2.196	(3.31)	[−4.30, 8.69]	74.05	12,651
2010	2005		−1.242	(3.66)	[−8.42, 5.94]	75.96	12,290
2012	2005		−2.952	(4.97)	[−12.69, 6.78]	69.94	12,205
2010–12	2005–06		0.528	(3.19)	[−5.73, 6.79]	73.30	12,801
2011	2005	VD school	0.825	(3.48)	[−6.00, 7.65]	66.21	8,837
2011	2005	PIN code	−1.449	(4.49)	[−10.25, 7.35]	73.81	9,560

Note. — This table conducts sensitivity analysis on the RD results for total school enrollment, at the village level. The first row reproduces our preferred specification from the top row of Table 1.5.5. The next three rows report results for the same regression, using outcomes and controls from adjacent years. Due to the unbalanced nature of DISE school panel dataset, nonmissing village observation counts are sensitive to the choice of outcome/control year. The fifth row averages village enrollment across 2010–2012 nonmissing values and controls across 2005–2006 nonmissing enrollment, which allows us to include 4 percent more schools than our preferred specification. The bottom two rows restrict our sample of village-schools matches to include only villages reported to have schools in the 2011 Village Directory; and only villages with Pincodes that match those reported in the DISE dataset. The fourth column shows the RD point estimate ($\hat{\beta}_1$) for each regression. All specifications control for total enrollment from 2005 (or 2006) and state fixed effects. The RD bandwidth includes single-habitation villages in 10th-Plan districts with 2001 populations between 150 and 450. Standard errors are clustered at the district level. Significance: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

A.2.9 Spatial spillovers

Villages economies do not exist in isolation, and it is important to consider potential spillover effects of electrification on neighboring villages. Below, we test for spatial spillovers by modifying Equation (1.1) such that the dependent variable is the average of each 2011 outcome across villages within 10-, 20-, and 50-km radii of each village in our RD sample. These regressions still treat single-habitation 10th-Plan villages with 2001 populations between 150–450 as the unit of analysis; they simply test for effects on RGGVY eligible on outcomes in other surrounding villages.[61]

Table A.2.25 reports these results, where we find little evidence of spatial spillovers. Compared to our main specification (labeled "0 km"), spillover results attenuate at 10 km. Weakly significant effects within a 10km radius would not necessarily provide evidence of spillovers, given the measurement error inherent to assigning villages to shapefiles and calculating spatial averages in GIS. As we have no reason to suspect different degrees of measurement error across outcomes, we conclude that economically meaningful spatial spillovers are unlikely.

[61] We exclude RD-sample villages when taking these spatial averages, in order to ensure that no village is simultaneously represented on the left-hand side and the right-hand side of these regressions.

Table A.2.25: Spatial Spillovers to Adjacent Villages

2011 Outcome Variable	Radius around within-bandwidth village			
	0 km	10 km	20 km	50 km
2011 nighttime brightness	0.1493**	0.0529	0.0475**	0.0178
	(0.0603)	(0.0360)	(0.0241)	(0.0189)
Male agricultural workers / male population	−0.0065**	−0.0004	−0.0004	0.0001
	(0.0033)	(0.0009)	(0.0007)	(0.0004)
Female agricultural workers / female population	−0.0051	−0.0012	−0.0004	−0.0002
	(0.0049)	(0.0015)	(0.0011)	(0.0006)
Male other workers / male population	0.0056**	0.0009	0.0007	0.0003
	(0.0023)	(0.0007)	(0.0005)	(0.0003)
Female other workers / female population	−0.0006	0.0010	0.0007	0.0002
	(0.0026)	(0.0009)	(0.0007)	(0.0004)

Note. — This table estimates Equation (1.1) on main RD samples, using the average 2011 outcomes of adjacent villages. For each single-habitation, 10th-Plan district within our 150–450 bandwidth, we calculate the average level of each outcome varaible for all adjacent villages *not* in the RD sample, within a 10-, 20-, and 50-kilometer radius of the village centroid. In each row, we present RD point estimates ($\hat{\beta}_1$) from four separate regressions — three regressions of the average outcome for adjacent villages within a given radius, and the main specification (i.e. 0 km) in the first column. Both nighttime brightness and labor regressions include 18,686 village observations, restricting the RD sample to the 12 states with available shapefiles that correlate with village areas. All specifications control for the 2001 level of the outcome variable and state fixed effects. Standard errors are clustered at the district level. Significance: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

A.2.10 Heterogeneous RGGVY implementation

One possible explanation for the small magnitudes of our RD results could be that RGGVY only impacted a subset of villages/districts/states in our sample. If this is were the case, and our RD estimates pooled villages with strong treatment effects and villages with no treatment effects, this would produce small average treatment effects. Below, we employ two strategies of subsampling our RD sample, by isolating districts and states most likely to demonstrate economically significant impacts from RGGVY.

First, we exploit the gradual rollout of RGGVY implementation under the 10th Plan. As shown in Table A.1.2, 10th-Plan districts received RGGVY funding as early as 2005 and as late as 2010. Even though the latest 10th-Plan funding predates our 2011 outcome data, it is quite possible that these 2011 data do not reflect the full impacts of electrification in districts where RGGVY implementation began in 2009

or 2010. This is especially likely to be true for medium-run economic outcomes. Table A.2.26 estimates Equation (1.1) on subsamples of RGGVY 10th-Plan districts that received RGGVY funds before 2007, 2008, 2009, and 2010. We see that while over half of 10th-Plan districts received funding in 2005 or 2006 (i.e. before 2007), this subsample yields RD point estimates very close to the full-sample averages. Variation in the timing of RGGVY rollout is unlikely to be obscuring large effects of electrification in districts with early RGGVY implementation.

Second, we consider heterogeneous power quality across states. Poor power quality could help to explain the small magnitudes of our economic results — newly electrified villages can only benefit from electricity infrastructure if power reliably flows through the grid. As a proxy for power quality at the state level, we use electricity demand surpluses/deficits as reported in the 2011–2012 Load Generation Balance Report (Central Electricity Authority (2011)). This defines demand surplus/deficit as the percentage of required electricity load that is available to the state. For the 12 states in our RD sample for nighttime brightness, this measure ranged from a 17.3 percent surplus in Chhattisgarh to a 19.4 percent deficit in Madhya Pradesh, with a 2011 national average shortfall of 10.3 percent of total load.

To test for heterogeneous effects in power quality, we estimate Equation (1.1) on the 7 states in our 12-state RD sample with 2011–2012 shortfalls that were better than the national average. These states are, in order of lowest-to-highest deficit (i.e. negative surplus): Chhattisgarh (-17.3 percent), Orissa (-15.4), Karnataka (-4.8), West Bengal (0.0), Gujarat (1.6), Haryana (6.0), and Rajasthan (7.0). Table A.2.27 shows that restricting the RD sample to these 7 states increases our RD point estimate for nighttime brightness by a factor of 1.7, without sacrificing precision. However, applying the same restriction to RD regressions on labor outcomes if anything attenuates our point estimates while barely affecting their confidence intervals. This demonstrates that even in states where RGGVY investments were likely coupled with above-average power availability, we can still reject economically significant changes in employment. Table A.2.28 presents analogous results for asset ownership, housing outcomes, and village-wide outcomes, while Tables A.2.29 and A.2.30 do so for the outcomes in Tables 1.5.4 and 1.5.5, respectively. The confidence intervals for the low-deficit sample remain broadly similar to those of both full 22-state RD sample and the 12-state lights sample, even at 32 percent smaller sample size.[62]

[62]For simplicity, A.2.27–A.2.30 use the 12-state RD lights sample for all outcome regressions. This excludes 11,079 villages in 10 states with low-quality or missing shapefiles, which are included in all other RD regressions on Census outcomes. 93 percent of these excluded villages are in Uttar Pradesh and Uttarakhand, two states with deficits well above the national average. Hence, the 12 vs. 22 state distinction is unimportant in this split-sample exercise, because the excluded states with below-average deficits are all very small.

There are three exceptions. First, the share of households owning bicycles in Table A.2.28 is negative and statistically significant in the low-deficit sample. This result appears to be spurious, as it is not robust across other assets that require electricity. Second, the adult male other employment result in Table A.2.29 is larger and statistically significant in the low-deficit sample. While the 0.034 upper bound of the 95 percent confidence interval is twice as large as with the full sample of 22 states, 0.034 still represents a relatively small change in the share of adult male workers.[63] Third, the five school enrollment results in Table A.2.30 are negative and statistically significant for low-deficit states, suggesting that RGGVY led to *decreases* in student enrollment. We report six corresponding RD pictures in Figure A.2.21, and only the SECC adult male other employment result reveals visual evidence of a discontinuity at the 300-person threshold.

RGGVY implementation may have been heterogeneous in other ways that limited its effectiveness. Given that district-specific RGGVY projects were often carried out by decentralized implementing agencies (i.e. state electricity boards, local distribution companies), the efficacy of project implementation might have varied widely across states/districts.[64] Even if all implementation efforts were identical, enforcement of the 300-person eligibility cutoff might have varied across implementers, which could reduce the power of our RD design. However, we lack a strong prior as to which states/districts are likely to have most strongly enforced the 300-person rule and most effectively implemented RGGVY projects. Interestingly, if we cherry-pick the 7 states with the largest RD point estimates for 2011 nighttime brightness, 6 of the 7 cherry-picked states also had below-average electricity deficits in 2011. This suggests that *if* poor implementation prevented barely eligible RGGVY villages in certain states from exhibiting increased nighttime brightness, these states were likely to have been states with above-average power shortfalls.

[63] Our labor share results from Table 1.5.3 use the village's full male population as the denominator. For comparison, a 3.4 percentage point increase in adult men in households with at least one poverty indicator translates to less than a 1 percentage point increase in the full male population. This is because for the average village in our sample, 67 percent of the male population is 16 or older, and only 44 percent of the male population is included in our SECC dataset.

[64] Apart from implementer-specific effects, this could reflect socioeconomic or political differences across states. For example, newspaper stories have cited ethnic conflict as having caused RGGVY implementation delays in the northeastern states of Bihar, Assam, and Jharkhand.

Table A.2.26: Subsample – Districts Receiving Early RGGVY Funding

2011 Outcome Variable	Received RGGVY funding before			
	2007	2008	2009	2010
2011 nighttime brightness	0.1414	0.1172*	0.1172*	0.1493**
	(0.0977)	(0.0609)	(0.0609)	(0.0603)
Number of villages	10,833	17,960	17,960	18,686
Number of districts	90	126	126	130
Number of states	10	12	12	12
Male agricultural workers / male pop.	−0.0081**	−0.0063**	−0.0070**	−0.0073***
	(0.0035)	(0.0029)	(0.0028)	(0.0028)
Female agricultural workers / female pop.	−0.0056	−0.0036	−0.0042	−0.0050
	(0.0049)	(0.0041)	(0.0041)	(0.0041)
Male other workers / male pop.	0.0047**	0.0041**	0.0046**	0.0046**
	(0.0024)	(0.0020)	(0.0020)	(0.0019)
Female other workers / female pop.	0.0016	−0.0004	−0.0003	−0.0004
	(0.0022)	(0.0021)	(0.0020)	(0.0020)
Number of villages	20,958	28,489	28,973	29,703
Number of districts	167	211	217	223
Number of states	12	18	20	21

Note. — This table estimates Equation (1.1) on subsamples of villages, in districts that received RGGVY 10th-Plan funding *before* a given year. All 10th-Plan districts received funding before the end of 2010, and only 2 districts in Nagaland received funding after January 1, 2010 (and are not included in the rightmost column above). In each row, we present RD point estimates ($\hat{\beta}_1$) from four separate regressions on subsamples of single-habitation, 10th-Plan villages within our RD bandwith. The number of villages/districts/states differ across nighttime brightness and labor regressions, because our nighttime lights sample includes only the 12 states with available shapefiles that correlate with village areas. All specifications control for the 2001 level of the outcome variable and state fixed effects. Standard errors are clustered at the district level. Significance: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

Table A.2.27: Subsample – States with Low Power Deficits (Lights and Labor)

2011 Outcome Variable	Full RD Sample	Full Lights Sample	Low-Deficit States
2011 nighttime brightness		0.1493^{**}	0.2481^{***}
		(0.0603)	(0.0737)
		[0.031, 0.268]	[0.104, 0.393]
Number of villages		18,686	12,679
Number of districts		130	67
Number of states		12	7
Male agricultural workers / male pop.	-0.0071^{**}	-0.0065^{**}	-0.0043
	(0.0028)	(0.0033)	(0.0039)
	[−0.013, −0.002]	[−0.013, −0.000]	[−0.012, 0.003]
Female agricultural workers / female pop.	-0.0049	-0.0051	-0.0040
	(0.0040)	(0.0049)	(0.0061)
	[−0.013, 0.003]	[−0.015, 0.005]	[−0.016, 0.008]
Male other workers / male pop.	0.0046^{**}	0.0056^{**}	0.0054^{**}
	(0.0019)	(0.0023)	(0.0025)
	[0.001, 0.008]	[0.001, 0.010]	[0.000, 0.010]
Female other workers / female pop.	-0.0004	-0.0006	-0.0006
	(0.0020)	(0.0026)	(0.0036)
	[−0.004, 0.004]	[−0.006, 0.005]	[−0.008, 0.006]
Number of villages	29,765	18,686	12,679
Number of districts	225	130	67
Number of states	22	12	7

Note. — This table estimates Equation (1.1) on our full RD sample (in the first column), our 12-state lights RD sample (in the second column), and a subsample of 7 states with the lowest reported electricity demand shortfalls for 2011 (in the third column). We define this demand shortfall as the percent of total electricity demand not met by each state (Central Electricity Authority (2011)). These 7 states are (in order of increasing demand shortfall) Chhattisgarh, Orissa, Karnataka, West Bengal, Gujarat, Haryana, and Rajasthan. In each row, we present RD point estimates ($\hat{\beta}_1$) from three separate regressions, include all villages in the full RD sample, the nighttime lights RD sample or the 7-state subset of the nighttime lights RD sample (the latter two samples exclude states with shapefiles that are either unavailable or uncorrelated with village areas). All specifications control for the 2001 level of the outcome variable and state fixed effects. We report 95 percent confidence intervals in brackets. Standard errors (in parentheses) are clustered at the district level. Significance: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

Table A.2.28: Subsample – States with Low Power Deficits (Assets, Housing, Public Goods)

2011 Outcome Variable	Full RD Sample	Full Lights Sample	Low-Deficit States
C. Asset ownership			
Share of households with telephone	0.0025	−0.0007	−0.0010
	[−0.008, 0.013]	[−0.014, 0.013]	[−0.018, 0.016]
Share of households with TV	0.0026	0.0041	0.0016
	[−0.005, 0.010]	[−0.004, 0.013]	[−0.008, 0.011]
Share of households with bicycle	−0.0015	−0.0048	−0.0116**
	[−0.010, 0.007]	[−0.016, 0.006]	[−0.023, −0.000]
Share of households with motorcycle	−0.0008	−0.0008	−0.0020
	[−0.006, 0.004]	[−0.007, 0.005]	[−0.010, 0.006]
Share of households without assets	0.0039	0.0060	0.0098
	[−0.004, 0.012]	[−0.005, 0.017]	[−0.003, 0.022]
D. Housing stock			
Share of households with elec/gas cooking	0.0005	−0.0013	−0.0045
	[−0.005, 0.006]	[−0.006, 0.003]	[−0.010, 0.001]
Share of households with kerosene lighting	0.0029	−0.0047	0.0027
	[−0.009, 0.015]	[−0.020, 0.011]	[−0.009, 0.014]
Share of households with mud floors	0.0043	0.0046	0.0040
	[−0.003, 0.012]	[−0.004, 0.013]	[−0.007, 0.015]
Share of households with thatched roof	−0.0034	−0.0045	0.0037
	[−0.013, 0.007]	[−0.016, 0.007]	[−0.008, 0.015]
Share of households dilapidated	−0.0031	−0.0060*	−0.0060
	[−0.009, 0.002]	[−0.013, 0.001]	[−0.014, 0.003]
E. Village-wide outcomes			
1/0 Mobile phone coverage in village	−0.0008	−0.0006	−0.0193
	[−0.023, 0.021]	[−0.030, 0.029]	[−0.049, 0.011]
1/0 Post office in village	0.0017	0.0002	0.0001
	[−0.005, 0.009]	[−0.008, 0.008]	[−0.009, 0.009]
1/0 Ag credit societies in village	0.0013	−0.0032	−0.0035
	[−0.006, 0.008]	[−0.011, 0.005]	[−0.011, 0.004]
1/0 Water from tubewell in village	−0.0075	−0.0170	−0.0290
	[−0.036, 0.021]	[−0.050, 0.016]	[−0.065, 0.007]
Share of village area irrigated	−0.0057	−0.0033	−0.0051
	[−0.016, 0.004]	[−0.011, 0.004]	[−0.014, 0.003]
Share of village area planted	0.0015	0.0064	0.0036
	[−0.010, 0.013]	[−0.008, 0.021]	[−0.014, 0.022]

Note. — This table is exactly analogous to Table A.2.27, except that it reports full sample, lights sample, and low-deficit-states results for outcomes in Panels C, D, and E of Table 1.5.3. Standard errors are omitted for brevity, but 95 percent confidence intervals are still presented in brackets. Please refer to the notes below Table A.2.27.

Table A.2.29: Subsample – States with Low Power Deficits (SECC Outcomes)

2011 Outcome Variable	Full RD Sample	Full Lights Sample	Low-Deficit States
A. Share of households			
At least one poverty indicator	0.0006	−0.0034	−0.0021
	[−0.011, 0.012]	[−0.019, 0.012]	[−0.022, 0.018]
Monthly income greater than Rs 5,000	0.0043	0.0069	0.0059
	[−0.004, 0.013]	[−0.003, 0.017]	[−0.006, 0.018]
One member holding salaried job	0.0030	0.0046**	0.0029
	[−0.002, 0.008]	[0.000, 0.009]	[−0.002, 0.008]
Owning any land	−0.0005	−0.0007	0.0068
	[−0.017, 0.016]	[−0.015, 0.014]	[−0.011, 0.024]
B. Adult employment			
Male agricultural workers / adult men	−0.0091*	−0.0078	−0.0047
	[−0.019, 0.001]	[−0.021, 0.005]	[−0.022, 0.013]
Female agricultural workers / adult women	−0.0039	−0.0002	−0.0001
	[−0.013, 0.006]	[−0.012, 0.012]	[−0.017, 0.017]
Male household workers / adult men	0.0008	0.0013	0.0014
	[−0.002, 0.004]	[−0.002, 0.005]	[−0.004, 0.007]
Female household workers / adult women	−0.0015	0.0065	0.0111
	[−0.016, 0.013]	[−0.014, 0.027]	[−0.019, 0.041]
Male other workers / adult men	0.0052	0.0124*	0.0177**
	[−0.007, 0.017]	[−0.002, 0.027]	[0.001, 0.034]
Female other workers / adult women	0.0054	0.0017	0.0033
	[−0.005, 0.016]	[−0.011, 0.014]	[−0.010, 0.017]

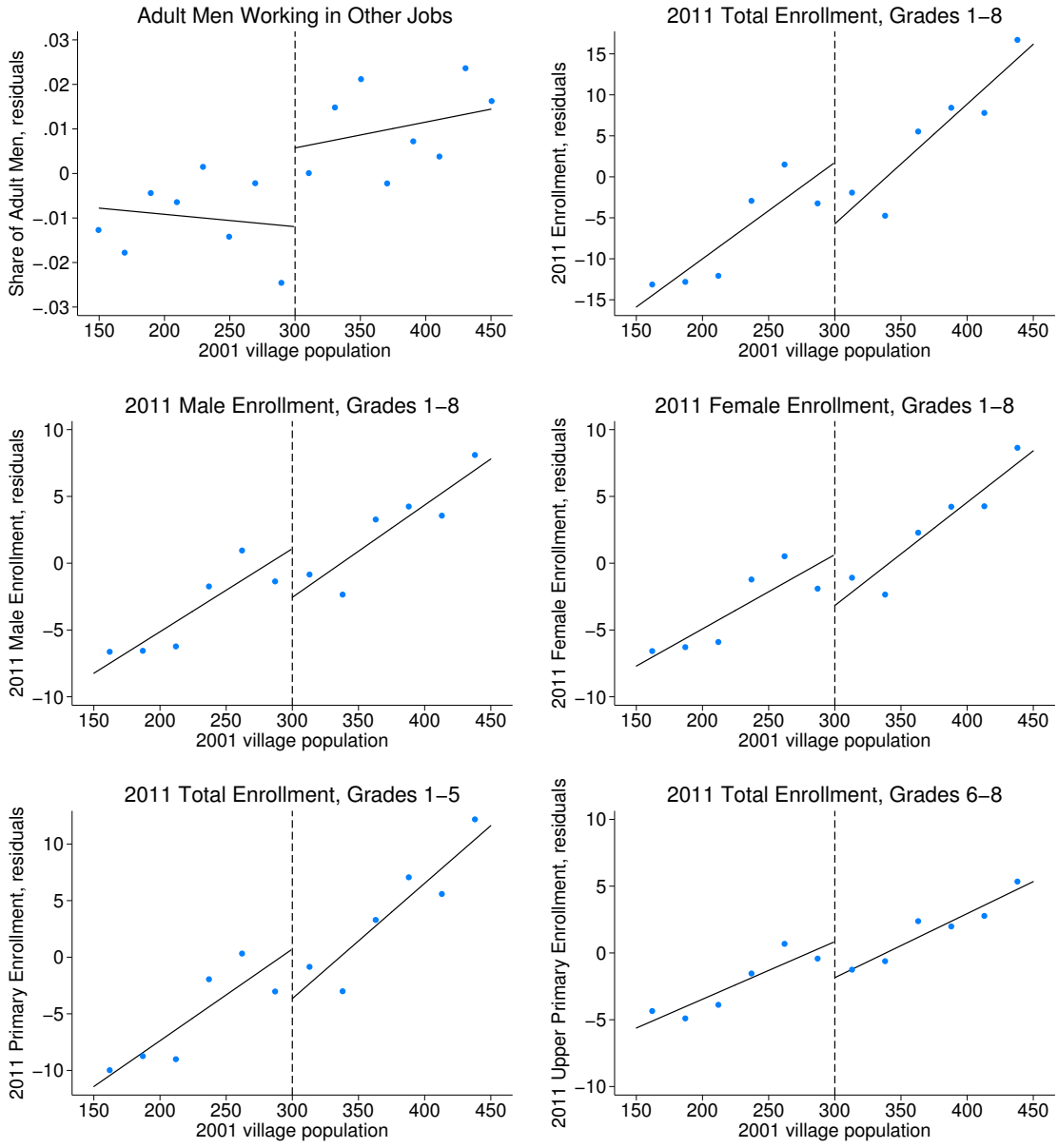
Note. — This table is exactly analogous to Table A.2.27, except that it reports full RD sample, lights RD sample, and low-deficit-state results for outcomes in Table 1.5.4. Regressions in the first column contain 25,942 village observations (as in Table 1.5.4), while regressions in the second and third columns contain 16,240 and 11,027 village observations, respectively. All regressions control for state fixed effects, but do not include any additional controls. Standard errors are omitted for brevity, but 95 percent confidence intervals are still presented in brackets. Please refer to the notes below Table A.2.27.

Table A.2.30: Subsample – States with Low Power Deficits (DISE Outcomes)

2011 Outcome Variable	Full RD Sample	Full Lights Sample	Low-Deficit States
Total enrollment, grades 1–8	−0.4725	−2.7663	−7.4681**
	[−8.182, 7.237]	[−8.174, 2.642]	[−13.325, −1.611]
Male enrollment, grades 1–8	0.1966	−1.5136	−3.6321**
	[−3.715, 4.108]	[−4.304, 1.277]	[−6.737, −0.527]
Female enrollment, grades 1–8	−0.6504	−1.1691	−3.7973***
	[−4.612, 3.311]	[−3.910, 1.572]	[−6.678, −0.917]
Total enrollment, grades 1–5	−0.4080	−0.8513	−4.3510**
	[−6.191, 5.375]	[−4.848, 3.146]	[−8.133, −0.569]
Total enrollment, grades 6–8	0.0513	−1.4067	−2.6964*
	[−2.892, 2.994]	[−4.083, 1.270]	[−5.815, 0.423]

Note. — This table is exactly analogous to Table A.2.27, except that it reports full RD sample, lights RD sample, and low-deficit-state results for outcomes in Table 1.5.5. We report the corresponding RD figures in A.2.21. Regressions in the first column contain 12,251 village observations (as in Table 1.5.5), while regressions in the second and third columns contain 8,569 and 5,482 village observations, respectively. All regressions control for state fixed effects and the 2005 level of the outcome variable. Standard errors are omitted for brevity, but 95 percent confidence intervals are still presented in brackets. Please refer to the notes below Table A.2.27.

Figure A.2.21: RD Sensitivity – Selected Regressions, Low-Deficit States



Note. — This figure presents RD results for the subset of states with above-average power quality (i.e., below-average deficits). They correspond to the adult male other workers regression in Table A.2.29, and all five village-level enrollment regressions in Table A.2.30. Only the upper-left RD plot for the share of adult men working in "other" job reveals visual evidence of a discontinuity, even though all six RD point estimates are statistically significant. Blue dots show average residuals form regressing the 2011 outcome on state fixed effects and the 2005 level of the outcome (for enrollment variables only). Each dot in the upper-left plot contains approximately 700 villages, averaged in 20-person population bins. For the five enrollment plots, each dot contains approximately 450, averaged in 25-person population bins.

A.2.11 Difference-in-differences results

We estimate a two-period difference-in-differences (DD) model in Section 1.6.3, as an alternative to our RD strategy. While invoking much stronger identifying assumptions (i.e. parallel trends, no time-varying unobservables, as-good-as-random selection into 10th-Plan treatment), this model allows us to include larger villages that are far from our 300-person RD cutoff.[65] Despite estimating treatment effects on a larger sample with a different counterfactuals (comparing 10th- vs. 11th-Plan villages, as opposed to barely eligible 10th-Plan vs. barely ineligible 10th-Plan villages), our DD estimates for nighttime brightness and male agricultural employment are quite comparable to our (preferred) RD estimates (see Figure 1.6.11).

Figure A.2.22 reports the analogous DD results for female agricultural, male other, and female other employment. We see that the DD point estimates are quite close to our RD estimates, except for female other employment.[66] This DD figure also shows relatively constant treatment effects for labor outcomes across the population support, which is consistent with the DD results for male agricultural employment in Figure 1.6.11. By contrast, DD treatment effects for nighttime brightness are positive and increase monotonically in village population.

Together, these results suggest that the small magnitude of our RD labor results is not simply an artifact of restricting our sample to villages smaller than 450 people. Even though the effects of electrification on nighttime brightness are increasing in population, labor effects are constant. This corroborates our RD evidence that electrification does not transform labor markets.

Table A.2.31 compares our preferred RD point estimates to DD estimates estimated using a pooled version of Equation (1.2):

$$Y_{vst} = \gamma_0 + \gamma_1 \mathbf{1}[\text{10th} \times \text{Post}]_{vt} + \delta_t + \eta_v + \varepsilon_{vt} \tag{A.2}$$

Taken at face value, the pooled DD point estimates (in the second column) are fairly consistent with the RD point estimates, except for female other employment. Importantly, for all four labor outcomes, 95 percent confidence intervals reject effects larger than 2.2 percent for both RD and DD models. The third column of Table

[65] Our difference-in-differences analysis also allows us to incorporate multi-habitation villages and villages we could not match to the habitation dataset, since these restrictions are not necessary for identification in the DD context.

[66] Our 300-person DD bins estimate separate treatments for villages with 2001 populations between 0–300 and 300–600. Because our 150–450 RD window spans two bins, the RD estimate lies between these two DD binned estimates. We construct population bins this way in order to allow for heterogeneous DD effects on either side of the 300-person cutoff.

Figure A.2.22: Difference-in-Differences Results

Note. — This figure compares the reduced form effects from our preferred RD specification (Equation (1.1)) to the results from our DD specification (Equation (1.2)), using 300-person population bins. Navy blue dots show the RD coefficients, and whiskers the RD 95 percent confidence interval. Light blue dots and dashed lines show the DD point estimates and 95 percent confidence intervals. From left to right, the three panels show effects for female agricultural, male other, and female other employment. Only the RD results for male employment is statistically significant at the 10 percent level. Table A.2.31 reports pooled DD results. DD regressions include 994,802 village-year observations.

A.2.31 tests the parallel trends assumption by estimating a district-level model on pre-RGGVY employment:

$$Y_{dt} = \zeta_0 + \zeta_1 \mathbf{1}[10\text{th} \times 2001]_{dt} + \delta_t + \eta_d + \varepsilon_{dt}\,, \quad \text{for } t \in \{1991, 2001\} \tag{A.3}$$

The parallel trends assumption necessary to identify Equations (1.2) and (A.2) requires that $\zeta_1 = 0$. However, Table A.2.31 reveals that we can reject parallel trends for three out of four labor outcomes, as $\hat{\zeta}_1$ is statistically different from zero. Hence, we interpret our DD results with caution.

Table A.2.31: RD vs. Difference-in-Differences Results

Outcome Variable	RD Point Estimate	DD Point Estimate	1991–2001 District Pre-Trend
Nighttime brightness	0.1493^{**}	0.4540^{*}	
	(0.0603)	(0.2659)	
	[0.031, 0.268]	[−0.067, 0.975]	
Number of villages	18,686	314,889	
Number of districts	130	307	
Male ag workers / male pop.	-0.0071^{**}	-0.0136^{***}	-0.0112^{**}
	(0.0028)	(0.0042)	(0.0049)
	[−0.013, −0.002]	[−0.022, −0.005]	[−0.021, −0.002]
Female ag workers / female pop.	−0.0049	−0.0069	0.0024
	(0.0040)	(0.0059)	(0.0101)
	[−0.013, 0.003]	[−0.019, 0.005]	[−0.017, 0.022]
Male other workers / male pop.	0.0046^{**}	0.0013	-0.0076^{**}
	(0.0019)	(0.0028)	(0.0033)
	[0.001, 0.008]	[−0.004, 0.007]	[−0.014, −0.001]
Female other workers / female pop.	−0.0004	0.0086^{***}	-0.0130^{***}
	(0.0020)	(0.0028)	(0.0039)
	[−0.004, 0.004]	[0.003, 0.014]	[−0.021, −0.005]
Number of villages	29,765	497,401	
Number of districts	225	499	499

Note. — This table compares our main RD regression results (from estimating Equation (1.1)) to results from a difference-in-differences model. The first column reproduces the RD results from Tables 1.5.2 and 1.5.3. The second column reports $\hat{\gamma}_1$ from separate regressions of Equation (A.2) on each outcome. The third column reports $\hat{\zeta}_1$ from separate district-level regressions of Equation (A.3) on each outcome. (weight districts by their 2001 rural populations). The number of villages and districts differs across nighttime brightness and labor regressions, because our nighttime lights sample includes only states with available shapefiles that correlate with village areas. We report 95 percent confidence intervals in brackets. All standard errors (in parentheses) are clustered at the district level. Significance: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

A.3 Electrification in India: A (More) Detailed History

A.3.1 Before RGGVY

India has a long history of rural electrification programs.[67] Upon Independence in 1947, rural electricity access was virtually nonexistent; by 2012, 92 percent of India's villages were electrified, based on the government's official definition. This dramatic increase is attributable to a series of rural electrification schemes. Each of India's Five Year Plans has funded some sort of rural electrification program, beginning with the 1st Plan (1951–1956) and continuing through the current 12th Plan (2012–2017).

The 1st Plan (1951–1956) focused its electrification efforts on agricultural production and irrigation. A village was legally considered electrified if any electricity was used within its boundaries for any purpose. During these early years, the government's goal was to provide electricity in every 200th village. Ultimately, 4,231 villages were electrified. Under the 2nd Plan (1956–1961), the government's goals shifted towards providing electricity as a "social amenity." This accelerated electrification efforts, bringing 14,458 villages and 350 towns online by the end of the 2nd Plan. The 3rd Plan (1961–1966) motivated electrification as an anti-poverty tactic, with an additional 25,955 villages receiving access to electricity.

With rural electrification becoming increasingly expensive, the All India Rural Credit Review Committee recommended forming a financing agency focused on energy access. At the start of the 4th Plan in 1969, the Rural Electrification Corporation (REC) opened its doors. As a Public Sector Undertaking with a significant degree of fiscal autonomy, the REC funded rural electrification with the joint goals of reducing poverty and promoting productive activity. During that period, India's Green Revolution was increasing the economic returns to electricity in rural areas: electrified pumps enabled the irrigation systems necessary to support new high yield variety grains. The REC had a mandate to promote electrified pumpsets, and it targeted villages with populations of at least 5,000. In 1974, the beginning of the 5th Plan, the Minimum Needs Programme was begun in order to improve standards of living and provide for basic needs. This scheme targeted states with village electrification rates below than the national average and subsidized short distance connections between villages and the existing grid. Broader access to electricity, beyond for agriculture alone, was a key component of this legislation. As a result, over 200,000 villages gained access to electricity between 1969 and 1979.

[67] The information from this section comes from a combination of Rural Electrification Corporation (2010) and Banerjee et al. (2014).

Between 1980 and 1990, the 6th and 7th Plans funded a variety of schemes to promote access to electricity, including the Integrated Rural Energy Program and Kutir Jyoti Yojana.[68] These programs had strong distributional motivations, and were designed to decrease energy poverty amongst India's poorest households. The 1980s saw 237,371 newly electrified villages.

Under the 8th Plan (1992–1997), the government created both the Ministry of Power and the Ministry of New and Renewable Energy. However, funding challenges forced rural electrification efforts to slow dramatically, with only 11,666 villages electrified in this five year period. The 9th Plan (1997–2002) once again sought to promote electrification as an economic development program. In keeping with this goal, the Ministry of Power released a new definition of electrification in 1997. Villages were now only considered electrified if electricity was being used in the *inhabited* areas of the village. This exemplified the shift away from electrification solely for agriculture's sake. At the same time, new government programs, including Pradhan Mantri Gramodaya Yojana, began providing subsidies for electricity services. New provisions in the Minimum Needs Programme, as well as the launch of the Accelerated Rural Electrification Program, helped to provide individuals and states with the financing necessary to increase rural energy access. During the 9th Plan, 13,317 villages were electrified. By 2001, 86 percent of all villages were deemed electrified.

The 10th Plan (2002–2007) spanned several major changes to India's rural electrification efforts. The Electricity Act, 2003, codified the government's commitment to rural electrification, stating that "The Central Government shall ... formulate a national policy...for rural electrification and for bulk purchase of power and management of local distribution in rural areas." The Act also requires the government to "endeavour to supply electricity to all areas including villages and hamlets" (Ministry of Law and Justice (2003)). In 2004, the Ministry of Power created a new, stricter definition of electrification, which is still in use today. A village is now officially considered electrified only if basic infrastructure, including transformers and distribution lines exist in that village and in its constituent habitations; if public locations such as schools, government offices, health centers, and others have electricity; and if at least 10 percent of the village's households are electrified.

In 2004's National Electricity Policy, the Ministry of Power laid the groundwork for the future of rural electrification in India. Invoking the 2003 Electricity Act, the National Electricity Policy states that "The key development objective of the power sector is supply of electricity to all areas including rural areas...governments would jointly endeavour to achieve this objective at the earliest" (Ministry of Power

[68]Kutir Jyoti Yojana provided 100 percent subsidies for single point connections to BPL households.

(2005a)). In particular, subsequent rural electrification programs are supposed to create a "Rural Electrification Distribution Backbone" (REDB) of at a minimum one 33/11 kV or 66/11 kV substation in each Census block, with higher-load regions supplied with additional substations. These substations are to be connected to the state transmission grid. In addition, each village should have supply feeders and at least one distribution transformer, such that every household may be connected on demand to the grid via that transformer. Every household should be connected on demand to the village's transformer.[69] In keeping with the theme of electrification as a tool for development, the Policy requires that electricity infrastructure be able to support the load from agriculture, textiles and other industries, small and medium enterprises, cold-chain (refrigeration) services, and other public services such as health and education. The Policy also stipulates that priority be given to electrification in "economically backwards" regions. Finally, the Policy makes the REC (now a division of the Ministry of Power) the nodal agency in charge of implementing the country's rural electrification goals.

A.3.2 RGGVY

In 2005, the REC initiated Rajiv Gandhi Grameen Vidyutikaran Yojana (RGGVY), the Prime Minister's Rural Electrification Plan. RGGVY was the flagship Indian rural electrification program, created with the National Common Minimum Programme goal of universal electricity access in mind.[70] Upon launch, RGGVY enveloped the remaining ongoing electrification schemes, including the Accelerated Electrification of One Lakh Villages and One Crore Households and the Minimum Needs Programme.[71] The RGGVY scheme was detailed in a Ministry of Power Office Memorandum from March of 2005 (Ministry of Power (2005b)), and its rules followed directly from the National Electricity Policy.

