IT'S ABOUT TIME (SERIES): A SIMPLE CORRECTION FOR DIFFERENCE-IN-DIFFERENCES ESTIMATORS

WORKING PAPER

Gary Cornwall[†] **and Scott Wentland**[††*]

December 31, 2025

ABSTRACT

This paper reconsiders the difference-in-differences (DiD) research design for panel data, particularly when serial correlation stems from first-order model misspecification (i.e., dependence in y_t rather than exclusively in ϵ_t). When time-series issues like this are overlooked, the traditional parallel trends assumption is insufficient. In fact, for most panel applications ($T > 2$ periods), DiD designs will misidentify and inflate a time-invariant treatment effect. To correct this, we show that DiD assumptions should be modified for dynamic panels and how explicitly accounting for temporal dependence in the design can recover the true, dynamically-robust effect. We evaluate a simple modification to DiD designs through Monte Carlo simulations and then explore its implications with empirical examples. Two examples leverage a policy shock used in recent literature to reevaluate the impact of household credit constraints on outcomes like state-level GDP growth and labor market participation. When we implement the proposed modification, which can be as simple as incorporating a lagged outcome and group interaction into a DiD model, the results illustrate a reduction in bias predicted by theory, yielding a more generalizable estimator for most applications. Finally, we find synthetic DiD and synthetic control methods do not remedy this particular issue, as similar modifications (e.g., pre-whitening) are needed to address temporal dependence in the outcome.

JEL Classifications: C21, C22, C23, E44, E65, J21, O46, R11

Keywords: difference-in-differences, panel data, parallel trends, time-series, autoregressive process, serial correlation, credit constraints, labor supply, GDP growth, labor force participation, synthetic control method, synthetic difference-in-differences

† : U.S. Bureau of Economic Analysis (BEA), 4600 Silverhill Road, Suitland, MD 20746, gary.cornwall@bea.gov.
†† : U.S. Bureau of Economic Analysis (BEA), 4600 Silverhill Road, Suitland, MD 20746, scott.wentland@bea.gov.

[*]Any views expressed here are those of the authors and not necessarily those of the BEA or the U.S. Department of Commerce. For helpful comments on prior drafts, we thank Tara Sinclair, Carter Bryson, Marina Gindelsky, Bob Martin, and Brian Quistorff. We would also like to thank participants at the 2025 Midwest Econometrics Group Conference, 2025 GWU Federal Forecasters Conference, 2026 ASSA Annual Meeting, and our ASSA discussant Doug Miller. All errors are our own.

This thing all things devours:
Birds, beasts, trees, flowers;
Gnaws iron, bites steel;
Grinds hard stones to meal;
Slays king, ruins town,
And beats high mountain down.
- *The Hobbit*, J.R.R. Tolkien

1 Introduction

Difference-in-differences (DiD) is one of the most common methods used by economists to conduct causal inference research (Currie et al., 2020; De Chaisemartin and d'Haultfoeuille, 2023). One reason for its popularity is its simplicity; at its core, a DiD research design compares two groups, treatment and control, over time.[2] A simple two-group, two-period (2×2) design can be readily adapted to a regression model and expanded to accommodate more intricate specifications, weighting schemes, and data structures. A second reason for its popularity stems from its accessibility. Researchers can use increasingly accessible observational (non-experimental) data to study a wide array of economic and policy phenomena, rather than relying on a laboratory or field experiment to assign treatment. Social scientists instead exploit a quasi-natural experiment like an unexpected economic or policy shock as a source for exogenous variation, requiring relatively few assumptions to derive plausibly causal estimates. These assumptions include, for example, that the treatment and control groups would have evolved through time similarly (i.e., parallel trends) and that the timing of the shock is neither anticipated nor contemporaneous with other shocks. But, what if the standard assumptions are incomplete? Can we (still) trust DiD estimates?[3]

In this paper, we revisit the assumptions of DiD from a time-series perspective, showing that first-order temporal dependence is critical for practitioners to consider in some of the most common panel data research designs.[4] While second-order temporal dependence, often referred to as arbitrary serial correlation, has been previously shown to affect standard errors (Bertrand et al., 2004), we show how autocorrelation in the outcome itself, or first-order dependence, can introduce an identification issue when panels extend beyond two periods. Under the standard identifying assumptions in DiD, a constant treatment effect (one-time shifter) will be misidentified as a function of this treatment effect and its propagation through time, dynamically inflating the measured effect. At the heart of the issue, with only some abuse of an analogy, is a simple idea that an experiment measuring how far a golf ball can be hit with different clubs, for example, would need to account for gravitational force, particularly if the experiments were conducted on Earth and the Moon. We provide both theory and evidence for this, arguing that the parallel trends assumption, which is central to identification in DiD, should be modified to account for other forces like first-order temporal dependence in most panel applications. Through Monte Carlo simulations, we illustrate both the extent of the problem and a simple remedy, which we explore further by replicating two recent DiD studies that employ a commonly used panel structure (i.e., a two-way fixed effects (TWFE) DiD model estimated on a panel of U.S. states). Overall, our results suggest that, yes, we *can* trust DiD estimates, provided that the relevant assumptions are satisfied *and* practitioners account for time-series issues that come into play when the scope of analysis expands beyond a 2×2 DiD construction.

[2] This can be as simple as 1) measuring an average outcome, $\bar{Y}_{g,t}$, for the treated group T across two time periods ($t \in \{0, 1\}$) before and after a treatment ($\bar{Y}_{T,1}$ - $\bar{Y}_{T,0}$) such as a policy change or economic shock, and 2) comparing this difference to another difference in outcomes over the same timeframe for a control group C (($\bar{Y}_{T,1}$ - $\bar{Y}_{T,0}$) - ($\bar{Y}_{C,1}$ - $\bar{Y}_{C,0}$)).

[3] We use the term 'trust' here in the same vein as the methodological literature concerned with statistical inferences made from DiD estimators that do not adequately account for a particular issue, such as arbitrary serial correlation (Bertrand et al., 2004), staggered adoption of a treatment (Baker et al., 2022), or spatial correlation Ferman (2023). Each of these papers asks some variation of "how much should we trust DiD estimates" in its title.

[4] The answer to Gollum's final riddle in the epigraph is, of course, time. This is more than a nod to the paper's general theme, but rather it foreshadows a few ideas central to the paper. First, to the extent the riddle is effective, it plays on our intuition to overlook time itself as a powerful, momentum-driving force. Second, time's effect is not uniform; it becomes a more binding constraint depending on the context or object, devouring birds and flowers more quickly than chewing through iron and steel. Finally, only after realizing time is the answer does it seem obvious all along.

For practitioners using panel data with multiple ($T > 2$) periods, there are a few practical takeaways from our analysis below. First, researchers should evaluate whether the outcome of interest follows an autoregressive process; and, if so, explicitly incorporate this into a dynamic panel design. A failure to incorporate this structure leads to an identification bias whereby a constant treatment effect will be properly evaluated only in the initial treatment period, but incrementally inflated in successive periods until converging to a long-run equilibrium state. In this circumstance, accounting for the autocorrelation directly in the model can effectively decompose the treatment estimate into two multiplicative elements: its time-invariant structural component and dynamic pathing (time-dependent) component. The proposed remedy is straightforward to implement, which can be as simple as incorporating lags of the outcome as regressors (or a single lag in the case of an AR(1) process).[5] Further, to allow for group-level heterogeneity in the autoregressive parameter, simply add interaction terms between the lagged outcome(s) and group assignment indicator(s). The resulting DiD modification yields an unbiased estimate of a time-invariant treatment value with better statistical properties in most panel applications, even in relatively small ($T = 10$) panel settings.[6]

To illustrate this DiD modification using real data, we replicate two recently published papers (Kumar and Liang (2019) and Kumar and Liang (2024)) that employ a commonly used TWFE design on a panel of state-level data.[7] Both papers leverage a shock to household credit constraints that occurred in a single treatment state, Texas, but on different state-level outcomes: GDP growth and labor force participation. This pair of papers allows us to compare a DiD outcome like labor force participation, which has higher time-dependence, and a differenced outcome like GDP growth (i.e., expressed in terms of growth rates) with lower dependence.[8] Consistent with theory, we find that the reported TWFE estimates were generally inflated in proportion to its outcome's time-dependence; and, the modified DiD estimator (which we refer to as a dynamically-robust estimator) is smaller in absolute value and substantially more precise.

A second practical takeaway, which is arguably more important than improving model specification and precision, concerns the (mis)interpretation of an average DiD effect itself. Without decomposing the effect of a treatment into its core time-invariant and dynamic elements, users of empirical DiD research (e.g., policymakers) can mistakenly generalize an inflated effect of a policy or shock. To see why, consider an example from Kumar and Liang (2024), where policy changes in Texas in 1997 and 2003 lowered the labor force participation rate (LFPR) by about 1 and 2 percentage points, respectively, relative to a control group of states.[9] A policymaker from another state interested in implementing a similar policy as Texas in 1997 might predict that its LFPR would also decline by about 1 percentage point (or, alternatively, reversing such a policy to increase LFPR by 1 percentage point). This prediction, of course, requires a number of assumptions; but one, which our analysis underscores, is that this prediction would need to assume no dynamic effects (or, at least, this state shares the same autoregressive process as Texas over this period) to generalize this result. Yet, empirically we find state-level labor force participation rates exhibit moderate-to-high autocorrelation along with heterogeneity across states. In the case of Kumar and Liang (2024) we estimate an autoregressive parameter (i.e., coefficient on the lagged outcome term) to be about 0.5, suggesting that about half of the

[5] Using lagged regressors in combination with individual fixed effects has a well-known tradeoff, a bias known as "Nickell Bias", which is relatively straightforward to compensate for using a general method of moments estimator (Arellano and Bond, 1991). We return to this later and explore a Whitening procedure as a pre-estimation step.

[6] Because the inclusion of the lagged outcomes as regressors facilitates a well-identified estimate of the treatment, we can be confident in the structure even when panel exhibits heterogeneous treatment effects and staggered adoption. We cover more complex issues related to staggered and synthetic DiD designs later in the paper.

[7] We replicate a third paper, Arkhangelsky et al. (2021), to examine our remedy in the context of synthetic DiD.

[8] Another reason why we chose these two papers is that they employ a very common data structure and research design in applied microeconomics, relying on a policy change at the state level and conducting a panel data analysis on state-level outcomes over many time periods. Bertrand et al. (2004) and Moulton (1990), for example, provide similar reasoning for using state-level panel data for illustrating econometrics issues. Moreover, their results are easily replicable thanks to the clarity of their papers and replication code, which we are grateful for.

[9] Kumar and Liang (2024) cite their preferred estimates as -0.8 and -1.8, respectively, for the 1997 and 2003 policy changes that increased access to home-equity loans, cash-out refinancing, and home-equity lines of credit in Texas. We round up in this example, but the rounding is also consistent with the average across specifications for their preferred control group (energy-producing states). Later, we return to the details of our replication of this paper and summarize its different specifications in greater depth.

inflated, unpurged DiD effect can be interpreted as an average treatment effect and the remaining portion attributable to its time-dependence. In other words, the policy change had a smaller (and still significant) generalizable effect, but its (potentially Texas-specific) momentum or time-dependence drove the larger effect locally.[10] Thus, by accounting for the autoregressive process more explicitly in the model, the dynamically-robust modification will produce a more generalizable estimator in practice.

A third takeaway for practitioners is that the concept of parallel trends in the pre-treatment period should also account for a process that is first-order temporally dependent. This is important given the prevalence of researchers illustrating how the outcome for the treatment and control groups evolved similarly prior to the treatment period, either in the raw summary statistics, conditioned on covariates, or as part of an event study analysis. Although parallel trends in the pre-treatment period does not guarantee the assumption has been satisfied (Cunningham (2021)),[11] if we do not find evidence that the treatment and control groups evolved similarly at some point, practitioners will generally interpret this as evidence that the control group is not an appropriate counterfactual. That is, once practitioners observe non-parallel trends in the pre-treatment period, they will either abandon the DiD specification or search for another control group that does pass this test (usually in anticipation of editors/reviewers requiring such evidence for publication). Our Monte Carlo simulations show that this approach is mistaken, at least for panel data with moderate to severe autocorrelation in the outcome of interest. Indeed, two groups can have an identical trend, but in the presence of an autoregressive process the traditional event study or "eyeball test" would reveal non-parallel trends in the pre-treatment period. Recent innovations in DiD such as the Callaway and Sant'Anna (2021) design and the Arkhangelsky et al. (2021) synthetic DiD design will not remedy this particular issue. Heeding this takeaway could help practitioners avoid discarding a viable control group by simply accounting for the autoregressive process directly as a dynamic panel or conducting a whitening procedure prior to estimation, including in synthetic DiD designs.

Given the growth in the methodological DiD literature, which is well-summarized by recent reviews such as Roth et al. (2023), De Chaisemartin and d'Haultfoeuille (2023), Callaway (2023), Arkhangelsky and Imbens (2024), and Baker et al. (2025), how is it possible that relatively low-hanging fruit has been overlooked and is still ripe for picking in this literature? Moreover, have these time-series issues not already been tackled by Bertrand et al. (2004) and others? In the next section, we answer these questions by summarizing the relevant literature through the lens of a well-known empirical example, as we attempt to clarify how this paper fits into this growing DiD methodological literature. The paper proceeds with analysis that more explicitly integrates the autoregressive process into panel DiD designs. Finally, we report results from simulations and replications that illustrate several lessons for practitioners, along with further discussion to conclude.

2 Background and Contribution

2.1 Gap DiD versus Trend DiD - Two Distinct Perspectives in the Literature

The 2×2 DiD setup serves as a convenient simplification for both applied and methodological work. Card and Krueger (1994), for example, employ a variation of this design in their study of employment in the fast food industry, where New Jersey was "treated" with a policy change to its minimum wage and a bordering state, Pennsylvania, functioned as the control labor market.[12] While their paper contains numerous regression specifications, survey articles

[10] Rather than an average treatment effect (ATE), one might interpret the unpurged results from Kumar and Liang (2019) or Kumar and Liang (2024) (or numerous other studies that do not account for the autoregressive process explicitly) as analogous to, or perhaps a special case of, a local average treatment effect (LATE).

[11] See Cunningham's (2021) "rant about parallel pre-treatment DD coefficients," as he describes it. He argues: "Assuming that the future is like the past is a form of the gambler's fallacy called the 'reverse position.' Just because a coin came up heads three times in a row does not mean it will come up heads the fourth time—not without further assumptions. Likewise, we are not obligated to believe that that counterfactual trends would be the same post-treatment because they had been similar pre-treatment without further assumptions...". See Roth (2022) for further discussion of additional issues with pre-treatment tests.

[12] New Jersey raised its minimum wage from \$4.25 to \$5.05 per hour in November 1992, while Pennsylvania had held it at \$4.25. The authors conducted a survey of fast food restaurants near the border of these two states before and after the policy change.

and econometrics texts (e.g., Angrist and Pischke (2009) or Cunningham (2021)) often reproduce Card and Krueger (1994)'s presentation of DiD via group averages before and after the policy implementation as an illustrative example of a 2×2 case.[13] We reproduce these estimates in Table 1 below, showing their well-known result that New Jersey's minimum wage hike increased employment in the fast food industry by 2.76 full time equivalent (FTE) employees per store (bottom right cell) relative to Pennsylvania.