Under RGGVY, states were required to "make adequate arrangements for supply of electricity," and to serve rural and urban customers equally. As stipulated by the National Electricity Policy, RGGVY was mandated to create the Rural Electricity Distribution Backbone, to electrify unelectrified habitations and villages, and to provide adequate distribution infrastructure in these newly electrified areas. This in-

[69] Rural Indian households typically pay for their own electricity connections, unless they are specifically subsidized through an electrification program.

[70] The program was designed to cover the entire country, but one state (Goa) and all of the union territories have been left out, since they had already achieved 100 percent village electrification by 2005.

[71] These other schemes had, by this point, been discontinued for financial reasons. In the South Asian numbering system, 1 lakh = 100,000 and 1 crore = 10,000,000.

frastructure was supposed to be able to support household load, as well as load from irrigation pumpsets, various industrial activity, cold chains, health care, education, and information technology. A small Decentralized Distributed Generation (DDG) provision was put in place for villages where grid connection would be infeasible or prohibitively expensive. RGGVY was specifically intended to "facilitate overall rural development, employment generation and poverty alleviation," and the policy specifically supported the poor by providing 100 percent subsidies for grid connections for below-poverty line (BPL) households.[72] As part of India's national anti-corruption efforts, details of proposed and completed electrification under RGGVY are available online.[73]

The Rural Electrification Corporation, serving as a nodal agency to the Ministry of Power, has been the main implementing agency for RGGVY, providing 90 percent of the capital needs as direct grants to states and loaning them the remaining 10 percent. The REC was responsible for providing detailed program guidelines, including requirements and standards for materials, equipment, and construction. Within each state that was eligible for RGGVY, the state government power utilities designated implementing agencies. These implementing agencies could be state power distribution companies, state electricity boards, state government power departments, central power sector undertakings (appointed by the state government), or co-operative societies.[74] These implementing agencies prepared Detailed Project Reports (DPRs) for each district under their jurisdiction, by carrying out surveys in every village. A DPR listed each village's electrification status, population, number of households (above/below the poverty line, and with/without electricity), and number of public places (with/without electricity). These reports proposed village-by-village RGGVY implementation plans for eligible villages, which included details on new electricity infrastructure and household connections to be installed (Ministry of Power (2014b)). These DPRs were then submitted to the REC after approval from the state government. After the REC conducted a comprehensive review of each DPR, it passed them on to the Ministry of Power for final approval. Once the

[72]Note that this does not include free power - RGGVY only provides free connections to BPL households. Above-poverty line households pay their own connection charges.

[73]The new program website with these details is `http://www.ddugjy.in/`.

[74]The choice of implementing agency was left to the states, and depended on the administrative structure and relative capacity amongst different state agencies. Importantly, these were not local agencies: the REC specifically prohibited Gram Panchayats (local governments) from implementing RGGVY projects.

Ministry of Power approved a DPR, it sanctioned that district's RGGVY project, and the REC disbursed funds to the implementing agency in charge of the project.[75]

Under the 10th Plan, RGGVY limited program eligibility to villages with constituent habitations of 300 people or above. It justified this population cutoff on the grounds of keeping program costs low. In all, the REC reports that RGGVY electrified 64,091 villages under the 10th Plan. This number included both "unelectrified" villages that did not meet the 2004 definition of electrification, as well as "de-electrified" that had previously been deemed electrified but no longer meet the official definition of electrification.[76]

In 2008, the Ministry of Power ordered the continuation of RGGVY in the 11th plan (2007–2012).[77] This second wave of RGGVY continued to target electrification for all, with the goal electrifying 115,000 un-electrified villages and providing free connections to 23.4 million BPL households. The REC continued as the nodal agency, with the same the 90/10 percent subsidy/loan capital split. Under the 11th Plan, RGGVY provided the same infrastructure as under the 10th Plan. However, states were now required to guarantee a minimum 6–8 hours of power supply for RGGVY villages before the REC would approved a given DPR proposal. The 11th Plan also included guidelines for electrifying villages where grid extensions would be cost-prohibitive with microgrids, under the small Decentralized Distributed Generation (DDG) carve-out.[78]

A three-tier quality monitoring mechanism was also put in place with the 11th plan. First, the state or sub-state level implementing agencies would conduct inspections to ensure that workmanship on RGGVY projects was up to standards. These agencies were mandated to randomly inspect 50 percent of RGGVY villages. Next, the REC was instructed to inspect materials before shipment to RGGVY sites, as well as a random subsample 10 percent of villages. Finally, the Ministry of Power hired National Quality Monitors to inspect 1 percent of villages. Amid complaints from state and local governments, the 300-person constraint was relaxed in the 11th plan. All villages with habitations of 100 people or more were supposed to be covered in RGGVY during this period. Between 2007 and 2012, the REC reports that an additional 46,206 villages were electrified.

[75] All implementing agencies were required to bring their own teams to villages for RGGVY electrification; no hiring of local labor was permitted.

[76] These "de-electrified" village either had access to electricity and then lost this access due to infrastructure breakdown, or moved out of official electrified status when the definition became more stringent in 2004.

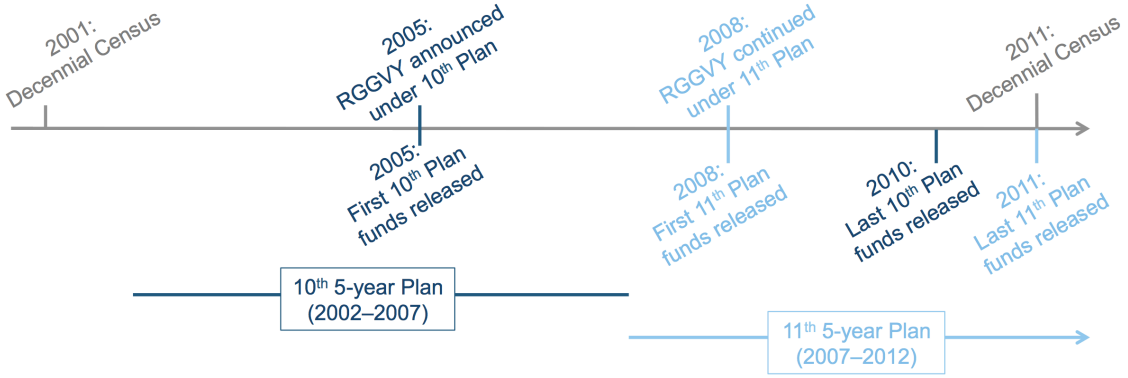
[77] A new Office Memorandum contains the details (Ministry of Power (2008)).

[78] For more details on this DDG carve-out, see: `http://powermin.nic.in/sites/default/files/uploads/Guidelines_for_Village_Electrification_DDG_under_RGGVY_0.pdf`

India is currently in its 12th Five-Year Plan (2012-2017). In September 2013, the Ministry of Power again extended RGGVY, this time to be covered under both the 12th and 13th Plans (forthcoming in 2017–2022).[79] This time, the scheme was sanctioned to complete unfinished projects from the 10th and 11th Plans and to cover all remaining habitations and villages with populations above 100. The third wave of RGGVY also continues to subsidize BPL connections. As of 2012, 92 percent of villages in India were officially classified as electrified.

Rural electrification work continues in India. In 2014, RGGVY was subsumed into a new program, Deendayal Upadhyaya Gram Jyoti Yojana (DDUGJY), the Deendayal Upadhyaya Village Light Plan.[80] This program is slated to carry out RGGVY's works as under the 2013 continuation document. Under DDUGJY, however, all villages and habitations are now eligible for rural electrification, regardless of population (Ministry of Power (2015)). DDUGJY also provides for the creation of new feeder lines to separate agricultural and non-agricultural consumers. It aims to strengthen sub-transmission and distribution infrastructure in the rural areas, with a particular focus on metering.

Figure A.3.1: RGGVY Implementation Timeline



Note. — This figure shows the timing of the Indian decennial census, the 10th and 11th Five-Year Plans, and RGGVY under these Plans. Our RD analysis uses data from the 2001 and 2011 Census years, described in detail in Section 1.4.

[79] Once again, an Office Memorandum marked the occasion (Ministry of Power (2013)).

[80] The Ministry of Power announced this decision in an Office Memorandum in December (Ministry of Power (2014a)).

Appendix B

Panel Data and Experimental Design – Appendix

B.1 Derivations and proofs

We are the first to derive analytic power calculation formulas for panel data models under non-i.i.d. error structures. In this section, we derive power calculation analytics for cross-sectional, difference-in-differences, and collapsed data estimators. We present the resulting equations in Sections 3.2 and 3.3 of the main text.

We first re-derive well-known power calculation formulas, for both a cross-sectional experiment and a panel experiment that applies the difference-in-differences estimator, under the assumption that error terms are uncorrelated. We then relax this assumption to consider the difference-in-differences estimator applied to a panel experiment in which the error structure of the data exhibits an arbitrary form of serial correlation. We show that the previously reported power calculation formula is incorrect in this case, and we derive the first power calculation formula that correctly incorporates arbitrary serial correlation in a panel data setting. We then consider a collapsed data research design, and we again show that the previously reported power calculation formula is incorrect in the presence of serial correlation, whereas our new power calculation formula gives the correct analytic results.

We then provide the proofs to the lemmas presented in the main text. These lemmas show that the above power calculation formulas can be applied even in the presence of cross-sectional correlations, so long as treatment is randomly assigned at the unit level. In particular, Lemma 2 states that the variance estimator we derive, which accounts for serial correlation, gives an unbiased estimate of the true variance under unit-level randomization, even when cross-sectional correlations exist.

In other words, our newly derived power calculation formula can be applied to any panel experiment setting, regardless of the true error structure of the data, so long as treatment is randomly assigned to units.

B.1.1 Cross section

B.1.1.1 Independent error structure

Model There are J units randomly assigned a treatment status D_i, with proportion P in treatment ($D_i = 1$) and proportion $(1 - P)$ in control ($D_i = 0$). The units are indexed so $i \in [1, PJ]$ is treated and $j \in [PJ + 1, J]$ is a control. We make standard assumptions for randomized trials:

Assumption 11 (Data generating process)**.** *The data are generated according to the following model:*

$$Y_i = \beta + \tau D_i + \varepsilon_i \tag{B.1}$$

where Y_i is the outcome of interest for unit i; β is the expected outcome of non-treated units; τ is the treatment effect which is homogenous across all units; D_i is a treatment indicator; and ε_i is an idiosyncratic error term distributed i.i.d. $\mathcal{N}(0, \sigma_\varepsilon^2)$.

Assumption 12 (Strict exogeneity)**.** $\mathrm{E}[\varepsilon_i \mid \mathbf{X}] = 0$*, where* $\mathbf{X} = [\beta \ D]$*. In practice, this follows from random assignment of D_i.*

Coefficient estimate The coefficient estimates from an OLS regression are

$$\begin{aligned}
\begin{pmatrix} \widehat{\beta} \\ \widehat{\tau} \end{pmatrix} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'Y \\
&= \begin{pmatrix} J & PJ \\ PJ & PJ \end{pmatrix}^{-1} \begin{pmatrix} \sum_{i=1}^{J} Y_i \\ \sum_{i=1}^{J} D_i Y_i \end{pmatrix} \\
&= \frac{1}{P(1-P)J^2} \begin{pmatrix} PJ\left(\sum_{i=1}^{J} Y_i - \sum_{i=1}^{J} D_i Y_i\right) \\ J\left(\sum_{i=1}^{J} D_i Y_i - P\sum_{i=1}^{J} Y_i\right) \end{pmatrix} \\
&= \begin{pmatrix} \frac{1}{(1-P)J}\sum_{i=1}^{J}(1 - D_i)Y_i \\ \frac{1}{PJ}\sum_{i=1}^{J} D_i Y_i - \frac{1}{(1-P)J}\sum_{i=1}^{J}(1 - D_i)Y_i \end{pmatrix} \\
&= \begin{pmatrix} \frac{1}{(1-P)J}\sum_{i=PJ+1}^{J} Y_i \\ \frac{1}{PJ}\sum_{i=1}^{PJ} Y_i - \frac{1}{(1-P)J}\sum_{i=PJ+1}^{J} Y_i \end{pmatrix}
\end{aligned}$$

Defining the mean outcome in the treatment and control groups, respectively, as

$$\overline{Y}_T = \frac{1}{PJ}\sum_{i=1}^{PJ} Y_i$$

$$\overline{Y}_C = \frac{1}{(1-P)J} \sum_{i=PJ+1}^{J} Y_i$$

gives coefficient estimates of

$$\begin{aligned}\widehat{\beta} &= \overline{Y}_C \\ \widehat{\tau} &= \overline{Y}_T - \overline{Y}_C\end{aligned}$$

Variance of coefficient estimate The variance of the estimate of the treatment effect, $\widehat{\tau}$, is

$$\begin{aligned}\operatorname{Var}\left(\widehat{\tau} \mid \mathbf{X}\right) &= \operatorname{Var}\left(\overline{Y}_T \mid \mathbf{X}\right) + \operatorname{Var}\left(\overline{Y}_C \mid \mathbf{X}\right) - 2\operatorname{Cov}\left(\overline{Y}_T, \overline{Y}_C \mid \mathbf{X}\right) \\ &= \operatorname{Var}\left(\overline{Y}_T \mid \mathbf{X}\right) + \operatorname{Var}\left(\overline{Y}_C \mid \mathbf{X}\right) \qquad \text{(B.2)}\end{aligned}$$

The first term of Equation (B.2) is

$$\begin{aligned}\operatorname{Var}\left(\overline{Y}_T \mid \mathbf{X}\right) &= \operatorname{Var}\left(\frac{1}{PJ}\sum_{i=1}^{PJ} Y_i \mid \mathbf{X}\right) \\ &= \frac{1}{(PJ)^2}\sum_{i=1}^{PJ} \operatorname{Var}(Y_i \mid \mathbf{X}) \\ &= \frac{\sigma_\varepsilon^2}{PJ} \qquad \text{(B.3)}\end{aligned}$$

Similarly, the second term of Equation (B.2) is

$$\operatorname{Var}\left(\overline{Y}_C \mid \mathbf{X}\right) = \frac{\sigma_\varepsilon^2}{(1-P)J} \qquad \text{(B.4)}$$

Substituting Equations (B.3) and (B.4) into Equation (B.2) gives

$$\begin{aligned}\operatorname{Var}(\widehat{\tau} \mid \mathbf{X}) &= \frac{\sigma_\varepsilon^2}{PJ} + \frac{\sigma_\varepsilon^2}{(1-P)J} \\ &= \frac{\sigma_\varepsilon^2}{P(1-P)J} \qquad \text{(B.5)}\end{aligned}$$

This is equal to the variance estimator produced by an OLS regression of Equation (B.1).

Minimum detectable effect The minimum detectable effect (MDE), or the smallest treatment effect we have the power to detect, is

$$MDE = \left(t_{1-\kappa}^{J-2} + t_{\alpha/2}^{J-2}\right)\sqrt{\operatorname{Var}\left(\widehat{\tau} \mid \mathbf{X}\right)}$$

$$= \left(t_{1-\kappa}^{J-2} + t_{\alpha/2}^{J-2}\right)\sqrt{\frac{\sigma_\varepsilon^2}{P(1-P)J}} \tag{B.6}$$

where κ is the power of the hypothesis test, α is the significance level, and the critical values are drawn from t-distributions with $J-2$ degrees of freedom. We present this well-known result as Equation (3.2) in the main text.

B.1.2 Difference-in-differences

B.1.2.1 Independent error structure

Model 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 13 (Data generating process)**.** *The data are generated according to the following model:*

$$\begin{aligned} Y_{it} &= \beta + \tau D_{it} + \varepsilon_{it} \\ &= \beta + \tau D_{it} + \upsilon_i + \delta_t + \omega_{it} \end{aligned} \tag{B.7}$$

where Y_{it} is the outcome of interest for unit i at time t; β is the expected outcome of non-treated observations; τ is the treatment effect that is homogenous across all units and all time periods; D_{it} is a time-varying treatment indicator; υ_i is a time-invariant unit effect distributed i.i.d. $\mathcal{N}(0, \sigma_\upsilon^2)$; δ_t is a common time effect distributed i.i.d. $\mathcal{N}(0, \sigma_\delta^2)$; and ω_{it} is an idiosyncratic error term distributed i.i.d. $\mathcal{N}(0, \sigma_\omega^2)$.

Assumption 14 (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 D, $J-1$ unit fixed effects, and $(m + r) - 1$ time fixed effects. This again follows from random assignment of D_{it}.*

Assumption 15 (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.*

Coefficient estimate The treatment effect, τ, can be estimated by OLS with unit and time fixed effects. In a balanced panel, this is equivalent to de-meaning at both the unit and time levels. Define

$$\ddot{Y}_{it} = Y_{it} - \overline{Y}_{i} - \overline{Y}_{t} + \overline{\overline{Y}} \tag{B.8}$$

$$\ddot{D}_{it} = D_{it} - \overline{D}_i - \overline{D}_t + \overline{\overline{D}} \tag{B.9}$$

$$\ddot{\omega}_{it} = \omega_{it} - \overline{\omega}_i - \overline{\omega}_t + \overline{\overline{\omega}} \tag{B.10}$$

where

$$\overline{Y}_i = \frac{1}{m+r} \sum_{t=-m+1}^{r} Y_{it}$$

$$\overline{Y}_t = \frac{1}{J} \sum_{i=1}^{J} Y_{it}$$

$$\overline{\overline{Y}} = \frac{1}{J(m+r)} \sum_{t=-m+1}^{r} \sum_{i=1}^{J} Y_{it}$$

with $\overline{D}_i$, $\overline{D}_t$, $\overline{\overline{D}}$, $\overline{\omega}_i$, $\overline{\omega}_t$, and $\overline{\overline{\omega}}$ defined analogously. Substituting Equations (B.8)–(B.10) into Equation (B.7) and simplifying gives the de-meaned DGP,

$$\ddot{Y}_{it} = \tau \ddot{D}_{it} + \ddot{\omega}_{it}$$

The estimate of the treatment effect is

$$\begin{aligned}
\widehat{\tau} &= (\ddot{D}'\ddot{D})^{-1}\ddot{D}'\ddot{Y} \\
&= \left(\sum_{i=1}^{J} \sum_{t=-m+1}^{r} \ddot{D}_{it}^2 \right)^{-1} \sum_{i=1}^{J} \sum_{t=-m+1}^{r} \ddot{D}_{it}\ddot{Y}_{it} \\
&= \frac{m+r}{P(1-P)Jmr} \left[\sum_{i=1}^{J} \sum_{t=-m+1}^{r} \ddot{D}_{it} Y_{it} - \sum_{i=1}^{J} \overline{Y}_i \sum_{t=-m+1}^{r} \ddot{D}_{it} \right. \\
&\qquad \left. - \sum_{t=-m+1}^{r} \overline{Y}_t \sum_{i=1}^{J} \ddot{D}_{it} + \overline{Y} \sum_{i=1}^{J} \sum_{t=-m+1}^{r} \ddot{D}_{it} \right] \\
&= \frac{m+r}{P(1-P)Jmr} \sum_{i=1}^{J} \sum_{t=-m+1}^{r} \ddot{D}_{it} Y_{it} \\
&= \frac{m+r}{P(1-P)Jmr} \left[\sum_{i=1}^{PJ} \left(\frac{-(1-P)r}{m+r} \sum_{t=-m+1}^{0} Y_{it} + \frac{(1-P)m}{m+r} \sum_{t=1}^{r} Y_{it} \right) \right. \\
&\qquad \left. + \sum_{i=PJ+1}^{J} \left(\frac{Pr}{m+r} \sum_{t=-m+1}^{0} Y_{it} + \frac{-Pm}{m+r} \sum_{t=1}^{r} Y_{it} \right) \right] \\
&= \frac{1}{PJ} \sum_{i=1}^{PJ} \left[\frac{-1}{m} \sum_{t=-m+1}^{0} Y_{it} + \frac{1}{r} \sum_{t=1}^{r} Y_{it} \right] \\
&- \frac{1}{(1-P)J} \sum_{i=PJ+1}^{J} \left[\frac{-1}{m} \sum_{t=-m+1}^{0} Y_{it} + \frac{1}{r} \sum_{t=1}^{r} Y_{it} \right]
\end{aligned}$$

$$= \left(\overline{Y}_T^A - \overline{Y}_T^B\right) - \left(\overline{Y}_C^A - \overline{Y}_C^B\right)$$

where

$$\overline{Y}_T^A = \frac{1}{PJr} \sum_{i=1}^{PJ} \sum_{t=1}^{r} Y_{it}$$

$$\overline{Y}_T^B = \frac{1}{PJm} \sum_{i=1}^{PJ} \sum_{t=-m+1}^{0} Y_{it}$$

$$\overline{Y}_C^A = \frac{1}{(1-P)Jr} \sum_{i=PJ+1}^{J} \sum_{t=1}^{r} Y_{it}$$

$$\overline{Y}_C^B = \frac{1}{(1-P)Jm} \sum_{i=PJ+1}^{J} \sum_{t=-m+1}^{0} Y_{it}$$

Variance of coefficient estimate The variance of the estimate of the treatment effect, $\widehat{\tau}$, is

$$\begin{aligned}
\operatorname{Var}(\widehat{\tau} \mid \mathbf{X}) &= \operatorname{Var}\left(\overline{Y}_T^A \mid \mathbf{X}\right) + \operatorname{Var}\left(\overline{Y}_T^B \mid \mathbf{X}\right) + \operatorname{Var}\left(\overline{Y}_C^A \mid \mathbf{X}\right) + \operatorname{Var}\left(\overline{Y}_C^B \mid \mathbf{X}\right) \\
&\quad - 2\operatorname{Cov}\left(\overline{Y}_T^A, \overline{Y}_T^B \mid \mathbf{X}\right) - 2\operatorname{Cov}\left(\overline{Y}_T^A, \overline{Y}_C^A \mid \mathbf{X}\right) + 2\operatorname{Cov}\left(\overline{Y}_T^A, \overline{Y}_C^B \mid \mathbf{X}\right) \\
&\quad + 2\operatorname{Cov}\left(\overline{Y}_T^B, \overline{Y}_C^A \mid \mathbf{X}\right) - 2\operatorname{Cov}\left(\overline{Y}_T^B, \overline{Y}_C^B \mid \mathbf{X}\right) - 2\operatorname{Cov}\left(\overline{Y}_C^A, \overline{Y}_C^B \mid \mathbf{X}\right) \\
&= \operatorname{Var}\left(\overline{Y}_T^A \mid \mathbf{X}\right) + \operatorname{Var}\left(\overline{Y}_T^B \mid \mathbf{X}\right) + \operatorname{Var}\left(\overline{Y}_C^A \mid \mathbf{X}\right) + \operatorname{Var}\left(\overline{Y}_C^B \mid \mathbf{X}\right) \\
&\quad - 2\Bigg[\operatorname{Cov}\left(\overline{Y}_T^A, \overline{Y}_T^B \mid \mathbf{X}\right) + \operatorname{Cov}\left(\overline{Y}_T^A, \overline{Y}_C^A \mid \mathbf{X}\right) \\
&\quad + \operatorname{Cov}\left(\overline{Y}_T^B, \overline{Y}_C^B \mid \mathbf{X}\right) + \operatorname{Cov}\left(\overline{Y}_C^A, \overline{Y}_C^B \mid \mathbf{X}\right)\Bigg]
\end{aligned} \tag{B.11}$$

The first term of Equation (B.11) is

$$\begin{aligned}
\operatorname{Var}\left(\overline{Y}_T^A \mid \mathbf{X}\right) &= \operatorname{Var}\left(\frac{1}{PJr} \sum_{i=1}^{PJ} \sum_{t=1}^{r} Y_{it}\right) \\
&= \frac{1}{(PJr)^2} \operatorname{Var}\left(\sum_{i=1}^{PJ} \sum_{t=1}^{r} Y_{it} \mid \mathbf{X}\right) \\
&= \frac{1}{PJr} \left[\operatorname{Var}(Y_{it} \mid \mathbf{X})\right. \qquad \text{(B.12)} \\
&\quad \left. + (r-1)\operatorname{Cov}(Y_{it}, Y_{is} \mid \mathbf{X}) + (PJ-1)\operatorname{Cov}(Y_{it}, Y_{jt} \mid \mathbf{X})\right] \\
&= \frac{1}{PJr}\left(r\sigma_v^2 + PJ\sigma_\delta^2 + \sigma_\omega^2\right) \qquad \text{(B.13)}
\end{aligned}$$

Similarly, the remaining variance terms of Equation (B.11) are

$$\operatorname{Var}\left(\overline{Y}_T^B \mid \mathbf{X}\right) = \frac{1}{PJm}\left(m\sigma_v^2 + PJ\sigma_\delta^2 + \sigma_\omega^2\right) \tag{B.14}$$

$$\text{Var}\left(\overline{Y}_C^A \mid \mathbf{X}\right) = \frac{1}{(1-P)Jr}\left(r\sigma_\upsilon^2 + (1-P)J\sigma_\delta^2 + \sigma_\omega^2\right) \tag{B.15}$$

$$\text{Var}\left(\overline{Y}_C^B \mid \mathbf{X}\right) = \frac{1}{(1-P)Jm}\left(m\sigma_\upsilon^2 + (1-P)J\sigma_\delta^2 + \sigma_\omega^2\right) \tag{B.16}$$

The first covariance component of Equation (B.11) is

$$\begin{aligned}
\text{Cov}\left(\overline{Y}_T^A, \overline{Y}_T^B \mid \mathbf{X}\right) &= \text{E}\left[\overline{Y}_T^A \overline{Y}_T^B \mid \mathbf{X}\right] - \text{E}\left[\overline{Y}_T^A \mid \mathbf{X}\right]\text{E}\left[\overline{Y}_T^B \mid \mathbf{X}\right] \\
&= \text{E}\left[\left(\frac{1}{PJr}\sum_{i=1}^{PJ}\sum_{t=1}^{r}(\beta + \tau + \upsilon_i + \delta_t + \omega_{it})\right)\right. \\
&\qquad \left.\times\left(\frac{1}{PJm}\sum_{i=1}^{PJ}\sum_{t=-m+1}^{0}(\beta + \upsilon_i + \delta_t + \omega_{it})\right) \mid \mathbf{X}\right] \\
&\quad - \text{E}\left[\frac{1}{PJr}\sum_{i=1}^{PJ}\sum_{t=1}^{r}(\beta + \tau + \upsilon_i + \delta_t + \omega_{it}) \mid \mathbf{X}\right] \\
&\qquad \times \text{E}\left[\frac{1}{PJm}\sum_{i=1}^{PJ}\sum_{t=-m+1}^{0}(\beta + \upsilon_i + \delta_t + \ \omega_{it}) \mid \mathbf{X}\right] \\
&= \left[\beta(\beta+\tau) + \frac{1}{PJ}\text{E}\left[\upsilon_i^2 \mid \mathbf{X}\right]\right] - \beta(\beta+\tau) \\
&= \frac{1}{PJ}\left(\text{Var}(\upsilon_i) - \text{E}\left[\upsilon_i \mid \mathbf{X}\right]^2\right) \\
&= \frac{\sigma_\upsilon^2}{PJ} \qquad \text{(B.17)}
\end{aligned}$$

Similarly, the remaining covariance terms of Equation (B.11) are

$$\text{Cov}\left(\overline{Y}_T^A, \overline{Y}_C^A \mid \mathbf{X}\right) = \frac{\sigma_\delta^2}{r} \tag{B.18}$$

$$\text{Cov}\left(\overline{Y}_T^B, \overline{Y}_C^B \mid \mathbf{X}\right) = \frac{\sigma_\delta^2}{m} \tag{B.19}$$

$$\text{Cov}\left(\overline{Y}_C^A, \overline{Y}_C^B \mid \mathbf{X}\right) = \frac{\sigma_\upsilon^2}{(1-P)J} \tag{B.20}$$

Substituting Equations (B.13)–(B.20) into Equation (B.11) gives

$$\begin{aligned}
\text{Var}\left(\widehat{\tau} \mid \mathbf{X}\right) &= 2\frac{\sigma_\upsilon^2}{P(1-P)J} + 2\frac{(m+r)\sigma_\delta^2}{mr} + \frac{(m+r)\sigma_\omega^2}{P(1-P)Jmr} \\
&\quad - 2\left[\frac{\sigma_\upsilon^2}{P(1-P)J} + \frac{(m+r)\sigma_\delta^2}{mr}\right] \qquad \text{(B.21)} \\
&= \left(\frac{\sigma_\omega^2}{P(1-P)J}\right)\left(\frac{m+r}{mr}\right) \qquad \text{(B.22)}
\end{aligned}$$

This is equal to the variance estimator produced by an OLS regression of Equation (B.7), except that it is scaled by σ_ω^2 instead of σ_ε^2.

Minimum detectable effect The MDE is

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

This is the standard Frison and Pocock (1992) result, also referenced by McKenzie (2012), and is shown as Equation (3.3) in the main text. We assume J degrees of freedom to be consistent with the assumptions of the CRVE; alternatively, $J(m+r)-(J+m+r)$ degrees of freedom would be consistent with the assumptions of OLS standard errors. Note that this model is, in fact, an extension of the result in these papers, as both Frison and Pocock (1992) and McKenzie (2012) assume $\sigma^2_\delta = 0$.

B.1.2.2 Serially correlated error structure

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.

Assumption 16 (Data generating process)**.** *The data are generated according to the following model:*

$$Y_{it} = \beta + \tau D_{it} + \upsilon_i + \delta_t + \omega_{it} \tag{B.23}$$

where Y_{it} is the outcome of interest for unit i at time t; β is the expected outcome of non-treated observations; τ is the treatment effect that is homogenous across all units and all time periods; D_{it} is a time-varying treatment indicator; υ_i is a unit-specific disturbance distributed i.i.d. $\mathcal{N}(0, \sigma^2_\upsilon)$; δ_t is a time-specific disturbance distributed i.i.d. $\mathcal{N}(0, \sigma^2_\delta)$; and ω_{it} is an idiosyncratic error term distributed (not necessarily i.i.d.) $\mathcal{N}(0, \sigma^2_\omega)$.

Assumption 17 (Strict exogeneity)**.** $\text{E}[\omega_{it} \mid \mathbf{X}] = 0$*, where* $\mathbf{X}$ *is a full rank matrix of regressors, including a constant, the treatment indicator D, $J-1$ unit fixed effects, and $(m+r)-1$ time fixed effects. This again follows from random assignment of D_{it}.*

Assumption 18 (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 19 (Independence across units)**.** $\mathrm{E}[\omega_{it}\omega_{js} \mid \mathbf{X}] = 0,\ \forall\ i \neq j,\ \forall\ t, s.$

Assumption 20 (Symmetric covariance structures)**.** *Define:*

$$\psi_T^B \equiv \frac{2}{PJm(m-1)} \sum_{i=1}^{PJ} \sum_{t=-m+1}^{-1} \sum_{s=t+1}^{0} \mathrm{Cov}\left(\omega_{it}, \omega_{is} \mid \mathbf{X}\right)$$

$$\psi_T^A \equiv \frac{2}{PJr(r-1)} \sum_{i=1}^{PJ} \sum_{t=1}^{r-1} \sum_{s=t+1}^{r} \mathrm{Cov}\left(\omega_{it}, \omega_{is} \mid \mathbf{X}\right)$$

$$\psi_T^X \equiv \frac{1}{PJmr} \sum_{i=1}^{PJ} \sum_{t=-m+1}^{0} \sum_{s=1}^{r} \mathrm{Cov}\left(\omega_{it}, \omega_{is} \mid \mathbf{X}\right)$$

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

Coefficient estimate The deterministic portion of this model is the same as that in Equation (B.7), so the estimate of the treatment effect is again

$$\widehat{\tau} = \left(\overline{Y}_T^A - \overline{Y}_T^B\right) - \left(\overline{Y}_C^A - \overline{Y}_C^B\right)$$

Variance of coefficient estimate With a serially correlated error structure, Equation (B.11) is still correct because of independence between observations that do not correspond to the same unit, but Equations (B.13)–(B.20) no longer hold. With serially correlated errors, the first term of Equation (B.11) is

$$\begin{aligned}
\mathrm{Var}\left(\overline{Y}_T^A \mid \mathbf{X}\right) &= \mathrm{Var}\left(\frac{1}{PJr} \sum_{i=1}^{PJ} \sum_{t=1}^{r} Y_{it} \mid \mathbf{X}\right) \\
&= \frac{1}{(PJr)^2} \mathrm{Var}\left(\sum_{i=1}^{PJ} \sum_{t=1}^{r} Y_{it} \mid \mathbf{X}\right) \\
&= \frac{1}{(PJr)^2} \left[PJr\, \mathrm{Var}(Y_{it} \mid \mathbf{X}) + (PJ-1)r\, \mathrm{Cov}(Y_{it}, Y_{jt} \mid \mathbf{X})\right. \qquad \text{(B.24)}
\end{aligned}$$

[1] 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, $\mathrm{E}[\psi_T^B] = \mathrm{E}[\psi_C^B]$, $\mathrm{E}[\psi_T^A] = \mathrm{E}[\psi_C^A]$, and $\mathrm{E}[\psi_T^X] = \mathrm{E}[\psi_C^X]$, making this a reasonable assumption *ex ante*. However, it is possible for treatment to alter the covariance structure of treated units only.

$$+2\sum_{i=1}^{PJ}\sum_{t=1}^{r-1}\sum_{s=t+1}^{r}\operatorname{Cov}(Y_{it},Y_{is}\mid\mathbf{X})\Bigg]$$

$$=\frac{\sigma_\varepsilon^2}{PJr}+\frac{(PJ-1)\sigma_\delta^2}{(PJ)^2r}+\frac{2}{(PJr)^2}\sum_{i=1}^{PJ}\sum_{t=1}^{r-1}\sum_{s=t+1}^{r}\left(\sigma_\upsilon^2+\operatorname{Cov}(\omega_{it},\omega_{is}\mid\mathbf{X})\right)$$

$$=\frac{1}{PJr}\left(r\sigma_\upsilon^2+PJ\sigma_\delta^2+\sigma_\omega^2\right)+\frac{2}{(PJr)^2}\sum_{i=1}^{PJ}\sum_{t=1}^{r-1}\sum_{s=t+1}^{r}\operatorname{Cov}(\omega_{it},\omega_{is}\mid\mathbf{X})$$

$$=\frac{1}{PJr}\left(r\sigma_\upsilon^2+PJ\sigma_\delta^2+\sigma_\omega^2+(r-1)\psi_T^A\right) \tag{B.25}$$

Similarly, with serial correlation, the remaining variance terms of Equation (B.11) are

$$\operatorname{Var}\left(\overline{Y}_T^B\mid\mathbf{X}\right)=\frac{1}{PJm}\left(m\sigma_\upsilon^2+PJ\sigma_\delta^2+\sigma_\omega^2+(m-1)\psi_T^B\right) \tag{B.26}$$

$$\operatorname{Var}\left(\overline{Y}_C^A\mid\mathbf{X}\right)=\frac{1}{(1-P)Jr}\left(r\sigma_\upsilon^2+(1-P)J\sigma_\delta^2+\sigma_\omega^2+(r-1)\psi_C^A\right) \tag{B.27}$$

$$\operatorname{Var}\left(\overline{Y}_C^B\mid\mathbf{X}\right)=\frac{1}{(1-P)Jm}\left(m\sigma_\upsilon^2+(1-P)J\sigma_\delta^2+\sigma_\omega^2+(m-1)\psi_C^B\right) \tag{B.28}$$

With serial correlation, the first covariance component of Equation (B.11) is

$$\operatorname{Cov}\left(\overline{Y}_T^A,\overline{Y}_T^B\mid\mathbf{X}\right)=\operatorname{E}\left[\overline{Y}_T^A\overline{Y}_T^B\mid\mathbf{X}\right]-\operatorname{E}\left[\overline{Y}_T^A\mid\mathbf{X}\right]\operatorname{E}\left[\overline{Y}_T^B\mid\mathbf{X}\right]$$

$$=\Bigg[\beta(\beta+\tau)+\frac{1}{PJ}\operatorname{E}\left[\upsilon_i^2\mid\mathbf{X}\right]$$

$$+\frac{1}{(PJ)^2mr}\sum_{i=1}^{PJ}\sum_{t=-m+1}^{0}\sum_{s=1}^{r}\operatorname{E}[\omega_{it}\omega_{is}\mid\mathbf{X}]\Bigg]-\beta(\beta+\tau) \tag{B.29}$$

$$=\frac{\sigma_\upsilon^2}{PJ}+\frac{1}{(PJ)^2mr}\sum_{i=1}^{PJ}\sum_{t=-m+1}^{0}\sum_{s=1}^{r}\operatorname{Cov}(\omega_{it},\omega_{is}\mid\mathbf{X})$$

$$=\frac{1}{PJ}\left(\sigma_\upsilon^2+\psi_T^X\right) \tag{B.30}$$

Similarly, with serial correlation, the remaining covariance terms of Equation (B.11) are

$$\operatorname{Cov}\left(\overline{Y}_T^A,\overline{Y}_C^A\mid\mathbf{X}\right)=\frac{\sigma_\delta^2}{r} \tag{B.31}$$

$$\operatorname{Cov}\left(\overline{Y}_T^B,\overline{Y}_C^B\mid\mathbf{X}\right)=\frac{\sigma_\delta^2}{m} \tag{B.32}$$

$$\operatorname{Cov}\left(\overline{Y}_C^A,\overline{Y}_C^B\mid\mathbf{X}\right)=\frac{1}{(1-P)J}\left(\sigma_\upsilon^2+\psi_C^X\right) \tag{B.33}$$

Substituting Equations (B.25)–(B.33) into Equation (B.11) and simplifying gives

$$\operatorname{Var}\left(\widehat{\tau}\mid\mathbf{X}\right)=\frac{1}{PJr}\left(r\sigma_\upsilon^2+PJ\sigma_\delta^2+\sigma_\omega^2+(r-1)\psi_T^A\right)$$

$$
\begin{aligned}
&+ \frac{1}{PJm}\left(m\sigma_{\upsilon}^2 + PJ\sigma_{\delta}^2 + \sigma_{\omega}^2 + (m-1)\psi_T^B\right) \\
&+ \frac{1}{(1-P)Jr}\left(r\sigma_{\upsilon}^2 + (1-P)J\sigma_{\delta}^2 + \sigma_{\omega}^2 + (r-1)\psi_C^A\right) \\
&+ \frac{1}{(1-P)Jm}\left(m\sigma_{\upsilon}^2 + (1-P)J\sigma_{\delta}^2 + \sigma_{\omega}^2 + (m-1)\psi_C^B\right) \\
&- 2\left[\frac{1}{PJ}\left(\sigma_{\upsilon}^2 + \psi_T^X\right) + \frac{\sigma_{\delta}^2}{r} + \frac{\sigma_{\delta}^2}{m} + \frac{1}{(1-P)J}\left(\sigma_{\upsilon}^2 + \psi_C^X\right)\right] \\
&= \left(\frac{m+r}{P(1-P)Jmr}\right)\sigma_{\omega}^2 + \left(\frac{m-1}{PJm}\right)\psi_T^B \qquad \text{(B.34)} \\
&+ \left(\frac{r-1}{PJr}\right)\psi_T^A + \left(\frac{m-1}{(1-P)Jm}\right)\psi_C^B \\
&+ \left(\frac{r-1}{(1-P)Jr}\right)\psi_C^A - \frac{2}{PJ}\psi_T^X - \frac{2}{(1-P)J}\psi_C^X \qquad \text{(B.35)}
\end{aligned}
$$

To further simplify, we set $\psi^B = \psi_T^B = \psi_C^B$, $\psi^A = \psi_T^A = \psi_C^A$, and $\psi^X = \psi_T^X = \psi_C^X$. This follows from Assumption 20 above, whereby the average covariance term of the treated group is equal to the comparable covariance term of the control group.[2]

Then the variance of the estimate of the treatment effect is

$$
\text{Var}\left(\widehat{\tau} \mid \mathbf{X}\right) = \left(\frac{1}{P(1-P)J}\right)\left[\left(\frac{m+r}{mr}\right)\sigma_{\omega}^2 + \left(\frac{m-1}{m}\right)\psi^B + \left(\frac{r-1}{r}\right)\psi^A - 2\psi^X\right] \qquad \text{(B.36)}
$$

Neither Equation (B.35) nor Equation (B.36) is equal to Equation (B.22), the variance of $\widehat{\tau}$ that is estimated by an OLS regression. Instead, the serially correlated error structure alters the true variance of the estimator such that the OLS estimator of the variance is not correct. The cluster-robust variance estimator must be used for correct inference.