[Insert Table 1 Here]

Economists use this type of table pedagogically to arrive at the 2.76 FTE result from two different perspectives. On one side of the coin, we can view DiD as a series of cross-sectional group comparisons. Prior to the minimum wage hike New Jersey had a 2.89 lower FTE employees per store compared to Pennsylvania (i.e., $20.44 - 23.33$). If this "gap" would remain the same in the second period, the difference-in-differences estimate would be zero. But, as this cross-group gap widens or narrows, we observe a nonzero DiD estimate. In this case, the gap narrowed to a difference of -0.14 (i.e., $21.03 - 21.17$), yielding the DiD estimate of 2.76.[14] In an instructive survey of the DiD literature, Baker et al. (2025) refer to these cross-group gap comparisons as a Gap DiD.[15] On the other side of the same coin, we can (and should) think of DiD as a series of time-series comparisons, which Baker et al. (2025) refer to as a Trend DiD. New Jersey's employment rose from 20.44 to 21.03 (or, +0.59) FTE employees per store over this timeframe. When we compare New Jersey's moderate upward trend to Pennsylvania, which trended downward over the two periods by 2.15 (or, 21.17 - 23.33), we arrive at the same final DiD estimate of 2.76 FTE employee increase in New Jersey.

Beyond its pedagogical utility, and more germane to this paper, Table 1 also serves as useful tool for understanding the streams of methodological literature on DiD and how to view the contribution of this paper. In the Card and Krueger (1994) example, the Gap DiD perspective begins with a NJ-PA gap prior to the treatment implementation.[16] This naturally shifts a practitioner's focus toward *ex ante* group differences that could be relevant for parallel trends and thus the suitability of the control group to serve as the counterfactual.[17] In other words, what is different about Pennsylvania that could have contributed to this gap in the first place, potentially making it an unsuitable counterfactual after the treatment? Card and Krueger (1994), of course, consider this issue in greater depth and incorporate covariates into their DiD regression specifications, which is another way of thinking about parallel trends as a conditional concept. The literature has since expanded on this idea of parallel trends conditional on covariates,[18] as well as weighting, selection, and covariate balance issues that address cross-sectional differences across treatment and control groups.[19]

Another way of thinking about the DiD literature to this point (at the cost of mixing metaphors) is that much of the focus has gravitated toward plucking the fruit from one (Gap DiD) side of the tree. If we walk around to the other side of this "tree" to view it from the Trend DiD perspective, low-hanging fruit may still be ripe for picking for at least a couple reasons. One reason is rooted in the canonical 2×2 construction of DiD: two periods barely constitute a

[13] Litigating the findings and details of a 30+ year old study falls outside the scope of this paper. Regardless of one's assessment of Card and Krueger (1994), the paper is undoubtedly well-known and familiar to economists, which is the primary purpose of using this example.

[14] Whether we interpret the DiD estimate as a causal effect depends on the context and that the research design satisfies a set of assumptions, which we discuss at greater length in Section 3.

[15] Note that the Gap DiD calculation is simply rearranging terms in the DiD calculation footnoted in the first paragraph above: ($\bar{Y}_{T,1}$ - $\bar{Y}_{C,1}$) - ($\bar{Y}_{T,0}$ - $\bar{Y}_{C,0}$).

[16] In a regression DiD specification, this gap would be estimated as the coefficient on the treatment group indicator or as an individual/group fixed effect in a TWFE specification.

[17] In a footnote, Baker et al. (2025) note that for a specification with a time invariant mean conditional on the fixed effects, "some researchers may find easier to understand these as "parallel changes" rather than "parallel trends". However, the use of "parallel trends" is now firmly established in the literature..." (fn. 6).

[18] See, for example, Abadie (2005), Athey and Imbens (2006), Sant'Anna and Zhao (2020), Caetano et al. (2022), Caetano and Callaway (2024).

[19] See Ho et al. (2007), Khan and Tamer (2010), Cefalu et al. (2020), Arkhangelsky et al. (2021), Słoczyński (2022), Sant'Anna and Xu (2023), Wooldridge (2023), Ye et al. (2024), and Goldsmith-Pinkham et al. (2024) for further analysis and discussion of weighting, composition, and balance issues in DiD and quasi-experimental methods more generally. For additional issues related to parallel trends, see Bilinski and Hatfield (2018), Marcus and Sant'Anna (2021), Roth and Sant'Anna (2023), and Rambachan and Roth (2023). For a more comprehensive review, see Roth et al. (2023) or Baker et al. (2025).

time-series. In the New Jersey and Pennsylvania example from Card and Krueger (1994), the two period case does not allow for an assessment of serial correlation or autoregressive processes. Hence, many time-series issues simply do not come into play or are easily overlooked when the analysis centers on two periods. A second reason is that, once the setup expands beyond two periods, it introduces *new* cross-sectional comparison issues. Recent literature explores issues related to time, like anticipation,[20] or staggered implementation of a treatment over multiple time periods,[21] but they still tend to don a Gap DiD lens when they go beyond the 2×2 case. The staggered implementation DiD literature, for example, primarily uses multiple time periods as a vehicle to re-examine a special case of cross-sectional comparisons over time (i.e., the "forbidden comparisons" issue) rather than time-series issues per se.[22]

One notable exception, which *does* in fact focus on time-series-specific issues of DiD, is the work by Bertrand et al. (2004). Among its many contributions to this literature, Bertrand et al. (2004) point out that the vast majority of empirical papers using DiD go beyond two periods in their analysis, and they show that arbitrary serial correlation generates inconsistent standard errors in multi-period panel applications. Failure to account for this leads to higher rejection rates of the null hypothesis; or, put simply, standard errors tend to be "too small" in most panel applications. Indeed, they point out that 75 percent of DiD studies (published in the decade prior) had more than two periods of data,[23] thus motivating further exploration of time-series issues.

2.2 Contribution - where does this paper fit in?

This paper adds to the DiD literature by revisiting time-series issues initially raised by Bertrand et al. (2004), expanding the analysis to parameter estimates and the assumptions of DiD more generally. Bertrand et al. (2004) explicitly avoid issues concerning bias, as they are clear from the outset that they "assume away biases in estimating the intervention's effect and instead focus on issues relating to the *standard error* of the estimate" (Bertrand et al. (2004), p. 250, with emphasis in the original). The insights from their paper are well-documented in econometric textbooks covering DiD, motivating practitioners to reexamine how they handle standard errors (e.g., block bootstrapping, clustering at the group level) and their data structure (e.g., aggregation) to address arbitrary serial correlation and standard errors that are "too small" (Cunningham (2021)). Because their objective was not focused on bias, it should not be surprising that their remedies do not mitigate the inflation of the treatment effect or autoregressive bias, which we confirm analytically and through simulations in the proceeding sections.

Though other papers have explored various time-series issues with DiD,[24] Arkhangelsky and Imbens (2024)'s survey of recent literature points out a critical gap, which is that the DiD literature "often pay(s) little explicit attention to dynamics and time-series structure in potential outcomes" despite their observation that "the absence of dynamic effects is unlikely to ever hold exactly" (Arkhangelsky and Imbens (2024), p. C2). Thus, we proceed by expanding on Bertrand et al. (2004) and answering the call from Arkhangelsky and Imbens (2024) to further explore dynamic effects in DiD, as we diagnose and offer some remedies for time-series problems associated with misspecifying DiD designs when the outcome follows an autoregressive process.

[20] For a review and further analysis of anticipation issues, see Piccininni et al. (2025).

[21] See, for example, Callaway and Sant'Anna (2021), Goodman-Bacon (2021), and Dube et al. (2025).

[22] For a helpful summary of forbidden comparisons in staggered DiD settings, see Callaway (2023)'s review.

[23] A more recent survey of the literature by Currie et al. (2020) shows that DiD continues to be one of the dominant methods used by economists. Paul Goldsmith-Pinkham's dashboard confirms these trends for NBER Working Papers and papers published in AEA journals. Neither source, unfortunately, provide a breakdown of how many of these studies use more than two periods of data in their analysis.

[24] For recent examples, see Hsiao and Zhou (2024); Dube et al. (2025); Piger and Stockwell (2025). See also Burlig et al. (2020), which explores a related idea on serial correlation in field experiments and its impact on power calculations for research designs using panel data.

3 Identification in Dynamic Settings

Much of the modern difference-in-differences (DiD) literature formalizes identification using the potential outcomes framework and expresses assumptions in terms of conditional expectations (e.g., Callaway and Sant'Anna (2021); Arkhangelsky and Imbens (2024)). However, as emphasized by Arkhangelsky and Imbens (2024), the literature pays relatively little attention to the *time-series structure* of potential outcomes.[25] In many applications, outcomes for both treatment and control units evolve dynamically and exhibit serial dependence well beyond what is assumed in canonical multi-period DiD designs. This section revisits identification in such environments, clarifying what the standard DiD estimator identifies and introducing a dynamically robust version of the parallel trends assumption that recovers a time-invariant, structural treatment effect in the presence of first-order temporal dependence.

3.1 The Basic DiD Framework

In general, we observe N units, $i = 1, \ldots, N$ over periods $t = 1, \ldots, T$ in a panel. Each unit belongs to one of K mutually exclusive groups $g_i \in \{1, \ldots, K\}$, where group $k = 1$ is never treated and $k > 1$ receives treatment in period $\tau_k (1 < \tau_k < T)$. Treatment is absorbing, meaning once treated the group will remain treated, and occurs only once for each group. Let $y_{i,t}(0)$ denote the untreated potential outcome for unit i in period t and $y_{i,t}(1)$ the treated potential outcome. The realized outcome is

$$y_{i,t} = y_{i,t}(0) + d_{i,t}\big(y_{i,t}(1) - y_{i,t}(0)\big), \tag{1}$$

where $d_{i,t} = \mathbf{1}\{t \geq \tau_{g_i}\}$ is the treatment indicator. The object of interest is often the average treatment effect on the treated (ATT)

$$\text{ATT}_{k,t} = \mathbb{E}[y_{i,t}(1) - y_{i,t}(0) | g_i = k], \text{ for } t \geq \tau_k. \tag{2}$$

Difference-in-differences is built upon two primary assumptions to plausibly identify the counterfactual group: *No Anticipation* and *Parallel Trends*. These serve to impute, in expectation, the "what if the treatment never occurred" counterfactual argument. More formally, in the 2×2 case, we express the these as

$$\mathbb{E}_2[y_{i,1}(1)] = \mathbb{E}_2[y_{i,1}(0)] \tag{A1a}$$

$$\mathbb{E}_2[y_{i,2}(0)] - \mathbb{E}_2[y_{i,1}(0)] = \mathbb{E}_1[y_{i,2}(0)] - \mathbb{E}_1[y_{i,1}(0)] \tag{A2a}$$

where going forward we will notate $\mathbb{E}_k[\cdot] = \mathbb{E}[\cdot \mid i \in k]$ so as limit the notational burden. The first states that, prior to the intervention or shock, the expected outcomes for those in a treated group will be as though they were in a world absent the intervention. The second, and perhaps the most controversial (see Bilinski and Hatfield (2018); Rambachan and Roth (2023); Roth and Sant'Anna (2023); Caetano and Callaway (2024) for example) states that, absent the intervention, the treated group would have evolved over the two periods in question identically to the untreated group.[26]

More generally, for $K > 2$ groups and $T > 2$ periods we might express these assumptions as

$$\mathbb{E}_k[y_{i,t}(1)] = \mathbb{E}_k[y_{i,t}(0)] \text{ for all } t \leq \tau_k \tag{A1b}$$

$$\mathbb{E}_k[y_{i,t}(0)] - \mathbb{E}_k[y_{i,s}(0)] = \mathbb{E}_1[y_{i,t}(0)] - \mathbb{E}_1[y_{i,s}(0)] \text{ for all } t \geq s \text{ and } k > 1 \tag{A2b}$$

[25] As noted in the prior section, a well-known exception to this is Bertrand et al. (2004), which showed that serial correlation in the error term leads to a block-diagonal covariance matrix that violates assumptions of standard two-way fixed effects estimators using OLS. The result is an empirical Type I error rate that far surpasses the nominal and negatively impacts inference. Their proposed correction included either collapsing the time dimension to two periods, pre- and post-treatment, or using a more general covariance estimator (e.g., clustering).

[26] There are other assumptions, and variations of these assumptions, that are important for identification. For example, one variation is the absence of relevant contemporaneous shocks that occur with the treatment timing.

where A1b states that, for any pre-intervention period, the treated group will have expected outcomes consistent with a world in which there was no intervention.[27] Meanwhile A2b states that the difference in expected outcomes for any pair of periods (t, s) is equivalent between any treated group and the never treated group.[28] In a static world, these assumptions are well understood. However, as we show next, when untreated potential outcomes evolve dynamically, they imply additional restrictions that are rarely acknowledged.

3.2 Parallel Trends: Second Moment Implications

To see how Assumption A2b implies additional restrictions, first let us define the autocovariance for unit i in group k between a specific period t and a reference period $h < t$ as:

$$\gamma_k(t, h) = \mathbb{E}_k\big[y_{i,t}(0)y_{i,h}(0)\big] - \mathbb{E}_k\big[y_{i,t}(0)\big]\mathbb{E}_k\big[y_{i,h}(0)\big]. \tag{3}$$

The standard DiD setup identifies the counterfactual for the treated group k, in period t as

$$\hat{Y}_{k,t}(0) = \mathbb{E}_k[y_{i,s}] + \underbrace{(\mathbb{E}_1[y_{i,t}] - \mathbb{E}_1[y_{i,s}])}_{\text{Control Group Trend}}. \tag{4}$$

We can then rearrange Equation 3 and substitute it into the control group trend such that

$$\hat{Y}_{k,t}(0) = \mathbb{E}_k[y_{i,s}] + \left(\frac{\mathbb{E}_1[y_{i,t}y_{i,s}]}{\mathbb{E}_1[y_{i,s}]} - \frac{\gamma_1(t,s)}{\mathbb{E}_1[y_{i,s}]} - \mathbb{E}_1[y_{i,s}]\right)$$

From here we note that the imputed outcome, $\hat{Y}_{k,t}(0)$ relies on $\gamma_1(t, s)$ explicitly. However, we know that the true counterfactual relies on $\gamma_k(t, s)$ through similar methods.

By applying the same decomposition to the unobserved true counterfactual, $\mathbb{E}_k(y_{i,t}(0)]$, we can express the "bias" of the standard DiD estimator as the difference between the true and imputed outcomes

$$\begin{aligned}\mathbb{E}_k[y_{i,t}(0)] - \hat{Y}_{k,t}(0) \;&= \left[\frac{\mathbb{E}_k[y_{i,t}y_{i,s}] - \gamma_k(t,s)}{\mathbb{E}_k[y_{i,s}]}\right] \\ &\quad - \left[\mathbb{E}_k[y_{i,s}] + \left(\frac{\mathbb{E}_1[y_{i,t}y_{i,s}] - \gamma_1(t,s)}{\mathbb{E}_1[y_{i,s}]} - \mathbb{E}_1[y_{i,s}]\right)\right].\end{aligned} \tag{5}$$

At first glance, this expression is seemingly complicated, though the implication is relatively straightforward. For the estimator to be unbiased, it is not enough for the groups to share a common trend, they must also share a common memory structure such that $\gamma_1(t, s) = \gamma_k(t, s)$. If the treated group is more persistent than the control group, the standard DiD estimator will fail to correctly predict the mean reversion of the treated group, leading to a biased estimate of the treatment effect. Consequently, while Assumption A2b is stated in terms of means, it implicitly enforces a restriction of **Dynamic Homogeneity**. When this restriction is violated parallel trends fails and standard identification breaks down.

Now let's suppose the Dynamic Homogeneity condition is satisfied, $\gamma_k(t, s) = \gamma_1(t, s) = \gamma(t, s)$, what does this mean for identification under Assumptions A1b and A2b? Substituting the decomposition into the definition of the average treatment on the treated yields

$$\text{ATT}_{k,t} = \mathbb{E}_k\left[\frac{\mathbb{E}_k\big[y_{i,t}(1)y_{i,s}(1)\big] - \gamma(t,s)}{\mathbb{E}_k\big[y_{i,s}(1)\big]}\right] - \mathbb{E}_k\left[\frac{\mathbb{E}_k\big[y_{i,t}(0)y_{i,s}(0)\big] - \gamma(t,s)}{\mathbb{E}_k\big[y_{i,s}(0)\big]}\right]. \tag{6}$$

[27] For now we will assume there is no anticipation though, conditional on the anticipation being known one could express this in a slightly less restrictive way as done in Callaway and Sant'Anna (2021).