Minimum detectable effect With serial correlation, the MDE is

$$
\begin{aligned}
MDE &= \left(t_{1-\kappa}^J + t_{\alpha/2}^J\right)\sqrt{\text{Var}\left(\widehat{\tau} \mid \mathbf{X}\right)} \\
&= \left(t_{1-\kappa}^J + t_{\alpha/2}^J\right)\left[\left(\frac{m+r}{P(1-P)Jmr}\right)\sigma_{\omega}^2 + \left(\frac{m-1}{PJm}\right)\psi_T^B + \left(\frac{r-1}{PJr}\right)\psi_T^A\right. \\
&\qquad + \left(\frac{m-1}{(1-P)Jm}\right)\psi_C^B + \left(\frac{r-1}{(1-P)Jr}\right)\psi_C^A \\
&\qquad \left. - \frac{2}{PJ}\psi_T^X - \frac{2}{(1-P)J}\psi_C^X\right]^{1/2}
\end{aligned}
$$

[2]Under random assignment into treatment, this will hold in expectation, as shown in Section B.1.3.

Following the same assumption as in Equation (B.36), the MDE simplifies to

$$MDE = (t^J_{1-\kappa} + t^J_{\alpha/2}) \sqrt{\left(\frac{1}{P(1-P)J}\right)\left[\left(\frac{m+r}{mr}\right)\sigma^2_\omega \right. \tag{B.37}$$

$$\left. + \left(\frac{m-1}{m}\right)\psi^B + \left(\frac{r-1}{r}\right)\psi^A - 2\psi^X\right] \tag{B.38}$$

where

$$\psi^B \equiv \frac{2}{Jm(m-1)} \sum_{i=1}^{J} \sum_{t=-m+1}^{-1} \sum_{s=t+1}^{0} \text{Cov}\left(\omega_{it}, \omega_{is} \mid \mathbf{X}\right) \tag{B.39}$$

$$\psi^A \equiv \frac{2}{Jr(r-1)} \sum_{i=1}^{J} \sum_{t=0}^{r-1} \sum_{s=t+1}^{r} \text{Cov}\left(\omega_{it}, \omega_{is} \mid \mathbf{X}\right) \tag{B.40}$$

$$\psi^X \equiv \frac{1}{Jmr} \sum_{i=1}^{J} \sum_{t=-m+1}^{0} \sum_{s=1}^{r} \text{Cov}\left(\omega_{it}, \omega_{is} \mid \mathbf{X}\right) \tag{B.41}$$

This is the serial-correlation-robust (SCR) power calculation formula, found in Equation (3.6) in the main text. Note that ψ^B, ψ^A, and ψ^X are likely to depend on the length of the pre- and post- treatment periods, m and r; serial correlation often diminishes as time periods become further apart, so larger values of m and r will result in less correlation on average and smaller parameter values. These parameters do not depend in a systematic way on the number of experimental units, J, however.

B.1.2.3 Serially and cross-sectionally correlated error structure

While randomization at the unit level eliminates cross-sectional correlation from Equation (B.36) in expectation, we can also characterize the full variance without this random assignment assumption. This highlights the type of correlation structures that someone wanting to perform a power calculation for a quasi-experimental design might face.

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.

Assumption 21 (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 Y_{it} is the outcome of interest for unit i at time t; β is the expected outcome of non-treated observations; τ is the treatment effect that is homogenous across all units and all time periods; D_{it} is a time-varying treatment indicator; 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. $\mathcal{N}(0, \sigma_\delta^2)$; and ω_{it} is an idiosyncratic error term distributed (not necessarily i.i.d.) $\mathcal{N}(0, \sigma_\omega^2)$.

Assumption 22 (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 D, $J-1$ unit fixed effects, and $(m + r) - 1$ time fixed effects. This again follows from random assignment of D_{it}.*

Assumption 23 (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 24 (Independence across units at different times)**.** $\mathrm{E}[\omega_{it}\omega_{js} \mid \mathbf{X}] = 0, \ \forall \ i \neq j, \ t \neq s.$

Define ψ_i to be the average serial correlation parameters previously defined in Section B.1.2.2, with the subscript i denoting the correlation is within unit i. Also define ψ_t to be the comparable parameters characterizing cross-sectional correlations, with the subscript t denoting the correlation is within time t. For example, the average cross-sectional covariance among the treated group post-treatment is

$$\psi_{t,T}^{A} = \frac{2}{PJ(PJ-1)r} \sum_{i=1}^{PJ-1} \sum_{j=i+1}^{PJ} \sum_{t=1}^{r} \mathrm{Cov}(\omega_{it}, \omega_{jt} \mid \mathbf{X})$$

Coefficient estimate The deterministic portion of this model is the same as that in Equation (B.7), so the estimate of the treatment effect is again

$$\widehat{\tau} = \left(\overline{Y}_{T}^{A} - \overline{Y}_{T}^{B}\right) - \left(\overline{Y}_{C}^{A} - \overline{Y}_{C}^{B}\right)$$

Variance of coefficient estimate As with the serially correlated error structure in Section B.1.2.2, Equation (B.11) is still correct because of independence between observations that do not correspond to the same unit or time period, but Equations (B.13)–(B.20) no longer hold. With these arbitrary correlations, the first term of Equation (B.11) is

$$\mathrm{Var}\left(\overline{Y}_{T}^{A} \mid \mathbf{X}\right) = \mathrm{Var}\left(\frac{1}{PJr} \sum_{i=1}^{PJ} \sum_{t=1}^{r} Y_{it} \mid \mathbf{X}\right)$$

$$= \frac{1}{(PJr)^2} \operatorname{Var}\left(\sum_{i=1}^{PJ}\sum_{t=1}^{r} Y_{it} \mid \mathbf{X}\right)$$

$$= \frac{1}{(PJr)^2}\left[PJr \operatorname{Var}(Y_{it} \mid \mathbf{X}) + 2\sum_{i=1}^{PJ}\sum_{t=1}^{r-1}\sum_{s=t+1}^{r} \operatorname{Cov}(Y_{it}, Y_{is} \mid \mathbf{X})\right. \tag{B.42}$$

$$\left. + 2\sum_{t=1}^{r}\sum_{i=1}^{PJ-1}\sum_{j=i+1}^{PJ} \operatorname{Cov}(Y_{it}, Y_{jt} \mid \mathbf{X})\right]$$

$$= \frac{\sigma_\varepsilon^2}{PJr} + \frac{2}{(PJr)^2}\sum_{i=1}^{PJ}\sum_{t=1}^{r-1}\sum_{s=t+1}^{r}\left(\sigma_\upsilon^2 + \operatorname{Cov}(\omega_{it}, \omega_{is} \mid \mathbf{X})\right)$$

$$+ \frac{2}{(PJr)^2}\sum_{t=1}^{r}\sum_{i=1}^{PJ-1}\sum_{j=i+1}^{PJ}\left(\sigma_\delta^2 + \operatorname{Cov}(\omega_{it}, \omega_{jt} \mid \mathbf{X})\right)$$

$$= \frac{1}{PJr}\left(r\sigma_\upsilon^2 + PJ\sigma_\delta^2 + \sigma_\omega^2\right) + \frac{2}{(PJr)^2}\sum_{i=1}^{PJ}\sum_{t=1}^{r-1}\sum_{s=t+1}^{r} \operatorname{Cov}(\omega_{it}, \omega_{is} \mid \mathbf{X})$$

$$+ \frac{2}{(PJr)^2}\sum_{t=1}^{r}\sum_{i=1}^{PJ-1}\sum_{j=i+1}^{PJ} \operatorname{Cov}(\omega_{it}, \omega_{jt} \mid \mathbf{X})$$

$$= \frac{1}{PJr}\left(r\sigma_\upsilon^2 + PJ\sigma_\delta^2 + \sigma_\omega^2 + (r-1)\psi_{i,T}^A + (PJ-1)\psi_{t,T}^A\right) \tag{B.43}$$

Similarly, with arbitrary correlations, the remaining variance terms of Equation (B.11) are

$$\operatorname{Var}\left(\overline{Y}_T^B \mid \mathbf{X}\right) = \frac{1}{PJm}\left(m\sigma_\upsilon^2 + PJ\sigma_\delta^2 + \sigma_\omega^2 + (m-1)\psi_{i,T}^B + (PJ-1)\psi_{t,T}^B\right) \tag{B.44}$$

$$\operatorname{Var}\left(\overline{Y}_C^A \mid \mathbf{X}\right) = \frac{1}{(1-P)Jr}\left(r\sigma_\upsilon^2 + (1-P)J\sigma_\delta^2\right.$$

$$\left. + \sigma_\omega^2 + (r-1)\psi_{i,C}^A + ((1-P)J-1)\psi_{t,C}^A\right) \tag{B.45}$$

$$\operatorname{Var}\left(\overline{Y}_C^A \mid \mathbf{X}\right) = \frac{1}{(1-P)Jm}\left(m\sigma_\upsilon^2 + (1-P)J\sigma_\delta^2\right.$$

$$\left. + \sigma_\omega^2 + (m-1)\psi_{i,C}^B + ((1-P)J-1)\psi_{t,C}^B\right) \tag{B.46}$$

With arbitrary correlations, the first covariance component of Equation (B.11) is

$$\operatorname{Cov}\left(\overline{Y}_T^A, \overline{Y}_T^B \mid \mathbf{X}\right) = \operatorname{E}\left[\overline{Y}_T^A\overline{Y}_T^B \mid \mathbf{X}\right] - \operatorname{E}\left[\overline{Y}_T^A \mid \mathbf{X}\right]\operatorname{E}\left[\overline{Y}_T^B \mid \mathbf{X}\right]$$

$$= \left[\beta(\beta+\tau) + \frac{1}{PJ}\operatorname{E}\left[\upsilon_i^2 \mid \mathbf{X}\right]\right. \tag{B.47}$$

$$\left. + \frac{1}{(PJ)^2 mr}\sum_{i=1}^{PJ}\sum_{t=-m+1}^{0}\sum_{s=1}^{r}\operatorname{E}[\omega_{it}\omega_{is} \mid \mathbf{X}]\right] - \beta(\beta+\tau)$$

$$= \frac{\sigma_\upsilon^2}{PJ} + \frac{1}{(PJ)^2 mr}\sum_{i=1}^{PJ}\sum_{t=-m+1}^{0}\sum_{s=1}^{r}\operatorname{Cov}(\omega_{it}, \omega_{is} \mid \mathbf{X})$$

$$= \frac{1}{PJ}\left(\sigma_v^2 + \psi_{i,T}^X\right) \tag{B.48}$$

Similarly, with arbitrary correlations, the remaining covariance terms of Equation (B.11) are

$$\text{Cov}\left(\overline{Y}_T^A, \overline{Y}_C^A \mid \mathbf{X}\right) = \frac{1}{r}\left(\sigma_\delta^2 + \psi_{t,X}^A\right) \tag{B.49}$$

$$\text{Cov}\left(\overline{Y}_T^B, \overline{Y}_C^B \mid \mathbf{X}\right) = \frac{1}{m}\left(\sigma_\delta^2 + \psi_{t,X}^B\right) \tag{B.50}$$

$$\text{Cov}\left(\overline{Y}_C^A, \overline{Y}_C^B \mid \mathbf{X}\right) = \frac{1}{(1-P)J}\left(\sigma_v^2 + \psi_{i,C}^X\right) \tag{B.51}$$

Substituting Equations (B.43)–(B.51) into Equation (B.11) and simplifying gives

$$\begin{aligned}
\text{Var}\left(\widehat{\tau} \mid \mathbf{X}\right) = {} & \frac{1}{PJr}\left(r\sigma_v^2 + PJ\sigma_\delta^2 + \sigma_\omega^2 + (r-1)\psi_{i,T}^A + (PJ-1)\psi_{t,T}^A\right) \\
& + \frac{1}{PJm}\left(m\sigma_v^2 + PJ\sigma_\delta^2 + \sigma_\omega^2 + (m-1)\psi_{i,T}^B + (PJ-1)\psi_{t,T}^B\right) \\
& + \frac{1}{(1-P)Jr}\left(r\sigma_v^2 + (1-P)J\sigma_\delta^2 + \sigma_\omega^2 + (r-1)\psi_{i,C}^A + ((1-P)J-1)\psi_{t,C}^A\right) \\
& + \frac{1}{(1-P)Jm}\left(m\sigma_v^2 + (1-P)J\sigma_\delta^2 + \sigma_\omega^2 + (m-1)\psi_{i,C}^B + ((1-P)J-1)\psi_{t,C}^B\right) \\
& - 2\left[\frac{1}{PJ}\left(\sigma_v^2 + \psi_{i,T}^X\right) + \frac{1}{r}\left(\sigma_\delta^2 + \psi_{t,X}^A\right) + \frac{1}{m}\left(\sigma_\delta^2 + \psi_{t,X}^B\right) + \frac{1}{(1-P)J}\left(\sigma_v^2 + \psi_{i,C}^X\right)\right] \\
= {} & \left(\frac{m+r}{P(1-P)Jmr}\right)\sigma_\omega^2 + \left(\frac{m-1}{PJm}\right)\psi_{i,T}^B + \left(\frac{PJ-1}{PJm}\right)\psi_{t,T}^B \\
& + \left(\frac{r-1}{PJr}\right)\psi_{i,T}^A + \left(\frac{PJ-1}{PJr}\right)\psi_{t,T}^A + \left(\frac{m-1}{(1-P)Jm}\right)\psi_{i,C}^B \\
& + \left(\frac{(1-P)J-1}{(1-P)Jm}\right)\psi_{t,C}^B + \left(\frac{r-1}{(1-P)Jr}\right)\psi_{i,C}^A + \left(\frac{(1-P)J-1}{(1-P)Jr}\right)\psi_{t,C}^A \\
& - \frac{2}{PJ}\psi_{i,T}^X - \frac{2}{m}\psi_{t,X}^B - \frac{2}{r}\psi_{t,X}^A - \frac{2}{(1-P)J}\psi_{i,C}^X
\end{aligned}$$

Minimum detectable effect With arbitrary correlations, the MDE is

$$\begin{aligned}
MDE = {} & \left(t_{1-\kappa}^J + t_{\alpha/2}^J\right)\sqrt{\text{Var}\left(\widehat{\tau} \mid \mathbf{X}\right)} \\
= {} & \left(t_{1-\kappa}^J + t_{\alpha/2}^J\right)\left[\left(\frac{m+r}{P(1-P)Jmr}\right)\sigma_\omega^2 + \left(\frac{m-1}{PJm}\right)\psi_{i,T}^B + \left(\frac{PJ-1}{PJm}\right)\psi_{t,T}^B\right. \\
& + \left(\frac{r-1}{PJr}\right)\psi_{i,T}^A + \left(\frac{PJ-1}{PJr}\right)\psi_{t,T}^A + \left(\frac{m-1}{(1-P)Jm}\right)\psi_{i,C}^B \\
& + \left(\frac{(1-P)J-1}{(1-P)Jm}\right)\psi_{t,C}^B + \left(\frac{r-1}{(1-P)Jr}\right)\psi_{i,C}^A + \left(\frac{(1-P)J-1}{(1-P)Jr}\right)\psi_{t,C}^A \\
& \left. - \frac{2}{PJ}\psi_{i,T}^X - \frac{2}{m}\psi_{t,X}^B - \frac{2}{r}\psi_{t,X}^A - \frac{2}{(1-P)J}\psi_{i,C}^X\right]^{1/2}
\end{aligned} \tag{B.52}$$

We show in Section B.1.3 that Equation (B.52) collapses (in expectation) to the serial-correlation-robust power calculation formula when treatment is randomly assigned.

B.1.2.4 Collapsed dataset

Bertrand, Duflo, and Mullainathan (2004) (henceforth BDM) suggest an alternative to the CRVE in order to achieve the correct false rejection rates in the presence of serial correlation: ignore the time-series structure of the data by averaging the pre-treatment data and the post-treatment data for each unit, then estimate a panel DD regression on this two-period collapsed dataset and apply the OLS variance estimator. While this does yield the desired false rejection rate, simply applying the FP formula to a collapsed dataset will *not* yield the desired power.

Consider again the model presented in Section B.1.2.2 under Assumptions 16–20, but now consider the case in which, prior to estimation, the data are collapsed to one pre- and one post-treatment observation per unit (to eliminate serial correlation, as suggested in BDM). The resulting DGP for the collapsed dataset is

$$Y_{ip}^{C} = \beta + \tau D_{ip}^{C} + \upsilon_i + \delta_p^{C} + \omega_{ip}^{C} \tag{B.53}$$

where Y_{ip}^{C} is the average outcome for unit i for collapsed period p and the other variables are as defined in Section B.1.2.2 or are the collapsed analogs. Note that the τ in Equation (B.53) is equivalent to that in Equation (B.7). These models will yield the same estimate of the treatment effect, $\widehat{\tau}$, but different estimates of its variance in the presence of a (pre-collapsed) serially correlated error structure.

Equivalence of first-difference model BDM show that applying the OLS variance estimator to Equation (B.53) achieves the correct false rejection rate. To see why this is the case, note that this collapsed model can alternatively be expressed as a first-difference model by subtracting each unit's collapsed pre-treatment data from its collapsed post-treatment data. Let ΔY_i^{C} be this difference for the outcome of interest, which gives

$$\begin{aligned}
\Delta Y_i^{C} &= Y_{iA}^{C} - Y_{iB}^{C} \\
&= \left(\beta + \tau D_{iA}^{C} + \upsilon_i + \delta_A^{C} + \omega_{iA}^{C}\right) - \left(\beta + \tau D_{iB}^{C} + \upsilon_i + \delta_B^{C} + \omega_{iB}^{C}\right) \\
&= \tau\left(D_{iA}^{C} - D_{iB}^{C}\right) + \left(\delta_A^{C} - \delta_B^{C}\right) + \left(\omega_{iA}^{C} - \omega_{iB}^{C}\right)
\end{aligned}$$

Defining the other differences variables similarly gives the first-difference DGP of

$$\Delta Y_i^{C} = \tau \Delta D_i^{C} + \Delta \delta^{C} + \Delta \omega_i^{C} \tag{B.54}$$

Equations (B.53) and (B.54) are equivalent, so estimating these models yields not only the same estimate of the treatment effect, $\widehat{\tau}$, but also the same estimate of its variance. Note that the first-difference model, Equation (B.54), is cross-sectional, so the error terms are i.i.d.[3] and the model meets the assumptions of OLS. As a result, the OLS variance estimator is unbiased for the first-difference model, as well as the equivalent collapsed model of Equation (B.53).

Power Although using a collapsed dataset yields the correct false rejection rate, experiments will not be correctly powered if the FP formula is applied to a collapsed dataset. To see this, first consider the MDE of an experiment based on the first-difference model of Equation (B.54). This is a cross-sectional model, so applying Equation (B.6) yields:

$$\text{Var}\left(\hat{\tau} \mid \mathbf{X}\right) = \frac{\sigma^2_{\Delta\omega^C}}{P(1-P)J} \tag{B.55}$$

$$MDE = \left(t^{J-2}_{1-\kappa} + t^{J-2}_{\alpha/2}\right)\sqrt{\frac{\sigma^2_{\Delta\omega^C}}{P(1-P)J}} \tag{B.56}$$

where $\sigma^2_{\Delta\omega^C}$ is the variance of the error term in the collapsed, first-difference model. This variance can be expressed as a function of the parameters that define the error structure of the collapsed data

$$\begin{aligned}
\sigma^2_{\Delta\omega^C} &= \text{Var}\left(\omega^C_{iA} - \omega^C_{iB} \mid \mathbf{X}\right) \\
&= \text{Var}\left(\omega^C_{iA} \mid \mathbf{X}\right) + \text{Var}\left(\omega^C_{iB} \mid \mathbf{X}\right) - 2\,\text{Cov}\left(\omega^C_{iA}, \omega^C_{iB} \mid \mathbf{X}\right) \\
&= 2\,\text{Var}\left(\omega^C_{ip} \mid \mathbf{X}\right) - 2\,\text{Cov}\left(\omega^C_{iA}, \omega^C_{iB} \mid \mathbf{X}\right) \\
&= 2\left(\sigma^2_{\omega^C} - \psi^{CX}\right) \qquad \text{(B.57)}
\end{aligned}$$

where $\sigma^2_{\omega^C}$ is the variance of the error term of the collapsed model, and ψ^{CX} is the average covariance between error terms for the same unit in the collapsed model. Substituting Equation (B.57) into Equation (B.55) gives the variance of $\hat{\tau}$ in terms of parameters of the collapsed data:

$$\text{Var}\left(\hat{\tau} \mid \mathbf{X}\right) = \left(\frac{2}{P(1-P)J}\right)\left(\sigma^2_{\omega^C} - \psi^{CX}\right) \tag{B.58}$$

This formula is equal to the variance of $\hat{\tau}$ from the SCR formula applied to collapsed data, where $m = r = 1$. Applying the FP formula to collapsed data, however, gives the incorrect variance:

$$\text{Var}\left(\hat{\tau} \mid \mathbf{X}\right) = \left(\frac{2}{P(1-P)J}\right)\sigma^2_{\omega^C} \tag{B.59}$$

[3]here are no cross-sectional error correlations due to Assumption B.5.2, because randomization obviates the need to account for this kind of correlation.

Equations (B.58) and (B.59) differ by the ψ^{CX} term that characterizes the covariance between the pre- and post-treatment error terms in the collapsed data. This term is omitted from the FP formula that (incorrectly) assumes no serial correlation in the error structure.

The error structure parameters of the collapsed data in Equation (B.58) can also be expressed as functions of the parameters that define the error structure of the original, uncollapsed data

$$\begin{aligned}
\sigma^2_{\omega^C} &= \frac{1}{2}\operatorname{Var}\left(\omega^C_{iA} \mid \mathbf{X}\right) + \frac{1}{2}\operatorname{Var}\left(\omega^C_{iB} \mid \mathbf{X}\right) \\
&= \frac{1}{2}\operatorname{Var}\left(\frac{1}{r}\sum_{t=1}^{r}\omega_{it} \mid \mathbf{X}\right) + \frac{1}{2}\operatorname{Var}\left(\frac{1}{m}\sum_{s=-m+1}^{0}\omega_{is} \mid \mathbf{X}\right) \\
&= \frac{1}{2r}\left[\sigma^2_\omega + (r-1)\psi^A\right] + \frac{1}{2m}\left[\sigma^2_\omega + (m-1)\psi^B\right] \\
&= \frac{1}{2}\left[\left(\frac{m+r}{mr}\right)\sigma^2_\omega + \left(\frac{r-1}{r}\right)\psi^A + \left(\frac{m-1}{m}\right)\psi^B\right] \qquad \text{(B.60)}
\end{aligned}$$

$$\begin{aligned}
\psi^{CX} &= \frac{1}{J}\sum_{i=1}^{J}\operatorname{Cov}\left(\omega^C_{iA}, \omega^C_{iB} \mid \mathbf{X}\right) \\
&= \frac{1}{J}\sum_{i=1}^{J}\operatorname{Cov}\left(\frac{1}{r}\sum_{t=1}^{r}\omega_{it}, \frac{1}{m}\sum_{s=-m+1}^{0}\omega_{is} \mid \mathbf{X}\right) \\
&= \frac{1}{Jmr}\sum_{i=1}^{J}\sum_{t=1}^{r}\sum_{s=-m+1}^{0}\operatorname{Cov}(\omega_{it}, \omega_{is} \mid \mathbf{X}) \\
&= \psi^X \qquad \text{(B.61)}
\end{aligned}$$

Substituting Equations (B.60) and (B.61) into Equation (B.58) gives

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

which is equivalent to the variance of $\hat{\tau}$ given in Equation (B.36). Hence, the uncollapsed, collapsed, and first-difference models yield (virtually) equivalent MDEs when using the appropriate power calculation formula.[4] By contrast, applying the FP formula ignores the across-period serial correlation that remains after collapsing serially correlated data.

[4]The only difference between these three MDE calculations is the critical values $t^d_{1-\kappa}$ and $t^d_{\alpha/2}$. The uncollapsed model will apply the CRVE *ex post*, which implies $d = J$ degrees of freedom *ex ante*. The collapsed model will apply the OLS variance estimator *ex post* to a panel with $2J$ observations and $J+2$ regressors, which implies $d = J-2$ degrees of freedom *ex ante*. The first-difference model will apply the OLS variance estimator *ex post* to a cross-sectional specification with J observations and 2 regressors, which implies $d = J-2$ degrees of freedom *ex ante*.

B.1.3 Arbitrary cross-sectional correlations

In this section, we provide proofs of Lemma 1 and Lemma 2.

Lemma 3. *In a cross-sectional model with random assignment to treatment,* $\frac{\sigma_\varepsilon^2}{P(1-P)J}$ *is an unbiased estimator of* $\text{Var}(\widehat{\tau} \mid \mathbf{X})$ *even if* $\text{E}[\varepsilon_i \varepsilon_j \mid \mathbf{X}] \neq 0$ *for some* $i \neq j$.

Proof We wish to demonstrate that the OLS variance estimator, which assumes that errors are independent across units, is an unbiased estimator of the true variance under non-i.i.d. errors when units are randomly assigned to treatment. To do this, consider the following setup:

There are J units, indexed by i. P fraction of these units is randomized into treatment, and $(1-P)$ fraction remain in the control group, such that $i \in [i, PJ]$ are treated, and $i \in [PJ+1, J]$ are control. We observe each unit only once. This yields a DGP of:

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

where Y_i is the outcome of interest for unit i, β is the expected outcome for control observations, τ is a treatment effect that is constant across all units, and ω_i is an idiosyncratic error term distributed $\mathcal{N}(0, \sigma_\omega^2)$. These errors need not be independently drawn.

Coefficient estimate The coefficient estimates from an OLS regression are

$$\begin{aligned}
\begin{pmatrix} \widehat{\beta} \\ \widehat{\tau} \end{pmatrix} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'Y \\
&= \begin{pmatrix} J & PJ \\ PJ & PJ \end{pmatrix}^{-1} \begin{pmatrix} \sum_{i=1}^{J} Y_i \\ \sum_{i=1}^{J} D_i Y_i \end{pmatrix} \\
&= \frac{1}{P(1-P)J^2} \begin{pmatrix} PJ\left(\sum_{i=1}^{J} Y_i - \sum_{i=1}^{J} D_i Y_i\right) \\ J\left(\sum_{i=1}^{J} D_i Y_i - P\sum_{i=1}^{J} Y_i\right) \end{pmatrix} \\
&= \begin{pmatrix} \frac{1}{(1-P)J}\sum_{i=1}^{J}(1-D_i)Y_i \\ \frac{1}{PJ}\sum_{i=1}^{J} D_i Y_i - \frac{1}{(1-P)J}\sum_{i=1}^{J}(1-D_i)Y_i \end{pmatrix} \\
&= \begin{pmatrix} \frac{1}{(1-P)J}\sum_{i=PJ+1}^{J} Y_i \\ \frac{1}{PJ}\sum_{i=1}^{PJ} Y_i - \frac{1}{(1-P)J}\sum_{i=PJ+1}^{J} Y_i \end{pmatrix}
\end{aligned}$$

Defining the mean outcome in the treatment and control groups, respectively, as

$$\overline{Y}_T = \frac{1}{PJ}\sum_{i=1}^{PJ} Y_i$$
$$\overline{Y}_C = \frac{1}{(1-P)J}\sum_{i=PJ+1}^{J} Y_i$$

gives coefficient estimates of

$$\widehat{\beta} = \overline{Y}_C$$
$$\widehat{\tau} = \overline{Y}_T - \overline{Y}_C$$

Variance of coefficient estimate Note first that the OLS variance estimator is:

$$\widehat{\text{Var}}_{OLS}\left(\widehat{\tau} \mid \mathbf{X}\right) = \frac{\sigma_\omega^2}{P(1-P)J}$$

The variance of the estimate of the treatment effect, $\hat{\tau}$, is:

$$\text{Var}\left(\widehat{\tau} \mid \mathbf{X}\right) = \text{Var}\left(\overline{Y}_T \mid \mathbf{X}\right) + \text{Var}\left(\overline{Y}_C \mid \mathbf{X}\right) - 2\,\text{Cov}\left(\overline{Y}_T, \overline{Y}_C \mid \mathbf{X}\right) \tag{B.62}$$

where the first term of Equation (B.62) is:

$$\begin{aligned}
\text{Var}\left(\overline{Y}_T \mid \mathbf{X}\right) &= \text{Var}\left(\frac{1}{PJ}\sum_{i=1}^{PJ} Y_i \mid \mathbf{X}\right) \\
&= \frac{1}{(PJ)^2}\,\text{Var}\left(\sum_{i=1}^{PJ} Y_i \mid \mathbf{X}\right) \\
&= \frac{1}{(PJ)^2}\sum_{i=1}^{PJ}\text{Var}(Y_i \mid \mathbf{X}) + \frac{2}{(PJ)^2}\sum_{i=1}^{PJ-1}\sum_{j=i+1}^{PJ}\text{Cov}(Y_i, Y_j \mid \mathbf{X}) \\
&= \frac{1}{(PJ)^2}\sum_{i=1}^{PJ}\text{Var}(\omega_i \mid \mathbf{X}) + \frac{2}{(PJ)^2}\sum_{i=1}^{PJ-1}\sum_{j=i+1}^{PJ}\text{Cov}(\omega_i, \omega_j \mid \mathbf{X}) \\
&= \frac{\sigma_\omega^2}{PJ} + \frac{PJ-1}{PJ}\psi_T
\end{aligned} \tag{B.63}$$

where

$$\psi_T \equiv \frac{2}{PJ(PJ-1)}\sum_{i=1}^{PJ-1}\sum_{j=i+1}^{PJ}\text{Cov}\left(\omega_i, \omega_j \mid \mathbf{X}\right)$$

is the average covariance between treated units. The second term of Equation (B.62) is:

$$\text{Var}\left(\overline{Y}_C \mid \mathbf{X}\right) = \frac{\sigma_\omega^2}{(1-P)J} + \frac{(1-P)J-1}{(1-P)J}\psi_C \tag{B.64}$$

where

$$\psi_C \equiv \frac{2}{(1-P)J((1-P)J-1)} \sum_{i=PJ+1}^{J-1} \sum_{j=i+1}^{J} \operatorname{Cov}\left(\omega_i, \omega_j \mid \mathbf{X}\right)$$

is the average covariance between control units. The third term of Equation (B.62) is:

$$\begin{aligned}
-2\operatorname{Cov}\left(\overline{Y}_T, \overline{Y}_C \mid \mathbf{X}\right) &= -2\operatorname{Cov}\left(\frac{1}{PJ}\sum_{i=1}^{PJ} Y_i, \frac{1}{(1-P)J}\sum_{j=PJ+1}^{J} Y_j \mid \mathbf{X}\right) \\
&= -2\sum_{i=1}^{PJ}\sum_{j=PJ+1}^{J} \operatorname{Cov}\left(\frac{1}{PJ}Y_i, \frac{1}{(1-P)J}Y_j \mid \mathbf{X}\right) \\
&= \frac{-2}{P(1-P)J^2}\sum_{i=1}^{PJ}\sum_{j=PJ+1}^{J} \operatorname{Cov}(\omega_i, \omega_j \mid \mathbf{X}) \\
&= -2\psi_{TC} \qquad \text{(B.65)}
\end{aligned}$$

where

$$\psi_{TC} \equiv \frac{1}{P(1-P)J^2}\sum_{i=1}^{PJ}\sum_{j=PJ+1}^{J} \operatorname{Cov}\left(\omega_i, \omega_j \mid \mathbf{X}\right)$$

is the average covariance between treatment and control units. Substituting Equations (B.63)–(B.65) into Equation (B.62) yields:

$$\begin{aligned}
\operatorname{Var}\left(\widehat{\tau} \mid \mathbf{X}\right) &= \left[\frac{\sigma_\omega^2}{PJ} + \frac{PJ-1}{PJ}\psi_T\right] + \left[\frac{\sigma_\omega^2}{(1-P)J} + \frac{(1-P)J-1}{(1-P)J}\psi_C\right] - 2\psi_{TC} \\
&= \frac{\sigma_\omega^2}{P(1-P)J} + \frac{PJ-1}{PJ}\psi_T + \frac{(1-P)J-1}{(1-P)J}\psi_C - 2\psi_{TC}
\end{aligned}$$

Note that σ_ω^2, P, and J are constant population or design parameters. With this in mind, taking expectations yields:

$$\begin{aligned}
\operatorname{E}\left[\operatorname{Var}\left(\widehat{\tau} \mid \mathbf{X}\right)\right] &= \frac{\sigma_\omega^2}{P(1-P)J} + \frac{PJ-1}{PJ}\operatorname{E}\left[\psi_T \mid \mathbf{X}\right] \\
&\quad + \frac{(1-P)J-1}{(1-P)J}\operatorname{E}\left[\psi_C \mid \mathbf{X}\right] - 2\operatorname{E}\left[\psi_{TC} \mid \mathbf{X}\right] \qquad \text{(B.66)}
\end{aligned}$$

where, by random assignment to treatment

$$\begin{aligned}
\operatorname{E}\left[\psi_T \mid \mathbf{X}\right] &= \frac{2}{PJ(PJ-1)}\operatorname{E}\left[\sum_{i=1}^{PJ-1}\sum_{j=i+1}^{PJ} \operatorname{Cov}\left(\omega_i, \omega_j \mid \mathbf{X}\right)\right] \\
&= \frac{2}{PJ(PJ-1)}\operatorname{E}\left[\sum_{i=1}^{J-1}\sum_{j=i+1}^{J} \operatorname{Cov}\left(\omega_i, \omega_j \mid \mathbf{X}\right)\mathbf{1}\{i \in T, j \in T\}\right]
\end{aligned}$$

$$= \frac{2}{PJ(PJ-1)} \sum_{i=1}^{J-1} \sum_{j=i+1}^{J} \operatorname{Cov}(\omega_i, \omega_j \mid \mathbf{X}) \operatorname{E}[\mathbf{1}\{i \in T, j \in T\}]$$

$$= \frac{2}{PJ(PJ-1)} \sum_{i=1}^{J-1} \sum_{j=i+1}^{J} \operatorname{Cov}(\omega_i, \omega_j \mid \mathbf{X}) P^2$$

$$= \frac{2P}{J(PJ-1)} \sum_{i=1}^{J-1} \sum_{j=i+1}^{J} \operatorname{Cov}(\omega_i, \omega_j \mid \mathbf{X}) \tag{B.67}$$

$$\operatorname{E}[\psi_C \mid \mathbf{X}] = \frac{2(1-P)}{J((1-P)J-1)} \sum_{i=1}^{J-1} \sum_{j=i+1}^{J} \operatorname{Cov}(\omega_i, \omega_j \mid \mathbf{X}) \tag{B.68}$$

$$\operatorname{E}[\psi_{TC} \mid \mathbf{X}] = \frac{2}{J^2} \sum_{i=1}^{J-1} \sum_{j=i+1}^{J} \operatorname{Cov}(\omega_i, \omega_j \mid \mathbf{X}) \tag{B.69}$$

Substituting Equations (B.67)–(B.69) into Equation (B.66) yields:

$$\begin{aligned}
\operatorname{E}[\operatorname{Var}(\widehat{\tau} \mid \mathbf{X})] &= \frac{\sigma_\omega^2}{P(1-P)J} + \frac{2}{J^2} \sum_{i=1}^{J-1} \sum_{j=i+1}^{J} \operatorname{Cov}(\omega_i, \omega_j \mid \mathbf{X}) \\
&\quad + \frac{2}{J^2} \sum_{i=1}^{J-1} \sum_{j=i+1}^{J} \operatorname{Cov}(\omega_i, \omega_j \mid \mathbf{X}) - \frac{4}{J^2} \sum_{i=1}^{J-1} \sum_{j=i+1}^{J} \operatorname{Cov}(\omega_i, \omega_j \mid \mathbf{X}) \\
&= \frac{\sigma_\omega^2}{P(1-P)J} \\
&= \widehat{\operatorname{Var}}_{OLS}(\widehat{\tau} \mid \mathbf{X})
\end{aligned}$$

Therefore, the OLS variance estimator is an unbiased estimate of the true variance under random assignment to treatment, even under non-i.i.d. errors. □

Lemma 4. *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* $\operatorname{Var}(\hat{\tau} \mid \mathbf{X})$*, even in the presence of arbitrary within-period cross-sectional correlations.*

Proof We wish to demonstrate that the serial-correlation-robust variance, which assumes that errors are independent across units, is an unbiased estimator of the true variance under arbitrary within-period correlations, when units are randomly assigned to treatment. To do this, consider the following setup:

There are J units, indexed by i. P fraction of these units is randomized into treatment, and $(1-P)$ fraction remain in the control group, such that $i \in [i, PJ]$ are treated, and $i \in [PJ+1, J]$ are control. We observe each unit m times in the pre-treatment period, and r times in the post-treatment period, such that time periods $t \in [-m+1, 0]$ are in the pre-treatment period, and time periods $s \in [1, r]$ are in the post-treatment period. Treatment begins at the same time for all treated units. There are unit-specific shocks that affect all time periods, as well as time-varying shocks that are common across units. Treatment begins at the same time for all treated units. This yields a DGP of:

$$\begin{aligned} Y_{it} &= \beta + \tau D_{it} + \varepsilon_{it} \\ &= \beta + \tau D_{it} + \upsilon_i + \delta_t + \omega_{it} \end{aligned}$$

where Y_{it} is the outcome of interest for unit i at time t; β is the expected outcome of non-treated observations; τ is a treatment effect that is constant across all units and time periods; D_{it} is a time-varying treatment indicator, equal to one if and only if a unit is in the treatment group and $t \geq 1$; υ_i is a time-invariant unit-specific effect distributed i.i.d. $\mathcal{N}(0, \sigma_\upsilon^2)$; δ_t is a time-specific effect distributed i.i.d. $\mathcal{N}(0, \sigma_\delta^2)$; and ω_{it} is an idiosyncratic error term distributed $\mathcal{N}(0, \sigma_\omega^2)$. These errors need not be independently drawn, such that ω_{it} may exhibit arbitrary within-unit and within-period correlations. However, this still assumes $\mathrm{E}\left[\omega_{it}, \omega_{jt} \mid \mathbf{X}\right] = 0$ in all cases where $i \neq j$ and $t \neq s$.

Coefficient estimate Under this model, the estimate of the treatment effect is

$$\widehat{\tau} = \left(\overline{Y}_T^A - \overline{Y}_T^B\right) - \left(\overline{Y}_C^A - \overline{Y}_C^B\right)$$

where

$$\begin{aligned} \overline{Y}_T^A &= \frac{1}{PJr} \sum_{i=1}^{PJ} \sum_{t=1}^{r} Y_{it} \\ \overline{Y}_T^B &= \frac{1}{PJm} \sum_{i=1}^{PJ} \sum_{t=-m+1}^{0} Y_{it} \\ \overline{Y}_C^A &= \frac{1}{(1-P)Jr} \sum_{i=PJ+1}^{J} \sum_{t=1}^{r} Y_{it} \\ \overline{Y}_C^B &= \frac{1}{(1-P)Jm} \sum_{i=PJ+1}^{J} \sum_{t=-m+1}^{0} Y_{it} \end{aligned}$$

Variance of coefficient estimate Note first that the serial-correlation-robust variance estimator is

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

We wish to demonstrate that this is an unbiased estimator of the variance of $\hat{\tau}$ when units have been randomly assigned to treatment but may exhibit between-unit error correlations with a given time period.

The variance of the treatment effect estimator is:

$$\begin{aligned}\text{(B.70)}\quad \text{Var}\left(\widehat{\tau} \mid \mathbf{X}\right) &= \text{Var}\left(\overline{Y}_{T}^{A} \mid \mathbf{X}\right) + \text{Var}\left(\overline{Y}_{T}^{B} \mid \mathbf{X}\right) + \text{Var}\left(\overline{Y}_{C}^{A} \mid \mathbf{X}\right) + \text{Var}\left(\overline{Y}_{C}^{B} \mid \mathbf{X}\right) \\ &\quad - 2\Bigg[\text{Cov}\left(\overline{Y}_{T}^{A}, \overline{Y}_{T}^{B} \mid \mathbf{X}\right) + \text{Cov}\left(\overline{Y}_{T}^{A}, \overline{Y}_{C}^{A} \mid \mathbf{X}\right) \\ \text{(B.71)}\quad &\quad + \text{Cov}\left(\overline{Y}_{T}^{B}, \overline{Y}_{C}^{B} \mid \mathbf{X}\right) + \text{Cov}\left(\overline{Y}_{C}^{A}, \overline{Y}_{C}^{B} \mid \mathbf{X}\right)\Bigg]\end{aligned}$$

This equation can be broken up into components, such that:

$$\begin{aligned}\text{Var}\left(\overline{Y}_{T}^{A} \mid \mathbf{X}\right) &= \text{Var}\left(\frac{1}{PJr}\sum_{i=1}^{PJ}\sum_{t=1}^{r} Y_{it} \mid \mathbf{X}\right) \\ &= \frac{1}{(PJr)^{2}}\,\text{Var}\left(\sum_{i=1}^{PJ}\sum_{t=1}^{r} Y_{it} \mid \mathbf{X}\right) \\ \text{(B.72)}\quad &= \frac{1}{(PJr)^{2}}\Bigg[PJr\,\text{Var}(Y_{it} \mid \mathbf{X}) + 2\sum_{i=1}^{PJ}\sum_{t=1}^{r-1}\sum_{s=t+1}^{r}\text{Cov}(Y_{it}, Y_{is} \mid \mathbf{X}) \\ &\quad + 2\sum_{t=1}^{r}\sum_{i=1}^{PJ-1}\sum_{j=i+1}^{PJ}\text{Cov}(Y_{it}, Y_{jt} \mid \mathbf{X})\Bigg] \\ &= \frac{\sigma_{\varepsilon}^{2}}{PJr} + \frac{2}{(PJr)^{2}}\sum_{i=1}^{PJ}\sum_{t=1}^{r-1}\sum_{s=t+1}^{r}\left(\sigma_{v}^{2} + \text{Cov}(\omega_{it}, \omega_{is} \mid \mathbf{X})\right) \\ &\quad + \frac{2}{(PJr)^{2}}\sum_{t=1}^{r}\sum_{i=1}^{PJ-1}\sum_{j=i+1}^{PJ}\left(\sigma_{\delta}^{2} + \text{Cov}(\omega_{it}, \omega_{jt} \mid \mathbf{X})\right) \\ &= \frac{1}{PJr}\left(r\sigma_{v}^{2} + PJ\sigma_{\delta}^{2} + \sigma_{\omega}^{2}\right) + \frac{2}{(PJr)^{2}}\sum_{i=1}^{PJ}\sum_{t=1}^{r-1}\sum_{s=t+1}^{r}\text{Cov}(\omega_{it}, \omega_{is} \mid \mathbf{X}) \\ &\quad + \frac{2}{(PJr)^{2}}\sum_{t=1}^{r}\sum_{i=1}^{PJ-1}\sum_{j=i+1}^{PJ}\text{Cov}(\omega_{it}, \omega_{jt} \mid \mathbf{X})\end{aligned}$$

(B.73) $$= \frac{1}{PJr}\left(r\sigma_v^2 + PJ\sigma_\delta^2 + \sigma_\omega^2 + (r-1)\psi_{i,T}^A + (PJ-1)\psi_{t,T}^A\right)$$

where

$$\psi_{i,T}^A \equiv \frac{2}{PJr(r-1)} \sum_{i=1}^{PJ} \sum_{t=1}^{r-1} \sum_{s=t+1}^{r} \text{Cov}\left(\omega_{it}, \omega_{is} \mid \mathbf{X}\right)$$

$$\psi_{t,T}^A \equiv \frac{2}{PJ(PJ-1)r} \sum_{i=1}^{PJ-1} \sum_{j=i+1}^{PJ} \sum_{t=1}^{r} \text{Cov}\left(\omega_{it}, \omega_{jt} \mid \mathbf{X}\right)$$

Similarly, with arbitrary correlations, the remaining variance terms of Equation (B.71) are

(B.74)
$$\text{Var}\left(\overline{Y}_T^B \mid \mathbf{X}\right) = \frac{1}{PJm}\left(m\sigma_v^2 + PJ\sigma_\delta^2 + \sigma_\omega^2 + (m-1)\psi_{i,T}^B + (PJ-1)\psi_{t,T}^B\right)$$

where

$$\psi_{i,T}^B \equiv \frac{2}{PJm(m-1)} \sum_{i=1}^{PJ} \sum_{t=-m+1}^{-1} \sum_{s=t+1}^{0} \text{Cov}\left(\omega_{it}, \omega_{is} \mid \mathbf{X}\right)$$

$$\psi_{t,T}^B \equiv \frac{2}{PJ(PJ-1)m} \sum_{i=1}^{PJ-1} \sum_{j=i+1}^{PJ} \sum_{t=-m+1}^{0} \text{Cov}\left(\omega_{it}, \omega_{jt} \mid \mathbf{X}\right)$$

(B.75)
$$\text{Var}\left(\overline{Y}_C^A \mid \mathbf{X}\right) = \frac{1}{(1-P)Jr}\left(r\sigma_v^2 + (1-P)J\sigma_\delta^2 + \sigma_\omega^2 + (r-1)\psi_{i,C}^A + ((1-P)J-1)\psi_{t,C}^A\right)$$

where

$$\psi_{i,C}^A \equiv \frac{2}{(1-P)Jr(r-1)} \sum_{i=PJ+1}^{J} \sum_{t=1}^{r-1} \sum_{s=t+1}^{r} \text{Cov}\left(\omega_{it}, \omega_{is} \mid \mathbf{X}\right)$$

$$\psi_{t,C}^A \equiv \frac{2}{(1-P)J((1-P)J-1)r} \sum_{i=PJ+1}^{J-1} \sum_{j=i+1}^{J} \sum_{t=1}^{r} \text{Cov}\left(\omega_{it}, \omega_{jt} \mid \mathbf{X}\right)$$

(B.76)
$$\text{Var}\left(\overline{Y}_C^B \mid \mathbf{X}\right) = \frac{1}{(1-P)Jm}\left(m\sigma_v^2 + (1-P)J\sigma_\delta^2 + \sigma_\omega^2 + (m-1)\psi_{i,C}^B + ((1-P)J-1)\psi_{t,C}^B\right)$$

where

$$\psi_{i,C}^B \equiv \frac{2}{(1-P)Jm(m-1)} \sum_{i=PJ+1}^{J} \sum_{t=-m+1}^{-1} \sum_{s=t+1}^{0} \text{Cov}\left(\omega_{it}, \omega_{is} \mid \mathbf{X}\right)$$

$$\psi_{t,C}^B \equiv \frac{2}{(1-P)J((1-P)J-1)m} \sum_{i=PJ+1}^{J-1} \sum_{j=i+1}^{J} \sum_{t=-m+1}^{0} \text{Cov}\left(\omega_{it}, \omega_{jt} \mid \mathbf{X}\right)$$

(B.77)

With arbitrary correlations, the first covariance component of Equation (B.71) is

$$\begin{aligned}
\operatorname{Cov}\left(\overline{Y}_T^A, \overline{Y}_T^B \mid \mathbf{X}\right) &= \mathrm{E}\left[\overline{Y}_T^A \overline{Y}_T^B \mid \mathbf{X}\right] - \mathrm{E}\left[\overline{Y}_T^A \mid \mathbf{X}\right] \mathrm{E}\left[\overline{Y}_T^B \mid \mathbf{X}\right] \\
&= \left[\beta(\beta+\tau) + \frac{1}{PJ}\,\mathrm{E}\left[\upsilon_i^2 \mid \mathbf{X}\right]\right. \qquad \text{(B.78)} \\
&\quad \left. + \frac{1}{(PJ)^2 mr} \sum_{i=1}^{PJ} \sum_{t=-m+1}^{0} \sum_{s=1}^{r} \mathrm{E}[\omega_{it}\omega_{is} \mid \mathbf{X}]\right] - \beta(\beta+\tau) \\
&= \frac{\sigma_v^2}{PJ} + \frac{1}{(PJ)^2 mr} \sum_{i=1}^{PJ} \sum_{t=-m+1}^{0} \sum_{s=1}^{r} \operatorname{Cov}(\omega_{it}, \omega_{is} \mid \mathbf{X}) \\
&= \frac{1}{PJ}\left(\sigma_v^2 + \psi_{i,T}^X\right) \qquad \text{(B.79)}
\end{aligned}$$

where

$$\psi_{i,T}^X \equiv \frac{1}{PJmr} \sum_{i=1}^{PJ} \sum_{t=-m+1}^{0} \sum_{s=1}^{r} \operatorname{Cov}\left(\omega_{it}, \omega_{is} \mid \mathbf{X}\right) \qquad \text{(B.80)}$$

Similarly, with arbitrary correlations, the remaining covariance terms of Equation (B.71) are

$$\operatorname{Cov}\left(\overline{Y}_T^A, \overline{Y}_C^A \mid \mathbf{X}\right) = \frac{1}{r}\left(\sigma_\delta^2 + \psi_{t,TC}^A\right) \qquad \text{(B.81)}$$

where

$$\psi_{t,TC}^A \equiv \frac{1}{P(1-P)J^2 r} \sum_{i=1}^{PJ} \sum_{j=PJ+1}^{J} \sum_{t=1}^{r} \operatorname{Cov}\left(\omega_{it}, \omega_{jt} \mid \mathbf{X}\right)$$

$$\operatorname{Cov}\left(\overline{Y}_T^B, \overline{Y}_C^B \mid \mathbf{X}\right) = \frac{1}{m}\left(\sigma_\delta^2 + \psi_{t,TC}^B\right) \qquad \text{(B.82)}$$

where

$$\psi_{t,TC}^B \equiv \frac{1}{P(1-P)J^2 m} \sum_{i=1}^{PJ} \sum_{j=PJ+1}^{J} \sum_{t=-m+1}^{0} \operatorname{Cov}\left(\omega_{it}, \omega_{jt} \mid \mathbf{X}\right)$$

$$\operatorname{Cov}\left(\overline{Y}_C^A, \overline{Y}_C^B \mid \mathbf{X}\right) = \frac{1}{(1-P)J}\left(\sigma_v^2 + \psi_{i,C}^X\right) \qquad \text{(B.83)}$$

where

$$\psi_{i,C}^X \equiv \frac{1}{(1-P)Jmr} \sum_{i=PJ+1}^{J} \sum_{t=-m+1}^{0} \sum_{s=1}^{r} \operatorname{Cov}\left(\omega_{it}, \omega_{is} \mid \mathbf{X}\right)$$

(B.84)

Substituting Equations (B.73)–(B.83) into Equation (B.71) and simplifying gives

$$
\begin{aligned}
\operatorname{Var}(\widehat{\tau} \mid \mathbf{X}) &= \frac{1}{PJr}\left(r\sigma_{\upsilon}^2 + PJ\sigma_{\delta}^2 + \sigma_{\omega}^2 + (r-1)\psi_{i,T}^A + (PJ-1)\psi_{t,T}^A\right) \\
&\quad + \frac{1}{PJm}\left(m\sigma_{\upsilon}^2 + PJ\sigma_{\delta}^2 + \sigma_{\omega}^2 + (m-1)\psi_{i,T}^B + (PJ-1)\psi_{t,T}^B\right) \\
&\quad + \frac{1}{(1-P)Jr}\left(r\sigma_{\upsilon}^2 + (1-P)J\sigma_{\delta}^2 + \sigma_{\omega}^2 + (r-1)\psi_{i,C}^A + ((1-P)J-1)\psi_{t,C}^A\right) \\
&\quad + \frac{1}{(1-P)Jm}\left(m\sigma_{\upsilon}^2 + (1-P)J\sigma_{\delta}^2 + \sigma_{\omega}^2 + (m-1)\psi_{i,C}^B + ((1-P)J-1)\psi_{t,C}^B\right) \\
&\quad - 2\left[\frac{1}{PJ}\left(\sigma_{\upsilon}^2 + \psi_{i,T}^X\right) + \frac{1}{r}\left(\sigma_{\delta}^2 + \psi_{t,X}^A\right) + \frac{1}{m}\left(\sigma_{\delta}^2 + \psi_{t,X}^B\right) + \frac{1}{(1-P)J}\left(\sigma_{\upsilon}^2 + \psi_{i,C}^X\right)\right] \\
&= \left(\frac{m+r}{P(1-P)Jmr}\right)\sigma_{\omega}^2 + \left(\frac{m-1}{PJm}\right)\psi_{i,T}^B + \left(\frac{PJ-1}{PJm}\right)\psi_{t,T}^B \\
&\quad + \left(\frac{r-1}{PJr}\right)\psi_{i,T}^A + \left(\frac{PJ-1}{PJr}\right)\psi_{t,T}^A + \left(\frac{m-1}{(1-P)Jm}\right)\psi_{i,C}^B \\
&\quad + \left(\frac{(1-P)J-1}{(1-P)Jm}\right)\psi_{t,C}^B + \left(\frac{r-1}{(1-P)Jr}\right)\psi_{i,C}^A + \left(\frac{(1-P)J-1}{(1-P)Jr}\right)\psi_{t,C}^A \\
&\quad - \frac{2}{PJ}\psi_{i,T}^X - \frac{2}{m}\psi_{t,X}^B - \frac{2}{r}\psi_{t,X}^A - \frac{2}{(1-P)J}\psi_{i,C}^X
\end{aligned}
$$

Next, we show that, in expectation, this is equal to $\widehat{\operatorname{Var}}_{SCR}(\widehat{\tau} \mid \mathbf{X})$. We begin by taking expectations:

$$
\begin{aligned}
\operatorname{E}[\operatorname{Var}(\widehat{\tau} \mid \mathbf{X})] &= \left(\frac{m+r}{P(1-P)Jmr}\right)\sigma_{\omega}^2 + \left(\frac{m-1}{PJm}\right)\operatorname{E}\left[\psi_{i,T}^B \mid \mathbf{X}\right] + \left(\frac{PJ-1}{PJm}\right)\operatorname{E}\left[\psi_{t,T}^B \mid \mathbf{X}\right] \\
&\quad + \left(\frac{r-1}{PJr}\right)\operatorname{E}\left[\psi_{i,T}^A \mid \mathbf{X}\right] + \left(\frac{PJ-1}{PJr}\right)\operatorname{E}\left[\psi_{t,T}^A \mid \mathbf{X}\right] + \left(\frac{m-1}{(1-P)Jm}\right)\operatorname{E}\left[\psi_{i,C}^B \mid \mathbf{X}\right] \\
&\quad + \left(\frac{(1-P)J-1}{(1-P)Jm}\right)\operatorname{E}\left[\psi_{t,C}^B \mid \mathbf{X}\right] + \left(\frac{r-1}{(1-P)Jr}\right)\operatorname{E}\left[\psi_{i,C}^A \mid \mathbf{X}\right] + \\
&\quad + \left(\frac{(1-P)J-1}{(1-P)Jr}\right)\operatorname{E}\left[\psi_{t,C}^A \mid \mathbf{X}\right] - \frac{2}{PJ}\operatorname{E}\left[\psi_{i,T}^X \mid \mathbf{X}\right] - \frac{2}{m}\operatorname{E}\left[\psi_{t,X}^B \mid \mathbf{X}\right] \\
&\quad - \frac{2}{r}\operatorname{E}\left[\psi_{t,X}^A \mid \mathbf{X}\right] - \frac{2}{(1-P)J}\operatorname{E}\left[\psi_{i,C}^X \mid \mathbf{X}\right]
\end{aligned}
$$

where

$$
\operatorname{E}\left[\psi_{i,T}^B \mid \mathbf{X}\right] = \frac{2}{Jm(m-1)}\sum_{i=1}^{J}\sum_{t=-m+1}^{-1}\sum_{s=t+1}^{0}\operatorname{Cov}(\omega_{it}, \omega_{is} \mid \mathbf{X})
$$

$$
\operatorname{E}\left[\psi_{t,T}^B \mid \mathbf{X}\right] = \frac{2P}{J(PJ-1)m}\sum_{i=1}^{J-1}\sum_{j=i+1}^{J}\sum_{t=-m+1}^{0}\operatorname{Cov}(\omega_{it}, \omega_{jt} \mid \mathbf{X})
$$

$$\mathrm{E}\left[\psi^A_{i,T} \mid \mathbf{X}\right] = \frac{2}{Jr(r-1)} \sum_{i=1}^{J} \sum_{t=1}^{r-1} \sum_{s=t+1}^{r} \mathrm{Cov}\left(\omega_{it}, \omega_{is} \mid \mathbf{X}\right)$$

$$\mathrm{E}\left[\psi^A_{t,T} \mid \mathbf{X}\right] = \frac{2P}{J(PJ-1)r} \sum_{i=1}^{J-1} \sum_{j=i+1}^{J} \sum_{t=1}^{r} \mathrm{Cov}\left(\omega_{it}, \omega_{jt} \mid \mathbf{X}\right)$$

$$\mathrm{E}\left[\psi^B_{i,C} \mid \mathbf{X}\right] = \frac{2}{Jm(m-1)} \sum_{i=1}^{J} \sum_{t=-m+1}^{-1} \sum_{s=t+1}^{0} \mathrm{Cov}\left(\omega_{it}, \omega_{is} \mid \mathbf{X}\right)$$

$$\mathrm{E}\left[\psi^B_{t,C} \mid \mathbf{X}\right] = \frac{2(1-P)}{J((1-P)J-1)m} \sum_{i=1}^{J-1} \sum_{j=i+1}^{J} \sum_{t=-m+1}^{0} \mathrm{Cov}\left(\omega_{it}, \omega_{jt} \mid \mathbf{X}\right)$$

$$\mathrm{E}\left[\psi^A_{i,C} \mid \mathbf{X}\right] = \frac{2}{Jr(r-1)} \sum_{i=1}^{J} \sum_{t=1}^{r-1} \sum_{s=t+1}^{r} \mathrm{Cov}\left(\omega_{it}, \omega_{is} \mid \mathbf{X}\right)$$

$$\mathrm{E}\left[\psi^A_{t,C} \mid \mathbf{X}\right] = \frac{2(1-P)}{J((1-P)J-1)r} \sum_{i=1}^{J-1} \sum_{j=i+1}^{J} \sum_{t=1}^{r} \mathrm{Cov}\left(\omega_{it}, \omega_{jt} \mid \mathbf{X}\right)$$

$$\mathrm{E}\left[\psi^X_{i,T} \mid \mathbf{X}\right] = \frac{1}{Jmr} \sum_{i=1}^{J} \sum_{t=-m+1}^{0} \sum_{s=1}^{r} \mathrm{Cov}\left(\omega_{it}, \omega_{is} \mid \mathbf{X}\right)$$

$$\mathrm{E}\left[\psi^B_{t,TC} \mid \mathbf{X}\right] = \frac{2}{J^2 m} \sum_{i=1}^{J-1} \sum_{j=i+1}^{J} \sum_{t=-m+1}^{0} \mathrm{Cov}\left(\omega_{it}, \omega_{jt} \mid \mathbf{X}\right)$$

$$\mathrm{E}\left[\psi^A_{t,TC} \mid \mathbf{X}\right] = \frac{2}{J^2 r} \sum_{i=1}^{J-1} \sum_{j=i+1}^{J} \sum_{t=1}^{r} \mathrm{Cov}\left(\omega_{it}, \omega_{jt} \mid \mathbf{X}\right)$$

$$\mathrm{E}\left[\psi^X_{i,C} \mid \mathbf{X}\right] = \frac{1}{Jmr} \sum_{i=1}^{J} \sum_{t=-m+1}^{0} \sum_{s=1}^{r} \mathrm{Cov}\left(\omega_{it}, \omega_{is} \mid \mathbf{X}\right)$$

Furthermore, define

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

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

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

Using these terms, we can rewrite:

$$\mathrm{E}\left[\mathrm{Var}\left(\widehat{\tau} \mid \mathbf{X}\right)\right] = \left(\frac{m+r}{P(1-P)Jmr}\right)\sigma^2_\omega + \left(\frac{m-1}{PJm}\right)\psi^B$$

$$
\begin{aligned}
&+\left(\frac{pJ-1}{PJm}\right)\left(\frac{2P}{J(PJ-1)m}\right)\sum_{i=1}^{J-1}\sum_{j=i+1}^{J}\sum_{t=-m+1}^{0}\mathrm{Cov}\left(\omega_{it},\omega_{jt}\mid\mathbf{X}\right)\\
&+\left(\frac{r-1}{PJr}\right)\psi^{A}+\left(\frac{PJ-1}{PJr}\right)\left(\frac{2P}{J(PJ-1)r}\right)\sum_{i=1}^{J-1}\sum_{j=i+1}^{J}\sum_{t=1}^{r}\mathrm{Cov}\left(\omega_{it},\omega_{jt}\mid\mathbf{X}\right)\\
&+\left(\frac{m-1}{(1-P)Jm}\right)\psi^{B}\\
&+\left(\frac{(1-P)J-1}{(1-P)Jm}\right)\left(\frac{2(1-P)}{J((1-P)J-1)m}\right)\sum_{i=1}^{J-1}\sum_{j=i+1}^{J}\sum_{t=-m+1}^{0}\mathrm{Cov}\left(\omega_{it},\omega_{jt}\mid\mathbf{X}\right)\\
&+\left(\frac{r-1}{(1-P)Jr}\right)\psi^{A}\\
&+\left(\frac{(1-P)J-1}{(1-P)Jr}\right)\left(\frac{2(1-P)}{J((1-P)J-1)r}\right)\sum_{i=1}^{J-1}\sum_{j=i+1}^{J}\sum_{t=1}^{r}\mathrm{Cov}\left(\omega_{it},\omega_{jt}\mid\mathbf{X}\right)\\
&-\left(\frac{2}{PJ}\right)\psi^{X}-\left(\frac{2}{m}\right)\left(\frac{2}{J^{2}m}\right)\sum_{i=1}^{J-1}\sum_{j=i+1}^{J}\sum_{t=-m+1}^{0}\mathrm{Cov}\left(\omega_{it},\omega_{jt}\mid\mathbf{X}\right)\\
&-\left(\frac{2}{(1-P)J}\right)\psi^{X}-\left(\frac{2}{r}\right)\left(\frac{2}{J^{2}r}\right)\sum_{i=1}^{J-1}\sum_{j=i+1}^{J}\sum_{t=1}^{r}\mathrm{Cov}\left(\omega_{it},\omega_{jt}\mid\mathbf{X}\right)\\
&=\left(\frac{m+r}{P(1-P)Jmr}\right)\sigma_{\omega}^{2}+\left[\frac{m-1}{PJm}+\frac{m-1}{(1-P)Jm}\right]\psi^{B}+\left[\frac{r-1}{PJr}+\frac{r-1}{(1-P)Jr}\right]\psi^{A}\\
&\quad+\left[\frac{-2}{PJ}-\frac{2}{(1-P)J}\right]\psi^{X}\\
&+\left[\frac{2}{J^{2}m^{2}}+\frac{2}{J^{2}m^{2}}-\frac{4}{J^{2}m^{2}}\right]\sum_{i=1}^{J-1}\sum_{j=i+1}^{J}\sum_{t=-m+1}^{0}\mathrm{Cov}\left(\omega_{it},\omega_{jt}\mid\mathbf{X}\right)\\
&\quad+\left[\frac{2}{J^{2}r^{2}}+\frac{2}{J^{2}r^{2}}-\frac{4}{J^{2}r^{2}}\right]\sum_{i=1}^{J-1}\sum_{j=i+1}^{J}\sum_{t=1}^{r}\mathrm{Cov}\left(\omega_{it},\omega_{jt}\mid\mathbf{X}\right)\\
&=\left(\frac{m+r}{P(1-P)Jmr}\right)\sigma_{\omega}^{2}+\left(\frac{m-1}{P(1-P)Jm}\right)\psi^{B}\\
&+\left(\frac{r-1}{P(1-P)Jr}\right)\psi^{A}-\frac{2}{P(1-P)J}\psi^{X}\\
&=\frac{1}{P(1-P)J}\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]\\
&=\widehat{\mathrm{Var}}_{SCR}\left(\widehat{\tau}\mid\mathbf{X}\right)
\end{aligned}
$$

And therefore the SCR variance estimator is an unbiased estimate of the true variance under random assignment to treatment, even under non-i.i.d. errors. □

B.2 Figures in main text

This section provides further detail on the simulations and power calculations referenced in the main text. We discuss the algorithms and assumptions behind each of the simulation plots, as well as the two analytical power calculation figures.

B.2.1 Simulated AR(1) data

In Figure 3.3.2, we run Monte Carlo simulations where each iteration generates a new simulated dataset with an idiosyncratic error term (ω_{it}) that evolves via an AR(1) process. We vary the following three parameters across sets of 10,000 simulations: the number of pre-treatment periods (m), the number of post-treatment periods (r), and the AR(1) dependece parameter (γ). The remaining parameters (J, P, α, κ, β, μ_υ, σ^2_υ, μ_δ, σ^2_δ, σ^2_ω) are fixed across all simulations.

Step 1: We calculate τ^{FP} and τ^{SCR} for each set of simulations, given a set of parameters values for m, r, and γ. These values are functions of the number of pre-treatment periods m, the number of post-treatment periods r, the number of units J, the proportion of units randomized into treatment P, the desired Type-I error rate α, the desired power κ, and the idiosyncratic variance σ^2_ω:

(B.1)
$$\tau^{FP} = \left(t^J_{1-\kappa} + t^J_{\alpha/2}\right)\sqrt{\left(\frac{\sigma^2_\omega}{P(1-P)J}\right)\left(\frac{m+r}{mr}\right)}$$

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

Note that for both τ^{FP} and τ^{SCR}, we calculate the critical values $t^J_{1-\kappa}$ and $t^J_{\alpha/2}$ assuming J degrees of freedom, which is consistent with applying the CRVE *ex post*, clustering at the unit level with J units. Note also that ψ^B, ψ^A, and ψ^X depend on the correlation structure of the errors, and the AR(1) process enables us to derive closed form expressions for these covariances in terms of γ, σ^2_ω, m, and r. Because we set $m = r$ across all simulations, we can write ψ^B, ψ^A, and ψ^X as:

(B.3)
$$\psi^B = \frac{2\sigma^2_\omega}{(m-1)m}\sum_{z=1}^{m-1}(m-z)\gamma^z$$

(B.4)
$$\psi^A = \psi^B$$

(B.5)
$$\psi^X = \frac{\sigma^2_\omega}{m^2}\left[\sum_{z=1}^{m} z\gamma^z + \sum_{z=m+1}^{2m-1}(2m-z)\gamma^z\right]$$

Step 2: For each simulation, we generate a dataset as specified by the data generating process:

$$Y_{it} = \beta + \upsilon_i + \delta_t + \omega_{it} \tag{B.6}$$

To do this, we draw J independent values of υ_i from the distribution $N(\mu_\upsilon, \sigma_\upsilon^2)$, and draw $m + r$ independent values of δ_t from the distribution $N(\mu_\delta, \sigma_\delta^2)$. We create the idiosyncratic error $\omega_{it} = \gamma\omega_{i(t-1)} + \xi_{it}$ by simulating an AR(1) process with serial correlation γ and a white noise term ξ_{it} drawn from the distribution $N(0, \sigma_\xi^2)$, where $\sigma_\xi^2 = \sigma_\omega^2(1 - \gamma^2)$.[5]

Step 3: We randomly assign treatment to PJ units. This involves randomly scrambling a vector of PJ ones and $(1 - P)J$ zeros and assigning each unit i either a 1 indicating treatment or a 0 indicating control.[6] This allows us to construct a time-varying treatment indicator D_{it}, where $D_{it} = 1$ for all treated units in post-treatment periods only, and $D_{it} = 0$ otherwise. We then create three outcome variables by adding treatment effects to the data generated in the previous step:

$$Y_{it}^{FP} \equiv Y_{it} + \tau^{FP} D_{it} \tag{B.7}$$

$$Y_{it}^{SCR} \equiv Y_{it} + \tau^{SCR} D_{it} \tag{B.8}$$

$$Y_{it}^{0} \equiv Y_{it} + \tau^{0} D_{it}, \tag{B.9}$$

where $\tau^0 = 0$ is a placebo treatment effect.

Step 4: We estimate the following three OLS-fixed effects regressions:

$$Y_{it}^{FP} = \beta + \tau^{FP} D_{it} + \upsilon_i + \delta_t + \omega_{it} \tag{B.10}$$

$$Y_{it}^{SCR} = \beta + \tau^{SCR} D_{it} + \upsilon_i + \delta_t + \omega_{it} \tag{B.11}$$

$$Y_{it}^{0} = \beta + \tau^{0} D_{it} + \upsilon_i + \delta_t + \omega_{it} \tag{B.12}$$

Step 5: For each estimated $\hat{\tau}^{FP}$, $\hat{\tau}^{SCR}$, and $\hat{\tau}^0$, we compute both OLS standard errors and CRVE standard errors, clustered at the unit level.

We repeat Steps 2–5 10,000 times, for values of $m = r \in \{1, \ldots, 20\}$ and for values of $\gamma \in \{0, 0.3, 0.5, 0.7, 0.9\}$.[7] After each set of 10,000 simulations, we calculate

[5] We allow for a sufficiently long "burn-in period" in this AR(1) process, so that the process starts to evolve many periods before the first period of simulated data.

[6] Our code rounds PJ to the nearest integer value, however PJ is already an integer in our main parameterization. Note that for the τ^{FP} and τ^{SCR} to be precisely calibrated, round$(PJ)/J$ needs to equal the parameter value P.

[7] We set $m = r$ only for simplicity. However, the results are very similar if fix $m = 3$ and vary $r \in \{1, \ldots, 20\}$, or vice versa. In Appendix B.3, we present results that vary m and r separately.

the percent of simulations where $\hat{\tau}^{FP}$, $\hat{\tau}^{SCR}$, and $\hat{\tau}^{0}$ reject null hypothesis of $\tau = 0$ at significance level α, under both OLS and CRVE standard errors.

Figure 3.3.2 reports these rejection rates on the vertical axes, with the number of pre- and post-treatment periods ($m = r$) on the horizontal axes, for each value of γ. Reading the top row left to right, we report the rejection rates for τ^{FP} under OLS standard errors, for τ^{FP} under CRVE standard errors, and for τ^{SCR} under CRVE standard errors, respectively. Because these are rejection rates of true effects, we interpret these curves as realized statistical power. Reading the bottom row left to right, we report the rejection rates for τ^{0} under OLS standard errors, for τ^{0} under CRVE standard errors, and for τ^{0} under CRVE standard errors, respectively. Because these are rejection rates of placebo effects, we report these curves as realized false rejection rates. (The bottom-center and bottom-right panels report identical rejection rates, because the center and right columns have the same test for false rejection rates.)