[28] Note that this is a slightly different formulation than that found in Callaway and Sant'Anna (2021) which uses a fixed reference point of the last untreated period, $\tau_k - 1$. In practice, it isn't just the last untreated period that is of interest in convincing a reader that parallel trends holds.

Let the treatment be a structural parameter λ that shifts the potential outcome such that $y_{i,t}(1) = y_{i,t}(0) + \lambda$. Because the process is persistent (governed by γ), this structural shock does not remain isolated in period t; it interacts with the history of the process. The term $\mathbb{E}_k[y_{i,t}(1)y_{i,s}(1)]$ captures not just the shift λ, but the propagation of that shift through the covariance structure $\gamma(t,s)$. Consequently, the estimator $\widehat{\text{ATT}}_{k,t}$ does not recover the structural parameter λ in isolation. Instead, it identifies a function $f(\lambda, \gamma)$ representing the *cumulative dynamic path* of the treatment. The structural parameter itself is only recovered in the special case where the process is memory-less (or lacks temporal dependence):

$$\widehat{\text{ATT}}_{k,t} = \lambda \iff \gamma(t,s) = 0. \tag{7}$$

In all other cases where $\gamma(t,s) \neq 0$, the standard DiD structure conflates the magnitude of the intervention with the persistence of the outcome variable. This distinction is vital for external validity: an estimate derived from a highly persistent setting cannot be directly transported to a less persistent setting, even if the structural impact (one-time shifter) of the policy λ is identical in both.

3.3 Dynamically Robust Difference-in-Differences

To resolve these limitations, it is necessary to decouple the structural treatment effect from the group-specific propagation. While the previous section characterized the bias using the autocovariance function, $\gamma_k(t,s)$, we can equivalently represent this structure using a linear filter.

Assume that the untreated potential outcome $y_{i,\cdot}(0)$ is weakly stationary around a deterministic trend. By the Wold Decomposition Theorem (see Brockwell and Davis (2009), Theorem 5.7.1), the autocovariance structure γ_k implies a group-specific, invertible linear filter $B_k(L) = 1 - \sum_{j=1}^{\infty} b_{k,j} L^j$. Applying this filter to the outcome $y_{i,t}$ transforming the process into a sequence of white-noise innovations, $y^*_{i,\cdot}$, that is free of first-order temporal dependence.

We define the "pre-whitened" untreated potential outcome by applying this group-specific filter

$$y^*_{i,t}(0) = B_k(L) y_{i,t}(0) \tag{8}$$

where $y^*_{i,t}(0)$ represents the variation left in period t after accounting for the unit's history. By moving from levels to innovations we can restate the identification assumptions in a way that is robust to dynamic heterogeneity.

First, we restate the No Anticipation assumption in terms of the pre-whitened outcomes

$$\mathbb{E}_k[y^*_{i,t}(1)] = \mathbb{E}_k[y^*_{i,t}(0)] \text{ for all } t < \tau_k. \tag{A3a}$$

This ensures that the innovations are unaffected by the treatment prior to its implementation. Next, we introduce the **Dynamically Robust Parallel Trends** assumption

$$\mathbb{E}_k[y^*_{i,t}(0)] - \mathbb{E}_k[y^*_{i,s}(0)] = \mathbb{E}_1[y^*_{i,t}(0)] - \mathbb{E}_1[y^*_{i,s}(0)] \text{ for all } t, s \text{ and } k > 1. \tag{A3b}$$

This assumption is weaker than the standard parallel trends assumption since the pre-whiteneing filter removes first-order temporal dependence and the resulting innovations are memory-less by construction.

We can now define a new estimand, $\nu^*_{k,t,s}$ as the standard DiD estimand applied to the dynamically-adjusted data

$$\nu^*_{k,t,s} = \big(\mathbb{E}_k[y^*_{i,t}] - \mathbb{E}_k[y^*_{i,s}]\big) - \big(\mathbb{E}_1[y^*_{i,t}] - \mathbb{E}_1[y^*_{i,s}]\big). \tag{9}$$

Under Assumptions A3a and A3b, this estimand identifies the structural shock λ rather than the cumulative path:

$$\nu^*_{k,s,t} = \mathbb{E}_k[y^*_{i,t}(1) - y^*_{i,t}(0)] = \lambda_k. \tag{10}$$

The key advantage of this formulation is twofold. First, it is robust to heterogeneous dynamics; Equation A3b remains plausible even when the original series have $\gamma_k \neq \gamma_1$, a setting that invalidates standard DiD. Second, it isolates the

structural parameter λ_k from its cumulative propagation. By decoupling the immediate structural impact (i.e., shifter) from the series' temporal persistence, this estimand facilitates external validity. This enables researchers to better compare effects across studies or help inform predictions for how the same treatment effect would propagate in a new environment with different dynamics.[29]

3.4 An Example: The AR(1) Linear Process

Suppose we have a structural model where the treatment is an additive, time-invariant shock λ_k that enters the innovation of the untreated potential outcome process exhibiting first-order temporal dependence. Let the potential outcomes for unit i in period t be defined as,

$$\Phi_k(L)y_{i,t} = \alpha_i + \delta t + \epsilon_{i,t} \text{ for } i \in k = 1 \tag{11}$$

$$\Phi_k(L)y_{i,t} = \alpha_i + \delta t + \lambda_k \mathbf{1}\{i \in k\}\mathbf{1}\{t \geq \tau_k\} + \epsilon_{i,t} \text{ for } i \in k > 1 \tag{12}$$

where δ is a common trend term, α_i a unit specific intercept, $\Phi_k(L) = 1 - \phi_{k,1}L$ with $|\phi_{k,1}| < 1$ for all k, and $\epsilon_{i,t}$ a mean-zero white noise process with variance σ_k^2.[30] Further assume that the observations are independent such that this process can be written equivalently as

$$y_{i,t} = \alpha_i + \delta t + \phi_{k,1}y_{i,t-1} + \epsilon_{i,t} \tag{13}$$

$$y_{i,t} = \alpha_i + \delta t + \phi_{k,1}y_{i,t-1} + \lambda_k \mathbf{1}\{i \in k\}\mathbf{1}\{t \geq \tau_k\} + \epsilon_{i,t}. \tag{14}$$

From this, we know that the unconditional (on past history) expectations for the untreated group can be expressed as $\mathbb{E}_{1,t} = \mathbb{E}[\alpha]/(1-\phi_1) + \delta t/(1-\phi_1)$ in both $Y(0)$ and $Y(1)$. For any treated group, $k > 1$ and $t < \tau_k$, we know that $\mathbb{E}_{k,t} = \mathbb{E}[\alpha]/(1-\phi_k) + \delta t/(1-\phi_k)$ and for $t \geq \tau_k$ and $\mathbb{E}_{k,t} = \mathbb{E}[\alpha]/(1-\phi_{k,1}) + \delta t/(1-\phi_{k,1}) + \sum_{j=0}^{t-\tau_k} \phi_k^j \lambda_k$.

If we use Assumptions A1b and A2b to form the standard difference-in-differences estimand, we find

$$\nu_{k,t,s} = \underbrace{\lambda_k \sum_{j=0}^{t-\tau_k} \phi_k^j}_{\text{Cumulative Dynamic Path}} + \underbrace{(t-s)\delta\left(\frac{1}{1-\phi_k} - \frac{1}{1-\phi_1}\right)}_{\text{Heterogeneous Autocorrelation Bias}}. \tag{15}$$

The first term is the constant treatment effect and its propagation over post treatment periods. Since we know the process is weakly stationary as $j \to \infty$ this will converge to $\lambda_k/(1-\phi_k)$. The second term is the bias potentially incurred by heterogeneous covariance structures between the treated and control groups; as $\phi_k \to \phi_1$ this term disappears, otherwise it will grow linearly with time at a rate consistent with the difference between the two autocovariance values. This second term is critical for most panel data applications, because if the autocovariance functions are not accounted for, the implied effect of the treatment grows in time and absent any additional future changes the treatment and control group would diverge *ad infinitum*. Meanwhile, using Assumptions A3a and A3b we find,

$$\nu^*_{k,t} = \lambda_k$$

which corresponds to the original time-invariant shock, which should ultimately be the DiD estimator of interest.

[29] We would like to note that in many cases where data appears to fail the standard parallel trends assumption, practitioners opt for other methods such as synthetic control estimators. Synthetic control methods re-weight control units to match the pre-treatment trends of the treated unit. While this constructs a control group that satisfies parallel trends in levels, it does not necessarily equate the autocovariance structures ($\gamma_k = \gamma_{synthetic}$). We return to this in Section 6.1, where we demonstrate that synthetic control methods can still suffer from dynamic bias if the matched units do not share the same persistence parameters.

[30] In Bertrand et al. (2004) it is assumed that the error term is the source of the temporal dependence and that if follows an AR(p) process. Our setup is not mutually exclusive with that and as a result we still recommend practitioners consider clustering standard errors where appropriate.

4 What Does This Look Like From the Practitioner's Point of View?

In this section we show what first-order temporal dependence actually does to the data and to standard DiD estimators in a simple, simulated panel. The goal is to answer: if you are a practitioner fitting standard DiD/event-study specification in data that look dynamic, what are you really estimating? In the previous section we showed how the identifying assumptions we currently use produce an estimand with properties similar to the underlying structural parameter *if and only if* there is no first-order temporal dependence in the generating process. However, we want to be more specific when it comes to describing what that means for practitioners hoping to identify a causal effect using a DiD framework with a balanced panel.

Let's start with the simplest case where there is both a single treated unit and a single untreated unit; that is we will set $K = N = 2$. We generate data using

$$\begin{aligned}
y_{1,t} &= \alpha_1 + \delta t + \phi_1 y_{1,t-1} + \epsilon_{1,t} \\
y_{2,t} &= \alpha_2 + \delta t + \phi_2 y_{2,t-1} + \lambda_2 \mathbf{1}\{t > \tau_2\} + \epsilon_{2,t} \\
\alpha_1 &= 0 \\
\alpha_2 &= 1.5 \\
\delta &= 0.25 \\
\epsilon_{1,t} \perp \epsilon_{2,t} &\sim \mathcal{N}(0, 0.1^2)
\end{aligned}$$

setting $T = 20$, $\lambda_2 = (0, 1)$.[31] Following our discussion from the previous section we will consider three cases: a no dynamic setting ($\phi_1 = \phi_2 = 0$), a homogeneous dynamic setting $\phi_1 = \phi_2 = 0.50$, and a heterogeneous dynamic setting $\phi_1 = 0.20, \phi_2 = 0.70$.

[Insert Figure 1 Here]

Figure 1 plots the time series of each individual with $y_{1,\cdot}$ corresponding to the control group. Panel (i) shows the sequence of innovations, independent across both individual and time, for each observation prior to the introduction of any additional elements. Note they are separated by $\alpha_2 - \alpha_1$ based on their individual intercept. Panel (ii) introduces the trend, but no first-order temporal dependence. Due to this lack of first-order dependence, the treatment introduced in period eleven results in an immediate shift of the post-treatment time series, increasing the distance between the treated and control groups post treatment by λ_2. We will get to estimation of that effect in a moment but it should be clear that the new distance between the two groups appears to be constant for all post treatment periods.

Panel (iii) introduces both the trend and homogeneous first-order dependence. Contrasting this with Panel (ii) we would like to note three important differences. First, the distance between the treated and control groups is larger than $\alpha_2 - \alpha_1$ in the pre-treatment periods. This is because, unconditional on past outcomes, the distance is a function of α_2, α_1, and ϕ, in fact, this distance is equal to the long-run equilibrium value of $(\alpha_2 - \alpha_1)/(1 - \phi) = 3$. Second, unlike in Panel (ii) which showed an instantaneous shift in the treated group at period eleven with no additional changes (relatively speaking) in subsequent periods, the transition from pre-treatment equilibrium to post-treatment equilibrium is much smoother for the treated group with early post-treatment periods showing a clear increase in the distance relative to the control group and the assumed counterfactual. Finally, we note that even at $T = 20$ it appears that the distance between the observed outcomes for the treated group and the assumed counterfactual is continuing to grow, albeit at a slower rate than in the early post-treatment periods. This may lead us to the conclusion that these groups, conditional upon the observed periods alone, will continue to diverge in subsequent post-treatment periods. In fact, since the process for the treatment group is trend-stationary, it is known that the distance will converge to the cumulative path of the treatment: $(\alpha_2 - \alpha_1 + \lambda_2)/(1 - \phi)$.

[31] We have limited the noise for illustrative purposes so as to make clear the impact of the dynamic setting on the outcome paths.

Panel (iv) introduces both the trend and heterogeneous first-order dependence to the innovations. As mentioned in Section 3, the parallel trends assumption, at least visually when we compare the pre-trends, appears to fail since the treatment and control groups are clearly diverging in the pre-treatment periods. Unconditional plots like these often lead researchers to reject the plausibility of a difference-in-differences analysis because of the clear lack of pre-treatment parallel trends. Like Panel (iii), the transition to the post-treatment equilibrium distance is smooth with an increasing distance in the early post-treatment periods. However, it is even more clear that in subsequent periods post-treatment, these two groups are likely to continue to diverge going forward. Note that Panels (ii - iv) are all based on the same sequence of innovations and the only change is the strength of dependence over time experienced by each.

Now that we know what the introduction of first-order dependence looks like in the simplest $N = K = 2$ case we open it up to a more general panel structure with $n_k = 500$; though for simplicity we still limit $K = 2$. Again, we generate data using

$$
\begin{aligned}
y_{i,t} &= \alpha_1 + \delta t + \phi_1 y_{i,t-1} + \epsilon_{i,t}, \text{ for } i \in k = 1 \\
y_{i,t} &= \alpha_2 + \delta t + \phi_2 y_{i,t-1} + \lambda_2 \mathbf{1}\{t > \tau_2\} + \epsilon_{i,t}, \text{ for } i \in k = 2 \\
\alpha_1 &= 0 \\
\alpha_2 &= 2 \\
\delta &= 0.25 \\
\epsilon_{i,t} &\overset{iid}{\sim} \mathcal{N}(0, 1)
\end{aligned}
$$

setting $T = 20$, $\lambda_2 = 2$. Identical to the $N = K = 2$ case we will keep the values of ϕ_1 and ϕ_2 consistent across the three different cases though we have increased the variance for the individual innovations primarily for illustrative purposes.

[Insert Figure 2 Here]

Figure 2 plots this panel of individuals in the same sequence of plots found in Figure 1. The bold lines represent $E_k(y_{i,t})$ for each $t \in (1, \ldots, 20)$. As before the treatment takes place in period eleven for the treated group. Panel (i) is similar to the same panel in Figure 1 with the average distance between the two sequence of innovations corresponding to the difference in intercepts. The large sample size for each group, $n_k = 500$, means that the average innovation for each period is very close to the intercept value despite the hundred-fold increase in variance. Likewise, Panels (ii-iv) when viewed through either the individual or the period average exhibit the same behavior as that found in Figure 2. Of course this is largely expected since the generating process for both the treated and control group is largely the same between the two figures, however we would like to point out that at least unconditional on past values, a panel does not seem to change the story being told by the raw data. Furthermore, one can obtain visual evidence of this through an examination of the Autocorrelation Function and Partial Autocorrelation Function consistent with common time series analysis techniques, a point we will return to in Section 6.

To illustrate the impact of the identification issues outlined in Section 3 we use two different estimation strategies. The first is a simple two-way fixed effects (TWFE) estimating equation given by,

$$
y_{i,t} = \eta_i + \eta_t + \underbrace{\hat{\rho}_1 y_{i,t-1} + \hat{\rho}_2 \mathbf{1}\{i \in k = 2\} y_{i,t-1}}_{\text{Dynamic Terms}} + \hat{\lambda} \mathbf{1}\{i \in k = 2\} \mathbf{1}\{t \geq 11\} + \epsilon_{i,t}. \tag{16}
$$

which we will implement in the fashion of Bertrand et al. (2004) by omitting the dynamic terms and clustering the standard errors at the individual level. Recognizing that a simple TWFE estimation strategy may not be as popular as it once was, we also employ the Synthetic Difference-in-Differences estimator first outlined by Arkhangelsky et al. (2021) which uses a regularization procedure to construct individual weights, to align pre-exposure trends in the

outcome of untreated units with those that are treated, and time weights to balance pre-and post treatment periods.[32] Following recommendations by the authors we calculate the standard errors using the bootstrap.

To implement Assumption A3b (Dynamically-Robust Parallel Trends) we include the dynamic terms in all three cases. Note that when there is no first-order dependence or homogeneous first order dependence, this model is mispecified since it has additional parameters which by construction in the generating process are zero. This should lead to slightly larger standard errors for the treatment estimator than otherwise would be necessary if one were to correctly specify the dynamic panel equation. To estimate this equation we turn to Arellano and Bond (1991) to accommodate the well-known Nickell Bias that presents in dynamic panel models with individual fixed-effects (Nickell, 1981). This estimation procedure proceeds using a two-stage general method of moments (GMM) procedure using additional lagged outcomes as instruments. Note that $\hat{\rho}_1$ is a direct estimate of ϕ_1 while $\hat{\rho}_1 + \hat{\rho}_2$ is an estimate of ϕ_2.

[Insert Table 2 Here]

Table 2 provides the coefficients and standard errors of the three estimators using Equation 16. When the process is static, that is $\phi_1 = \phi_2 = 0$, all three estimators provide similar point estimates though the standard error using the DR-PT structure is larger. This is a combination of the additional parameters $\hat{\rho}_1$ and $\hat{\rho}_2$ introducing inefficiencies and the absence of cluster or heteroskedastic robust standard errors. When there is homogeneous first-order temporal dependence, $\phi_1 = \phi_2 = 0.50$, we see the point estimate for the BDM and SDID estimates overstate the treatment by nearly seventy-five percent. This is due to the fact that what is being identified under Assumptions A1b and A2b is not the structural parameter λ, rather it is the average over the post-treatment periods of the cumulative dynamic effect: $\lambda_k \sum_{j=0}^{T-\tau_k} \phi_k^j$. As the number of post treatment periods gets larger this point estimate will converge to the change in the long-run equilibrium. Finally, for completeness we include estimates of $\hat{\lambda}$ in the heterogeneous first-order dependence case, $\phi_1 = 0.20$ and $\phi_2 = 0.70$, even though we have shown Assumption A2b fails under such conditions. Here again we see an overstatement of the estimated treatment effect by nearly 500% with standard errors consistent with very large t-statistics.

Practitioners often, when faced with data in levels, decide to take a first difference to avoid issues of potential unit roots (see Kumar and Liang (2019), which we discuss in more detail later, as one such example). With this in mind Table 2 also provides the point estimates and standard errors across the three specifications using $\Delta y_{i,t}$ as the object of interest. Here we see two things, first, for the DR-PT specification the sign of $\hat{\rho}_1$ and $\hat{\rho}_2$ has flipped, a natural consequence of taking the first difference of outcomes with first-order temporal dependence. The point estimate however for the DR-PT specification has not experienced a meaningful change though in all three cases the standard error is larger due to reduced observation count (one entire time-period is lost due to the first-difference). Consistent with the statements earlier, the point estimates for both the BDM and SDID estimators understate the treatment by a large margin due to the change in sign of the dependence. This serves as evidence that issues stemming from the presence of first-order temporal dependence are not fully solved or avoided by the difference operator.

[Insert Table 3 Here]

Finally, we want to take a moment and examine what has become more common in difference-in-differences applications, the use of an event-study. We can write the estimating equation in this scenario as

$$y_{i,t} = \eta_i + \eta_t + \underbrace{\hat{\rho}_1 y_{i,t-1} + \hat{\rho}_2 \mathbf{1}\{i \in k = 2\} y_{i,t-1}}_{\text{Dynamic Terms}} + \sum_{l=-10}^{9} \mathbf{1}\{l = t - \tau_k\}\{i \in k = 2\} + \epsilon_{i,t}. \tag{17}$$

In this case we will use the BDM estimate as before using a standard fixed effects type approach and add the estimator detailed by Callaway and Sant'Anna (2021). Broadly speaking, this consists of a series of 2×2 comparisons for

[32] We implemented this estimator directly using the R package "synthdid" provided by the authors.

each time period relative to the reference period.[33] As before we turn to Arellano and Bond (1991) to estimate the dynamic panel structure while accommodating Nickell bias. For all three of these estimators the reference period is the last period prior to treatment, $t = 10$. Figure 3 visualizes the point estimates and 95% confidence intervals for the corresponding period-treatment interaction terms. In Panel 3i we see a "well-behaved" event study with coefficients prior to the reference period failing to reject zero, indicating that there is no evidence of a relative difference to the reference period between treatment and control prior to the intervention. Post-periods show a fixed effect with point estimates near the true value for all three estimators and comparable confidence intervals. Panel 3ii mirrors what we saw in the unconditional plots of the data. In the periods prior to the reference period we see no evidence of a relative difference between the treatment and control group but post-treatment we see a smooth transition for the BDM and CS event study estimates to a new relative equilibrium that settles near the long-run unconditional equilibrium, $\lambda/(1-\phi)$. Meanwhile, the DR-PT event study in the post-treatment periods have point estimates near λ with corresponding confidence intervals covering the true structural parameter. In Panel 3iii, when Assumption A2b does not hold, we see evidence of a clear difference the pre-treatment periods between the treatment and control group relative to the reference period for both the BDM and CS estimators; this is capturing the pre-treatment divergence between the two groups we saw in Figure 2 (iv). Similarly, we see a continuing divergence in the post-treatment periods for these estimators with the implication that these series will continue to diverge going forward. Recall however that both the treatment and control group share a common trend term and the DR-PT implementation, by conditioning on past values, is able to recognize that with pre-treatment estimates indicating no evidence of a difference between treatment and control groups relative to the reference period and point estimates with corresponding confidence intervals post-treatment that are near the true value for the structural parameter λ.

In Section 3 we acknowledged that, in some cases, a practitioner may actually be seeking to estimate the cumulative dynamic effect. If that is the object of interest then one need not change the Parallel Trends assumption *per se*, though failing to adopt the DR-PT ultimately leads to estimating points in a function rather than the function itself. The trade-off of obtaining the cumulative dynamic path of the treatment this way comes in the form of external validity. Specifically, imagine we are in a case where the original parallel trends assumption holds but there is first-order temporal dependence. For clarity we will stick with the example we have here and assume $\phi_1 = \phi_2 = 0.50$, results of which are illustrated in Figure 3ii. Now imagine that same policy is implemented in two new treatment areas, one with a much lower level of dependence, say $\phi_2^a = 0.20$ and one with a much higher level of dependence, $\phi_2^b = 0.80$. Given the results from Table 2, a researcher might argue that policymakers in the new regions can expect an average treatment effect of approximately 3.66 units over the course of ten post-treatment periods. For posterity sake, let us say that the newly treated units are compared to a similar counterfactual unit with the same autocovariance structure, that is we are always in the case where ϕ is common.

In Figure 4 we plot event study estimates for the three estimators: BDM, CS, and DR-PT. Recall that the structural parameter has not changed for any of these cases, λ is always fixed as as one-time, two unit treatment effect (shifter). When the study is replicated in a low dependence environment, $\phi_a = 0.30$, the estimates are near the structural parameter value but are still systematically overstated for both BDM and CS with post-treatment period averages of 2.26 (0.035) and 2.36 (0.071) units respectively. When replicated in a high dependence environment, $\phi_b = 0.80$, the estimates for BDM and CS are more exaggerated than in the original study with post-treatment period averages of 6.98 (0.091) and 6.45 (0.099) respectively. Meanwhile, in all three regimes, the post-treatment period point event study estimates are near the structural value of $\lambda = 2$ with corresponding confidence intervals. This means that in a new environment, a treatment with a known identical effect, is appropriately decomposed from the autocovariance structure of the groups. This leads to better inference and more trustworthy analysis from economists to policymakers more generally.

[Insert Figure 4 Here]

[33] We implement this estimator using the "did" package in R which has been provided by the authors.

For the practitioner, there are a few key takeaways from this section. First, the difference cases we discuss in this section illustrate how first-order temporal dependence can change the appearance of raw data when applied to a static set of sequences. As we have mentioned throughout the paper up to this point, the presence of this dependence can lead to policies that go unevaluated by conventional DiD methods (such as TWFE) due to the appearance of non-parallel trends (i.e., potentially resulting from heterogeneous group autocovariance structures). In cases where the original parallel trends assumption holds we wanted to highlight how even homogeneous first-order dependence can smooth the transition from pre-treatment equilibrium to post-treatment equilibrium and bias the distance between the expected outcomes. Second, when we expose different estimators (BDM, CS, SDID, DR-PT) in both a TWFE-style estimating equation and an event-study estimating equation, the results show how a failure to recognize the dynamic parts of the process can lead to incorrect inferences about the effect of the intervention. This is due to identification of the cumulative dynamic path of the treatment, the structural parameter plus its propagation, instead of the treatment effect (or time-invariant intercept shift) itself. This issue persists even if one transitions to evaluating the policies impact on the change in the outcome rather than the levels. Finally, a failure to account for the dynamics can also affect external validity when a policy is implemented in a new location even if a counterfactual group with equivalent autocovariance structure can be used for comparison. By no means does this fully enumerate all of the issues that may arise from dynamic processes in a panel type setting, however, this should be a reasonable benchmark for practitioners and theorists alike to explore further.

5 Monte Carlo Simulations

In this section we use a Monte Carlo study to highlight the impact of first-order temporal dependence on difference-in-differences estimators. In Section 3 we identified three main cases: no first-order dependence, homogeneous first-order dependence, and group level heterogeneous first-order dependence, all of which will be covered. The main results will focus on both the bias in the point estimate under the different identifying assumptions and the estimates corresponding mean-square-error. To supplement this we also provide the average standard error over the iterations and its standard deviation.[34]

For simplicity, we return to basic generating process from Section 4,

$$\begin{aligned} y_{i,t} &= \alpha_i + \delta t + \phi_k y_{i,t-1} + \epsilon_{i,t} \text{ for } i \in k = 1 \\ y_{i,t} &= \alpha_i + \delta t + \phi_k y_{i,t-1} + \lambda_k \mathbf{1}\{t > \tau_k\} + \epsilon_{i,t} \text{ for } i \in k = 2 \end{aligned} \tag{18}$$

where α_i will be an unit specific intercept assigned via a standard normal distribution, $\delta = 1$ is a common trend parameter, $\lambda_k = 1$ is the treatment effect, and $\epsilon_{i,t} \overset{\text{iid}}{\sim} N(0, \sigma_k^2)$, with $\sigma_k^2 = 1$ for all k. Sample size will be fixed at $N = 1,000$ with a slight asymmetry in the size of the treated and control groups; the latter of which will be the larger of the two. In each case there will be $10,000$ replications with parameters varying across case by the length of the panel, $T \in \{10, 20\}$ and the level of first-order dependence as controlled by ϕ_k. For simplicity we set $\tau_k = \lfloor T/2 \rfloor + 1$ for $k = 2$ where $\lfloor \cdot \rfloor$ is the floor operator. For example, if $T = 10$, then $\tau_2 = 6$. When working with homogeneous autocovariance functions let $\phi \in \{-0.90, -0.75, -0.50, -0.25, 0.00, 0.25, 0.50, 0.75, 0.90\}$ with bounds chosen to sufficiently avoid a unit root. Since $\gamma_k(s)$, and by extension ϕ_k, is continuous over the support we limit the heterogeneous cases to four representative pairs: $(\phi_0 = 0.00, \phi_1 = 0.60)$, $(\phi_0 = 0.30, \phi_1 = 0.80)$, $(\phi_0 = 0.80, \phi_1 = 0.30)$, $(\phi_0 = 0.50, \phi_1 = -0.50)$.

Overall, the study will compare four estimators across two different estimating equations. Estimates produced using the collapsed time-series structure from Bertrand et al. (2004) (BDM), synthetic difference-in-differences from Arkhangelsky et al. (2021) (SDID), and the DR-PT structure under Equation 16. These will represent the average treatment effect across the observed post-treatment periods. We also include the comparable group aggregate from

[34] All estimates will use the default standard error construction for the "fixest", "did", and "plm" packages in R so as to allow for apples-to-apples comparison.

Callaway and Sant'Anna (2021) (CS) using Equation 17 for comparison. This group aggregation and its standard error are constructed from the point estimates and standard errors provided by series of 2×2 comparisons directly.

[Insert Table 3 Here]

Table 3 provides an accounting of the results for scenarios in which the original Parallel Trends assumption holds; no first-order dependence and homogeneous first-order dependence. Recall that the parameter of interest, λ_k is a constant, time-invariant intercept shift. Using Assumption A2b for identification means that the resulting estimand no longer satisfies the analogy principle in full; it identifies the cumulative path of treatment which is a time-varying function of λ_k and not the structural parameter. As a result, the post-treatment periods represent both the causal effect of interest and its propagation. Using Assumption A3b produces an estimand and corresponding estimate that satisfies the analogy principle in full. The improvement in mean-square-error, over 90% for DR-PT over the alternatives, is being primarily being driven by the fact it fully satisfies this principle since it disentangles the effect of treatment and time. Note that when $\phi = 0$, the mean-square-error is considerably larger for DR-PT though this is being driven by variance in the estimator; a foreseeable impact from including two extraneous parameters, ρ_1 and ρ_2, in the estimating equation and the reliance on two-stage GMM. Note that, in addition to the improvements in mean-square-error, the standard errors for the DR-PT structure are relatively stable across the support of ϕ with an average standard error of 0.095 (0.015). In contrast, the standard errors for BDM, SDID, and CS all vary as the level of first-order dependence changes with both the average and standard deviation of the standard errors increasing as ϕ moves away from zero. The average standard error for the CS estimator, 0.0927, is broadly similar to that of DR-PT, though with double the overall variation in those standard errors, 0.0299 for CS versus 0.0147 for DR-PT.

Table 4 provides results when the autocovariance functions vary by group and thus we see heterogeneous first-order temporal dependence. Recall from our earlier discussion that Assumption A2b no longer holds in this scenario, even when the groups share a common deterministic trend. In Section 4 we highlighted this under a single example in Table 2 under estimating Equation 16 and Figure 3iii under estimating Equation 17. The results in Table 4 are supportive of this singular example and show that identification bias introduced by the use of Assumption A2b is large in absolute value and seemingly arbitrary in direction. Meanwhile, the estimates produced using Assumption A3b for identification produce bias and mean-square error values consistent with those from Table 3. Moreover, in untabulated results the results indicate that empirical coverage of the DR-PT backed estimates (0.9496) is consistent with a nominal $\alpha = 5\%$. Note that this is not the fault of the BDM, CS, and SDID estimators *per se*, rather it is a failure of Assumption A2b to hold under the heterogeneous autocovariance functions even when it is known that both treatment and control groups share a common, deterministic trend.