We fix $\alpha = 0.05$ and $\kappa = 0.80$ across all simulations, as these are the critical values commonly used in practice. However, they are essentially arbitrary, and our simulation results would look identical if we had chosen alternative tolerances for Type I vs. Type II errors. (The only difference would be that the vertical axes would change to reflect these alternative values.) All other fixed parameter values are arbitrary. We have set $J = 50$, $P = 0.5$, $\beta = 0$, $\mu_{\upsilon} = 100$, $\sigma^2_{\upsilon} = 80$, $\mu_{\delta} = 20$, $\sigma^2_{\delta} = 10$, and $\sigma^2_{\omega} = 10$. These values of σ^2_{υ}, σ^2_{δ}, and σ^2_{ω} imply an intracluster correlation coefficient of $\rho_{\upsilon} = 0.8$ and within-period correlation coefficient of $\rho_{\delta} = 0.1$. Importantly, our simulation results do not depend on any particular combination of these parameters values, because they rely on the internal consistency of the *ex ante* treatment effect calibration and the *ex post* estimation, conditional on a *given* set of parameter values. The only exceptions are for J and P: J must be larger enough to allow us to use the CRVE estimator (i.e., at least 40 clusters), and P must be within a reasonable range (i.e., between 0.1 and 0.9) such that there are a sufficient number of both treated clusters and control clusters.

The ANCOVA simulations in Figure 3.5.7 follow the exact same algorithm as those in Figure 3.3.2, except that after Step 3, we collapse all pre-treatment observations into a unit-specific unweighted average:

$$\overline{Y}^{FP}_{i,pre} \equiv \sum_{t \in pre} Y^{FP}_{it} \qquad \overline{Y}^{SCR}_{i,pre} \equiv \sum_{t \in pre} Y^{SCR}_{it} \qquad \overline{Y}^{0}_{i,pre} \equiv \sum_{t \in pre} Y^{0}_{it}$$

Step 4 then becomes a set of ANCOVA regressions with time fixed effects, but replacing unit fixed effects with a unit-specific pre-period control:

$$Y^{FP}_{it} = \beta + \tau^{FP} D_i + \theta \overline{Y}^{FP}_{i,pre} + \delta_t + \omega_{it} \tag{B.13}$$

$$Y_{it}^{SCR} = \beta + \tau^{SCR} D_i + \theta \overline{Y}_{i,pre}^{SCR} + \delta_t + \omega_{it} \tag{B.14}$$

$$Y_{it}^{0} = \beta + \tau^{0} D_i + \theta \overline{Y}_{i,pre}^{0} + \delta_t + \omega_{it} \tag{B.15}$$

We estimate these ANCOVA specifications on post-treatment observations only, meaning that the treatment indicator D_i is now time-invariant. In Step 5, we collect only CRVE standard errors, and Figure 3.5.7 reports the resulting rejection rates of the ANCOVA estimator. We do not report the false rejection rates from the Y_i^0 regression for the sake of brevity, and they do achieve the desired $\alpha = 0.05$ rejection rate in all cases.

An additional nuance with the ANCOVA simulations is that our simulation results are now sensitive to the intracluster correlation coefficient ρ_υ. This is because the proportion of variance that is unit-specific now affects the precision of the $\hat{\tau}$ estimator, because we have replaced the unit fixed effect (which directly controlled for this variance) with a linear control in the average pre-treatment level of the outcome variable. For a low ρ_υ, removing the unit fixed effect sacrifices little in terms of efficiency, because the between-unit differences are relatively small; for a high ρ_υ, removing the unit fixed effect can substantially reduce efficiency — enough to almost fully erode the efficiency gains of ANCOVA relative to DD. For this reason, our ANCOVA simulations in Figure 3.5.7 simulate over $m = r$, γ, and now ρ_υ. The left panel shows results for $\gamma \in \{0, 0.3, 0.5, 0.7, 0.9\}$ with $\rho_\upsilon = 0.5$; the right panel shows results for $\rho_\upsilon \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$ with $\gamma = 0.5$.[8]

B.2.2 Bloom et al. (2015) data

In Figure 3.4.3, we run Monte Carlo simulations using data from Bloom et al. (2015). These simulations are analogous to those described above, except that rather than simulating data, we use an actual dataset from a published panel RCT. We downloaded the paper's dataset from the *Quarterly Journal of Economics* website, and focused on the data used to estimate the paper's main results, reported in Column (1) of Table II of the paper. Consistent with the regression model that produced this result, we base our Monte Carlo analysis on the following DD specification:

$$Performance_{it} = \alpha Treat_i \times Experiment_t + \beta_t + \gamma_i + \varepsilon_{it}$$

Converting the original paper's notation to our notation, and substituting the outcome variable and fixed effects with the names of the variables in the Bloom et al. (2015) dataset, we have:

$$\underbrace{\text{perform1}_{it}}_{Y_{it}} = \tau D_{it} + \underbrace{\text{year\_week}_t}_{\delta_t} + \underbrace{\text{personid}_i}_{\upsilon_i} + \omega_{it}$$

[8] Mechanically, we vary ρ_υ by varying σ_υ^2 only, and leaving σ_δ^2 and σ_ω^2 unchanged. The values $\sigma_\upsilon^2 \in \{\frac{20}{9}, \frac{60}{7}, 20, \frac{140}{3}, 180\}$ translate to the values $\rho_\upsilon \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$.

We keep only units in the main sample (i.e. expgroup $\in \{0, 1\}$), only pre-treatment weeks of data (i.e. year_week < 201049), and only individuals with non-missing perform1_{it} values for all weeks of pre-treatment data. This leaves us with a balanced panel of $J = 79$ individuals across 48 weeks.

We conduct simulations on this dataset by varying the number of pre-treatment periods (m) and the number of post-treatment periods (r). As with the simulations described above, we vary the panel length for values $m = r \in \{1, \ldots, 20\}$, iterating 10,000 simulations for each value of $m = r$. We set the parameter values $\sigma^2_\omega = 0.507$ and $\gamma = 0.233$ by regressing Y_{it} on person and week fixed effects, calculating the variance of the resulting residuals $\hat{\omega}_{it}$, and then estimating $\hat{\gamma}$ by estimating

$$\hat{\omega}_{it} = \gamma \hat{\omega}_{i(t-1)} + \xi_{it} \tag{B.16}$$

Step 1: We calculate τ^{FP} and $\tau^{AR(1)}$ for each set of simulations, given $m = r$:

$$\tau^{FP} = \left(t^J_{1-\kappa} + t^J_{\alpha/2}\right) \sqrt{\left(\frac{\sigma^2_\omega}{P(1-P)J}\right)\left(\frac{m+r}{mr}\right)} \tag{B.17}$$

$$\tau^{AR(1)} = (t^J_{1-\kappa} + t^J_{\alpha/2}) \sqrt{\left(\frac{1}{P(1-P)J}\right)\left[\left(\frac{m+r}{mr}\right)\sigma^2_\omega + \left(\frac{m-1}{m}\right)\ddot{\psi}^B + \left(\frac{r-1}{r}\right)\ddot{\psi}^A - 2\ddot{\psi}^X\right]} \tag{B.18}$$

where

$$\ddot{\psi}^B = \frac{2\sigma^2_\omega}{(m-1)m} \sum_{z=1}^{m-1} (m-z)\gamma^z \tag{B.19}$$

$$\ddot{\psi}^A = \ddot{\psi}^B \tag{B.20}$$

$$\ddot{\psi}^X = \frac{\sigma^2_\omega}{m^2}\left[\sum_{z=1}^{m} z\gamma^z + \sum_{z=m+1}^{2m-1} (2m-z)\gamma^z\right] \tag{B.21}$$

We denote the covariance terms as $\ddot{\psi}^B$, $\ddot{\psi}^A$, and $\ddot{\psi}^X$ to indicate that the AR(1) error assumption is a (poor) representation of the more complex covariance structure of this dataset. For both τ^{FP} and $\tau^{AR(1)}$, we calculate the critical values $t^J_{1-\kappa}$ and $t^J_{\alpha/2}$ assuming J degrees of freedom, which is consistent with applying the CRVE *ex post*, clustering at the individual level with J individuals.

Step 2: We calculate τ^{SCR} given $m = r$, by non-parametrically estimating $\sigma^2_{\hat{\omega}}$, $\psi^B_{\hat{\omega}}$, $\psi^A_{\hat{\omega}}$, and $\psi^X_{\hat{\omega}}$ from residuals. Appendix B.4.1 provides step-by-step details of this estimation algorithm. Rather than impose an AR(1) structure on the serial correlation,

this method enables us to flexibly characterize the covariance structure of the Bloom et al. (2015) dataset with just three averaged parameters. This allows us to calculate τ^{SCR} as:

$$\tau^{SCR} = \left(t^J_{1-\kappa} + t^J_{\alpha/2}\right)\left\{\left(\frac{1}{P(1-P)J}\right)\left[\left(\frac{m+r}{mr}\right)k_\sigma\sigma^2_{\hat{\omega}} + \left(\frac{m-1}{m}\right)k_B\psi^B_{\hat{\omega}} + \left(\frac{r-1}{r}\right)k_A\psi^A_{\hat{\omega}} - 2k_X\psi^X_{\hat{\omega}}\right]\right\}^{1/2} \tag{B.22}$$

where

$$\begin{aligned} k_\sigma &= \frac{I(m+r)^2}{2(I-1)mr} \\ k_B &= \frac{I(m+r)^2}{2(I-1)r^2} \\ k_A &= \frac{I(m+r)^2}{2(I-1)m^2} \\ k_X &= 0 \end{aligned}$$

Appendix B.5 provides a derivation of the coefficients k_σ, k_B, k_A, and k_X, and it proves that this expression for τ^{SCR} as a function of *estimated* variance-covariance parameters is equal (in expectation) to the MDE as a function of the *true* variance-covariance parameters.[9]

Step 3: For each simulation, we randomly select a range of $m+r$ consecutive weeks in the dataset. This subset of weeks will become the $(m+r)$-period panel dataset used in this particular simulation. We randomly assign treatment to PJ individuals. This involves randomly scrambling a vector of PJ ones and $(1-P)J$ zeros and assigning each individual i either a 1 indicating treatment or a 0 indicating control.[10] This allows us to construct a time-varying treatment indicator D_{it}, where $D_{it} = 1$ for all treated units in post-treatment periods only, and $D_{it} = 0$ otherwise. We then create three outcome variables by adding treatment effects to the data generated in the previous step:

$$Y^{FP}_{it} \equiv Y_{it} + \tau^{FP}D_{it} \tag{B.23}$$

$$Y^{AR(1)}_{it} \equiv Y_{it} + \tau^{AR(1)}D_{it} \tag{B.24}$$

[9] I denotes the number of units used to estimate $\sigma^2_{\hat{\omega}}$, $\psi^B_{\hat{\omega}}$, $\psi^A_{\hat{\omega}}$ and $\psi^X_{\hat{\omega}}$. This is distinct from the sample size of the experiment J, however these simuations set $I = J = 79$ to include all units in the Bloom et al. (2015) dataset.

[10] Our code rounds PJ to the nearest integer value, even though PJ is already an integer in our main parameterization. Note that for the τ^{FP}, $\tau^{AR(1)}$, and τ^{SCR} to be precisely calibrated, the effective $\tilde{P}$ (where $\tilde{P} = \text{round}(PJ)/J$) needs to equal the actual parameter value P.

$$Y_{it}^{SCR} \equiv Y_{it} + \tau^{SCR} D_{it} \tag{B.25}$$

Step 4: We estimate the following three OLS-fixed effects regressions:

$$Y_{it}^{FP} = \beta + \tau^{FP} D_{it} + \upsilon_i + \delta_t + \omega_{it} \tag{B.26}$$

$$Y_{it}^{AR(1)} = \beta + \tau^{AR(1)} D_{it} + \upsilon_i + \delta_t + \omega_{it} \tag{B.27}$$

$$Y_{it}^{SCR} = \beta + \tau^{SCR} D_{it} + \upsilon_i + \delta_t + \omega_{it} \tag{B.28}$$

Step 5: For each estimated $\hat{\tau}^{FP}$, $\hat{\tau}^{AR(1)}$, and $\hat{\tau}^{SCR}$, we compute CRVE standard errors, clustered at the individual level.

As with the AR(1) simulations above, we repeat Steps 3–5 10,000 times, for values of $m = r \in \{1, \ldots, 20\}$.[11] After each set of 10,000 simulations, we calculate the percent of simulations where $\hat{\tau}^{FP}$, $\hat{\tau}^{AR(1)}$, and $\hat{\tau}^{SCR}$ reject null hypothesis of $\tau = 0$ at significance level α, under CRVE standard errors. Figure 3.4.3 reports these three rejection rates on the vertical axes, with the number of pre- and post-treatment periods ($m = r$) on the horizontal axes. We can interpret these curves as realized statistical power, just as in the top row of Figure 3.3.2. We fix $\alpha = 0.05$ and $\kappa = 0.80$ across all simulations, for the reasons discussed above. Besides our arbitrary choices of $P = 0.5$, all other parameters are determined by the Bloom et al. (2015) dataset: $J = I = 79$, $\sigma_\upsilon^2 = 0.243$, $\sigma_\delta^2 = 0.146$, and $\sigma_\omega^2 = 0.507$, implying $\rho_\upsilon = 0.271$ and $\rho_\delta = 0.163$. We do not estimate β, μ_υ, or μ_δ, as these parameters are no longer relevant when simulating on top of an existing dataset.

B.2.3 Pecan Street data

In Figure 3.4.4, we present analogous Monte Carlo results for simulations using the Pecan Street dataset of household electricity consumption (Pecan Street (2016)). These data are publicly available (with a researcher login) at `https://dataport.pecanstreet.org/data/interactive`, and they include 699 households over 26,888 hours. As with the Bloom et al. (2015) simulations, we construct a balanced panel of households and hours, by restricting the full Pecan Street dataset to a sample of households that report non-missing, non-zero electricity consumption for every hour between January 1, 2013 and December 31, 2014. This results in a balanced panel of $J = 97$ households over 17,520 hours, which we collapse to create daily, weekly, and monthly datasets.

These Pecan Street simulations follow an algorithm identical to the Bloom et al. (2015) simulations, and we describe this algorithm in detail above. We repeat the

[11] As with the AR(1) simulations, we set $m = r$ only for simplicity. However, the results are very similar if fix $m = 3$ and vary $r \in \{1, \ldots, 20\}$, or vice versa.

same set of simulations four times, estimating separate rejection rates for τ^{FP}, $\tau^{AR(1)}$, and τ^{SCR}, for each of the hourly, daily, weekly, and monthly datasets. We again set $\alpha = 0.05$, $\kappa = 0.80$, and $P = 0.5$. The other relevant parameters for each dataset are:

Table B.2.1: Pecan Street Simulation Parameters

Dataset	J	σ^2_υ	σ^2_δ	σ^2_ω	γ	ρ_υ	ρ_δ
Hourly	97	0.257	0.458	0.642	0.623	0.189	0.337
Daily	97	0.257	0.234	0.135	0.651	0.411	0.373
Weekly	97	0.256	0.211	0.083	0.713	0.465	0.384
Monthly	97	0.256	0.203	0.058	0.654	0.495	0.392

These values are estimated separately from each dataset used in the simulations.

B.2.4 Analytic Power Calculations

Figure 3.4.6 displays the results of analytic power calculations performed using the daily Pecan Street dataset. In other words, we calculate the number of units needed by applying the FP and SCR power calculation formulas by using the data to estimate $\sigma^2_{\hat{\omega}}$, $\psi^B_{\hat{\omega}}$, $\psi^A_{\hat{\omega}}$, and $\psi^X_{\hat{\omega}}$. For each experiment of length $m = r \in \{1, \ldots, 12\}$, we estimate the average values of $\sigma^2_{\hat{\omega}}$, $\psi^B_{\hat{\omega}}$, $\psi^A_{\hat{\omega}}$, and $\psi^X_{\hat{\omega}}$ over all possible panels of that length.[12] We assign half of the households to treatment ($P = 0.5$), allow for a 5 percent Type I error rate ($\alpha = 0.05$), and calibrate to 80 percent power ($\kappa = 0.80$). Finally, we rearrange Equations (3.3) and (B.2) to calculate the number of households required to detect MDEs that range from 0 to 15 percent of baseline electricity consumption.

Figure 3.5.8 also shows the results of analytic power calculations using the SCR formula. However, instead of parameterizing Equation (3.6) using estimates from a dataset, we now normalize $\sigma^2_\omega = 1$ and assume an AR(1) correlation structure with $\gamma \in \{0, 0.3, 0.5, 0.7, 0.9\}$. For panel lengths $m = r \in \{1, \ldots, 100\}$, we analytically

[12]We follow the algorithm outlined in Section B.4.1 below. For a given value of $m = r$, we consider each (consecutive) subset S of the daily Pecan Street data with length $2r$. We first residualize this subset of the data with household and day fixed effects, and we calculate $\sigma^2_{\hat{\omega},S}$ from these residuals. We then assign the first m residuals for each household to the pre-treatment period and the remaining r residuals to the post-treatment period, thereby estimating $\psi^B_{\hat{\omega}S}$, $\psi^A_{\hat{\omega}S}$, and $\psi^X_{\hat{\omega}S}$ (by averaging all pairwise covariances for subset S. Averaging $\sigma^2_{\hat{\omega}S}$, $\psi^B_{\hat{\omega}S}$, $\psi^A_{\hat{\omega}S}$, and $\psi^X_{\hat{\omega}S}$ over all subsets S, we arrive at estimates for $\sigma^2_{\hat{\omega}}$, $\psi^B_{\hat{\omega}}$, $\psi^A_{\hat{\omega}}$, and $\psi^X_{\hat{\omega}}$.

derive ψ^B, ψ^A, and ψ^X using the formulas from Equations (B.3)–(B.5). In the left panel, we fixe $P = 0.5$, $\alpha = 0.05$, $\kappa = 0.80$, and $J = 100$, and use Equation (3.6) to solve for MDE as a function of $m = r$ and γ. In the right panel, we fix $P = 0.5$, $\alpha = 0.05$, $\kappa = 0.80$, and $MDE = 1$, and rearrange Equation (3.6) to solve for J as a function of $m = r$ and γ.

B.3 Additional results

In this section, we present extensions of our simulation results from the main text. Here, we focus on short panels, as many existing randomized trials in economics contain few waves of data. We also allow m and r to vary separately in this section, to demonstrate the robustness of our results in contexts where m and r are unequal.

We begin by extending the realized power results from the middle column of Figure 3.3.2 in the main text, allowing the number of pre- and post-treatment periods to vary separately. We simulate data with serially correlated errors of varying AR(1) parameters γ, for different panel lengths, and we calibrate treatment effect sizes using the Frison and Pocock power calculation formula (Equation (3.3)). Figure B.3.1 displays the results for panels with $m \in \{1, \ldots, 6\}$ and $r \in \{1, \ldots, 6\}$.

Notably, across all levels of non-zero serial correlation ($\gamma > 0$), the "naive" FP power calculation formula yields over-powered experiments for panels with either one pre-treatment or one post-treatment period. This is striking, given the number of existing experiments in economics that follow a traditional "one baseline, one follow-up" model. Given that panel data typically exhibit serial correlation, it follows that, had they relied on the FP formula to conduct *ex ante* power calculations (and had the assumed MDE been equal to the true effect), most existing experiments are likely to have been overpowered.[13] In line with the results in Figure 3.3.2, realized power decreases monotonically as panel length increases, and the FP formula becomes increasingly likely to yield underpowered experiments.

In Figure B.3.2, we perform an analogous exercise using the Bloom et al. (2015) and Pecan Street datasets. These figures present results from the same algorithm used to produce Figures 3.4.3 and 3.4.4, except that we vary m and r separately, for $m \in \{1, \ldots, 6\}$ and $r \in \{1, \ldots, 6\}$. As with the simulated data, we find that the "naive" FP power calculation yields over-powered experiments with either one pre-treatment or one post-treatment period.

Next, rather than estimating a standard panel fixed effects model, we instead follow McKenzie (2012) and estimate treatment effects using ANCOVA methods. According to McKenzie (2012) and Frison and Pocock (1992), the analogous power calculation formula for an ANCOVA model with i.i.d. errors is:

$$MDE = \left(t_{1-\kappa} + t_{\alpha/2}\right) \sqrt{\left(\frac{\sigma_\varepsilon^2}{P(1-P)J}\right)\left[\frac{1+(r-1)\rho}{r} - \frac{m\rho^2}{1+(m-1)\rho}\right]} \tag{B.1}$$

[13] Technically, this also assumes that these *ex ante* power calculations parameterized the FP formula with a residual variance either calculated over a long time series, or inflated to correct for the downward bias from estimating residual variance in a short time series.

Converting this expression to our notation:

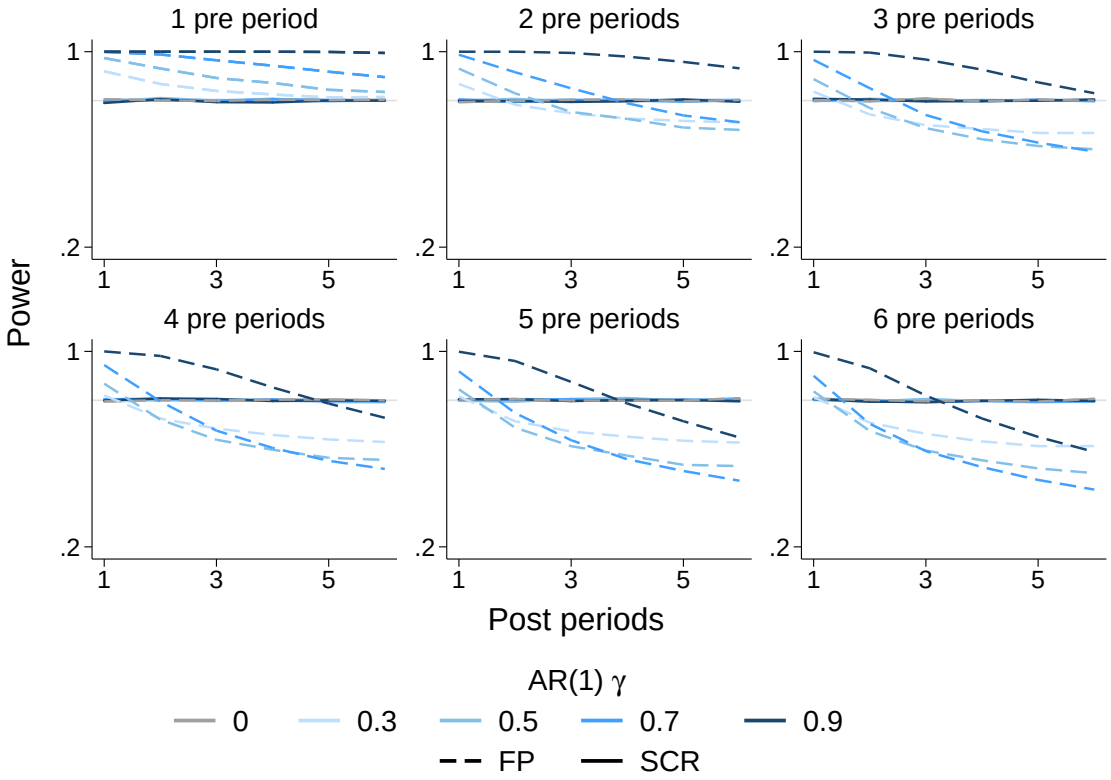
$$MDE = \left(t^J_{1-\kappa} + t^J_{\alpha/2}\right)\sqrt{\left(\frac{1}{P(1-P)J}\right)\left[(1-\theta)^2\sigma^2_\upsilon + \frac{m+\theta^2 r}{mr}\sigma^2_\omega\right]} \tag{B.2}$$

$$\text{where } \theta = \frac{m\sigma^2_\upsilon}{m\sigma^2_\upsilon + \sigma^2_\delta + \sigma^2_\omega} \tag{B.3}$$

As in Figure 3.5.7, we simulate data with varying levels of AR(1) serial correlation and estimate Equation (3.13). We vary the number of pre-treatment and post-treatment periods separately, for values of $m \in \{1, \ldots, 5\}$ and $r \in \{1, \ldots, 5\}$. We also vary the intracluster correlation coefficient $\rho_\upsilon \in \{0.1, 0.5, 0.9\}$. We calibrate treatment effect sizes equal to MDE in Equation (B.2), and apply CRVE standard errors *ex post* (clustered by unit). Figure B.3.3 displays the results of this exercise.

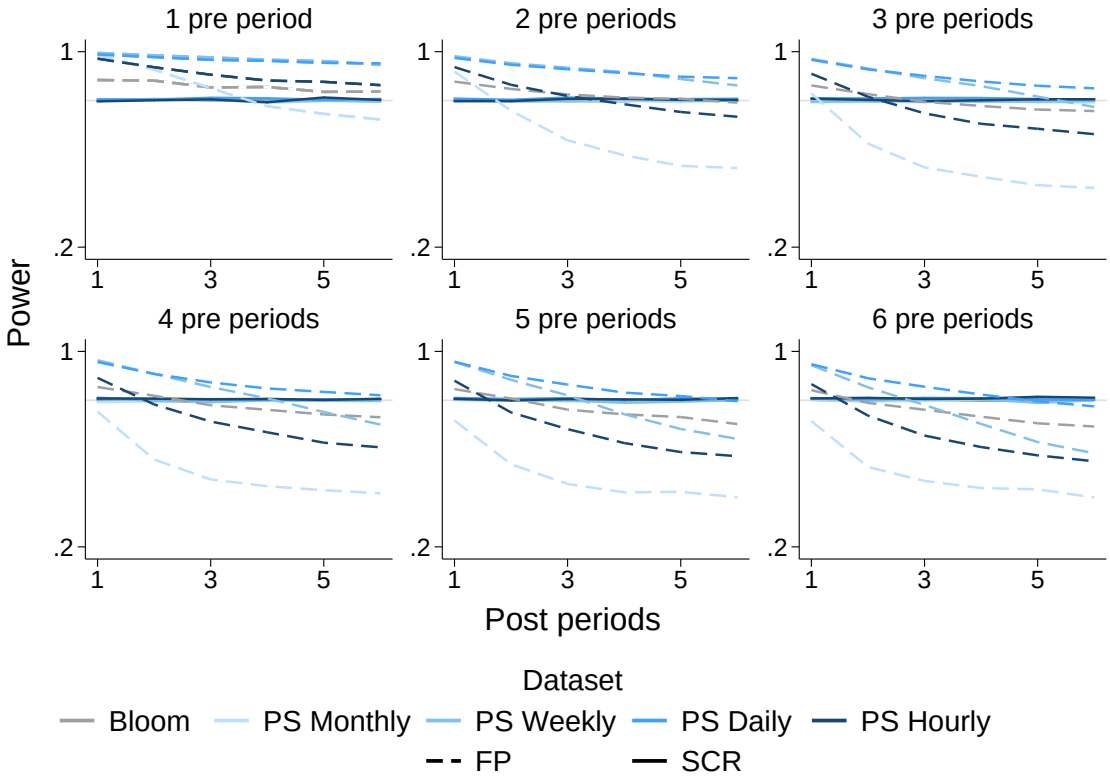
In the presence of non-zero serial correlation ($\gamma > 0$), the "naive" FP ANCOVA power calculation formula yields overpowered experiments in shorter panels, and underpowered experiments in longer panels. This is consistent with the panel DD results from Figure B.3.1, and demonstrates that simply implementing an ANCOVA design under the i.i.d. assumption will not ensure a correctly powered experiment.

Figure B.3.1: Power in short panels – AR(1) data



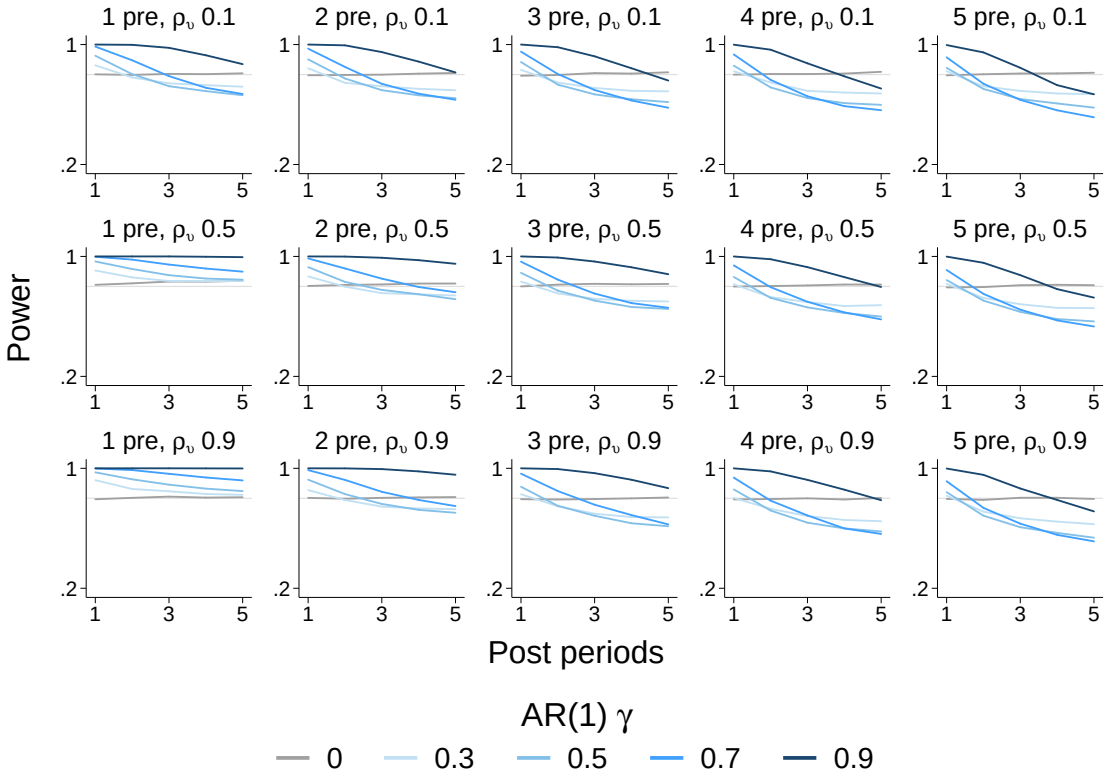
Notes: This figure displays realized power from performing power calculations using Frison and Pocock (1992)'s formula (Equation (3.3)) and the serial-correlation-robust formula (Equation (3.6)) to calibrate the treatment effect size. We cluster standard errors *ex post* in all cases, following Bertrand, Duflo, and Mullainathan (2004). We separately vary the number of the pre-treatment and post-treatment periods, where differing pre-period lengths are shown in different panels, and differing post-period lengths are shown along each panel's horizontal axis. For all cases with either one pre-treatment period or one post-treatment period, the FP formula yields over-powered experiments across the full range of positive AR(1) parameters. Experiments that follow the traditional "one baseline, one follow-up" structure will be overpowered, having calibrated an excessively large sample size. As the number of pre-treatment periods increases, power decreases monotonically for all $\gamma > 0$. At the same time, the serial-correlation-robust formula is properly powered in all cases.

Figure B.3.2: Power in short panels – Real data



Notes: This figure displays realized power from performing power calculations using Frison and Pocock (1992)'s formula (Equation (3.3)) and the serial-correlation-robust formula (Equation (3.6)) to calibrate the treatment effect size. We cluster standard errors *ex post* in all cases, following Bertrand, Duflo, and Mullainathan (2004). We separately vary the number of the pre-treatment and post-treatment periods, where differing pre-period lengths are shown in different panels, and differing post-period lengths are shown along each panel's horizontal axis. We show results for five different datasets: the Bloom et al. (2015) data, as well as monthly, weekly, daily, and hourly Pecan Street data. For nearly all cases with either one pre-treatment period or one post-treatment period, the FP formula yields over-powered experiments. Experiments that follow the traditional "one baseline, one follow-up" structure will be overpowered, having calibrated an excessively large sample size. As the number of pre-treatment periods increases, power decreases monotonically for each dataset. By contrast, the serial-correlation-robust formula is properly powered in all cases.

Figure B.3.3: Traditional ANCOVA methods fail to achieve desired power



Notes: This figure displays realized power from performing power calculations using Frison and Pocock (1992)'s ANCOVA formula (Equation (B.2)) to calibrate the treatment effect size. We cluster standard errors ex post in all cases, following Bertrand, Duflo, and Mullainathan (2004), and estimate treatment effects using the ANCOVA estimating Equation (3.13). We separately vary the number of pre-treatment periods (by column), the number of post-treatment periods (along each panel's horizontal axis), and the intra-cluster correlation of the pre-period, the length of the post-period, and the intracluster correlation coefficient (by by row). As with Figure B.3.1, applying the FP ANCOVA formula in the presence of serial correlation will lead to incorrectly experiments, and experiments with either one pre- or one post-treatment period will be overpowered.

B.4 A practical guide to power calculations

In this section, we address several practical considerations when conducting power calculations. Most of these challenges involve variance and covariance parameters that must be either estimated or assumed in order to operationalize a power calculation formula. We also outline steps for estimating power calculations via simulation, which is our preferred method but requires a representative pre-existing dataset.

B.4.1 Analytical power calculations

The most challenging aspect of analytical power calculations is parameterizing the variance and (if applicable) covariance terms that characterize the data's error structure. In the absence of a representative pre-existing dataset, researchers may struggle to even guess the order of magnitude of the error variance, let alone generate a precise estimate of this key parameter. Our theoretical results demonstrate that the error covariance structure is likewise key to determining the statistical power of panel RCTs. As in the paper, we denote the true parameters governing the data generating process as σ^2_ω, ψ^B, ψ^A, and ψ^X. We define $\sigma^2_{\hat{\omega}}$, $\psi^B_{\hat{\omega}}$, $\psi^A_{\hat{\omega}}$, and $\psi^X_{\hat{\omega}}$ to be the parameters that characterize the *residuals* (rather than real errors). If researchers do have access to a representative dataset when performing *ex ante* power calculations, they can directly estimate $\sigma^2_{\hat{\omega}}$, $\psi^B_{\hat{\omega}}$, $\psi^A_{\hat{\omega}}$, and $\psi^X_{\hat{\omega}}$, and use these values to parameterize power calculations. Section B.5.2 proves that researchers can recover the true MDE using these residual-based parameters. Nevertheless, this process is not trivial, for several reasons.

First, while the idiosyncratic variance σ^2_ω is a population parameter, the three ψ parameters are functions of both the full covariance structure of the population and the specific values of m and r.[14] For a given population and serially correlated outcome variable, experiments with small m and r are likely to exhibit larger ψ^B, ψ^A, and ψ^X parameters than experiments with large m and r. This is because as the number of pre-treatment (post-treatment) periods increases, ψ^B (ψ^A) averages across covariances of time periods that are farther apart. For example, compare ψ^B with $m = 3$ vs. $m = 30$, for an outcome with a covariance structure where adjacent periods are more positively correlated than distant periods. For $m = 3$, ψ^B averages $m(m-1)/2 = 3$ pairwise covariances, 2 of which are for adjacent periods; for $m = 30$, ψ^B averages $m(m-1)/2 = 435$ pairwise covariances, only 29 of which are for adjacent

[14] Deriving the residual-based parameters $\sigma^2_{\hat{\omega}}$, $\psi^B_{\hat{\omega}}$, $\psi^A_{\hat{\omega}}$, and $\psi^X_{\hat{\omega}}$ introduces an additional complexity, as these residual-based parameters are defined by the number of pre-treatment periods (m), post-treatment periods (r) and cross-sectional units (I) used to produce these residuals.

periods. Because ψ^X expands with both m and r, it attenuates relatively faster than ψ^B and ψ^A as panel length grows.

Second, estimating $\text{Cov}(\hat{\omega}_{it}, \hat{\omega}_{is})$ using residuals from an existing dataset is fundamentally impossible, given that each dataset contains only one realization of (Y_{it}, Y_{is}). However, researchers may treat the $(I \times 1)$ vectors of residuals $(\vec{\hat{\omega}}_t, \vec{\hat{\omega}}_s)$ as I draws from the distributions of residuals for periods (t,s) and estimate these distributions' covariance. The resulting estimates, which we denote $\tilde{\sigma}^2_{\hat{\omega}}$, $\tilde{\psi}^B_{\hat{\omega}}$, $\tilde{\psi}^A_{\hat{\omega}}$, and $\tilde{\psi}^X_{\hat{\omega}}$, are unbiased estimators of $\sigma^2_{\hat{\omega}}$, $\psi^B_{\hat{\omega}}$, $\psi^A_{\hat{\omega}}$, and $\psi^X_{\hat{\omega}}$.[15]

Third, if the representative dataset contains a long time series, the residual variance and covariance structure may change throughout the time series. This means if researchers estimate $\tilde{\sigma}^2_{\hat{\omega}}$, $\tilde{\psi}^B_{\hat{\omega}}$, $\tilde{\psi}^A_{\hat{\omega}}$, and $\tilde{\psi}^X_{\hat{\omega}}$ by averaging across the full time series, these estimated parameters may be less representative than if they were estimated from just the end of the time series.[16] Furthermore, because the residual variance is not a function of panel length, it may be tempting to estimate $\tilde{\sigma}^2_{\hat{\omega}}$ using a long vector of residuals, while estimating $\tilde{\psi}^B_{\hat{\omega}}$, $\tilde{\psi}^A_{\hat{\omega}}$, and $\tilde{\psi}^X_{\hat{\omega}}$ using only residuals within an $(m + r)$-period range. In a time series where the variance-covariance structure is changing, this would produce $\tilde{\psi}_{\hat{\omega}}$ estimates that are inconsistent with $\tilde{\sigma}^2_{\hat{\omega}}$.