[Insert Table 4 Here]

6 Empirical Examples - Replicating and Modifying Studies using State-level Panel Data

6.1 Texas's Natural Experiment in Expanding Credit Access

States in the U.S., often referred to as "laboratories of democracy," regularly provide economists with (quasi-)natural experiments to exploit as sources of exogenous variation in DiD models. To highlight the influence of the autoregressive component "in the wild," we replicate two recent papers that both 1) employ a commonly used data structure in economics (a balanced panel of U.S. state-level data) and 2) exploit a state-level policy shock for their DiD analysis. Specifically, in the cases of Kumar and Liang (2019) and Kumar and Liang (2024), they both rely on a natural experiment in Texas that changed credit access for homeowners in 1997 and once again in 2003 via state constitutional amendments. Prior to 1998, homeowners in all other U.S. states (except Texas) could access credit through home equity loans (HELs), cash-out refinancing, and home equity lines of credit (HELOCs). Texas reversed its outlier status in November 1997 by changing state policy, expanding homeowners' ability to borrow against their homes via home equity loans or cash-out refinancing. A 2003 amendment further expanded access to HELOCs to Texans.

Several papers in applied microeconomics have used Texas's shock to study the effect of credit constraints on a variety of outcomes, including retail spending (Abdallah and Lastrapes (2012)), mortgage defaults (Kumar (2018)), home prices (Zevelev (2021)), small business job creation (Lastrapes et al. (2022)), small business lending (Bahadir et al. (2024)), labor force participation (Kumar and Liang (2024)), and GDP growth (Kumar and Liang (2019)). The latter two papers by Kumar and Liang (2019; 2024) are particularly illustrative for a few reasons. First, both studies employ a familiar TWFE specification on a state-level panel that is reasonably representative of empirical work in economics over the last couple decades.[35] Second, they use multiple periods for their analysis - 16 years in their default TWFE specifications (1992-2007) - which is a reasonably representative time-series to investigate the influence of the autoregressive process on DiD results.[36] Finally, the authors assembled clean replication code for the general public,[37] which made it easy to replicate and convenient to re-examine outcomes with heterogeneous autoregressive processes.

Before replicating Kumar and Liang's (2019; 2024) results, we first examine the raw time series of their outcomes of interest. We plot the state-level labor force participation rate (LFPR), real GDP growth, and logged real GDP (in levels) in Texas and all other U.S. states (averaged) from 1985 through 2015 in panels (a), (d), and (g) in Figure 5. Though LFPRs have been generally declining in the U.S. since the mid-1990s, the fall is neither uniform across states or nor over time, suggesting that the series exhibits some persistence and heterogeneity across states. To explore time-dependence more formally, we plot the autocorrelation and partial autocorrelation functions of LFPR (for six lags) across the distribution of 50 states in panels (b) and (c) of Figure 5. We see strong evidence of autocorrelation with the first lag (i.e., an AR(1) process) for all states in the detrended series of LFPR. In contrast, state-level GDP in growth rates (as used by Kumar and Liang (2019)) and levels tell a somewhat different story. Specifically, in panels (h) and (i) in Figure 5, we can see that GDP in levels strongly exhibits autocorrelation, but in panels (e) and (f) we see the transformation into growth rates shows very little. Hence, based on this initial evidence and looking at the unconditioned, raw data, we might predict that a study of LFPRs (such as Kumar and Liang (2024)) would exhibit a greater autoregressive bias than one focusing on GDP growth (such as Kumar and Liang (2019)).[38] Hence, we have two useful real world cases, a low ϕ and a higher ϕ scenario, to explore how the results are impacted in the presence of varying degrees of autocorrelation.

[Insert Figure 5 Here]

The above prediction, of course, assumes a first-order misspecification problem. However, we should underscore that if the main source of the autocorrelation is second-order misspecification (i.e., confined to the standard errors), then the DiD coefficients of interest from *both* studies should not really change if we incorporate an DR-PT correction. Indeed, following lessons from Bertrand et al. (2004) and Cameron and Miller (2015), both Kumar and Liang (2019; 2024) studies adopt a common convention in the literature using cluster-robust standard errors that are clustered at the state level. Thus, one of the goals of this exercise is evaluate whether there is any evidence for a first-order or second-order misspecification problem.

[35] These papers do not only rely on TWFE; they use it as a starting point of their analysis. They subsequently modify DiD and explore other methods like the synthetic control method, but for the purposes of this paper we focus on replicating their TWFE DiD specifications.

[36] Although their review of the literature is now more than 20 years old, it is also worth pointing out that the average number of periods analyzed by the 92 DiD papers surveyed in Bertrand et al. (2004) was 16.5 periods, making the time-series used by Kumar and Liang (2019, 2024) representative of the papers sampled in Bertrand et al. (2004).

[37] Not all heroes wear capes.

[38] To be sure, examining the data series and autocorrelations is too simple. Both studies incorporate covariates into their DiD specifications, and it would be more appropriate to make a prediction based on the conditioned series or specification's residuals. For example, most statistical packages are equipped with autocorrelation or serial correlation diagnostic tests (e.g, Durbin-Watson, Breusch-Godfrey, or Wooldridge test). Also, see Born and Breitung (2016) for additional discussion of these tests as well as their own test statistic. In any case, we think it is good practice to first examine the raw series and conduct simple tests prior to a more formal analysis.

6.1.1 Re-examining the effect of credit constraints on GDP growth: Low ϕ scenario

We first replicate TWFE specifications from Kumar and Liang (2019), which estimates the effect of the Texas policy change on its GDP. The main finding in Kumar and Liang (2019) is that the 1997 policy change had a noisy, insignificant impact on real GDP growth, which is illustrated through the replicated TWFE estimates (in blue) in the top set of panels in Figure 6. Kumar and Liang (2019) focus primarily on the effect of the 1997 law change in Texas, which is labeled λ_1 and is the DiD estimator of interest. The DiD estimates for real GDP growth prove to be robust, as the point estimates for λ_1 are only slightly smaller in absolute terms and are still statistically insignificant when we incorporate an AR(1) term (i.e., one lag of the dependent variable) and an interaction term with the treatment indictor. The full results are tabulated in the appendix, Table 7, which shows a relatively low estimated value of ϕ for the GDP growth regressions. This is consistent with our expectation that a transformation of the outcome into either growth rates or first differences would have a low degree of residual time-dependence or serial correlation, thus the correction having little impact on the point estimates.

[Insert Figure 6 Here]

An alternative approach to Kumar and Liang (2019) TWFE growth rate specifications would be to evaluate the DiD specification in terms of (logged) levels, which we illustrate in the bottom set of panels in Figure 6 (with full results tabulated in appendix Table 8). First-differencing and growth rate transformations both address autocorrelation, but they come with specific assumptions about the nature of ϕ and time-dependence. If $\phi = 1$, then first-differencing makes sense; however, if $\phi < 1$ or states have heterogeneous ϕ, then we can run into issues of over-differencing. The time-series literature points out that one symptom of over-differencing is a less efficient estimator with changes in signs across specifications. When we re-estimate Kumar and Liang (2019)'s main TWFE specifications in logged real GDP, Figure 6 shows that the corrected (red) estimates are more "precise zeroes" and do not switch signs when the control group changes (from all states to energy states).[39] With a high autocorrelation variable like real GDP, the logged specification (in levels) coupled with the AR terms can offer a more flexible specification. To be clear, this specification would still confirm Kumar and Liang (2019)'s overall findings of no significant impact on changes in GDP, but it does so with greater precision and better model fit.

6.1.2 Re-examining the effect of credit constraints on labor force participation: Higher ϕ scenario

In our next set of estimates, we replicate results from Kumar and Liang (2024)'s study on LFPRs, illustrating a scenario where the main finding is non-zero and the outcome of interest has moderate to high autocorrelation. Recall that Kumar and Liang (2024)'s primary finding is that Texas's 1997 and 2003 policy changes reduced labor force participation rates, which can be seen in blue in Figure 7. In the 2024 study, they separately estimate the 1997 and 2003 policy effects (which we label λ_1 and λ_2), finding the 2003 policy (allowing HELOCs) had a larger impact than the 1997 policy (allowing HELs and cash-out refinancing). Like the 2019 study of GDP, Kumar and Liang (2024) estimate a simple TWFE DiD with no controls (panel (a) and (b)), but for two control groups (all states vs. energy states), and they progressively incorporate additional controls in the model. To simplify the presentation of the estimates, we reproduce only one other specification in Figure 7, which adds census-division-by-year fixed effects and state-specific linear time trends (i.e., state-trend interactions).[40]

Overall, we can see that once we explicitly control for an AR(1) process (i.e., a single lag of y) and its interaction with the control group, the corrected estimates in black show a smaller, but more generalizable effect for λ_1 and λ_2. In the simplest TWFE specification (panel (a)), the DR-PT corrected model yields an effect size of -0.30 and -1.02 for the 1997 and 2003 policy changes, respectively (compared to -1.08 and -2.07 in the original study). The energy-

[39] Table 8 in the appendix shows that while ϕ is high, it is about 0.9 overall with a significant interaction for a Texas-specific effect. The DR-PT corrected model shows large improvement in AIC and BIC, as we would expect.

[40] A full replication of Table 2 of their study can be found in the appendix, (Table 9 (Panel A replication) and Table 10 (Panel B replication))

intensive states (panel (b)),[41] which the authors view as a more plausible control group, have a similar AR-bias, where the corrected effect sizes are -0.45 and -1.51 versus -1.15 and -2.36 in the original. This multiplicative bias falls (approximately) in line with what we expect when we look at the coefficients on the lagged LFPR and its interaction with Texas across all specifications (ϕ_1 and ϕ_2, respectively) in Appendix Tables 9 and 10. For the specification that corresponds to panel (a) in Figure 7, we observe a ϕ_1 of 0.76, suggesting that the purged estimate of λ should be about one-quarter the size of the original. This is approximately in line with the comparison of -0.30 to -1.08, although the 2003 effect is closer to half the effect size. For the energy states specification, we observe a ϕ_1 of 0.73, but a statistically significant ϕ_2 of -0.23,[42] suggesting that the DR-PT corrected estimate should be about half of the original estimate. Together, this is approximately what we observe; the corrected estimates in panel (b) are about half the size of the original (i.e., -0.45/-1.15 = .39 and -1.51/-2.36 = .64). In fact, if we average over all specifications in Appendix Table 10 for the energy states, we observe an average $\phi_1 = 0.58$ (with ϕ_2 statistically insignificant in most specifications), which is approximately in line with the average DR-PT corrected λ's being roughly half the size of the original (-0.48 and -1.52 versus -1.02 and -1.99). Thus, the evidence points to a significant first-order misspecification problem in this scenario. About half of the effect we observe from the Texas natural experiment (or, a quarter of the 2003 effect) can be attributed to the autoregressive bias, with the remaining (DR-PT corrected) effect can be interpreted as a constant, generalizable effect.

[Insert Figure 7 Here]

A second advantage of correcting for the autoregressive bias in this way is that it can generate more precise estimates. If the issue is strictly a second-order misspecification problem, the usual prediction would be that in presence of serial correlation the standard errors will tend to be too small (Bertrand et al. (2004)). An uncorrected first-order misspecification issue, on the other hand, can come with higher standard errors and poor model fit. Specifically, in panel (c) of Figure 7 we compare the estimates of λ with DR-PT correction (in black) against the original estimates (in blue) with census-division-by-year fixed effects and state-specific linear time trends. The results still show a smaller effect for the DR-PT corrected estimates, although some of these coefficient estimates are much closer to their uncorrected counterpart. This is consistent with what we see in some of the remaining regressions in the Kumar and Liang (2024) replication in Appendix Tables 9 and 10). In virtually all cases, however, the standard errors of the DR-PT corrected estimates are substantially smaller and the model fit is improved (i.e., lower AIC and BIC) with this simple correction to the model. We highlight the specification in panels (c) in Figure 7 because the authors incorporate so many interactions (50 state-specific trends) and fixed effects (9 geographic divisions by 16 years) to "account for regional-specific macro shocks," which is a common approach in the empirical literature. These additional covariates are correlated with the autoregressive process to some extent, but only account for it indirectly and somewhat inefficiently. Indeed, the AIC and BIC are lower, for example, in the simplest TWFE DiD (no controls, but with a lag and interaction) compared to a much more saturated specification with state-specific trends and division-year interactions in Table 9. Thus, while Kumar and Liang (2024) follow common practices and norms in the DiD literature to incorporate additional X's to account for potentially relevant variation in the model, this example illustrates how moderate-to-high estimate of ϕ (i.e., 0.52 to 0.73 for the energy states) can yield a substantial improvement to the model with a simple addition of two terms (as opposed to dozens of fixed effects and interactions).[43]

Finally, it is important to note that Kumar and Liang (2019; 2024) exploit multiple methodologies to examine the effect of the Texas policy changes, including the synthetic control method (SCM). We replicate Figure 5 from Kumar and

[41] The "energy states" in both Kumar and Liang (2019; 2024) studies include 12 states with at least a one percent share of total employment in the mining sector.

[42] Recall that the Texas-specific coefficient is a combination of the two, $\phi = \phi_1 + \phi_2$.

[43] In addition to improvement in standard errors, t-statistics, AIC/BIC, we also observe a lower variance in the coefficients across specifications for the DR-PT corrected estimates. If we calculate the variance of all DiD estimators across the original specifications in both Kumar and Liang's (2019; 2024) papers, and we compare them to the variance in the DR-PT corrected coefficients, we see a substantial drop in variance with DR-PT corrected results. This is additional evidence of precision or a positive design attribute that practitioners care about.

Liang (2024) in panel (i) of Figure 8, which depicts the labor force participation rate over time. It compares Texas's LFPR to a "synthetic Texas" before and after the 2003 policy shock. The difference between "synthetic Texas" and Texas is graphed in the bottom-left panel (iii), showing the initial effect of the 2003 policy was about a 0.7 percentage point drop in the LFPR. By 2007, this drop grew to be about 2 percentage points relative to the control. Based on what we learned from the prior sections, it begs the question, is this treatment effect really heterogeneous and increasing (in absolute terms) over time, or is this a one-time, constant effect that appears multiplicatively inflated due to its temporal dependence?

[Figure 8 here]

The results from panels (ii) and (iv) of Figure 8 are more consistent with the smaller effect size we observe from the DR-PT TWFE estimates, and what we would expect to see if much of the effect in Kumar and Liang (2024) is due to temporal dependence. In particular, for the panels on the right side of Figure 8 (panels (ii) and (iv)), we made one modification to Kumar and Liang (2024). Before constructing the 'synthetic Texas' we pre-whitened the LFPR time-series for each individual state, essentially purging AR(1) dependence from panel observations.[44] We find that the effect in the pre-whitened series was a 0.7 percentage point drop in LFPR in 2004, and this effect remained approximately stable over the post-period (from approx. 0.5 to 0.8). The states that dominate the weighting of synthetic Texas also change when the state series are whitened.[45] Overall, the results from the SCM replication align with the broader takeaway from the TWFE estimates, that the more generalizable effect size is still significant, but substantially smaller than what would ordinarily be reported by most practitioners using DiD and SCM. More generally, our results suggest that SCM practitioners should exercise caution when the outcome's ϕ is nonzero and heterogeneous across panel units, as re-weighting the control group does not necessarily address the misidentification issue.

6.2 Re-examining the California Smoking Cessation Program and Synthetic Difference-in-Differences

Earlier in the paper (Section 4) we briefly touched on the synthetic difference-in-differences (Arkhangelsky et al., 2021) method, which shares (or augments) relevant assumptions of standard DiD estimators. In our final empirical example, we will briefly revisit the impact of California Proposition 99, which comes from a benchmark study by Abadie et al. (2010) that sought to estimate the impact of a cigarette tax increase on smoking in California. Their dataset consists of a complete panel of thirty-nine states covering the period from 1970 to 2000 at an annual frequency. California passed a tax increase in 1989 leading to nineteen pre-treatment periods and twelve post-treatment periods in the original study. Since California is the sole state to implement this policy in 1989, there is a single treated unit and thirty-eight control units.

To align more directly with the DR-PT assumption (A3b), we modified the (Arkhangelsky et al., 2021) method by pre-whitening the series, similar to what we had implemented in Section 6.1 with Kumar and Liang's (2024) SCM .[46]

[44] In this case, we pursued a very simple approach to whitening by regressing the outcome on its first lag for each state individually. The residuals become the whitened series in this case, as they represent the variation in LFPR not explained by the AR(1) process. The Kumar and Liang (2024) dataset extended back to 1980, which we use to obtain a more precise estimate of ϕ for each state. There are, of course, many ways to pre-whiten a time series such as variations of ARIMA models, which we explored in untabulated results, finding qualitatively similar estimates. However, another paper could be devoted entirely on evaluating the optimal ways to pre-whiten panel data for DiD, SCM, synthetic DiD, which we leave for future research.

[45] In the orginal Kumar and Liang (2024) study, the primary states selected for synthetic Texas were Colorado, New Mexico, and Iowa, while the whitened series selected New Mexico, Kansas, North Dakota, and Arkansas (with weights greater than 0.10).

[46] In this case, we implement a standard ARIMAX to whiten the each time-series (state) in the panel. That is, for each state we construct the whitened outcome as:

$$y^*_{s,t} = y_{s,t} - \hat{\rho}_s y_{s,t-1} \tag{19}$$

where $\hat{\rho}_s$ comes from the estimating equation

$$y_{s,t} = \hat{\alpha}_s + \hat{\beta}_2 \mathbf{1}\{\text{Year} > 1989\} + \hat{\beta}_2 \mathbf{1}\{\text{California}\}\mathbf{1}\{\text{Year} > 1989\} + \hat{\delta}_s t + \hat{\rho}_s y_{s,t-1} + \epsilon_{s,t}. \tag{20}$$

The result is that $y^*_{s,t}$ is the remaining variation in $y_{s,t}$ unexplained by past outcomes directly.

For the thirty-nine states in the sample period, we estimate first-order dependence (ϕ) to be, on average, 0.851 (with standard deviation 0.078). California happens to be on the higher end of this range (0.934). In Figure 9i we plot the outcome of interest by state, where the solid line in the forefront represents the California series. Like the other states, California has a downward trend in smoking over the sample period, ranking in the middle of the pack in 1970 and ending up with one of the lowest smoking levels by 2000. In Figure 9ii we instead plot the whitened outcome, which is the remaining variation in the outcome that is independent of its AR(1) process. Given that California has a relatively high ϕ compared to other states, it is easy to see from Figure 9ii that California ranks among the bottom of all states from 1970 to 2000, given that the remaining variation in the outcome is relatively small due to temporal dependence explaining California's smoking path quite well.

[Insert Figure 9 Here]

In Table 6 we recreate Table 1 from Arkhangelsky et al. (2021), showing the average effect of the increased cigarette taxes on per capita cigarette sales in California. The estimates corresponding to the traditional DiD assumptions (Assumption A2b) indicate that the cigarette reduction effect falls within approximately fifteen to twenty-five pack per capita (with standard errors indicating statistical significance marginally different from zero).[47] In contrast, after applying the DR-PT version of parallel trends via pre-whitening, the point estimates are much closer to zero with standard errors indicating a null effect under any standard significance threshold.[48]

[Insert Table 6 Here]

Alternatively, to visually compare these two approaches, we recreate Figure 1 from Arkhangelsky et al. (2021), as shown in Figure 10 below in two tri-panels. The left side illustrates the original result from Arkhangelsky et al. (2021), highlighting the lack of parallel trends in the pre-treatment periods for the DiD estimate and corresponding corrections via reweighting of the control units in the SCM and SDiD cases. The right side of Figure 9ii, however, tells a different story when we pre-whiten the outcome. The DiD plot in the top part of the right panel (ii) exhibits clear parallel trends during both the pre- and post-treatment periods. In fact, all three parts of the right (whitened) panel tell this story; the policy change in California had no measurable impact on packs per capita sales in the post-treatment periods once we account for temporal dependence in the whitened series. In other words, California smoking was falling in large part because it had been falling previously, and this temporal momentum or persistence had been more acute for California specifically. Once we account for this temporal dependence, California's trends largely track the control states before and after the policy change. Insofar as there is an effect from the 1989 policy, it is California-specific (i.e., not necessarily generalizable) as it is largely derived from its temporal momentum.[49]

[Insert Figure 10 Here]

More generally, the California smoking policy example provides a useful illustration of heterogeneous first-order temporal dependence that SDiD, by itself, does not remedy. SDiD corrects the outcomes in such a way as to allow the standard parallel trends assumption to hold through its weighting scheme. The SDiD correction does not, however, address the underlying second-moment restriction implied by the parallel trends assumption more broadly, which is

[47] We thank the authors for their excellent replication file that includes the "synthdid" package used in their analysis, in addition to the corresponding data files. Standard errors are computed in the same fashion (placebo method) as in the original paper. As an alternative to the pre-whitening process one can also include the lag of the outcome as a covariate in the "synthdid" package. This results in a point estimate of -1.42 with a standard error of 3.38. For this approach, note that only the lag was included and we did not include interaction terms by state or treatment group, which would mean that the estimate reflects an assumption of homogeneous autocovariance functions.

[48] Note that 1970 is excluded from this analysis based on the use of a lag in the whitening process, thus there is one less pre-treatment period than in the original analysis.

[49] We leave the interpretation for this momentum to future research. We speculate that this could be do to unobservables, like culture and norms around smoking that accelerated the descent of smoking in a way that is California-specific. The law change itself and its timing could be a product of or coincide with this unobservable momentum; but again, this is speculation worthy of merely a footnote. A more rigorous investigation falls outside the scope of this paper and could itself be its own paper.

a core idea from Sections 3 and 4 above. Thus, practitioners should consider modifications to SDiD such as pre-whitening to recover a more generalizable, dynamically robust effect.

7 Conclusion

For most panel applications of DiD, understanding whether (or to what extent) the outcome of interest follows an autoregressive process is critical for both estimation and interpretation of the results. When the outcome of interest exhibits significant autocorrelation, and this is rooted in a second-order misspecification issue, prior work has shown how arbitrary serial correlation will yield standard errors that are 'too small' and thus tendency for overrejection of the null (Bertrand et al. (2004)). However, if the serial correlation is a first-order misspecification problem, which is common among panel applications of DiD, this paper shows both analytically and through Monte Carlo simulations that the standard DiD estimator will be biased and likely misinterpreted. Indeed, without explicitly accounting for the autoregressive process in the design, practitioners will mistakenly interpret the DiD estimator as an (inflated) average treatment effect rather than a combination of a dynamic effect and a (smaller) constant treatment effect. Simulations show that directly controlling for the autoregressive process in the model, such as including a lagged dependent variable and an interaction with the treatment, can effectively mitigate autoregressive bias by decomposing the time-dependent effects from the constant treatment effect, yielding a more generalizable DiD estimator in most panel applications.

To demonstrate how this works in practice, we replicate two recent studies (Kumar and Liang (2019) and Kumar and Liang (2024)) that employ a common TWFE specification in their DiD panel data design. Both papers evaluate the same shock - a Texas policy change to homeowner credit instruments - on two different outcomes, GDP growth and labor force participation, which also differ in their degree of autocorrelation. We find evidence of substantially inflated or biased TWFE estimates when the outcome, labor force participation rate, has moderate-to-high autocorrelation. Our analysis confirms the directional findings of both studies; but, the modified results show that the more generalizable, corrected estimates are smaller by roughly half (or a quarter, in some specifications) than the preferred estimates in the original Kumar and Liang (2024) study. The dynamically-robust parallel trends (DR-PT) corrected estimates were more aligned with the uncorrected estimates when examining a low autocorrelation case (GDP growth) in Kumar and Liang (2019). Yet, the DR-PT correction substantially improved the precision or model fit in both papers. Moreover, we apply the same logic to synthetic control and synthetic DiD methods (replicating Arkhangelsky et al. (2021)), finding that a similar modification (pre-whitening) will recover the analogous DR-PT estimate.

Another takeaway for practitioners and policymakers is that the nature of this multiplicative bias requires extra care in interpreting the results. The results from Kumar and Liang (2019; 2024) and Arkhangelsky et al. (2021) are, in many ways, correct. Kumar and Liang (2019; 2024) account for arbitrary serial correlation in ways that follow the prevailing norms in the economics literature. They cluster standard errors at the appropriate level and, in the case of Kumar and Liang (2019), they address autocorrelation in the level of GDP by expressing the outcome in terms of growth rates. So, their inferences do not run afoul of the lessons from Bertrand et al. (2004), and the effect of increasing credit access to homeowners in Texas in 1997 and 2003 are precisely what is reported in these studies. The standard TWFE approach, however, narrows the interpretation of DiD results to be specific to the particular time and place of the study. Yet, without decomposing the AR(1) process from the DiD estimator, this approach makes it difficult for policymakers and the general public to discern the impact of a similar change to credit constraints. Our modified DR-PT DiD estimator, on the other hand, does purge the autoregressive process from the DiD estimator, which means that policymakers in another state or in another time (when the temporal dependence may be different) more effectively interpret the effect size in their context. More generally, we speculate that a better accounting of the autoregressive process in the model may also help economists reconcile results across studies, which may study a similar policy change (e.g., a change to the minimum wage, health insurance policy, or some other state-level regulation) that occurs in multiple states but produces results of differing magnitudes.

Overall, the results underscore that time-series issues merit further attention in the DiD literature and the causal inference literature more broadly. At least since Bertrand et al. (2004), practitioners have been aware that time-series issues such as arbitrary serial correlation can impact standard errors. Indeed, most econometrics textbooks cover serial correlation as a more general issue (as well as some specific applications relevant to practitioners), but the results from this paper (and the infrequent incorporation of lagged parameters in published DiD studies) suggest further integration of time-series concepts into causal inference methods is needed. We put forth a modification to the standard parallel trends assumption in DiD, which should help researchers using DiD better interpret their findings and understand what variation is being modeled. Fortunately, a correction that aligns the DiD estimator with these modified assumptions can be simple and easy to implement (without a new package to install into one's preferred statistical software), as we have shown using relatively straightforward examples and simple data generating processes.

As a final note, we readily acknowledge that we have not resolved all time-series issues for DiD research designs; and, perhaps we are just scratching the surface on dynamic panel issues. Indeed, additional time-series issues could be at play in the panel data applications we explore here, potentially complicating the diagnosis we examine in this paper and the simple remedies we investigate. Other data structures, such as repeated cross-sections, will have further complications beyond what we have covered here. The results in this paper do suggest, however, that even in common uses of DiD, SCM, and SDiD, there are still substantial improvements to be made by applying lessons from the time-series literature to causal inference methods. More generally, we hope the results from this paper signal that it is about time that the econometrics literature tackles more time-series topics in causal inference research, as helping practitioners resolve these issues can dramatically improve research designs that practitioners use everyday.

8 Tables and Figures

Table 1: Card and Krueger (1994) - Average Number of Full-time Equivalent (FTE) Employees per Store in the NJ-PA Fast Food Labor Markets

	PA	NJ	*Gap DiD*
FTE employment before	23.33	20.44	-2.89
	(1.35)	(0.51)	(1.44)
FTE employment after	21.17	21.03	-0.14
	(0.94)	(0.52)	(1.07)
Trend DiD	-2.15	0.59	2.76
	(1.25)	(0.54)	(1.36)

Note: Adapted from Table 3 of Card and Krueger (1994), showing changes in FTE in the treatment (NJ) and control (PA) states before and after NJ's change in its minimum wage law. Standard errors are in parenthesis.

Table 2: The Impact of First-Order Dependence on Estimation

Dependent Variable		$y_{i,t}$			$\Delta y_{i,t}$		
		BDM	SDID	DR-PT	BDM	SDID	DR-PT
$\phi_1 = \phi_2 = 0$	$\hat{\rho}_1$			0.002			-0.372
				(0.013)			(0.013)
	$\hat{\rho}_2$			-0.011			-0.316
				(0.021)			(0.038)
	$\hat{\lambda}$	2.037	2.045	2.103	0.200	0.205	2.109
		(0.030)	(0.027)	(0.081)	(0.015)	(0.012)	(0.165)
$\phi_1 = \phi_2 = \phi$	$\hat{\rho}_1$			0.519			-0.137
				(0.011)			(0.007)
	$\hat{\rho}_2$			-0.027			-0.148
				(0.019)			(0.013)
	$\hat{\lambda}$	3.662	3.664	2.097	0.409	0.416	2.026
		(0.053)	(0.055)	(0.079)	(0.019)	(0.019)	(0.109)
$\phi_1 \neq \phi_2$	$\hat{\rho}_1$			0.209			-0.196
				(0.012)			(0.008)
	$\hat{\rho}_2$			0.499			-0.108
				(0.094)			(0.008)
	$\hat{\lambda}$	10.507	10.473	2.023	0.650	0.656	2.080
		(0.060)	(0.105)	(0.094)	(0.020)	(0.020)	(0.111)

Note: These results are estimates from applying the estimator indicated to estimating equation 16 where the parameter of interest is the average treatment effect in post-treatment periods. Note that for the DR-PT based estimator the model always assumes first-order temporal dependence is present and is heterogeneous. This means that the model is misspecified except when there is heterogeneous dynamics and this misspecification will present in the standard errors for the parameter of interest.

Table 3: Monte Carlo Results: Case 1 and 2

		Avg. Bias				Mean-Square-Error			
T	ϕ	DR-PT	BDM	CS	SDID	DR-PT	BDM	CS	SDID
10	-0.90	**-0.0002**	-0.3940	-0.3936	-0.3934	**0.0045**	0.1580	0.1828	0.1556
10	-0.75	**-0.0009**	-0.3684	-0.3701	-0.3702	**0.0069**	0.1370	0.1485	0.1379
10	-0.50	**0.0009**	-0.2868	-0.2868	-0.2913	**0.0102**	0.0833	0.0888	0.0855
10	-0.25	**-0.0010**	-0.1675	-0.1681	-0.1702	**0.0107**	0.0293	0.0335	0.0300
10	0.00	**0.0000**	0.0001	0.0002	0.0049	0.0104	**0.0016**	0.0048	0.0019
10	0.25	**0.0004**	0.2447	0.2449	0.2452	**0.0104**	0.0623	0.0649	0.0627
10	0.50	**-0.0009**	0.6130	0.6126	0.6218	**0.0101**	0.3796	0.3809	0.3899
10	0.75	**-0.0022**	1.1688	1.1690	1.1746	**0.0102**	1.3732	1.3734	1.3868
10	0.90	**-0.0033**	1.6272	1.6290	1.6252	**0.0145**	2.6586	2.6616	2.6486
20	-0.90	**-0.0000**	-0.4580	-0.4586	-0.4578	**0.0051**	0.2102	0.2333	0.2096
20	-0.75	**-0.0017**	-0.4058	-0.4067	-0.4049	**0.0067**	0.1650	0.1757	0.1643
20	-0.50	**-0.0001**	-0.3110	-0.3101	-0.3113	**0.0087**	0.0971	0.1020	0.0973
20	-0.25	**-0.0000**	-0.1841	-0.1836	-0.189	**0.0092**	0.0344	0.0385	0.0363
20	0.00	-0.0012	**0.0000**	-0.0009	0.0025	0.0089	**0.0008**	0.0045	0.0010
20	0.25	**-0.0003**	0.2888	0.2895	0.2903	**0.0091**	0.0847	0.0885	0.0860
20	0.50	**-0.0002**	0.7993	0.7993	0.7955	**0.0089**	0.6415	0.6445	0.6355
20	0.75	**-0.0021**	1.8652	1.8653	1.8576	**0.0089**	3.4859	3.4878	3.4559
20	0.90	**-0.0022**	3.1368	2.1359	3.1396	**0.0146**	9.8546	9.8458	9.8691