Fourth, while $\tilde{\sigma}^2_{\hat{\omega}}$, $\tilde{\psi}^B_{\hat{\omega}}$, $\tilde{\psi}^A_{\hat{\omega}}$, and $\tilde{\psi}^X_{\hat{\omega}}$ are unbiased estimators of $\sigma^2_{\hat{\omega}}$, $\psi^B_{\hat{\omega}}$, $\psi^A_{\hat{\omega}}$, and $\psi^X_{\hat{\omega}}$, they are **n**ot unbiased estimators of σ^2_{ω}, ψ^B, ψ^A, and ψ^X. This is because the residuals from the regression $Y_{it} = \upsilon_i + \delta_t + \omega_{it}$ will have a variance less than the parameter σ^2_{ω} from the data generating process, by the properties of linear projection. In addition, when they are estimated using residuals from shorter panels, $\tilde{\sigma}^2_{\hat{\omega}}$, $\tilde{\psi}^B_{\hat{\omega}}$, $\tilde{\psi}^A_{\hat{\omega}}$, and $\tilde{\psi}^X_{\hat{\omega}}$ have a more severe bias, but these estimates converge to their true values (i.e., σ^2_{ω}, ψ^B, ψ^A, and ψ^X) as the panel length used to the estimate these residuals increases.[17] Importantly, for the purposes of power calculations using the SCR formula, we *can* recover an unbiased estimate of the minimum detectable effect with the true parameters using our parameter estimates. That is, $MDE^{est}(\sigma^2_{\hat{\omega}}, \psi^B_{\hat{\omega}}, \psi^A_{\hat{\omega}}, \psi^X_{\hat{\omega}}) = MDE(\sigma^2_{\omega}, \psi^B, \psi^A, \psi^X)$. As Section B.5.1 shows, $\text{E}[\tilde{\sigma}^2_{\hat{\omega}} \mid \mathbf{X}] = \sigma^2_{\hat{\omega}}$, $\text{E}[\tilde{\psi}^B_{\hat{\omega}} \mid \mathbf{X}] = \psi^B_{\hat{\omega}}$, $\text{E}[\tilde{\psi}^A_{\hat{\omega}} \mid \mathbf{X}] = \psi^A_{\hat{\omega}}$, and $\text{E}[\tilde{\psi}^X_{\hat{\omega}} \mid \mathbf{X}] = \psi^X_{\hat{\omega}}$. Combining these two proofs suggests that

[15] Appendix B.5.1 proves that $\text{E}[\tilde{\sigma}^2_{\hat{\omega}} \mid \mathbf{X}] = \sigma^2_{\hat{\omega}}$, $\text{E}[\tilde{\psi}^B_{\hat{\omega}} \mid \mathbf{X}] = \psi^B_{\hat{\omega}}$, $\text{E}[\tilde{\psi}^A_{\hat{\omega}} \mid \mathbf{X}] = \psi^A_{\hat{\omega}}$, and $\text{E}[\tilde{\psi}^X_{\hat{\omega}} \mid \mathbf{X}] = \psi^X_{\hat{\omega}}$.

[16] Of course, if the researcher knows that certain parts of her data are more likely to represent the experimental timeframe and data, it would be wise to perform power calculations on this subset alone.

[17] The estimated residuals include both the true idiosyncratic error, ω_{it}, and (attenuating) fixed-effect estimation error. Although both sets fixed effects, υ_i and δ_t, are unbiased and consistent in T and I, respectively, error in estimating these parameters will always yield residuals that are smaller on average, biasing the estimation of these parameters. The estimation error and resulting biases decrease in T and I.

$MDE^{est}(\mathrm{E}[\tilde{\sigma}^2_{\hat{\omega}} \mid \mathbf{X}], \mathrm{E}[\tilde{\psi}^B_{\hat{\omega}} \mid \mathbf{X}], \mathrm{E}[\tilde{\psi}^A_{\hat{\omega}} \mid \mathbf{X}], \mathrm{E}[\tilde{\psi}^X_{\hat{\omega}} \mid \mathbf{X}]) = MDE(\sigma^2_{\omega}, \psi^B, \psi^A, \psi^X)$. Therefore, for values of $\tilde{\sigma}^2_{\hat{\omega}}$, $\tilde{\psi}^B_{\hat{\omega}}$, $\tilde{\psi}^A_{\hat{\omega}}$, and $\tilde{\psi}^X_{\hat{\omega}}$ estimated from a pre-existing dataset with I cross-sectional units:

$$MDE^{est} = \left(t^J_{1-\kappa} + t^J_{\alpha/2}\right)\left\{\left(\frac{1}{P(1-P)J}\right)\left[\left(\frac{m+r}{mr}\right)k_\sigma \,\mathrm{E}\left[\tilde{\sigma}^2_{\hat{\omega}} \mid \mathbf{X}\right] + \left(\frac{m-1}{m}\right)k_B \,\mathrm{E}\left[\tilde{\psi}^B_{\hat{\omega}} \mid \mathbf{X}\right]\right.$$

$$\left. + \left(\frac{r-1}{r}\right)k_A \,\mathrm{E}\left[\tilde{\psi}^A_{\hat{\omega}} \mid \mathbf{X}\right] - 2k_X \,\mathrm{E}\left[\tilde{\psi}^X_{\hat{\omega}} \mid \mathbf{X}\right]\right]\right\}^{1/2} \tag{B.1}$$

$$= \left(t^J_{1-\kappa} + t^J_{\alpha/2}\right)\left\{\left(\frac{1}{P(1-P)J}\right)\left[\left(\frac{m+r}{mr}\right)k_\sigma \sigma^2_{\hat{\omega}} + \left(\frac{m-1}{m}\right)k_B \psi^B_{\hat{\omega}}\right.\right.$$

$$\left.\left. + \left(\frac{r-1}{r}\right)k_A \psi^A_{\hat{\omega}} - 2k_X \psi^X_{\hat{\omega}}\right]\right\}^{1/2} \tag{B.2}$$

where

$$k_\sigma = \frac{I(m+r)^2}{2(I-1)mr}$$
$$k_B = \frac{I(m+r)^2}{2(I-1)r^2}$$
$$k_A = \frac{I(m+r)^2}{2(I-1)m^2}$$
$$k_X = 0$$

and the expectation of parameters are taken over subsets of the dataset, as described in the next point. Appendix B.5.2 proves that Equations (B.1) and (B.2) are equivalent, and we derive the above expressions for the coefficients k_σ, k_B, k_A, and k_X in Appendix B.5.2.

Fifth, because the estimated variance-covariance terms enter the power calculation under a radical, researchers must be conscious of Jensen's Inequality. If the researcher is estimating $\sigma^2_{\hat{\omega}}$, $\psi^B_{\hat{\omega}}$, $\psi^A_{\hat{\omega}}$, and $\psi^X_{\hat{\omega}}$ by taking the expectation of $\tilde{\sigma}^2_{\hat{\omega}}$, $\tilde{\psi}^B_{\hat{\omega}}$, $\tilde{\psi}^A_{\hat{\omega}}$, and $\tilde{\psi}^X_{\hat{\omega}}$ across a range of $(m+r)$-period subsets, then the correct calculation is:

$$MDE^{est}\left(\mathrm{E}\left[\tilde{\sigma}^2_{\hat{\omega}} \mid \mathbf{X}\right], \mathrm{E}\left[\tilde{\psi}^B_{\hat{\omega}} \mid \mathbf{X}\right], \mathrm{E}\left[\tilde{\psi}^A_{\hat{\omega}} \mid \mathbf{X}\right], \mathrm{E}\left[\tilde{\psi}^X_{\hat{\omega}} \mid \mathbf{X}\right]\right),$$

not

$$E\left[MDE^{est}\left(\tilde{\sigma}^2_{\hat{\omega}}, \tilde{\psi}^B_{\hat{\omega}}, \tilde{\psi}^A_{\hat{\omega}}, \tilde{\psi}^X_{\hat{\omega}}\right) \mid \mathbf{X}\right].$$

Similarly, if Equation (3.6) is rearranged as a function of κ, it becomes convex in the variance-covariance parameters, and the correct calculation is:

$$\kappa\left(\mathrm{E}\left[\tilde{\sigma}^2_{\hat{\omega}} \mid \mathbf{X}\right], \mathrm{E}\left[\tilde{\psi}^B_{\hat{\omega}} \mid \mathbf{X}\right], \mathrm{E}\left[\tilde{\psi}^A_{\hat{\omega}} \mid \mathbf{X}\right], \mathrm{E}\left[\tilde{\psi}^X_{\hat{\omega}} \mid \mathbf{X}\right]\right),$$

not

$$\mathrm{E}\left[\kappa\left(\tilde{\sigma}^2_{\tilde{\omega}}, \tilde{\psi}^B_{\tilde{\omega}}, \tilde{\psi}^A_{\tilde{\omega}}, \tilde{\psi}^X_{\tilde{\omega}}\right) \mid \mathbf{X}\right].$$

When solving for sample size J, Equation (3.6) becomes linear in variance-covariance parameters, meaning that Jensen's Inequality does not affect the estimate of

$$J\left(\sigma^2_{\tilde{\omega}}, \psi^B_{\tilde{\omega}}, \psi^A_{\tilde{\omega}}, \psi^X_{\tilde{\omega}}\right)$$

.

In light of each of these issues, we recommend the following algorithm for estimating the MDE using a pre-existing panel dataset:[18]

1. Determine all feasible ranges of experiments with $(m + r)$ periods, given the number of time periods in the pre-existing dataset. For example, if this dataset contains 100 time periods indexed $t = \{1, \ldots, 100\}$, and $m = 5$ and $r = 6$, then there are 90 feasible ranges for an experiment with $(m + r) = 11$ periods (i.e., beginning in periods $t = \{1, \ldots, 90\}$).

2. For each feasible range S:

   a) Regress the outcome variable on unit and time-period fixed effects, $Y_{it} = \upsilon_i + \delta_t + \omega_{it}$, and store the residuals. (This regression includes all I available cross-sectional units, but only time periods with the specific range S.[19])

   b) Calculate the variance of the stored residuals, and save as $\tilde{\sigma}^2_{\tilde{\omega},S}$.

   c) For each pair of pre-treatment periods (i.e., the first m periods in range S), calculate the covariance between these periods' residuals. Take an unweighted average of these $m(m-1)/2$ covariances, and save as $\tilde{\psi}^B_{\tilde{\omega},S}$.

   For example, if $m = 4$, $r = 2$, and range S begins in period $t = 1$, sum $\mathrm{Cov}(\omega_{i1}, \omega_{i2})$, $\mathrm{Cov}(\omega_{i1}, \omega_{i3})$, $\mathrm{Cov}(\omega_{i1}, \omega_{i4})$, $\mathrm{Cov}(\omega_{i2}, \omega_{i3})$, $\mathrm{Cov}(\omega_{i2}, \omega_{i4})$, and $\mathrm{Cov}(\omega_{i3}, \omega_{i4})$, and divide by $m(m-1)/2 = 6$.

   d) For each pair of post-treatment periods (i.e., the last r periods in range S), calculate the covariance between these periods' residuals. Take an unweighted average of these $r(r-1)/2$ covariances, and save as $\tilde{\psi}^A_{\tilde{\omega},S}$.

   For example, if $m = 4$, $r = 2$, and range S begins in period $t = 1$, $\tilde{\psi}^A_{\tilde{\omega},S}$ is the average of a single post-period covariance, $\mathrm{Cov}(\omega_{i5}, \omega_{i6})$.

[18] Our accompanying software packages implement this algorithm using the programs `pc_dd_covar` (in STATA) and `PCDDCovar` (in `R`).

[19] This bears no relationship to the sample size J units to be included in the power calculation. Assuming that all I units in the pre-existing dataset represent the population to be included in the randomization, estimating the variance and covariances using all available units will provide the best estimates of $\tilde{\sigma}^2_{\tilde{\omega}}$, $\tilde{\psi}^B_{\tilde{\omega}}$, $\tilde{\psi}^A_{\tilde{\omega}}$, and $\tilde{\psi}^X_{\tilde{\omega}}$ (by the weak law of large numbers).

e) For each pair of pre- and post-treatment periods (i.e. the first m and the last r periods in range S), calculate the covariance between these periods' residuals. Take an unweighted average of these mr covariances, and save as $\tilde{\psi}^X_{\hat{\omega},S}$.

For example, if $m = 4$, $r = 2$, and range S begins in period $t = 1$, sum $\text{Cov}(\omega_{i1}, \omega_{i5})$, $\text{Cov}(\omega_{i1}, \omega_{i6})$, $\text{Cov}(\omega_{i2}, \omega_{i5})$, $\text{Cov}(\omega_{i2}, \omega_{i6})$, $\text{Cov}(\omega_{i3}, \omega_{i5})$, $\text{Cov}(\omega_{i3}, \omega_{i6})$, $\text{Cov}(\omega_{i4}, \omega_{i5})$, and $\text{Cov}(\omega_{i4}, \omega_{i6})$, and divide by $mr = 8$.

3. Calculate the average of $\tilde{\sigma}^2_{\hat{\omega},S}$, $\tilde{\psi}^B_{\hat{\omega},S}$, $\tilde{\psi}^A_{\hat{\omega},S}$, and $\tilde{\psi}^X_{\hat{\omega},S}$ across all ranges S, deflating $\tilde{\sigma}^2_{\hat{\omega},S}$ by $\frac{IT-1}{IT}$, and $\tilde{\psi}^B_{\hat{\omega},S}$, $\tilde{\psi}^A_{\hat{\omega},S}$, and $\tilde{\psi}^X_{\hat{\omega},S}$ by $\frac{I-1}{I}$. These averages are equal in expectation to $\sigma^2_{\hat{\omega}}$, $\psi^B_{\hat{\omega}}$, $\psi^A_{\hat{\omega}}$, and $\psi^X_{\hat{\omega}}$.

4. Plug these values into Equation (B.2) to produce MDE^{est}.

Figure B.4.1 shows the difference between true and estimated variance-covariance parameters in AR(1) data. In particular, we show true parameter values of $\sigma^2_\omega, \psi^B, \psi^A$, and ψ^X alongside estimated values of these same parameters, calculated according to the procedure outlined above. As expected, $\sigma^2_{\hat{\omega}}$ is biased downwards relative to σ^2_ω, but converges towards this value as the panel length increases. This convergence is slower for larger AR(1) parameters, as highly serially correlated errors make it harder to identify the unit fixed effects. Similarly, while the true ψ^X is positive across all panel lengths, $\psi^X_{\hat{\omega}}$ is negative everywhere, and ψ^B and ψ^A also differ from their estimated counterparts. Despite the differences between the true parameters and their estimated values, Section B.5.2 proves that we can recover the MDE based on true underlying parameters using residual-based parameters. In conjunction with the fact that we can estimate the residual-based parameters from real data, this confirms that researchers can use estimated parameters to calibrate power calculations.

Figure B.4.2 uses the Bloom et al. (2015) dataset to present an analogous comparison between actual vs. estimated σ^2_ω, $\psi^{B,A}$, and ψ^X parameters. Here, as in Figure B.4.1, the dotted lines estimate $\sigma^2_{\hat{\omega}}$, $\psi^{B,A}_{\hat{\omega}}$, and $\psi^X_{\hat{\omega}}$ using the above algorithm. However, unlike with simulated AR(1) datasets, the "true" parameters of the Bloom et al. (2015) data generating process are unknown. We estimate these "true" values using residuals from the full 48-period time series, which minimizes the fixed effect estimation error that biases $\sigma^2_{\hat{\omega}}$, $\psi^B_{\hat{\omega}}$, $\psi^A_{\hat{\omega}}$, and $\psi^X_{\hat{\omega}}$ in short panels.[20] This reveals a very similar pattern: "subsetted" $\sigma^2_{\hat{\omega}}$, $\psi^B_{\hat{\omega}}$, $\psi^A_{\hat{\omega}}$, and $\psi^X_{\hat{\omega}}$ estimates are systematically

[20]These $\sigma^2_{\hat{\omega}}$, $\psi^B_{\hat{\omega}}$, $\psi^A_{\hat{\omega}}$, and $\psi^X_{\hat{\omega}}$ estimates (represented by solid lines in Figure B.4.2) result from the same algorithm as detailed above, except omitting Step 2(a) and estimating a single 48-period set of residuals in Step 1. This provides the closest possible approximation to the "true" variance-covariance structure of these data, and hence the most apples-to-apples comparison to Figure B.4.1.

biased downward, but converge to their "full time-series" (i.e. closer to "true") values as panel length increases. As in Figure B.4.1, we show that both sets of estimated variance-covariance parameters yield (virtually) identical MDEs, as long as Equation (3.6) uses estimated parameters that are internally consistent.

Figure B.4.3 replicates Figure B.4.2 for all four Pecan Street datasets. We see that while the estimated values $\sigma^2_{\hat{\omega}}$, $\psi^B_{\hat{\omega}}$, $\psi^A_{\hat{\omega}}$, and $\psi^X_{\hat{\omega}}$ differ across different levels of aggregation, they follow the same pattern. The subsetted estimates are biased downward, but appear to converge to the full time series estimates (i.e. closer to truth) as panel length increases. In all four cases, the MDE is (virtually) identical when calculated using either all full time series estimates or all subsetted estimates.

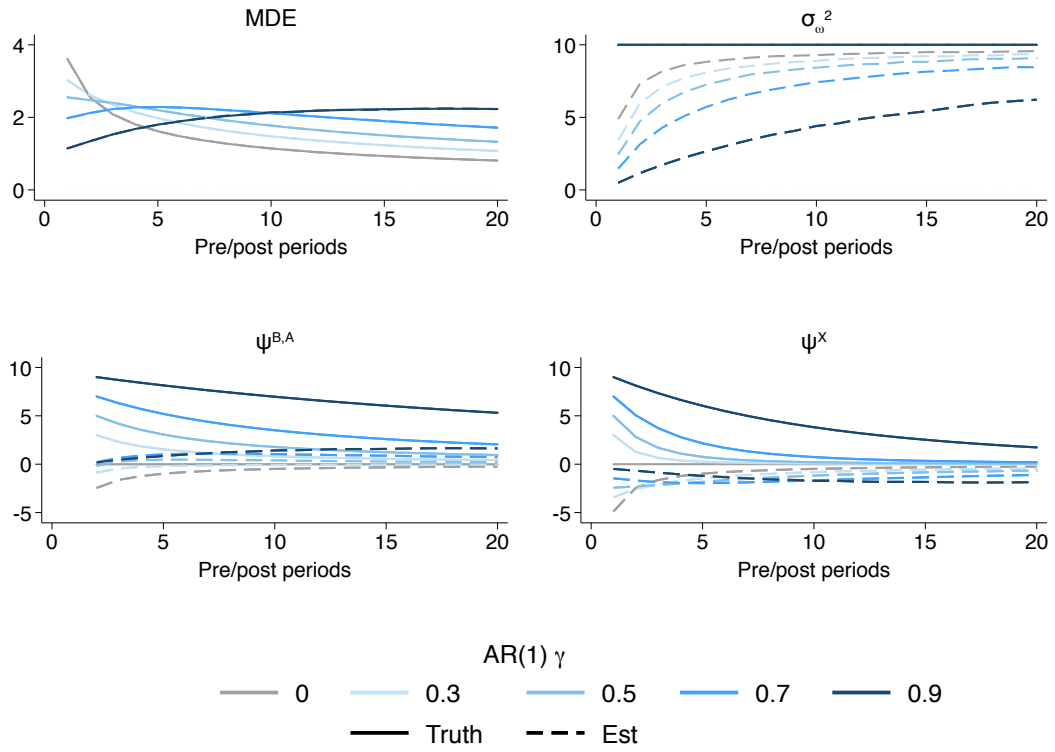
Two additional nuances that arise during analytical power calculations are worth noting. First, the critical values $t^d_{1-\kappa}$ and $t^d_{\alpha/2}$ should be drawn from an inverse t-distribution with the same degrees of freedom as the *ex post* regression model. This means that if researchers plan to use CRVE standard errors clustered by experimental unit, they should draw these critical values from an inverse t-distribution with J degrees of freedom. To be precise, these critical values should be sensitive to changes in the number of unit/clusters J, although the t degrees of freedom has a very small effect on MDE, relative to other parameters in Equation (3.6).

Finally, in panel RCTs with CRVE standard errors clustered by unit, the proportion of units treated P cannot be too large or too small. Our simulations have demonstrated that the Equation (3.6) performs poorly is $P < 0.1$ or $P > 0.9$, because the CRVE requires a sufficient number of clusters that are both treated and control.

This paper's accompanying software packages allow researchers to conduct analytical power calculations for difference-in-differences RCTs, via the functions `pc_dd_analytic` in STATA and `PCDDAnalytic` in `R`.[21] These programs implement the estimation version of the serial-correlation-robust power calculation formula, Equation (B.2), allowing researchers to solve for effect size MDE, sample size J, or power κ, as functions of all other parameters. Users may provide assumed variance-covariance parameters, or alternatively our software will estimate these parameters directly by applying the above procedure to a dataset stored in memory.

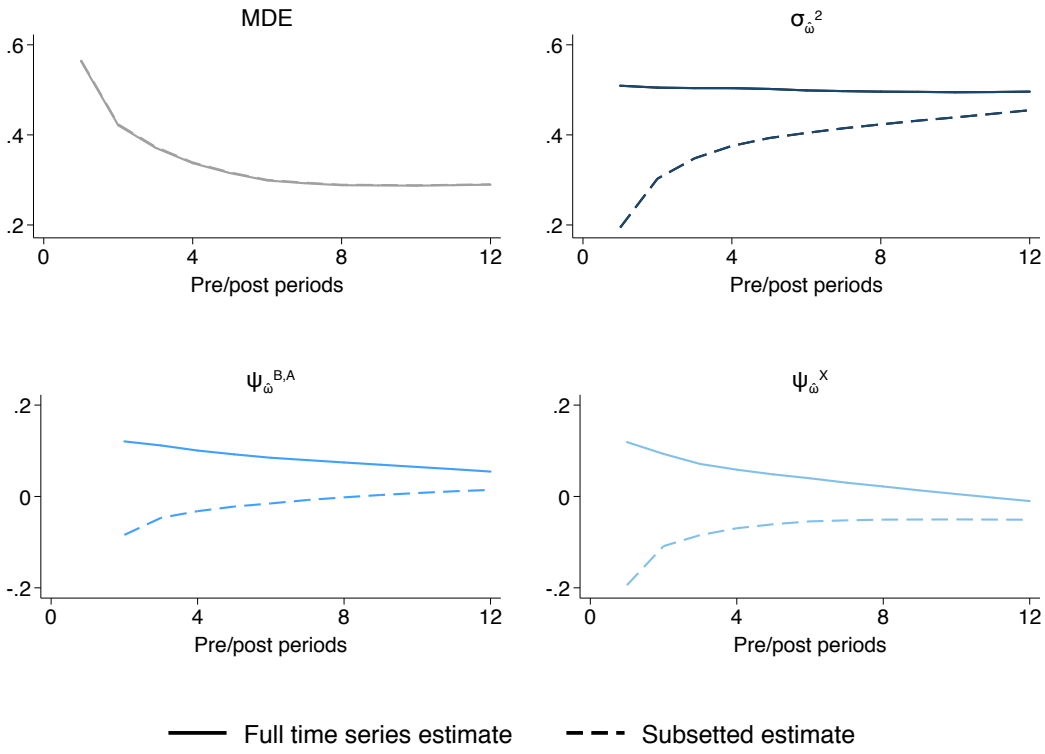
[21] See Section B.4.4 for details on acquiring this software.

Figure B.4.1: Actual vs. estimated parameters – AR(1) data



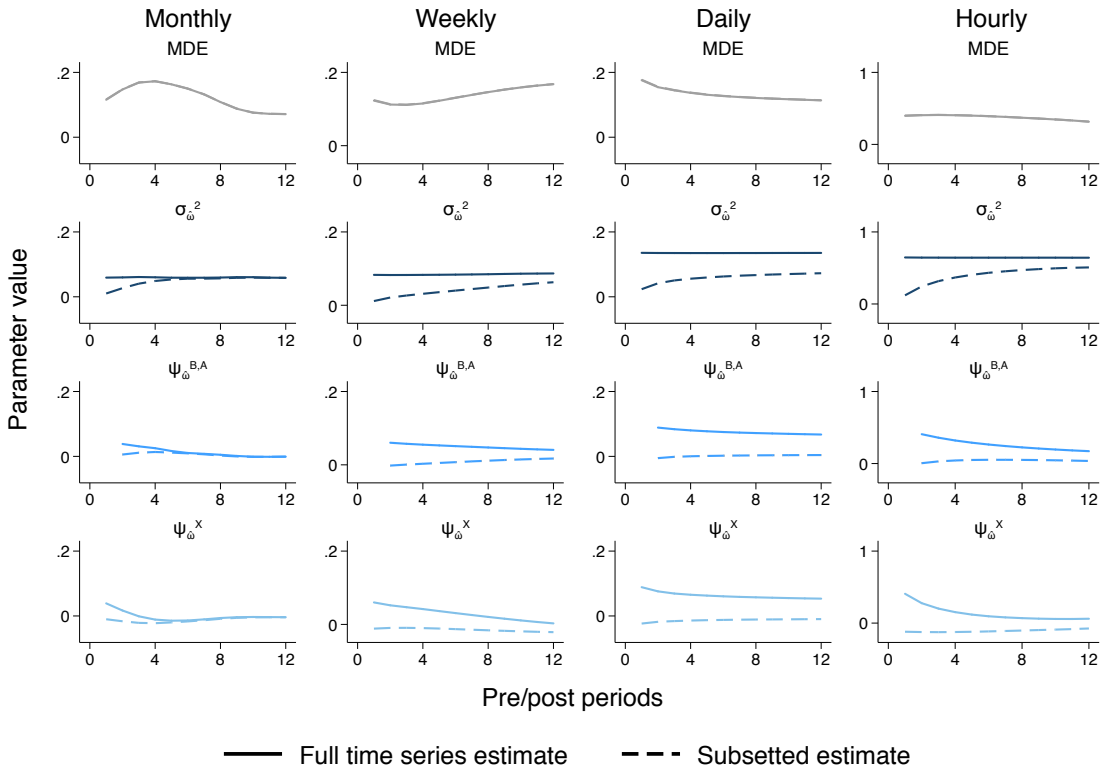
Notes: This figure displays the difference between the true residual variance (σ^2_ω), average pre (post)-period covariance ($\psi^{B,A}$), and average cross-period covariance (ψ^X), and their estimated counterparts over varying panel lengths. This figure also shows the resulting minimum detectable effect (MDE), calculated using the serial-correlation-robust power calculation formula (Equation (3.6)). These parameters and estimates come from a simulated datasets with AR(1) errors, generated identically to those presented in Figure 3.3.2. True parameters are displayed with solid lines (with ψ terms derived analytical using Equations (B.3)–(B.5)), and estimates are displayed with dashed lines (estimated according to the algorithm described above). As expected, the true σ^2_ω is constant across panel lengths, while $\psi^{B,A}$ declines with the number of pre (post) periods, and ψ^X declines more quickly than $\psi^{B,A}$ as the panel length increases. The higher the AR(1) parameter, the larger are the ψ terms. The estimated parameters behave quite differently from their true counterparts. In short panels, $\sigma^2_{\hat{\omega}}$ is biased downwards, because the regression model's individual fixed effects are inconsistently estimated, and capture more of the true error variance than they should explain in expectation. This has the effect of reducing the estimated covariances $\psi^{B,A}_{\hat{\omega}}$, which scale with $\sigma^2_{\hat{\omega}}$. At the same time, $\psi^X_{\hat{\omega}}$ is mechanically negative, as the estimated fixed effects yield residuals that are negatively correlated within individuals across pre/post-treatment time periods (and some of the variation that should be captured by $\sigma^2_{\hat{\omega}}$ is instead loaded on to the $\psi_{\hat{\omega}}$ terms). Note that as the panel length increases, the $\psi_{\hat{\omega}}$ terms converge to the true ψ values. Importantly, despite the fact that the estimated parameters are different from the true parameters, these differences offset such that both estimated and real parameters result in the same MDEs, as demonstrated by the top left panel.

Figure B.4.2: Estimated parameters - Bloom et al. (2015) data



Notes: This figure shows two different methods for estimating variance parameters, applied to the Bloom et al. (2015) data, and also depicts the resulting minimum detectable effect (MDE) resulting from both methods. The solid lines show parameters calculated by running a regression of the outcome variable on unit and time period fixed effects, estimated on the entire time series of data, and then using the residuals from this regression to calculate $\sigma^2_{\hat{\omega}}$, $\psi^B_{\hat{\omega}}$, $\psi^A_{\hat{\omega}}$, and $\psi^X_{\hat{\omega}}$ for the average panel of length $m+r$. We then plug these estimates into Equation (3.6) to calculate the minimum detectable effect. The dashed lines show parameters estimated using the procedure described above, where rather than use residuals from the full time series, we subset the dataset into shorter panels of length $m+r$, calculate the parameters using residuals only from this subset, and average across all possible subsets to arrive at $\sigma^2_{\hat{\omega}}$, $\psi^B_{\hat{\omega}}$, $\psi^A_{\hat{\omega}}$, and $\psi^X_{\hat{\omega}}$. We calculate the MDE by plugging these estimates into Equation (B.2). Note that these variance-covariance estimates converge as the panel length increases. Both procedures yield (virtually) identical MDEs, even though the underlying parameter estimates differ substantially.

Figure B.4.3: Estimated parameters - Pecan Street data



Notes: This figure shows two different methods for estimating variance parameters, applied to the Pecan Street data at the four levels of aggregation presented in the main text, and also depicts the resulting minimum detectable effect (MDE) resulting from both methods. Note that the y axis scale differs between the hourly data and the other three datasets; this is because the degree of residual variation left in the hourly data after removing time and individual fixed effects is much greater than in the other datasets, which is to be expected. The solid lines show parameters calculated by running a regression of the outcome variable on unit and time period fixed effects, estimated on the entire time series of data, and then using the residuals from this regression to calculate $\sigma^2_{\hat{\omega}}$, $\psi^B_{\hat{\omega}}$, $\psi^A_{\hat{\omega}}$, and $\psi^X_{\hat{\omega}}$ for the average panel of length $m+r$. We then plug these estimates into Equation (3.6) to calculate the minimum detectable effect. By contrast, the dashed lines show parameters estimated using the procedure described above, where rather than use residuals from the full time series, we subset the dataset into shorter panels of length $m+r$, calculate the parameters using residuals only from this subset, and average across all possible subsets to arrive at $\sigma^2_{\hat{\omega}}$, $\psi^B_{\hat{\omega}}$, $\psi^A_{\hat{\omega}}$, and $\psi^X_{\hat{\omega}}$. We calculate the MDE by plugging these estimates into Equation (B.2). Note that these variance-covariance estimates converge as the panel length increases. Both procedures yield (virtually) identical MDEs, even though the two procedures' method for estimating the underlying parameters differ substantially.

B.4.2 Simulation-based power calculations

In cases where researchers have access to a representative pre-existing dataset, we recommend that they perform power calculations via simulation. This obviates the need to estimate *ex ante* variance-covariance parameters, and it ensures that *ex ante* power calculations assume the same experimental design, regression model, and variance estimator expected to be used in *ex post* analysis. Our accompanying software packages perform simulation-based power calculations, using the programs `pc_simulate` in STATA and `PCSimulate` in `R`. These programs implement the following algorithm:

1. Choose the following candidate parameters: sample size J, pre-treatment periods m, and post-treatment periods r, treatment ratio P, minimum detectable effect MDE, and significance level α. Let X_{it} denote the outcome variable of interest in the pre-existing dataset.

2. Randomly select J units from the representative dataset, and randomly select a range of $(m + r)$ consecutive time periods. This will serve as a simulated experimental dataset, with sample size J, m pre-treatment periods, and r post-treatment periods.

3. Randomly scramble a $[J \times 1]$ vector of PJ ones and $(1-P)J$ zeros, rounding PJ to the nearest integer. Assign each of the J units to either treatment $(D = 1)$ or control $(D = 0)$, based on the order of this scrambled vector.

4. Construct an experimental outcome variable Y_{it}, where $Y_{it} = X_{it} + MDE$ for treated units in post-treatment periods, and $Y_{it} = X_{it}$ otherwise.

5. Using this simulated experimental dataset and the simulated outcome variable Y_{it}, implement the exact regression specification and variance estimator to be used in *ex post* analysis. Record whether this model rejects the null hypotheses of zero treatment effects with significance level α (i.e. $H_0 : \tau = 0$).

6. Repeat Steps 2–5 many times, and calculate the rejection rate across all simulations. This is the experiment's statistical power as a function of J, m, r, P, MDE, and α.

7. Repeat Steps 1–6 for a range of MDEs and design parameters, increasing the number of simulations after narrowing down this range of parameters to more precisely calibrate power.

This algorithm allows users to test alternative regression specifications and alternative standard error assumptions, without needing to formally derive a power calculation expression for each model. If the pre-existing dataset contains fewer cross-sectional units than the desired sample size J, our software allows users to simulate additional units by bootstrapping units with replacement from the existing dataset (using the functions `pc_bootstrap_units` in STATA and `PCBootstrapUnits` in `R`). Unfortunately, if the pre-existing dataset contains fewer time periods than the desired panel length $(m + r)$, an analogous bootstrapping procedure would be much less straightforward (because unlike cross-sectional units, time periods are ordered and have a ordered covariance structure that is not orthogonal to the treatment vector D).

Importantly, this simulation-based algorithm can only calibrate statistical power κ. Rather than rely on the critical value $t^d_{1-\kappa}$, the algorithm simply estimates realized power as the proportion of simulations where the treatment effect is statistically distinguishable from zero. (By contrast, users may algebraically rearrange (or invert) an analytical power calculation formula to solve for any one of its parameters.) Calibrating simulation-based power calculations for a parameter other than κ necessitates a grid search over candidate parameter values, as described in Step 7 above. For example, to calibrate sample size J by simulation, users may repeat Step 1–6 over a range of candidate J values, narrowing this range (while simultaneously increasing the number of simulations) to calibrate to the desired power.

B.4.3 Lack of (representative) pre-existing data

To perform accurate *ex ante* power calculations, researchers must either have access to data that is representative (in expectation) of their future experimental data, or be able to parameterize an analytical power calculation with accurate estimates of the variance and covariance of the error structure. We recommend that researchers conduct power calculations via simulation (as described above), in cases where they have a representative pre-existing dataset with (i) data for the desired outcome (and relevant control variables); (ii) at least as many unique cross-sectional units as the desired experimental sample size; and (iii) a time series at least as long the desired experimental panel length. Many candidate experiments likely satisfy these criteria, such as when researchers partner with organizations that maintain historical databases on the desired population of experimental subjects.

At the same time, there are many cases where researchers cannot obtain representative data *ex ante*. This problem is not unique to panel data, as even the simple cross-sectional power calculation formula hinges on (an estimate of) the variance σ^2_ε. However, power calculations for panel RCT designs require four variance-covariance

parameters: σ^2_ω, ψ^B, ψ^A, and ψ^X. While σ^2_ω is fixed in the population, the ψ (and ψ_ω) terms are endogenous to the panel length of the experiment, which underscores the importance of estimating $\psi^B_{\hat{\omega}}$, $\psi^A_{\hat{\omega}}$, and $\psi^X_{\hat{\omega}}$ from a representative time series.