		Avg. Standard Error				Std Dev Standard Error			
T	ϕ	DR-PT	BDM	CS	SDID	DR-PT	BDM	CS	SDID
10	-0.90	**0.0676**	0.2074	0.1677	0.0756	**0.0016**	0.0587	0.0072	0.0044
10	-0.75	0.0837	0.1473	0.1074	**0.0745**	**0.0022**	0.0336	0.0046	0.0042
10	-0.50	0.1009	0.0941	**0.0807**	0.0809	**0.0027**	0.0191	0.0034	0.0052
10	-0.25	0.1026	**0.0590**	0.0718	0.0938	**0.0025**	0.0133	0.0031	0.0056
10	0.00	0.1019	**0.0385**	0.0692	0.1087	**0.0024**	0.0098	0.0030	0.0060
10	0.25	0.1013	0.0720	**0.0704**	0.1310	**0.0024**	0.0137	0.0030	0.0075
10	0.50	0.1010	0.1729	**0.0751**	0.1623	**0.0025**	0.0169	0.0033	0.0099
10	0.75	0.1015	0.3641	**0.0829**	0.2043	**0.0031**	0.0199	0.0036	0.0116
10	0.90	0.1203	0.5450	**0.0898**	0.2348	0.0132	0.0218	**0.0039**	0.0134
20	-0.90	0.0715	0.1223	0.1497	**0.0140**	0.0017	0.0354	0.0065	**0.0008**
20	-0.75	0.0828	0.0792	0.1002	**0.0518**	**0.0020**	0.0170	0.0044	0.0029
20	-0.50	0.0924	**0.0514**	0.0767	0.0583	**0.0021**	0.0088	0.0033	0.0038
20	-0.25	0.0944	**0.0360**	0.0686	0.0674	**0.0021**	0.0060	0.0030	0.0039
20	0.00	0.0946	**0.0280**	0.0663	0.0814	**0.0021**	0.0046	0.0029	0.0044
20	0.25	0.0945	**0.0430**	0.0682	0.1018	**0.0021**	0.0064	0.0030	0.0063
20	0.50	0.0943	0.1054	**0.0751**	0.1356	**0.0021**	0.0092	0.0033	0.0074
20	0.75	0.0940	0.2930	**0.0910**	0.2040	**0.0023**	0.0140	0.0040	0.0126
20	0.90	0.1212	0.5855	**0.1085**	0.2799	**0.0203**	0.0181	0.0047	0.0162

Note: This table shows relevant measures of the estimate of λ_k for the treatment regime. Treatment period is $\lfloor T/2 \rfloor + 1$. DR-PT based estimates are produced using a two-stage GMM estimator robust to Nickell Bias. Minimum case values are bolded.

Table 4: Monte Carlo Results:: Case 3

			Avg. Bias				Mean-Square-Error			
T	ϕ_0	ϕ_1	DR-PT	BDM	CS	SDID	DR-PT	BDM	CS	SDID
10	0.00	0.60	**-0.0002**	8.3091	5.3114	8.3101	**0.0102**	69.0451	28.2165	69.0797
10	0.30	0.80	**-0.0008**	19.0813	11.9883	18.4644	**0.0103**	364.1008	143.7257	341.0916
10	0.80	0.30	**-0.0020**	-17.4655	-10.3716	-13.6595	**0.0100**	305.0488	107.5753	186.8121
10	0.50	-0.50	**0.0035**	-6.9532	-4.2855	-6.4349	**0.0105**	48.3490	18.3716	41.4340
20	0.00	0.60	**0.0046**	16.1265	9.3793	16.1223	**0.0089**	260.0652	87.9759	260.0083
20	0.30	0.80	**0.0001**	37.8524	21.8416	37.1788	**0.0083**	1432.8070	477.062	1382.7640
20	0.80	0.30	**-0.0068**	-35.2767	-19.2609	-29.5421	**0.0086**	1244.4490	370.9881	873.5982
20	0.50	-0.50	**-0.0048**	-13.6458	-7.6478	-13.0504	**0.0091**	186.2087	58.4950	170.3905

			Avg. Standard Error				Std Dev Standard Error			
T	ϕ_1	ϕ_0	DR-PT	BDM	CS	SDID	DR-PT	BDM	CS	SDID
10	0.00	0.60	0.1007	1.6582	**0.0738**	0.1098	**0.0023**	0.0154	0.0033	0.0064
10	0.30	0.80	0.1008	3.8350	**0.0785**	0.1370	**0.0024**	0.0168	0.0036	0.0081
10	0.80	0.30	0.1018	3.4837	**0.0781**	0.2135	**0.0024**	0.0165	0.0034	0.0122
10	0.50	-0.50	0.1016	1.3536	**0.0780**	0.1629	**0.0023**	0.0154	0.0032	0.0095
20	0.00	0.60	0.0943	2.1252	**0.0738**	0.0814	**0.0021**	0.0092	0.0032	0.0048
20	0.30	0.80	0.0943	5.0732	**0.0841**	0.1068	**0.0022**	0.0113	0.0037	0.0062
20	0.80	0.30	0.0946	4.7963	**0.0834**	0.2242	**0.0023**	0.0109	0.0036	0.0134
20	0.50	-0.50	0.0943	1.8121	**0.0760**	0.1363	**0.0021**	0.0080	0.0034	0.0081

Note: This table shows measures of the estimate of λ_k for the treatment regime under the condition that $\phi_0 \neq \phi_1 \neq 0$. Treatment period is $\lfloor T/2 \rfloor + 1$. DR-PT based estimates are produced using a two-stage GMM estimator robust to Nickell Bias. Empirical coverage for the DR-PT based estimates is close to the nominal value across the simulations, 94.97% (0.0071) but is zero for other estimators due to identification failure.

Table 5: Evidence of heterogeneity in ϕ for LFPR.

Coefficient	All States	P-value	Energy States	P-value	Non-Energy States	P-value
$\mathbb{E}[\hat{\rho}]$	0.8961	0.0263	0.8671	0.0736	0.9052	0.1920
	(0.0591)		(0.0842)		(0.0464)	
$\mathbb{E}[\hat{\delta}]$	-0.0756	0.0000	-0.1051	0.3624	-0.0662	0.000
	(0.0772)		(0.1058)		(0.0648)	

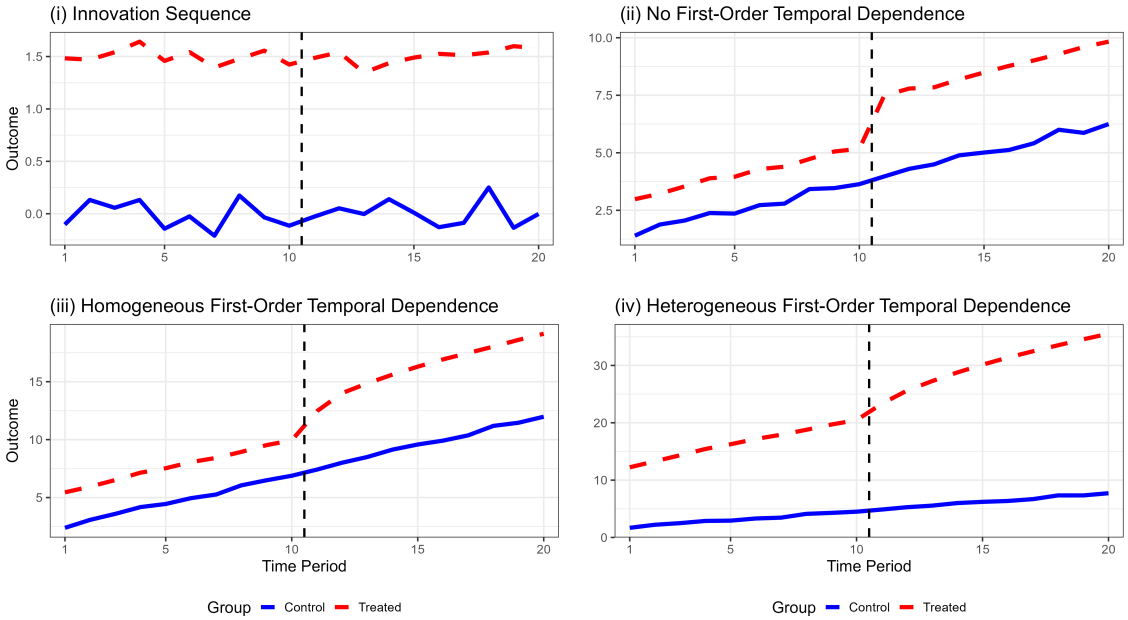
Note: This table presents the average of coefficients from the whitening procedure over all 50 states (standard deviations of those coefficients are included in parenthesis below). As a comparison, we conduct a two-sided Student T test comparing the coefficient for Texas (the treated state) against the sample of proposed comparison states. Results from this indicate that the AR(1) coefficient potentially differs between Texas and remaining "energy states" while the trend coefficient is statistically indistinguishable between the two. Texas, when compared against other energy states, will likely will fall into Case 3 from Section 3.

Table 6: Replicating the Effect of California's 1989 Tax on Smoking: Synthetic DiD vs. Pre-whitened Synthetic DiD

	SDID	SC	DID
Standard PT Assumptions (Assumption A2b)	-15.60	-19.62	-27.35
	(8.37)	(9.92)	(17.40)
DR-PT Corrected or Pre-whitened (Assumption A3b)	-0.54	-4.34	-0.08
	(5.17)	(2.73)	(4.10)

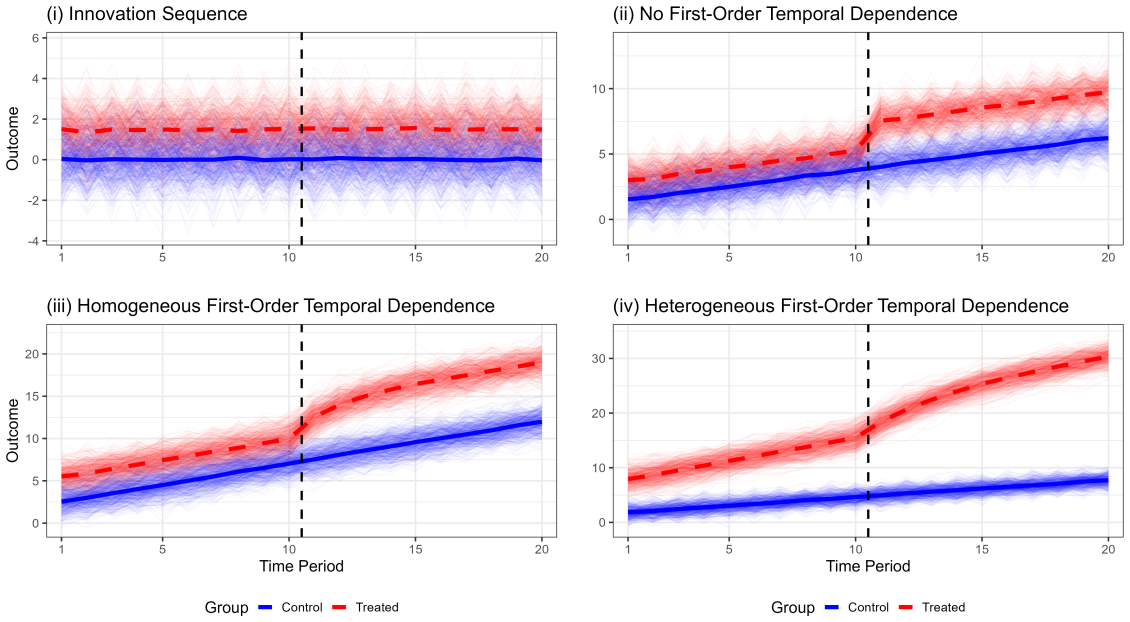
Note: The first row replicates Table 1 from Arkhangelsky et al. (2021) with "placebo method" standard errors in parentheses. The second row tabulates the corresponding estimators after implementing a pre-whitening procedure, which are neither economically or statistically significant at conventional thresholds.

Figure 1: The Impact of First-Order Temporal Dependence: $N = K = 2$



Note: This panel illustrates the impact of first-order temporal dependence on unconditional plots of the outcome. In this particular panel there a single treated unit and a single control unit. The time series of each unit is constructed off of a fixed set of innovations as illustrated in Panel (i). Panels (ii-iv) apply a common deterministic trend and different forms of first-order dependence (none, homogeneous, heterogeneous) to that set of innovations. The only element changing between Panels (ii-iv) is the group level autocovariance function, all other factors including the innovation sequence are held constant.

Figure 2: The Impact of First-Order Temporal Dependence: $N = 1000,\ K = 2$

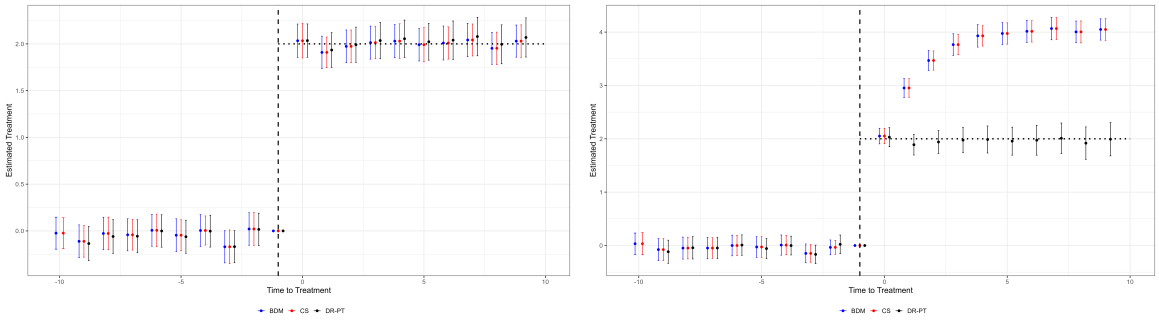


Note: This panel is an extention of Figure 1 to the panel setting where each group is expanded to include 500 units. Like before, Panel (i) displays the innovation sequence for each individual (transparent blue and red lines) and overlays the period-group average (solid blue and dashed red) line. All other factors except the group autocovariance functions are fixed as in Figure 1 with the exception of an increase in unit-level variance.

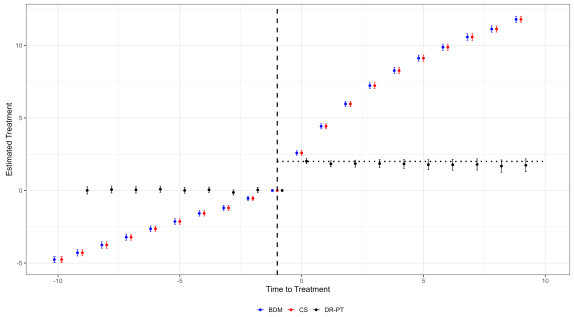
Figure 3: Event Study Estimates and First-Order Temporal Dependence

(i) Case 1

(ii) Case 2



(iii) Case 3

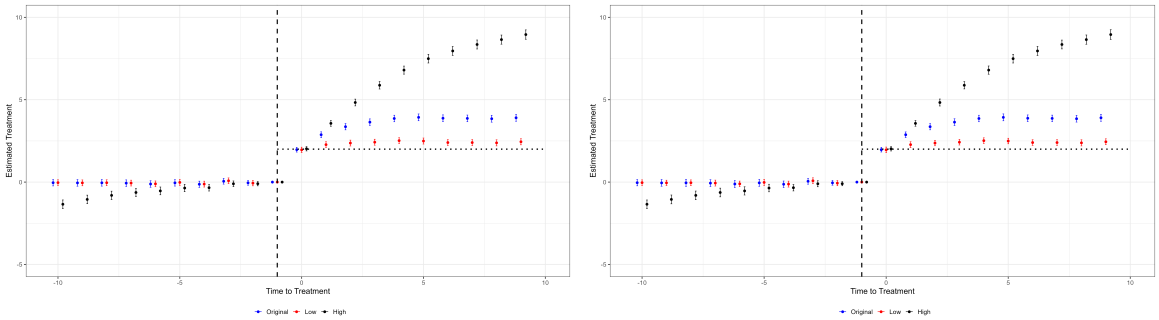


Note: These panels details the point estimates and 95% confidence intervals produced by evaluating Equation 17 with each estimator. Panel (i) shows that, in a static world, all three estimators produce point estimates consistent with the value of λ_k post-treatment. Moreover, all three fail-to-reject the null of a difference between treated and control groups relative to the reference period in the pre-treatment periods. Panel (ii) mirrors the smooth transition we saw in Figure 2 for the CS and BDM estimators with a time varying post-treatment estimate that converges to the long-run equilibrium value of $\lambda_k/(1-\phi)$. The DR-PT based estimator recovers a time-invariant estimate in the post treatment periods by separating the function into its constituent parts: time and treatment. Panel (iii) shows the effect of heterogeneous first-order dependence. As expected, when estimated without respect to past values Assumption A2b fails and a clear growth pattern is seen in both pre- and post-treatment periods for both the BDM and CS estimators. On the other hand the DR-PT based estimator correctly identifies that these groups share a common trend and correctly captures the structural effect of treatment.