In the absence of representative data, we generally recommend using analytical formulas in conjunction with appropriate sensitivity analyses.[22] Depending on the type of data that *is* available, approximating the parameters σ^2_ω, ψ^B, ψ^A, and ψ^X may be possible. We consider three cases:

1. *Too few units:* If researchers have access to a representative pre-existing dataset with too few cross-sectional units, they may still estimate $\sigma^2_{\hat{\omega}}$, $\psi^B_{\hat{\omega}}$, $\psi^A_{\hat{\omega}}$, and $\psi^X_{\hat{\omega}}$, and apply these values to the (estimation-specific) analytic formula.[23] These variance-covariance parameters do *not* depend on sample size J in the SCR power calculation formula, and estimates of $\sigma^2_{\hat{\omega}}$, $\psi^B_{\hat{\omega}}$, $\psi^A_{\hat{\omega}}$, and $\psi^X_{\hat{\omega}}$ derived from residuals are *not* sensitive in expectation to the number of panel units J to be used in the experiment.[24] Alternatively, we recommend that researchers bootstrap units by sampling existing units with replacement, and use this expanded dataset (including simulated units) to conduct power calculations by simulation, as described above.[25]

2. *Too few time periods:* If the pre-existing dataset contains too few time periods, researchers may still estimate $\tilde{\sigma}^2_{\hat{\omega}}$ using residuals from a regression with fewer than $m + r$ periods (because σ^2_ω does not depend on panel length). However, the ψ terms do depend on panel length, and they cannot be estimated directly from a dataset with fewer than $m + r$ periods. One strategy is to simply estimate $\psi^B_{\hat{\omega}}$, $\psi^A_{\hat{\omega}}$, and $\psi^X_{\hat{\omega}}$ using the longest possible panel (i.e., all available time periods in the pre-existing dataset), even if it is shorter than $m+r$ periods.

[22]The alternative would be to impose assumptions on some existing data to construct a simulated representative dataset, which could then be used to conduct power calculations by simulation. (This may involve structural assumptions on the data generating process and/or assumptions on the representativeness of the best available pre-existing dataset.) However, this process will be much more computationally intensive than simply applying an analytical formula with appropriate parameter sensitivities. The exception is the case where users simulate additional cross-sectional units by bootstrapping an existing dataset (Case 1 here), which facilitates simulation-based power calculations using both real and simulated experiment units.

[23]It is important that researchers using *estimated* parameters use Equation (B.2) rather than Equation (3.6).

[24]Estimates of $\sigma^2_{\hat{\omega}}$, $\psi^B_{\hat{\omega}}$, $\psi^A_{\hat{\omega}}$, and $\psi^X_{\hat{\omega}}$ are sensitive to the number of cross-sectional units I used to estimate $\hat{\omega}_{it}$, but this is not related to the sample size parameter J. In Equation (B.2), the comparative static $d\psi_{\hat{\omega}}/dJ = 0$.

[25]Our accompanying software packages allow users to perform this bootstrapping algorithm using the functions `pc_bootstrap_units` in Stata and `PCBootstrapUnits` in `R`.

The resulting $\psi_{\hat{\omega}}$ estimates are likely to be upper bounds (in absolute value) on the $\psi_{\hat{\omega}}$ estimates for longer panels, because as the panel length increases, the ψ terms incorporate more covariances between time periods that are further apart (which tend to become less correlated in distance). Another strategy is to attempt to extend the time series for each unit, analogous to the approach of bootstrapping units. As a rule of thumb, researchers often approximate time series data as an AR(k) process with $k \geq \sqrt[3]{T}$, where T is the full time series length. To extend short panels, researchers may estimate this AR(k) process using (residuals from) the existing dataset, and then simulate forward for each unit's outcome realization. Neither of these strategies is perfect, and we recommend conducting appropriate sensitivity analyses in either case.

3. *No data:* In the complete absence of data, power calculations will be challenging. At the very least, we recommend that researchers search for estimates of the residual variance in the existing literature, noting that panel fixed effects models are likely to yield lower residual variances than cross-sectional models with similar outcome data. If this is not possible, researchers may iterate analytical power calculations over a range of parameter choices. If researchers are able to guess a reasonable value of σ^2_ω, they my test a range of AR(1) parameters for plausible values of $\psi^{B,A,X}$. As a rule of thumb, $\psi^{B,A,X}$ are likely to be positive in the absence of a strong prior of negative serial correlation. In absolute value, $\psi^{B,A,X}$ should not exceed σ^2_ω, and they should decrease monotonically in panel length. To provide a sense of what plausible values of ψ^B, ψ^A, and ψ^X (and their residual-based counterparts) may be, we plot estimates from a range of panel lengths using simulated AR(1) data, the Bloom et al. (2015) data, and Pecan Street data, in Figures B.4.1, B.4.2, and B.4.3, respectively.

B.4.4 Software packages

To facilitate user implementation of the methods described above, we are releasing accompanying software packages in Stata and in `R`. These packages each contain the following four programs:

`pc_simulate` (Stata), `PCSimulate` (`R`): Performs simulation-based power calculations using a pre-existing dataset. Accomodates four types of RCTs:

- **o**ne-shot (one wave of post-treatment data)
- **p**ost-only (multiple waves of post-treatment data)
- **d**ifference-in-difference (pre-treatment and post-treatment data)

- **A**NCOVA (post-treatment data, conditioning on pre-treatment data)

Calculates realized power over a set of user-provided design parameters. Allows for arbitrary linear regression specifications, linear controls, stratified randomization, and estimation of collapsed models.

`pc_bootstrap_units` **(STATA),** `PCBootstrapUnits` **(R):** Bootstraps additional units from an existing dataset, to allow for simulation-based power calculations user more cross-sectional units than are included in the pre-existing dataset.

`pc_dd_analytic` **(STATA),** `PCDDAnalytic` **(R):** Performs analytical power calculations using the serial-correlation-robust power calculation formula (Equation (3.6)). Solves for either sample size J, minimum detectable effect MDE, or power κ, as a function of all other parameter values. Users may either input assumed values for variance-covariance parameters, or allow the program to estimate variance-covariance parameters from data stored in memory.

`pc_dd_covar` **(STATA),** `PCDDCovar` **(R):** Estimates $\sigma^2_{\hat{\omega}}$, $\psi^B_{\hat{\omega}}$, $\psi^A_{\hat{\omega}}$, and $\psi^X_{\hat{\omega}}$ from an existing dataset, for a given number of pre-treatment and post-treatment periods. Serves as subprogram within `pc_dd_analytic`/`PCDDAnalytic`.

All of these programs are available for download. The STATA packages are available via `-ssc install pcpanel-`. The `R` packages remain in development, but will be posted shortly.

B.5 Estimation-related proofs

In this section, we prove that researchers may calculate unbiased power calculations by estimating the variance-covariance parameters from a real dataset, where the parameters governing the data generating process is unknown.

B.5.1 Recovering estimated parameters

Here, we demonstrate that the procedure described in Appendix B.4.1 recovers unbiased estimates of the variance and covariance parameters governing the residuals $\hat{\omega}_{it}$ from a regression of Y_{it} on unit and time fixed effects, $\sigma^2_{\hat{\omega}}$, $\psi^B_{\hat{\omega}}$, $\psi^A_{\hat{\omega}}$, and $\psi^X_{\hat{\omega}}$ (though these do *not* represent unbiased estimates of the *true* parameters σ^2_{ω}, ψ^B, ψ^A, and ψ^X). We denote our procedure for computing these parameters with a $\sim$. Note that throughout this section, we are considering I units in the sample used to estimate $\hat{\omega}_{it}$, which may be distinct from the sample size J units used in the ensuing power calculations. Note also that because we are estimating the variance and covariance of a *population* of residuals, we use the population variance/covariance estimators as opposed to the (unbiased) sample variance/covariance estimators.[26]

In order to estimate the variance of the residuals, $\sigma^2_{\hat{\omega}}$, we simply calculate:

$$\begin{aligned}\tilde{\sigma}^2_{\hat{\omega}} &= \text{Var}\left(\hat{\omega}_{it} \mid \mathbf{X}\right) \\ &= \frac{1}{IT}\sum_{i=1}^{I}\sum_{t=-m+1}^{r}\left(\hat{\omega}_{it} - \bar{\bar{\hat{\omega}}}\right)^2\end{aligned}$$

where $\bar{\bar{\hat{\omega}}} = \frac{1}{IT}\sum_{i=1}^{I}\sum_{t=-m+1}^{r}\hat{\omega}_{it} = 0$. Taking expectations of both sides:

$$\begin{aligned}\text{E}\left[\tilde{\sigma}^2_{\hat{\omega}} \mid \mathbf{X}\right] &= \frac{1}{IT}\sum_{i=1}^{I}\sum_{t=-m+1}^{r}\text{E}\left[\hat{\omega}^2_{it} \mid \mathbf{X}\right] \\ &= \sigma^2_{\hat{\omega}}\end{aligned}$$

To estimate $\psi^B_{\hat{\omega}}$, $\psi^A_{\hat{\omega}}$, and $\psi^X_{\hat{\omega}}$, we define the $[I \times 1]$ vector of residuals for period t as $\vec{\hat{\omega}}_t$. This allows us to calculate:

$$\tilde{\psi}^B_{\hat{\omega}} = \frac{2}{m(m-1)}\sum_{t=-m+1}^{-1}\sum_{s=t+1}^{0}\text{Cov}\left(\vec{\hat{\omega}}_t, \vec{\hat{\omega}}_s \mid \mathbf{X}\right)$$

[26] This means that to calculate $\tilde{\sigma}^2_{\hat{\omega}}$, we deflate the sample variance estimate by $\frac{IT-1}{IT}$, and to calculate the $\tilde{\psi}_{\hat{\omega}}$ terms, we deflate the sample covariance estimates by $\frac{I-1}{I}$. This distinction is ultimately innocuous, and the following derivations simply rely on a consistent decision to use either the population or sample variance/covariance estimators.

$$= \frac{2}{Im(m-1)} \sum_{t=-m+1}^{-1} \sum_{s=t+1}^{0} \sum_{i=1}^{I} \left(\hat{\omega}_{it}\hat{\omega}_{is} - \hat{\omega}_{it}\bar{\hat{\omega}}_s - \hat{\omega}_{is}\bar{\hat{\omega}}_t + \bar{\hat{\omega}}_t\bar{\hat{\omega}}_s \right)$$

where $\bar{\hat{\omega}}_t = \frac{1}{I}\sum_{i=1}^{I} \hat{\omega}_{it} = 0$. Taking expectations yields:

$$\begin{aligned}
\mathrm{E}\left[\tilde{\psi}^B_{\hat{\omega}} \mid \mathbf{X}\right] &= \frac{2}{Im(m-1)} \sum_{t=-m+1}^{-1} \sum_{s=t+1}^{0} \sum_{i=1}^{I} \mathrm{E}\left[\hat{\omega}_{it}\hat{\omega}_{is} \mid \mathbf{X}\right] \\
&= \frac{2}{Im(m-1)} \sum_{t=-m+1}^{-1} \sum_{s=t+1}^{0} \sum_{i=1}^{I} \mathrm{Cov}(\hat{\omega}_{it}, \hat{\omega}_{is} \mid \mathbf{X}) \\
&= \psi^B_{\hat{\omega}}
\end{aligned}$$

Similarly:

$$\begin{aligned}
\tilde{\psi}^A_{\hat{\omega}} &= \frac{2}{r(r-1)} \sum_{t=1}^{r-1} \sum_{s=t+1}^{r} \mathrm{Cov}\left(\vec{\hat{\omega}}_t, \vec{\hat{\omega}}_s \mid \mathbf{X}\right) \\
&= \frac{2}{Ir(r-1)} \sum_{t=1}^{r-1} \sum_{s=t+1}^{r} \sum_{i=1}^{I} \left(\hat{\omega}_{it}\hat{\omega}_{is} - \hat{\omega}_{it}\bar{\hat{\omega}}_s - \hat{\omega}_{is}\bar{\hat{\omega}}_t + \bar{\hat{\omega}}_t\bar{\hat{\omega}}_s \right)
\end{aligned}$$

and therefore:

$$\mathrm{E}\left[\tilde{\psi}^A_{\hat{\omega}} \mid \mathbf{X}\right] = \psi^A_{\hat{\omega}}$$

Applying the same steps to $\psi^X_{\hat{\omega}}$:

$$\begin{aligned}
\tilde{\psi}^X_{\hat{\omega}} &= \frac{1}{mr} \sum_{t=-m+1}^{0} \sum_{s=1}^{r} \mathrm{Cov}\left(\vec{\hat{\omega}}_t, \vec{\hat{\omega}}_s \mid \mathbf{X}\right) \\
&= \frac{1}{Imr} \sum_{t=-m+1}^{0} \sum_{s=1}^{r} \sum_{i=1}^{I} \left(\hat{\omega}_{it}\hat{\omega}_{is} - \hat{\omega}_{it}\bar{\hat{\omega}}_s - \hat{\omega}_{is}\bar{\hat{\omega}}_t + \bar{\hat{\omega}}_t\bar{\hat{\omega}}_s \right)
\end{aligned}$$

Taking expectations of both sides:

$$\begin{aligned}
\mathrm{E}\left[\tilde{\psi}^X_{\hat{\omega}} \mid \mathbf{X}\right] &= \frac{1}{Imr} \sum_{t=-m+1}^{0} \sum_{s=1}^{r} \sum_{i=1}^{I} \mathrm{E}\left[\hat{\omega}_{it}\hat{\omega}_{is} \mid \mathbf{X}\right] \\
&= \frac{1}{Imr} \sum_{t=-m+1}^{0} \sum_{s=1}^{r} \sum_{i=1}^{I} \mathrm{Cov}\left(\hat{\omega}_{it}, \hat{\omega}_{is} \mid \mathbf{X}\right) \\
&= \psi^X_{\hat{\omega}}
\end{aligned}$$

Hence, we can recover unbiased estimates of the parameters $\sigma^2_{\hat{\omega}}$, $\psi^B_{\hat{\omega}}$, $\psi^A_{\hat{\omega}}$, and $\psi^X_{\hat{\omega}}$ (defined over residuals $\hat{\omega}_{it}$, rather than errors ω_{it}) by calculating the averages of the estimated $\tilde{\sigma}^2_{\hat{\omega}}$, $\tilde{\psi}^B_{\hat{\omega}}$, $\tilde{\psi}^A_{\hat{\omega}}$, and $\tilde{\psi}^X_{\hat{\omega}}$, respectively.

B.5.2 Estimating MDE from residual-based parameters

To calculate the MDE using the SCR formula, we must know the true parameters that characterize the variance and covariance of the error structure, σ^2_ω, ψ^B, ψ^A, and ψ^X. We cannot calculate these these parameters directly from a real dataset, however, because we do not observe the true error structure or data generating process. Instead, we estimate a residual for each observation and calculate the residual-based analogs of these parameters, $\sigma^2_{\hat{\omega}}$, $\psi^B_{\hat{\omega}}$, $\psi^A_{\hat{\omega}}$, and $\psi^X_{\hat{\omega}}$. In this section, we derive an expression for MDE^{est} in terms of these residual-based parameters that is equivalent to MDE from the SCR formula as defined in terms of true variance-covariance parameters:

$$MDE^{est}\left(\sigma^2_{\hat{\omega}}, \psi^B_{\hat{\omega}}, \psi^A_{\hat{\omega}}, \psi^X_{\hat{\omega}}\right) = MDE\left(\sigma^2_\omega, \psi^B, \psi^A, \psi^X\right)$$

Model While estimating the variance and covariance parameters of a dataset does not require a treatment, we assume that all other features of this model are identical to the model that generates the serial-correlation-robust power calculation formula, Equation (3.6).

That is, 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. We restate the remaining assumptions from Section B.1.2.2 here for convenience:

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

$$Y_{it} = \beta + \tau D_{it} + \upsilon_i + \delta_t + \omega_{it} \tag{B.3}$$

where Y_{it} is the outcome of interest for unit i at time t; β is the expected outcome of non-treated observations; τ is the treatment effect that is homogenous across all units and all time periods; D_{it} is a time-varying treatment indicator; υ_i is a unit-specific disturbance distributed i.i.d. $\mathcal{N}(0, \sigma^2_\upsilon)$; δ_t is a time-specific disturbance distributed i.i.d. $\mathcal{N}(0, \sigma^2_\delta)$; and ω_{it} is an idiosyncratic error term distributed (not necessarily i.i.d.) $\mathcal{N}(0, \sigma^2_\omega)$.

Assumption (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 D, $J-1$ unit fixed effects, and $(m+r)-1$ time fixed effects. This again follows from random assignment of D_{it}.*

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

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

$$\psi_T^B \equiv \frac{2}{PJm(m-1)} \sum_{i=1}^{PJ} \sum_{t=-m+1}^{-1} \sum_{s=t+1}^{0} \mathrm{Cov}\left(\omega_{it}, \omega_{is} \mid \mathbf{X}\right)$$

$$\psi_T^A \equiv \frac{2}{PJr(r-1)} \sum_{i=1}^{PJ} \sum_{t=1}^{r-1} \sum_{s=t+1}^{r} \mathrm{Cov}\left(\omega_{it}, \omega_{is} \mid \mathbf{X}\right)$$

$$\psi_T^X \equiv \frac{1}{PJmr} \sum_{i=1}^{PJ} \sum_{t=-m+1}^{0} \sum_{s=1}^{r} \mathrm{Cov}\left(\omega_{it}, \omega_{is} \mid \mathbf{X}\right)$$

to be the average pre-treatment, post-treatment, and across-period covariance between different error terms of the same treated unit, respectively. Define ψ_C^B, ψ_C^A, and ψ_C^X analogously, where we consider the $(1-P)J$ control units instead of the PJ treated units. Using these definitions, assume that $\psi^B = \psi_T^B = \psi_C^B$; $\psi^A = \psi_T^A = \psi_C^A$; and $\psi^X = \psi_T^X = \psi_C^X$.[27]

Estimates We first need to estimate the residuals of this model. To do this, we regress Y_{it} on a constant and fixed effects at the unit and time levels. For a balanced panel, the estimated coefficients are

$$\hat{\beta} = \frac{1}{IT} \sum_{i=1}^{I} \sum_{t=-m+1}^{r} Y_{it}$$

$$\hat{\upsilon}_i = \frac{1}{T} \sum_{t=-m+1}^{r} TY_{it} - \frac{1}{IT} \sum_{i=1}^{I} \sum_{t=-m+1}^{r} Y_{it}$$

$$\hat{\delta}_t = \frac{1}{I} \sum_{i=1}^{I} Y_{it} - \frac{1}{IT} \sum_{i=1}^{I} \sum_{t=-m+1}^{r} Y_{it}$$

Then the residual is defined as

$$\hat{\omega}_{it} = Y_{it} - \hat{Y}_{it}$$

[27] 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, $\mathrm{E}[\psi_T^B] = \mathrm{E}[\psi_C^B]$, $\mathrm{E}[\psi_T^A] = \mathrm{E}[\psi_C^A]$, and $\mathrm{E}[\psi_T^X] = \mathrm{E}[\psi_C^X]$, making this a reasonable assumption *ex ante*. However, it is possible for treatment to alter the covariance structure of treated units only.

$$
\begin{aligned}
&= (\beta + \upsilon_i + \delta_t + \omega_{it}) - \left(\hat{\beta} + \hat{\upsilon}_i + \hat{\delta}_t\right) \\
&= \omega_{it} - \bar{\omega}_i - \bar{\omega}_t + \bar{\bar{\omega}}
\end{aligned}
$$

where

$$
\begin{aligned}
\bar{\omega}_i &= \frac{1}{T} \sum_{s=-m+1}^{r} \omega_{is} \\
\bar{\omega}_t &= \frac{1}{I} \sum_{j=1}^{I} \omega_{jt} \\
\bar{\bar{\omega}} &= \frac{1}{IT} \sum_{j=1}^{I} \sum_{s=-m+1}^{r} \omega_{js}
\end{aligned}
$$

We can now use this definition of residuals to derive expressions for $\sigma^2_{\hat{\omega}}$, $\psi^B_{\hat{\omega}}$, $\psi^A_{\hat{\omega}}$, and $\psi^X_{\hat{\omega}}$. We first derive an expression for $\sigma^2_{\hat{\omega}}$, the average variance of a residual.

$$
\begin{aligned}
\sigma^2_{\hat{\omega}} &= \frac{1}{IT} \sum_{i=1}^{I} \sum_{t=-m+1}^{r} \operatorname{Var}(\hat{\omega}_{it} \mid \mathbf{X}) \\
&= \frac{1}{IT} \sum_{i=1}^{I} \sum_{t=-m+1}^{r} \operatorname{Var}\left(\omega_{it} - \bar{\omega}_i - \bar{\omega}_t + \bar{\bar{\omega}} \mid \mathbf{X}\right) \\
&= \frac{1}{IT} \sum_{i=1}^{I} \sum_{t=-m+1}^{r} [\operatorname{Var}(\omega_{it} \mid \mathbf{X}) + \operatorname{Var}\left(\bar{\omega}_i \mid \mathbf{X}\right) + \operatorname{Var}\left(\bar{\omega}_t \mid \mathbf{X}\right) + \operatorname{Var}\left(\bar{\bar{\omega}} \mid \mathbf{X}\right) \\
&\qquad - 2\operatorname{Cov}\left(\omega_{it}, \bar{\omega}_i \mid \mathbf{X}\right) - 2\operatorname{Cov}\left(\omega_{it}, \bar{\omega}_t \mid \mathbf{X}\right) + 2\operatorname{Cov}\left(\omega_{it}, \bar{\bar{\omega}} \mid \mathbf{X}\right) \\
&\qquad + 2\operatorname{Cov}\left(\bar{\omega}_i, \bar{\omega}_t \mid \mathbf{X}\right) - 2\operatorname{Cov}\left(\bar{\omega}_i, \bar{\bar{\omega}} \mid \mathbf{X}\right) - 2\operatorname{Cov}\left(\bar{\omega}_t, \bar{\bar{\omega}} \mid \mathbf{X}\right)]
\end{aligned}
$$

Calculating each of these terms, in turn, gives

$$
\begin{aligned}
\frac{1}{IT} \sum_{i=1}^{I} \sum_{t=-m+1}^{r} \operatorname{Var}(\omega_{it} \mid \mathbf{X}) &= \sigma^2_{\omega} \\
\frac{1}{IT} \sum_{i=1}^{I} \sum_{t=-m+1}^{r} \operatorname{Var}\left(\bar{\omega}_i \mid \mathbf{X}\right) &= \frac{1}{IT} \sum_{i=1}^{I} \sum_{t=-m+1}^{r} \operatorname{Var}\left(\frac{1}{T} \sum_{s=-m+1}^{r} \omega_{is} \mid \mathbf{X}\right) \\
&= \frac{1}{IT^3} \sum_{i=1}^{I} \sum_{t=-m+1}^{r} \sum_{s=-m+1}^{r} \operatorname{Var}\left(\omega_{is} \mid \mathbf{X}\right) \\
&\quad + \frac{2}{IT^3} \sum_{i=1}^{I} \sum_{t=-m+1}^{0} \sum_{s=-m+1}^{r-1} \sum_{p=s+1}^{r} \operatorname{Cov}(\omega_{is}, \omega_{ip} \mid \mathbf{X}) \\
&= \frac{1}{T}\sigma^2_{\omega} + \frac{m(m-1)}{T^2}\psi^B + \frac{r(r-1)}{T^2}\psi^A + \frac{2mr}{T^2}\psi^X \\
\frac{1}{IT} \sum_{i=1}^{I} \sum_{t=-m+1}^{r} \operatorname{Var}\left(\bar{\omega}_t \mid \mathbf{X}\right) &= \frac{1}{IT} \sum_{i=1}^{I} \sum_{t=-m+1}^{r} \operatorname{Var}\left(\frac{1}{I} \sum_{j=1}^{I} \omega_{jt} \mid \mathbf{X}\right)
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{1}{I}\sigma_\omega^2 \\
\frac{1}{IT}\sum_{i=1}^{I}\sum_{t=-m+1}^{r} \operatorname{Var}\left(\bar{\omega} \mid \mathbf{X}\right) &= \frac{1}{IT}\sum_{i=1}^{I}\sum_{t=-m+1}^{r} \operatorname{Var}\left(\frac{1}{IT}\sum_{j=1}^{I}\sum_{s=-m+1}^{r} \omega_{js} \mid \mathbf{X}\right) \\
&= \frac{1}{IT}\sigma_\omega^2 + \frac{m(m-1)}{IT^2}\psi^B + \frac{r(r-1)}{IT^2}\psi^A + \frac{2mr}{IT^2}\psi^X \\
\frac{1}{IT}\sum_{i=1}^{I}\sum_{t=-m+1}^{r} 2\operatorname{Cov}\left(\omega_{it}, \bar{\omega}_i \mid \mathbf{X}\right) &= \frac{2}{IT}\sum_{i=1}^{I}\sum_{t=-m+1}^{r} \operatorname{Cov}\left(\omega_{it}, \frac{1}{T}\sum_{s=-m+1}^{r} \omega_{is} \mid \mathbf{X}\right) \\
&= \frac{2}{T}\sigma_\omega^2 + \frac{2m(m-1)}{T^2}\psi^B + \frac{2r(r-1)}{T^2}\psi^A + \frac{4mr}{T^2}\psi^X \\
\frac{1}{IT}\sum_{i=1}^{I}\sum_{t=-m+1}^{r} 2\operatorname{Cov}\left(\omega_{it}, \bar{\omega}_t \mid \mathbf{X}\right) &= \frac{2}{IT}\sum_{i=1}^{I}\sum_{t=-m+1}^{r} \operatorname{Cov}\left(\omega_{it}, \frac{1}{I}\sum_{j=1}^{I} \omega_{jt} \mid \mathbf{X}\right) \\
&= \frac{2}{I}\sigma_\omega^2 \\
\frac{1}{IT}\sum_{i=1}^{I}\sum_{t=-m+1}^{r} 2\operatorname{Cov}\left(\omega_{it}, \bar{\bar{\omega}} \mid \mathbf{X}\right) &= \frac{2}{IT}\sum_{i=1}^{I}\sum_{t=-m+1}^{r} \operatorname{Cov}\left(\omega_{it}, \frac{1}{IT}\sum_{j=1}^{I}\sum_{s=-m+1}^{r} \omega_{js} \mid \mathbf{X}\right) \\
&= \frac{2}{IT}\sigma_\omega^2 + \frac{2m(m-1)}{IT^2}\psi^B + \frac{2r(r-1)}{IT^2}\psi^A + \frac{4mr}{IT^2}\psi^X \\
\frac{1}{IT}\sum_{i=1}^{I}\sum_{t=-m+1}^{r} 2\operatorname{Cov}\left(\bar{\omega}_i, \bar{\omega}_t \mid \mathbf{X}\right) &= \frac{2}{IT}\sum_{i=1}^{I}\sum_{t=-m+1}^{r} \operatorname{Cov}\left(\frac{1}{T}\sum_{s=-m+1}^{r} \omega_{is}, \frac{1}{I}\sum_{j=1}^{I} \omega_{jt} \mid \mathbf{X}\right) \\
&= \frac{2}{IT}\sigma_\omega^2 + \frac{2m(m-1)}{IT^2}\psi^B + \frac{2r(r-1)}{IT^2}\psi^A + \frac{4mr}{IT^2}\psi^X \\
\frac{1}{IT}\sum_{i=1}^{I}\sum_{t=-m+1}^{r} 2\operatorname{Cov}\left(\bar{\omega}_i, \bar{\bar{\omega}} \mid \mathbf{X}\right) &= \frac{2}{IT}\sum_{i=1}^{I}\sum_{t=-m+1}^{r} \operatorname{Cov}\left(\frac{1}{T}\sum_{s=-m+1}^{r} \omega_{is}, \frac{1}{IT}\sum_{j=1}^{I}\sum_{p=-m+1}^{r} \omega_{jp} \mid \mathbf{X}\right) \\
&= \frac{2}{IT}\sigma_\omega^2 + \frac{2m(m-1)}{IT^2}\psi^B + \frac{2r(r-1)}{IT^2}\psi^A + \frac{4mr}{IT^2}\psi^X \\
\frac{1}{IT}\sum_{i=1}^{I}\sum_{t=-m+1}^{r} 2\operatorname{Cov}\left(\bar{\omega}_t, \bar{\bar{\omega}} \mid \mathbf{X}\right) &= \frac{2}{IT}\sum_{i=1}^{I}\sum_{t=-m+1}^{r} \operatorname{Cov}\left(\frac{1}{I}\sum_{j=1}^{I} \omega_{jt}, \frac{1}{IT}\sum_{k=1}^{I}\sum_{s=-m+1}^{r} \omega_{ks} \mid \mathbf{X}\right) \\
&= \frac{2}{IT}\sigma_\omega^2 + \frac{2m(m-1)}{IT^2}\psi^B + \frac{2r(r-1)}{IT^2}\psi^A + \frac{4mr}{IT^2}\psi^X
\end{aligned}
$$

Combining these terms yields

$$
\begin{aligned}
\sigma_{\bar{\bar{\omega}}}^2 = \sigma_\omega^2 &+ \left[\frac{1}{T}\sigma_\omega^2 + \frac{m(m-1)}{T^2}\psi^B + \frac{r(r-1)}{T^2}\psi^A + \frac{2mr}{T^2}\psi^X\right] + \frac{1}{I}\sigma_\omega^2 \\
&+ \left[\frac{1}{IT}\sigma_\omega^2 + \frac{m(m-1)}{IT^2}\psi^B + \frac{r(r-1)}{IT^2}\psi^A + \frac{2mr}{IT^2}\psi^X\right]
\end{aligned}
$$

$$
\begin{aligned}
&- \left[\frac{2}{T}\sigma_\omega^2 + \frac{2m(m-1)}{T^2}\psi^B + \frac{2r(r-1)}{T^2}\psi^A + \frac{4mr}{T^2}\psi^X\right] - \frac{2}{I}\sigma_\omega^2 \\
&+ \left[\frac{2}{IT}\sigma_\omega^2 + \frac{2m(m-1)}{IT^2}\psi^B + \frac{2r(r-1)}{IT^2}\psi^A + \frac{4mr}{IT^2}\psi^X\right] \\
&+ \left[\frac{2}{IT}\sigma_\omega^2 + \frac{2m(m-1)}{IT^2}\psi^B + \frac{2r(r-1)}{IT^2}\psi^A + \frac{4mr}{IT^2}\psi^X\right] \\
&- \left[\frac{2}{IT}\sigma_\omega^2 + \frac{2m(m-1)}{IT^2}\psi^B + \frac{2r(r-1)}{IT^2}\psi^A + \frac{4mr}{IT^2}\psi^X\right] \\
&- \left[\frac{2}{IT}\sigma_\omega^2 + \frac{2m(m-1)}{IT^2}\psi^B + \frac{2r(r-1)}{IT^2}\psi^A + \frac{4mr}{IT^2}\psi^X\right] \\
&= \sigma_\omega^2 + \left[\frac{1}{T}\sigma_\omega^2 + \frac{m(m-1)}{T^2}\psi^B + \frac{r(r-1)}{T^2}\psi^A + \frac{2mr}{T^2}\psi^X\right] + \frac{1}{I}\sigma_\omega^2 \\
&+ \left[\frac{1}{IT}\sigma_\omega^2 + \frac{m(m-1)}{IT^2}\psi^B + \frac{r(r-1)}{IT^2}\psi^A + \frac{2mr}{IT^2}\psi^X\right] \\
&- \left[\frac{2}{T}\sigma_\omega^2 + \frac{2m(m-1)}{T^2}\psi^B + \frac{2r(r-1)}{T^2}\psi^A + \frac{4mr}{T^2}\psi^X\right] - \frac{2}{I}\sigma_\omega^2 \\
&= \left[1 + \frac{1}{T} + \frac{1}{I} + \frac{1}{IT} - \frac{2}{T}\right]\sigma_\omega^2 + \left[\frac{m(m-1)}{T^2} + \frac{m(m-1)}{IT^2} - \frac{2m(m-1)}{T^2}\right]\psi^B \\
&+ \left[\frac{r(r-1)}{T^2} + \frac{r(r-1)}{IT^2} - \frac{2r(r-1)}{T^2}\right]\psi^A + \left[\frac{2mr}{T^2} + \frac{2mr}{IT^2} - \frac{4mr}{T^2}\right]\psi^X \\
&= \left(\frac{(I-1)(T-1)}{IT}\right)\sigma_\omega^2 - \left(\frac{(I-1)m(m-1)}{IT^2}\right)\psi^B - \left(\frac{(I-1)r(r-1)}{IT^2}\right)\psi^A \\
&- \left(\frac{2(I-1)mr}{IT^2}\right)\psi^X
\end{aligned}
$$

We next derive an expression for $\psi_{\hat{\omega}}^A$.

$$
\begin{aligned}
\psi_{\hat{\omega}}^A &= \frac{2}{r(r-1)I}\sum_{i=1}^{I}\sum_{t=1}^{r-1}\sum_{s=t+1}^{r}\mathrm{Cov}(\hat{\omega}_{it}, \hat{\omega}_{is} \mid \mathbf{X}) \\
&= \frac{2}{r(r-1)I}\sum_{i=1}^{I}\sum_{t=1}^{r-1}\sum_{s=t+1}^{r}\mathrm{Cov}(\omega_{it} - \bar{\omega}_i - \bar{\omega}_t + \bar{\bar{\omega}}, \omega_{is} - \bar{\omega}_i - \bar{\omega}_s + \bar{\bar{\omega}} \mid \mathbf{X}) \\
&= \frac{2}{r(r-1)I}\sum_{i=1}^{I}\sum_{t=1}^{r-1}\sum_{s=t+1}^{r}[\mathrm{Cov}(\omega_{it}, \omega_{is} \mid \mathbf{X}) - \mathrm{Cov}(\omega_{it}, \bar{\omega}_i \mid \mathbf{X}) \\
&- \mathrm{Cov}(\omega_{it}, \bar{\omega}_s \mid \mathbf{X}) + \mathrm{Cov}(\omega_{it}, \bar{\bar{\omega}} \mid \mathbf{X}) \\
&\qquad - \mathrm{Cov}(\bar{\omega}_i, \omega_{is} \mid \mathbf{X}) + \mathrm{Cov}(\bar{\omega}_i, \bar{\omega}_i \mid \mathbf{X}) + \mathrm{Cov}(\bar{\omega}_i, \bar{\omega}_s \mid \mathbf{X}) - \mathrm{Cov}(\bar{\omega}_i, \bar{\bar{\omega}} \mid \mathbf{X}) \\
&\qquad - \mathrm{Cov}(\bar{\omega}_t, \omega_{is} \mid \mathbf{X}) + \mathrm{Cov}(\bar{\omega}_t, \bar{\omega}_i \mid \mathbf{X}) + \mathrm{Cov}(\bar{\omega}_t, \bar{\omega}_s \mid \mathbf{X}) - \mathrm{Cov}(\bar{\omega}_t, \bar{\bar{\omega}} \mid \mathbf{X}) \\
&\qquad + \mathrm{Cov}(\bar{\bar{\omega}}, \omega_{is} \mid \mathbf{X}) - \mathrm{Cov}(\bar{\bar{\omega}}, \bar{\omega}_i \mid \mathbf{X}) - \mathrm{Cov}(\bar{\bar{\omega}}, \bar{\omega}_s \mid \mathbf{X}) + \mathrm{Cov}(\bar{\bar{\omega}}, \bar{\bar{\omega}} \mid \mathbf{X})]
\end{aligned}
$$

We again calculate each term.

$$\frac{2}{r(r-1)I}\sum_{i=1}^{I}\sum_{t=1}^{r-1}\sum_{s=t+1}^{r}\mathrm{Cov}(\omega_{it},\omega_{is}\mid\mathbf{X})=\psi^{A}$$

$$\frac{2}{r(r-1)I}\sum_{i=1}^{I}\sum_{t=1}^{r-1}\sum_{s=t+1}^{r}\mathrm{Cov}(\omega_{it},\bar{\omega}_{i}\mid\mathbf{X})=\frac{2}{r(r-1)IT}\sum_{i=1}^{I}\sum_{t=1}^{r}\sum_{p=-m+1}^{r}(r-t)\,\mathrm{Cov}(\omega_{it},\omega_{ip}\mid\mathbf{X})$$

$$\frac{2}{r(r-1)I}\sum_{i=1}^{I}\sum_{t=1}^{r-1}\sum_{s=t+1}^{r}\mathrm{Cov}(\omega_{it},\bar{\omega}_{s}\mid\mathbf{X})=\frac{1}{I}\psi^{A}$$

$$\frac{2}{r(r-1)I}\sum_{i=1}^{I}\sum_{t=1}^{r-1}\sum_{s=t+1}^{r}\mathrm{Cov}(\omega_{it},\bar{\bar{\omega}}\mid\mathbf{X})$$

$$=\frac{2}{r(r-1)I^{2}T}\sum_{i=1}^{I}\sum_{t=1}^{r}\sum_{p=-m+1}^{r}(r-t)\,\mathrm{Cov}(\omega_{it},\omega_{ip}\mid\mathbf{X})$$

$$\frac{2}{r(r-1)I}\sum_{i=1}^{I}\sum_{t=1}^{r-1}\sum_{s=t+1}^{r}\mathrm{Cov}(\bar{\omega}_{i},\omega_{is}\mid\mathbf{X})=\frac{2}{r(r-1)IT}\sum_{i=1}^{I}\sum_{t=1}^{r}\sum_{p=-m+1}^{r}(t-1)\,\mathrm{Cov}\left(\omega_{it},\omega_{ip}\mid\mathbf{X}\right)$$

$$\frac{2}{r(r-1)I}\sum_{i=1}^{I}\sum_{t=1}^{r-1}\sum_{s=t+1}^{r}\mathrm{Cov}(\bar{\omega}_{i},\bar{\omega}_{i}\mid\mathbf{X})=\frac{1}{T}\sigma_{\omega}^{2}+\frac{m(m-1)}{T^{2}}\psi^{B}+\frac{r(r-1)}{T^{2}}\psi^{A}+\frac{2mr}{T^{2}}\psi^{X}$$

$$\frac{2}{r(r-1)I}\sum_{i=1}^{I}\sum_{t=1}^{r-1}\sum_{s=t+1}^{r}\mathrm{Cov}(\bar{\omega}_{i},\bar{\omega}_{s}\mid\mathbf{X})=\frac{2}{r(r-1)I^{2}T}\sum_{i=1}^{I}\sum_{t=1}^{r}\sum_{p=-m+1}^{r}(t-1)\,\mathrm{Cov}(\omega_{it},\omega_{ip}\mid\mathbf{X})$$

$$\frac{2}{r(r-1)I}\sum_{i=1}^{I}\sum_{t=1}^{r-1}\sum_{s=t+1}^{r}\mathrm{Cov}(\bar{\omega}_{i},\bar{\bar{\omega}}\mid\mathbf{X})=\frac{1}{IT}\sigma_{\omega}^{2}+\frac{m(m-1)}{IT^{2}}\psi^{B}+\frac{r(r-1)}{IT^{2}}\psi^{A}+\frac{2mr}{IT^{2}}\psi^{X}$$

$$\frac{2}{r(r-1)I}\sum_{i=1}^{I}\sum_{t=1}^{r-1}\sum_{s=t+1}^{r}\mathrm{Cov}(\bar{\omega}_{t},\omega_{is}\mid\mathbf{X})=\frac{1}{I}\psi^{A}$$

$$\frac{2}{r(r-1)I}\sum_{i=1}^{I}\sum_{t=1}^{r-1}\sum_{s=t+1}^{r}\mathrm{Cov}(\bar{\omega}_{t},\bar{\omega}_{i}\mid\mathbf{X})=\frac{2}{r(r-1)I^{2}T}\sum_{i=1}^{I}\sum_{t=1}^{r}\sum_{p=-m+1}^{r}(r-t)\,\mathrm{Cov}(\omega_{it},\omega_{ip}\mid\mathbf{X})$$

$$\frac{2}{r(r-1)I}\sum_{i=1}^{I}\sum_{t=1}^{r-1}\sum_{s=t+1}^{r}\mathrm{Cov}(\bar{\omega}_{t},\bar{\omega}_{s}\mid\mathbf{X})=\frac{1}{I}\psi^{A}$$

$$\frac{2}{r(r-1)I}\sum_{i=1}^{I}\sum_{t=1}^{r-1}\sum_{s=t+1}^{r}\mathrm{Cov}(\bar{\omega}_{t},\bar{\bar{\omega}}\mid\mathbf{X})=\frac{2}{r(r-1)I^{2}T}\sum_{i=1}^{I}\sum_{t=1}^{r}\sum_{p=-m+1}^{r}(r-t)\,\mathrm{Cov}(\omega_{it},\omega_{ip}\mid\mathbf{X})$$

$$\frac{2}{r(r-1)I}\sum_{i=1}^{I}\sum_{t=1}^{r-1}\sum_{s=t+1}^{r}\mathrm{Cov}(\bar{\bar{\omega}},\omega_{is}\mid\mathbf{X})=\frac{2}{r(r-1)I^{2}T}\sum_{i=1}^{I}\sum_{t=1}^{r}\sum_{p=-m+1}^{r}(t-1)\,\mathrm{Cov}(\omega_{it},\omega_{ip}\mid\mathbf{X})$$

$$\frac{2}{r(r-1)I}\sum_{i=1}^{I}\sum_{t=1}^{r-1}\sum_{s=t+1}^{r}\mathrm{Cov}(\bar{\bar{\omega}},\bar{\omega}_{i}\mid\mathbf{X})=\frac{1}{IT}\sigma_{\omega}^{2}+\frac{m(m-1)}{IT^{2}}\psi^{B}+\frac{r(r-1)}{IT^{2}}\psi^{A}+\frac{2mr}{IT^{2}}\psi^{X}$$

$$\frac{2}{r(r-1)I}\sum_{i=1}^{I}\sum_{t=1}^{r-1}\sum_{s=t+1}^{r}\operatorname{Cov}(\bar{\omega},\bar{\omega}_s \mid \mathbf{X}) = \frac{2}{r(r-1)I^2T}\sum_{i=1}^{I}\sum_{t=1}^{r}\sum_{p=-m+1}^{r}(t-1)\operatorname{Cov}(\omega_{it},\omega_{ip} \mid \mathbf{X})$$

$$\frac{2}{r(r-1)I}\sum_{i=1}^{I}\sum_{t=1}^{r-1}\sum_{s=t+1}^{r}\operatorname{Cov}(\bar{\omega},\bar{\omega} \mid \mathbf{X}) = \frac{1}{IT}\sigma_\omega^2 + \frac{m(m-1)}{IT^2}\psi^B + \frac{r(r-1)}{IT^2}\psi^A + \frac{2mr}{IT^2}\psi^X$$

Combining these terms yields

$$\begin{aligned}
\psi^A_{\hat{\omega}} &= \psi^A - \frac{2}{r(r-1)IT}\sum_{i=1}^{I}\sum_{t=1}^{r}\sum_{p=-m+1}^{r}(r-t)\operatorname{Cov}(\omega_{it},\omega_{ip} \mid \mathbf{X}) - \frac{1}{I}\psi^A \\
&\quad + \frac{2}{r(r-1)I^2T}\sum_{i=1}^{I}\sum_{t=1}^{r}\sum_{p=-m+1}^{r}(r-t)\operatorname{Cov}(\omega_{it},\omega_{ip} \mid \mathbf{X}) \\
&\quad - \frac{2}{r(r-1)IT}\sum_{i=1}^{I}\sum_{t=1}^{r}\sum_{p=-m+1}^{r}(t-1)\operatorname{Cov}(\omega_{it},\omega_{ip} \mid \mathbf{X}) \\
&\quad + \left[\frac{1}{T}\sigma_\omega^2 + \frac{m(m-1)}{T^2}\psi^B + \frac{r(r-1)}{T^2}\psi^A + \frac{2mr}{T^2}\psi^X\right] \\
&\quad + \frac{2}{r(r-1)I^2T}\sum_{i=1}^{I}\sum_{t=1}^{r}\sum_{p=-m+1}^{r}(t-1)\operatorname{Cov}(\omega_{it},\omega_{ip} \mid \mathbf{X}) \\
&\quad - \left[\frac{1}{IT}\sigma_\omega^2 + \frac{m(m-1)}{IT^2}\psi^B + \frac{r(r-1)}{IT^2}\psi^A + \frac{2mr}{IT^2}\psi^X\right] - \frac{1}{I}\psi^A \\
&\quad + \frac{2}{r(r-1)I^2T}\sum_{i=1}^{I}\sum_{t=1}^{r}\sum_{p=-m+1}^{r}(r-t)\operatorname{Cov}(\omega_{it},\omega_{ip} \mid \mathbf{X}) + \frac{1}{I}\psi^A \\
&\quad - \frac{2}{r(r-1)I^2T}\sum_{i=1}^{I}\sum_{t=1}^{r}\sum_{p=-m+1}^{r}(r-t)\operatorname{Cov}(\omega_{it},\omega_{ip} \mid \mathbf{X}) \\
&\quad + \frac{2}{r(r-1)I^2T}\sum_{i=1}^{I}\sum_{t=1}^{r}\sum_{p=-m+1}^{r}(t-1)\operatorname{Cov}(\omega_{it},\omega_{ip} \mid \mathbf{X}) \\
&\quad - \left[\frac{1}{IT}\sigma_\omega^2 + \frac{m(m-1)}{IT^2}\psi^B + \frac{r(r-1)}{IT^2}\psi^A + \frac{2mr}{IT^2}\psi^X\right] \\
&\quad - \frac{2}{r(r-1)I^2T}\sum_{i=1}^{I}\sum_{t=1}^{r}\sum_{p=-m+1}^{r}(t-1)\operatorname{Cov}(\omega_{it},\omega_{ip} \mid \mathbf{X}) \\
&\quad + \left[\frac{1}{IT}\sigma_\omega^2 + \frac{m(m-1)}{IT^2}\psi^B + \frac{r(r-1)}{IT^2}\psi^A + \frac{2mr}{IT^2}\psi^X\right] \\
&= \psi^A - \frac{2}{r(r-1)IT}\sum_{i=1}^{I}\sum_{t=1}^{r}\sum_{p=-m+1}^{r}(r-1)\operatorname{Cov}(\omega_{it},\omega_{ip} \mid \mathbf{X}) - \frac{1}{I}\psi^A
\end{aligned}$$

$$
\begin{aligned}
&+ \frac{2}{r(r-1)I^2T} \sum_{i=1}^{I} \sum_{t=1}^{r} \sum_{p=-m+1}^{r} (r-1) \operatorname{Cov}(\omega_{it}, \omega_{ip} \mid \mathbf{X}) \\
&+ \left[\frac{1}{T}\sigma_\omega^2 + \frac{m(m-1)}{T^2}\psi^B + \frac{r(r-1)}{T^2}\psi^A + \frac{2mr}{T^2}\psi^X \right] \\
&- \left[\frac{1}{IT}\sigma_\omega^2 + \frac{m(m-1)}{IT^2}\psi^B + \frac{r(r-1)}{IT^2}\psi^A + \frac{2mr}{IT^2}\psi^X \right] \\
&= \left(\frac{I-1}{I} \right) \psi^A + \left(\frac{I-1}{I} \right) \left[\frac{1}{T}\sigma_\omega^2 + \frac{m(m-1)}{T^2}\psi^B + \frac{r(r-1)}{T^2}\psi^A + \frac{2mr}{T^2}\psi^X \right] \\
&- \frac{2(I-1)}{rI^2T} \sum_{i=1}^{I} \sum_{t=1}^{r} \sum_{p=-m+1}^{r} \operatorname{Cov}(\omega_{it}, \omega_{ip} \mid \mathbf{X}) \\
&= \left(\frac{I-1}{I} \right) \psi^A + \left(\frac{I-1}{I} \right) \left[\frac{1}{T}\sigma_\omega^2 + \frac{m(m-1)}{T^2}\psi^B + \frac{r(r-1)}{T^2}\psi^A + \frac{2mr}{T^2}\psi^X \right] \\
&- \frac{2(I-1)}{rI^2T} \left[\sum_{i=1}^{I} \sum_{t=1}^{r} \operatorname{Var}(\omega_{it} \mid \mathbf{X}) + 2 \sum_{i=1}^{I} \sum_{t=1}^{r-1} \sum_{p=t+1}^{r} \operatorname{Cov}(\omega_{it}, \omega_{ip} \mid \mathbf{X}) \right. \\
&\left. + \sum_{i=1}^{I} \sum_{t=1}^{r} \sum_{p=-m+1}^{0} \operatorname{Cov}(\omega_{it}, \omega_{ip} \mid \mathbf{X}) \right] \\
&= \left(\frac{I-1}{I} \right) \psi^A + \left(\frac{I-1}{I} \right) \left[\frac{1}{T}\sigma_\omega^2 + \frac{m(m-1)}{T^2}\psi^B + \frac{r(r-1)}{T^2}\psi^A + \frac{2mr}{T^2}\psi^X \right] \\
&- \left(\frac{I-1}{I} \mid \right) \left[\frac{2}{T}\sigma_\omega^2 + \frac{2(r-1)}{T}\psi^A + \frac{2m}{T}\psi^X \right] \\
&= \left(\frac{I-1}{I} \right) \left[\frac{1}{T} - \frac{2}{T} \right] \sigma_\omega^2 + \left(\frac{I-1}{I} \right) \left[\frac{m(m-1)}{T^2} \right] \psi^B \\
&+ \left(\frac{I-1}{I} \right) \left[1 + \frac{r(r-1)}{T^2} - \frac{2(r-2)}{T} \right] \psi^A + \left(\frac{I-1}{I} \right) \left[\frac{2mr}{T^2} + \frac{2m}{T} \right] \psi^X \\
&= - \left(\frac{I-1}{IT} \right) \sigma_\omega^2 + \left(\frac{(I-1)m(m-1)}{IT^2} \right) \psi^B \\
&+ \left(\frac{(I-1)(m^2+2m+r)}{IT^2} \right) \psi^A - \left(\frac{2(I-1)m^2}{IT^2} \right) \psi^X
\end{aligned}
$$

By symmetry, the expression for $\psi_{\dot{\omega}}^B$ is

$$
\begin{aligned}
\psi_{\dot{\omega}}^B = &- \left(\frac{I-1}{IT} \right) \sigma_\omega^2 + \left(\frac{(I-1)(r^2+2r+m)}{IT^2} \right) \psi^B \\
&+ \left(\frac{(I-1)r(r-1)}{IT^2} \right) \psi^A - \left(\frac{2(I-1)r^2}{IT^2} \right) \psi^X
\end{aligned}
$$

We finally derive an expression for $\psi^X_{\hat{\omega}}$.

$$\begin{aligned}
\psi^X_{\hat{\omega}} &= \frac{1}{Imr}\sum_{i=1}^{I}\sum_{t=-m+1}^{0}\sum_{s=1}^{r}\mathrm{Cov}(\hat{\omega}_{it},\hat{\omega}_{is}\mid\mathbf{X}) \\
&= \frac{1}{Imr}\sum_{i=1}^{I}\sum_{t=-m+1}^{0}\sum_{s=1}^{r}\mathrm{Cov}(\omega_{it}-\bar{\omega}_i-\bar{\omega}_t+\bar{\bar{\omega}},\omega_{is}-\bar{\omega}_i-\bar{\omega}_s+\bar{\bar{\omega}}\mid\mathbf{X}) \\
&= \frac{1}{Imr}\sum_{i=1}^{I}\sum_{t=-m+1}^{0}\sum_{s=1}^{r}[\mathrm{Cov}(\omega_{it},\omega_{is}\mid\mathbf{X})-\mathrm{Cov}(\omega_{it},\bar{\omega}_i\mid\mathbf{X}) \\
&\quad -\mathrm{Cov}(\omega_{it},\bar{\omega}_s\mid\mathbf{X})+\mathrm{Cov}(\omega_{it},\bar{\bar{\omega}}\mid\mathbf{X})] \\
&\qquad\qquad -\mathrm{Cov}(\bar{\omega}_i,\omega_{is}\mid\mathbf{X})+\mathrm{Cov}(\bar{\omega}_i,\bar{\omega}_i\mid\mathbf{X}) \\
&\quad +\mathrm{Cov}(\bar{\omega}_i,\bar{\omega}_s\mid\mathbf{X})-\mathrm{Cov}(\bar{\omega}_i,\bar{\bar{\omega}}\mid\mathbf{X}) \\
&\qquad\qquad -\mathrm{Cov}(\bar{\omega}_t,\omega_{is}\mid\mathbf{X})+\mathrm{Cov}(\bar{\omega}_t,\bar{\omega}_i\mid\mathbf{X}) \\
&\quad +\mathrm{Cov}(\bar{\omega}_t,\bar{\omega}_s\mid\mathbf{X})-\mathrm{Cov}(\bar{\omega}_t,\bar{\bar{\omega}}\mid\mathbf{X}) \\
&\qquad\qquad +\mathrm{Cov}(\bar{\bar{\omega}},\omega_{is}\mid\mathbf{X})-\mathrm{Cov}(\bar{\bar{\omega}},\bar{\omega}_i\mid\mathbf{X})-\mathrm{Cov}(\bar{\bar{\omega}},\bar{\omega}_s\mid\mathbf{X}) \\
&\quad +\mathrm{Cov}(\bar{\bar{\omega}},\bar{\bar{\omega}}\mid\mathbf{X})]
\end{aligned}$$

We again calculate each term.

$$\frac{1}{Imr}\sum_{i=1}^{I}\sum_{t=-m+1}^{0}\sum_{s=1}^{r}\mathrm{Cov}(\omega_{is},\omega_{it}\mid\mathbf{X})=\psi^X$$

$$\frac{1}{Imr}\sum_{i=1}^{I}\sum_{t=-m+1}^{0}\sum_{s=1}^{r}\mathrm{Cov}(\omega_{it},\bar{\omega}_i\mid\mathbf{X})=\frac{1}{T}\sigma^2_\omega+\frac{m-1}{T}\psi^B+\frac{r}{T}\psi^X$$

$$\frac{1}{Imr}\sum_{i=1}^{I}\sum_{t=-m+1}^{0}\sum_{s=1}^{r}\mathrm{Cov}(\omega_{it},\bar{\omega}_s\mid\mathbf{X})=\frac{1}{I}\psi^X$$

$$\frac{1}{Imr}\sum_{i=1}^{I}\sum_{t=-m+1}^{0}\sum_{s=1}^{r}\mathrm{Cov}(\omega_{it},\bar{\bar{\omega}}\mid\mathbf{X})=\frac{1}{IT}\sigma^2_\omega+\frac{m-1}{IT}\psi^B+\frac{r}{IT}\psi^X$$

$$\frac{1}{Imr}\sum_{i=1}^{I}\sum_{t=-m+1}^{0}\sum_{s=1}^{r}\mathrm{Cov}(\bar{\omega}_i,\omega_{is}\mid\mathbf{X})=\frac{1}{T}\sigma^2_\omega+\frac{r-1}{T}\psi^A+\frac{m}{T}\psi^X$$

$$\frac{1}{Imr}\sum_{i=1}^{I}\sum_{t=-m+1}^{0}\sum_{s=1}^{r}\mathrm{Cov}(\bar{\omega}_i,\bar{\omega}_i\mid\mathbf{X})=\frac{1}{T}\sigma^2_\omega+\frac{m(m-1)}{T^2}\psi^B+\frac{r(r-1)}{T^2}\psi^A+\frac{2mr}{T^2}\psi^X$$

$$\frac{1}{Imr}\sum_{i=1}^{I}\sum_{t=-m+1}^{0}\sum_{s=1}^{r}\mathrm{Cov}(\bar{\omega}_i,\bar{\omega}_s\mid\mathbf{X})=\frac{1}{IT}\sigma^2_\omega+\frac{r-1}{IT}\psi^A+\frac{m}{IT}\psi^X$$

$$\frac{1}{Imr}\sum_{i=1}^{I}\sum_{t=-m+1}^{0}\sum_{s=1}^{r}\mathrm{Cov}(\bar{\omega}_i,\bar{\bar{\omega}}\mid\mathbf{X})=\frac{1}{IT}\sigma^2_\omega+\frac{m(m-1)}{IT^2}\psi^B+\frac{r(r-1)}{IT^2}\psi^A+\frac{2mr}{IT^2}\psi^X$$

$$\frac{1}{Imr}\sum_{i=1}^{I}\sum_{t=-m+1}^{0}\sum_{s=1}^{r}\mathrm{Cov}(\bar{\omega}_t,\omega_{is}\mid\mathbf{X})=\frac{1}{I}\psi^X$$

$$\frac{1}{Imr}\sum_{i=1}^{I}\sum_{t=-m+1}^{0}\sum_{s=1}^{r}\mathrm{Cov}(\bar{\omega}_t,\bar{\omega}_i\mid\mathbf{X})=\frac{1}{IT}\sigma_\omega^2+\frac{m-1}{IT}\psi^B+\frac{r}{IT}\psi^X$$

$$\frac{1}{Imr}\sum_{i=1}^{I}\sum_{t=-m+1}^{0}\sum_{s=1}^{r}\mathrm{Cov}(\bar{\omega}_t,\bar{\omega}_s\mid\mathbf{X})=\frac{1}{I}\psi^X$$

$$\frac{1}{Imr}\sum_{i=1}^{I}\sum_{t=-m+1}^{0}\sum_{s=1}^{r}\mathrm{Cov}(\bar{\omega}_t,\bar{\bar{\omega}}\mid\mathbf{X})=\frac{1}{IT}\sigma_\omega^2+\frac{m-1}{IT}\psi^B+\frac{r}{IT}\psi^X$$

$$\frac{1}{Imr}\sum_{i=1}^{I}\sum_{t=-m+1}^{0}\sum_{s=1}^{r}\mathrm{Cov}(\bar{\bar{\omega}},\omega_{is}\mid\mathbf{X})=\frac{1}{IT}\sigma_\omega^2+\frac{r-1}{IT}\psi^A+\frac{m}{IT}\psi^X$$

$$\frac{1}{Imr}\sum_{i=1}^{I}\sum_{t=-m+1}^{0}\sum_{s=1}^{r}\mathrm{Cov}(\bar{\bar{\omega}},\bar{\omega}_i\mid\mathbf{X})=\frac{1}{IT}\sigma_\omega^2+\frac{m(m-1)}{IT^2}\psi^B+\frac{r(r-1)}{IT^2}\psi^A+\frac{2mr}{IT^2}\psi^X$$

$$\frac{1}{Imr}\sum_{i=1}^{I}\sum_{t=-m+1}^{0}\sum_{s=1}^{r}\mathrm{Cov}(\bar{\bar{\omega}},\bar{\omega}_s\mid\mathbf{X})=\frac{1}{IT}\sigma_\omega^2+\frac{r-1}{IT}\psi^A+\frac{m}{IT}\psi^X$$

$$\frac{1}{Imr}\sum_{i=1}^{I}\sum_{t=-m+1}^{0}\sum_{s=1}^{r}\mathrm{Cov}(\bar{\bar{\omega}},\bar{\bar{\omega}}\mid\mathbf{X})=\frac{1}{IT}\sigma_\omega^2+\frac{m(m-1)}{IT^2}\psi^B+\frac{r(r-1)}{IT^2}\psi^A+\frac{2mr}{IT^2}\psi^X$$

Combining these terms yields

$$\begin{aligned}
\psi_{\hat{\omega}}^X &= \psi^X-\left[\frac{1}{T}\sigma_\omega^2+\frac{m-1}{T}\psi^B+\frac{r}{T}\psi^X\right]-\frac{1}{I}\psi^X+\left[\frac{1}{IT}\sigma_\omega^2+\frac{m-1}{IT}\psi^B+\frac{r}{IT}\psi^X\right]\\
&\quad-\left[\frac{1}{T}\sigma_\omega^2+\frac{r-1}{T}\psi^A+\frac{m}{T}\psi^X\right]+\left[\frac{1}{T}\sigma_\omega^2+\frac{m(m-1)}{T^2}\psi^B+\frac{r(r-1)}{T^2}\psi^A+\frac{2mr}{T^2}\psi^X\right]\\
&\quad+\left[\frac{1}{IT}\sigma_\omega^2+\frac{r-1}{IT}\psi^A+\frac{m}{IT}\psi^X\right]-\left[\frac{1}{IT}\sigma_\omega^2+\frac{m(m-1)}{IT^2}\psi^B+\frac{r(r-1)}{IT^2}\psi^A+\frac{2mr}{IT^2}\psi^X\right]\\
&\quad-\frac{1}{I}\psi^X+\left[\frac{1}{IT}\sigma_\omega^2+\frac{m-1}{IT}\psi^B+\frac{r}{IT}\psi^X\right]+\frac{1}{I}\psi^X-\left[\frac{1}{IT}\sigma_\omega^2+\frac{m-1}{IT}\psi^B+\frac{r}{IT}\psi^X\right]\\
&\quad+\left[\frac{1}{IT}\sigma_\omega^2+\frac{r-1}{IT}\psi^A+\frac{m}{IT}\psi^X\right]-\left[\frac{1}{IT}\sigma_\omega^2+\frac{m(m-1)}{IT^2}\psi^B+\frac{r(r-1)}{IT^2}\psi^A+\frac{2mr}{IT^2}\psi^X\right]\\
&\quad-\left[\frac{1}{IT}\sigma_\omega^2+\frac{r-1}{IT}\psi^A+\frac{m}{IT}\psi^X\right]+\left[\frac{1}{IT}\sigma_\omega^2+\frac{m(m-1)}{IT^2}\psi^B+\frac{r(r-1)}{IT^2}\psi^A+\frac{2mr}{IT^2}\psi^X\right]\\
&= \psi^X-\left[\frac{1}{T}\sigma_\omega^2+\frac{m-1}{T}\psi^B+\frac{r}{T}\psi^X\right]-\frac{1}{I}\psi^X+\left[\frac{1}{IT}\sigma_\omega^2+\frac{m-1}{IT}\psi^B+\frac{r}{IT}\psi^X\right]\\
&\quad-\left[\frac{1}{T}\sigma_\omega^2+\frac{r-1}{T}\psi^A+\frac{m}{T}\psi^X\right]+\left[\frac{1}{T}\sigma_\omega^2+\frac{m(m-1)}{T^2}\psi^B+\frac{r(r-1)}{T^2}\psi^A+\frac{2mr}{T^2}\psi^X\right]\\
&\quad+\left[\frac{1}{IT}\sigma_\omega^2+\frac{r-1}{IT}\psi^A+\frac{m}{IT}\psi^X\right]-\left[\frac{1}{IT}\sigma_\omega^2+\frac{m(m-1)}{IT^2}\psi^B+\frac{r(r-1)}{IT^2}\psi^A+\frac{2mr}{IT^2}\psi^X\right]
\end{aligned}$$

$$
\begin{aligned}
&= \left[-\frac{1}{T}+\frac{1}{IT}-\frac{1}{T}+\frac{1}{T}+\frac{1}{IT}-\frac{1}{IT}\right]\sigma^2_\omega \\
&\quad + \left[-\frac{m-1}{T}+\frac{m-1}{IT}+\frac{m(m-1)}{T^2}-\frac{m(m-1)}{IT^2}\right]\psi^B \\
&\quad + \left[-\frac{r-1}{T}+\frac{r(r-1)}{T^2}+\frac{r-1}{IT}+\frac{r(r-1)}{IT^2}\right]\psi^A \\
&\quad + \left[1-\frac{r}{T}-\frac{1}{I}+\frac{r}{IT}-\frac{m}{T}+\frac{2mr}{T^2}+\frac{m}{IT}-\frac{2mr}{IT^2}\right]\psi^X
\end{aligned}
$$

$$
\begin{aligned}
&= -\left(\frac{I-1}{IT}\right)\sigma^2_\omega - \left(\frac{(I-1)r(m-1)}{IT^2}\right)\psi^B \\
&\quad - \left(\frac{(I-1)m(r-1)}{IT^2}\right)\psi^A + \left(\frac{2(I-1)mr}{IT^2}\right)\psi^X
\end{aligned}
$$

To summarize, the residual-based variance and covariance parameters can be expressed as functions of the variance and covariance parameters that define the error structure.

$$
\begin{aligned}
\sigma^2_{\hat{\omega}} &= \left(\frac{(I-1)(T-1)}{IT}\right)\sigma^2_\omega - \left(\frac{(I-1)m(m-1)}{IT^2}\right)\psi^B - \left(\frac{(I-1)r(r-1)}{IT^2}\right)\psi^A \\
&\quad - \left(\frac{2(I-1)mr}{IT^2}\right)\psi^X \\
\psi^B_{\hat{\omega}} &= -\left(\frac{I-1}{IT}\right)\sigma^2_\omega + \left(\frac{(I-1)(r^2+2r+m)}{IT^2}\right)\psi^B \\
&\quad + \left(\frac{(I-1)r(r-1)}{IT^2}\right)\psi^A - \left(\frac{2(I-1)r^2}{IT^2}\right)\psi^X \\
\psi^A_{\hat{\omega}} &= -\left(\frac{I-1}{IT}\right)\sigma^2_\omega + \left(\frac{(I-1)m(m-1)}{IT^2}\right)\psi^B \\
&\quad + \left(\frac{(I-1)(m^2+2m+r)}{IT^2}\right)\psi^A - \left(\frac{2(I-1)m^2}{IT^2}\right)\psi^X \\
\psi^X_{\hat{\omega}} &= -\left(\frac{I-1}{IT}\right)\sigma^2_\omega - \left(\frac{(I-1)r(m-1)}{IT^2}\right)\psi^B \\
&\quad - \left(\frac{(I-1)m(r-1)}{IT^2}\right)\psi^A + \left(\frac{2(I-1)mr}{IT^2}\right)\psi^X
\end{aligned}
$$

Factoring $\frac{I-1}{IT^2}$ from each term, we have:

$$
\begin{aligned}
\sigma^2_{\hat{\omega}} &= \left(\frac{I-1}{IT^2}\right)\left(T(T-1)\sigma^2_\omega - m(m-1)\psi^B - r(r-1)\psi^A - 2mr\psi^X\right) \\
\psi^B_{\hat{\omega}} &= \left(\frac{I-1}{IT^2}\right)\left(-T\sigma^2_\omega + (r^2+2r+m)\psi^B + r(r-1)\psi^A - 2r^2\psi^X\right)
\end{aligned}
$$

$$\psi^A_{\hat{\omega}} = \left(\frac{I-1}{IT^2}\right)\Big(-T\sigma^2_\omega + m(m-1)\psi^B + (m^2+2m+r)\psi^A - 2m^2\psi^X\Big)$$
$$\psi^X_{\hat{\omega}} = \left(\frac{I-1}{IT^2}\right)\Big(-T\sigma^2_\omega - r(m-1)\psi^B - m(r-1)\psi^A + 2mr\psi^X\Big)$$

In matrix notation:

$$\begin{bmatrix}\sigma^2_{\hat{\omega}}\\ \psi^B_{\hat{\omega}}\\ \psi^A_{\hat{\omega}}\\ \psi^X_{\hat{\omega}}\end{bmatrix} = \mathbf{\Gamma}\begin{bmatrix}\sigma^2_\omega\\ \psi^B\\ \psi^A\\ \psi^X\end{bmatrix}$$

where

$$\mathbf{\Gamma} = \frac{I-1}{IT^2}\begin{bmatrix} T(T-1) & -m(m-1) & -r(r-1) & -2mr\\ -T & r^2+2r+m & r(r-1) & -2r^2\\ -T & m(m-1) & m^2+2m+r & -2m^2\\ -T & -r(m-1) & -m(r-1) & 2mr\end{bmatrix}$$

Minimum detectable effect We are ultimately interested in deriving an expression for the MDE of an experiment as a function of the residual-based parameters, $\sigma^2_{\hat{\omega}}$, $\psi^B_{\hat{\omega}}$, $\psi^A_{\hat{\omega}}$, and $\psi^X_{\hat{\omega}}$, rather than the true parameters, σ^2_ω, ψ^B, ψ^A, and ψ^X. Recall that:

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

Having solved for $\sigma^2_{\hat{\omega}}$, $\psi^B_{\hat{\omega}}$, $\psi^A_{\hat{\omega}}$, and $\psi^X_{\hat{\omega}}$ as linear functions of the true parameters σ^2_ω, ψ^B, ψ^A, and ψ^X, we can define k_σ, k_B, k_A, and k_X as coefficients on the residual-based parameters $\sigma^2_{\hat{\omega}}$, $\psi^B_{\hat{\omega}}$, $\psi^A_{\hat{\omega}}$, and $\psi^X_{\hat{\omega}}$. These coefficients will allow us to use residual-based parameters in the SCR formula in place of the true parameters. In other words, k_σ, k_B, k_A, and k_X must satisfy the following equation:[28]

$$\left(t^J_{1-\kappa} + t^J_{\alpha/2}\right)\sqrt{\left(\frac{1}{P(1-P)J}\right)\left[\left(\frac{m+r}{mr}\right)k_\sigma\sigma^2_{\hat{\omega}} + \left(\frac{m-1}{m}\right)k_B\psi^B_{\hat{\omega}} + \left(\frac{r-1}{r}\right)k_A\psi^A_{\hat{\omega}} - 2k_X\psi^X_{\hat{\omega}}\right]}$$
$$= \left(t^J_{1-\kappa} + t^J_{\alpha/2}\right)\sqrt{\left(\frac{1}{P(1-P)J}\right)\left[\left(\frac{m+r}{mr}\right)\sigma^2_\omega + \left(\frac{m-1}{m}\right)\psi^B + \left(\frac{r-1}{r}\right)\psi^A - 2\psi^X\right]}$$

[28]This assumes k_σ, k_B, k_A, and k_X are functions of m, r, and I only, and do not themselves depend on σ^2_ω, ψ^B, ψ^A, or ψ^X. We show this to be true below.

or more simply

$$\left(\frac{m+r}{mr}\right)k_\sigma\sigma^2_{\hat{\omega}} + \left(\frac{m-1}{m}\right)k_B\psi^B_{\hat{\omega}} + \left(\frac{r-1}{r}\right)k_A\psi^A_{\hat{\omega}} - 2k_X\psi^X_{\hat{\omega}}$$
$$= \left(\frac{m+r}{mr}\right)\sigma^2_{\omega} + \left(\frac{m-1}{m}\right)\psi^B + \left(\frac{r-1}{r}\right)\psi^A - 2\psi^X$$

We can express this equality in matrix form as[29]

$$\begin{bmatrix}\frac{m+r}{mr} & \frac{m-1}{m} & \frac{r-1}{r} & -2\end{bmatrix}\begin{bmatrix}\sigma^2_{\omega}\\ \psi^B\\ \psi^A\\ \psi^X\end{bmatrix} = \begin{bmatrix}\left(\frac{m+r}{mr}\right)k_\sigma & \left(\frac{m-1}{m}\right)k_B & \left(\frac{r-1}{r}\right)k_A & -2k_X\end{bmatrix}\begin{bmatrix}\sigma^2_{\hat{\omega}}\\ \psi^B_{\hat{\omega}}\\ \psi^A_{\hat{\omega}}\\ \psi^X_{\hat{\omega}}\end{bmatrix}$$
$$= \begin{bmatrix}\left(\frac{m+r}{mr}\right)k_\sigma & \left(\frac{m-1}{m}\right)k_B & \left(\frac{r-1}{r}\right)k_A & -2k_X\end{bmatrix}\mathbf{\Gamma}\begin{bmatrix}\sigma^2_{\omega}\\ \psi^B\\ \psi^A\\ \psi^X\end{bmatrix}$$

where $\mathbf{\Gamma}$ is defined as above:

$$\mathbf{\Gamma} = \frac{I-1}{IT^2}\begin{bmatrix}T(T-1) & -m(m-1) & -r(r-1) & -2mr\\ -T & r^2+2r+m & r(r-1) & -2r^2\\ -T & m(m-1) & m^2+2m+r & -2m^2\\ -T & -r(m-1) & -m(r-1) & 2mr\end{bmatrix}$$

$\mathbf{\Gamma}$ is a singular matrix and cannot be inverted. However, we can show that, rather than having no solutions, this system is instead overdetermined and there are an infinite number of solutions. We are simply interested in one such solution. To find one set of k_σ, k_B, k_A, and k_X coefficients that solve this system, we iteratively solve for each k coefficient as a function of the remaining k coefficients and then substitute into the subsequent equations. That is, we use the first equation of this system to solve for k_σ as a function of k_B, k_A, and k_X and substitute this expression in place of k_σ in the subsequent equations in the system, and we repeat for the remaining coefficients and equations. This iterative procedure yields:

$$k_\sigma = \frac{I(m+r)^2 + (I-1)[r(m-1)k_B + m(r-1)k_A - 2mrk_X]}{(I-1)(m+r)(m+r-1)}$$
$$k_B = \frac{I(m+r)^2(m-r+mr+r^2) + (I-1)[m^2(r-1)(2-m-r)k_A + 2mr(r-m-mr-r^2)k_X]}{(I-1)r^2(3m+r+mr+r^2-2)}$$
$$k_A = \frac{I(m+r)^2 - 2(I-1)mrk_X}{2(I-1)m^2}$$

[29]Note that if $m = 1$ (or $r = 1$), the corresponding ψ^B and $\psi^B_{\hat{\omega}}$ (or ψ^A and $\psi^A_{\hat{\omega}}$) parameters are undefined and no longer enter the system. Similarly, the corresponding row(s) and column(s) are removed from $\mathbf{\Gamma}$.

$k_X = 0$

We iteratively substitute each coefficient into the expressions for the remaining coefficients. That is, we first substitute this value of k_X into the expressions for the other three coefficients, then substitute k_A, and so on. This yields expressions for k_σ, k_B, k_A, and k_X in terms of m, r, and I.

$$\begin{aligned} k_\sigma &= \frac{I(m+r)^2}{2(I-1)mr} \\ k_B &= \frac{I(m+r)^2}{2(I-1)r^2} \\ k_A &= \frac{I(m+r)^2}{2(I-1)m^2} \\ k_X &= 0 \end{aligned}$$

We can now express the MDE of an experiment as a function of the residual-based parameters:

$$MDE^{est} = \left(t^J_{1-\kappa} + t^J_{\alpha/2}\right) \left\{ \left(\frac{1}{P(1-P)J}\right) \left[\left(\frac{m+r}{mr}\right) \left(\frac{I(m+r)^2}{2(I-1)mr}\right) \sigma^2_{\hat{\omega}} + \left(\frac{m-1}{m}\right) \left(\frac{I(m+r)^2}{2(I-1)r^2}\right) \psi^B_{\hat{\omega}} + \left(\frac{r-1}{r}\right) \left(\frac{I(m+r)^2}{2(I-1)m^2}\right) \psi^A_{\hat{\omega}} \right] \right\}^{1/2} \quad \square$$