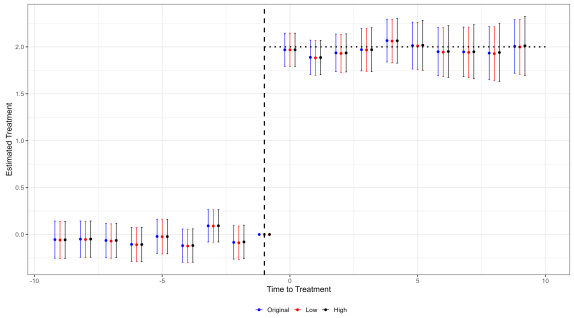
Figure 4: A Comparison Across Different Regimes

(i) BDM Comparison

(ii) CS Comparison

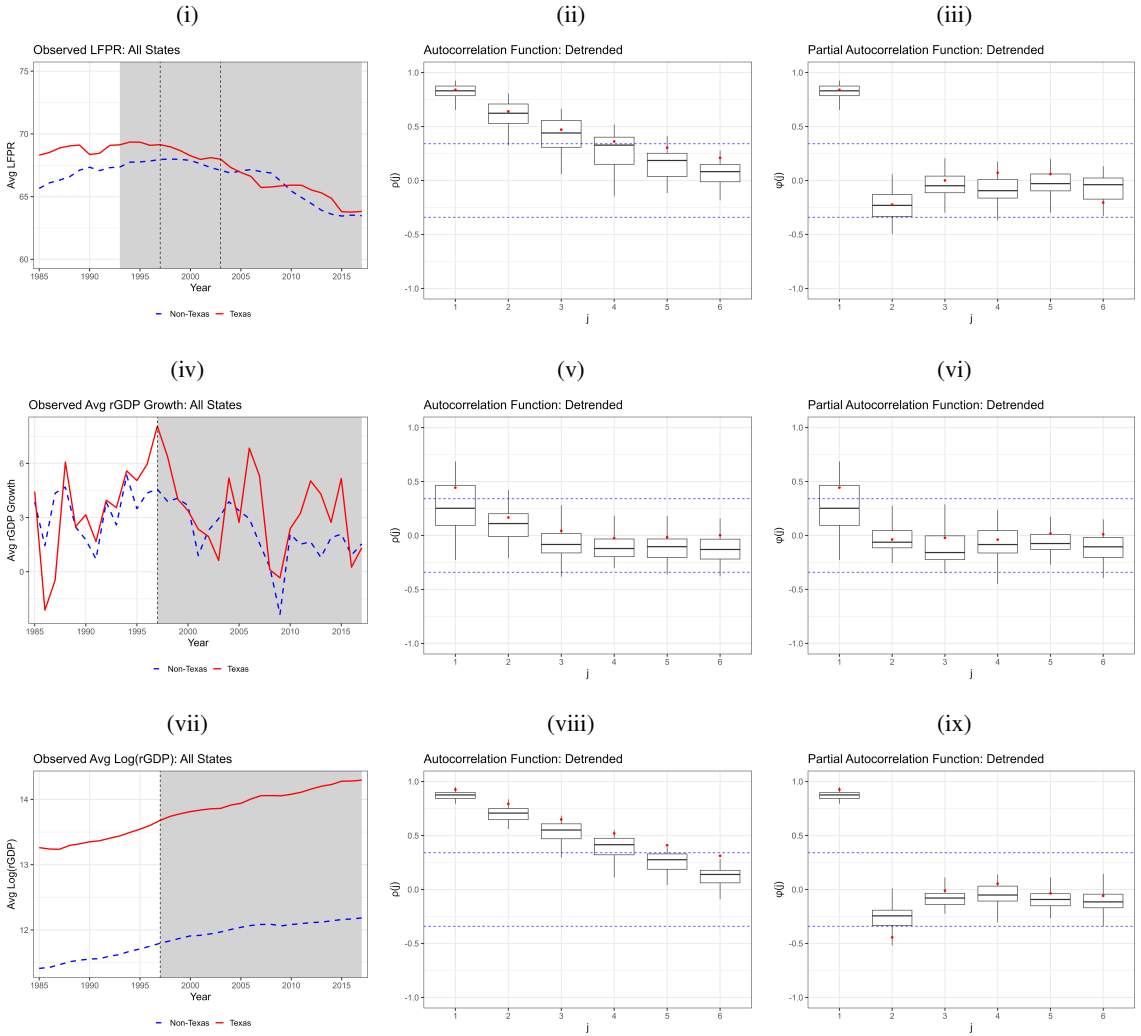


(iii) DR-PT Comparison



Note: This figure illustrates how external validity might be affected by failing to account for first-order temporal dependence. Panel (i) and (ii) show the BDM and CS estimator, when used in concert with Equation 17, can produce varying treatment effect sizes based on the dynamics of the new environment. If the new environment has a higher level of persistence then the same policy implemented with the same effect will be thought to have more of an effect. If the new environment has a lower level of dependence then this effect will be smaller. If the new environment has negative or oscillatory temporal dependence then the estimate will be biased towards zero. In all three cases the estimates produced by DR-PT are consistent with the true structural impact of the treatment.

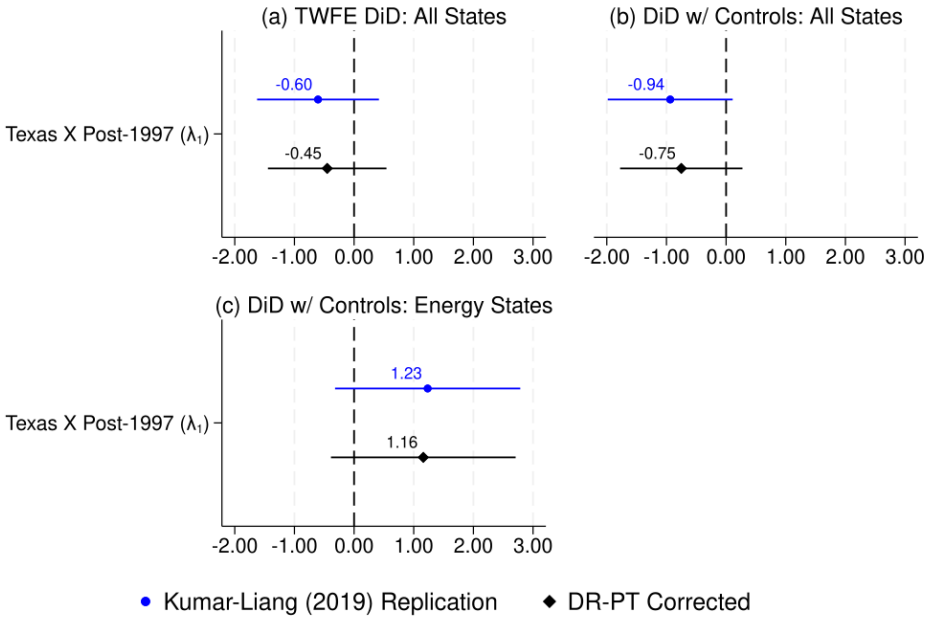
Figure 5: Visualizing the Effect of Autocorrelation in K-L Data



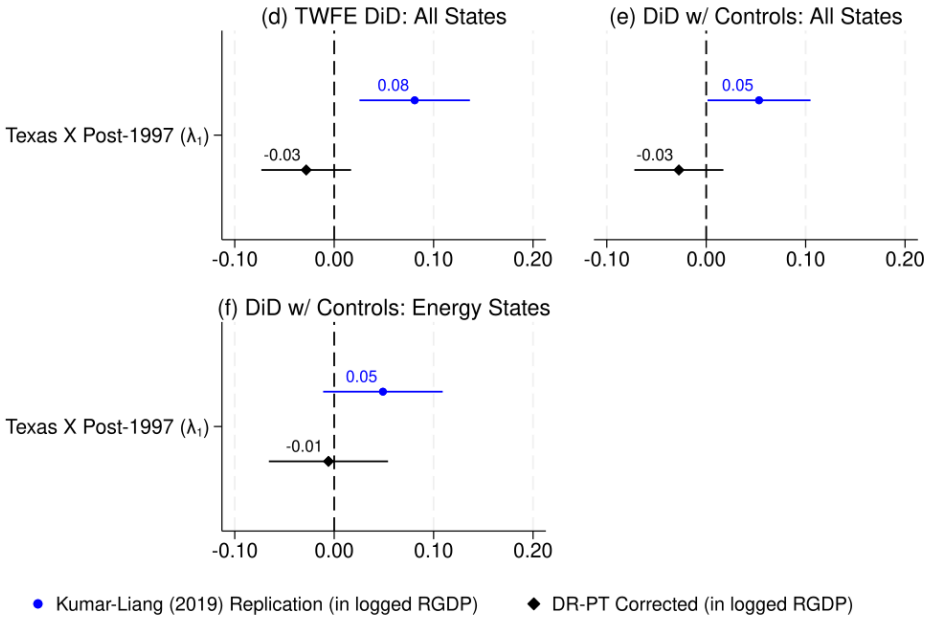
Note: In a time series workflow, broadly speaking, one would examine the raw outcome and determine its order of integration then examine the autocorrelation and partial autocorrelation function to determine a (starting) lag structure. A relatively slow to decay ACF when compared to the PACF is typically thought of as an indication of an autoregressive process where lags are initially chosen based on the number of significant lags in the PACF. The plots above provide a clear indication of an AR process.

Figure 6: Credit Access Effect on State Real GDP - Replicating Kumar and Liang (2019)

(i) Outcome: Real GDP Growth (low autocorrelation)



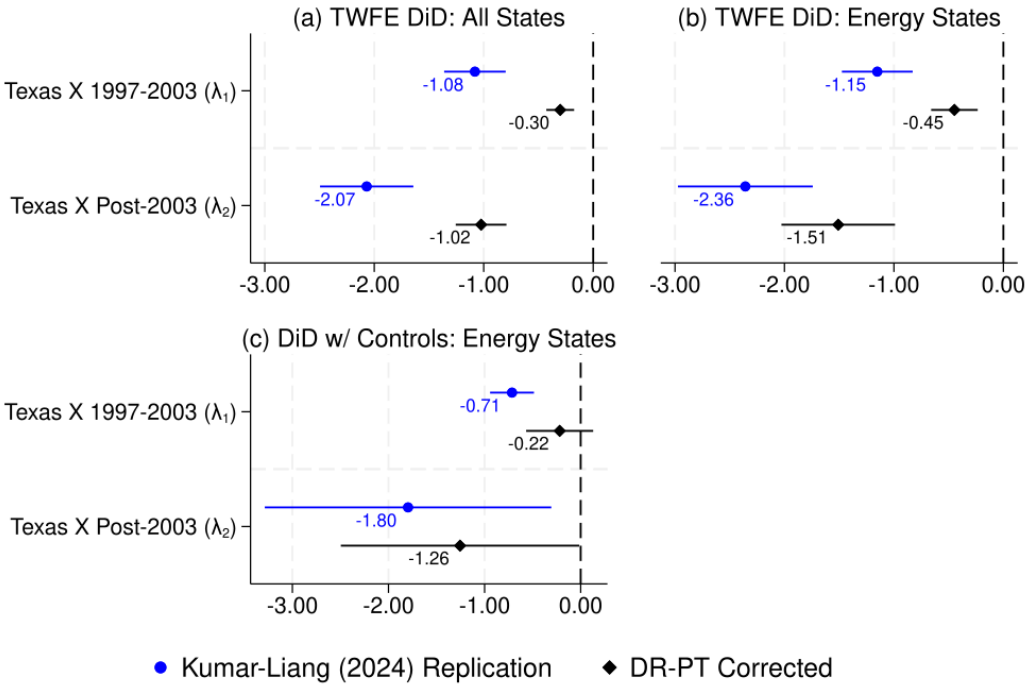
(ii) Outcome: ln(Real GDP) in levels (high autocorrelation)



Note: This figure replicates selected TWFE specifications from Kumar and Liang (2019). Panel (i) replicates their Table 1 in blue, which plots the point estimates of the DiD estimator for a TWFE regression with state and year fixed effects (panel (a)), adding in demographic controls (panel (b)), and confining the control group to other major energy producing states (panel (c)). The black estimates modify the model to incorporate an AR(1) term and its interaction with the treatment. Panel (ii) repeats this analysis, but with ln(real GDP) in levels. A comparison of Kumar and Liang (2019) full TWFE table and our AR-PT corrected estimates can be found in the appendix (Table 7 (full replication) and Table 8 (rGDP in levels).

Figure 7: Credit Access Effect on State-level LFPR - Replicating Kumar and Liang (2024)

(i) Outcome: Labor Force Participation Rate (moderate/high autocorrelation)

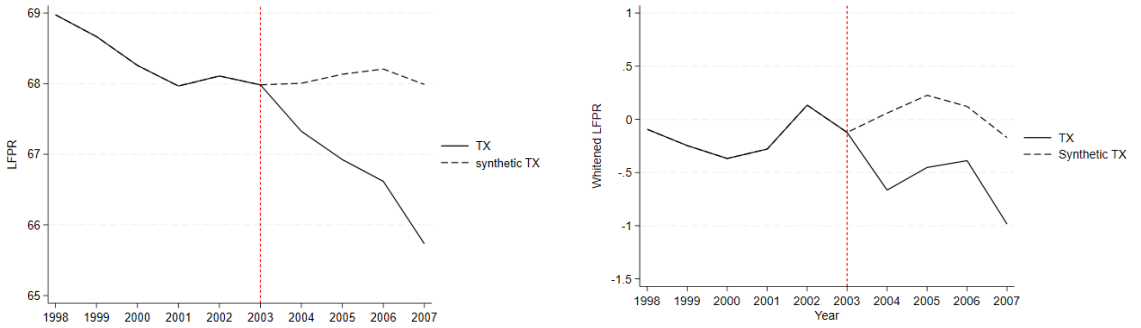


Note: This figure replicates selected TWFE specifications from Table 2 of Kumar and Liang (2024). Panel (a) is a basic TWFE specification with state and year fixed effects, while (b) is the same DiD specification constrained to energy states as the control group. These correspond to regressions 1a and 1b in Table 2 of their paper. To account for regional-specific shocks and state-specific trends, panel (c) incorporates census-division-by-year fixed effects and state-specific linear time trends (i.e., state-trend interactions). This corresponds to regression 3b in Table 2 of Kumar and Liang (2024). Their original point estimates and 95 percent confidence intervals appear in blue, while the black estimates incorporate an AR(1) term and its interaction with the treatment (i.e., DR-PT corrected). A comparison of Kumar and Liang (2024)'s full TWFE table and our AR-PT corrected estimates can be found in the appendix (Table 9 (Panel A replication) and Table 10 (Panel B replication)).

Figure 8: Credit Access Effect on State-level LFPR - Modifying the Synthetic Control Method

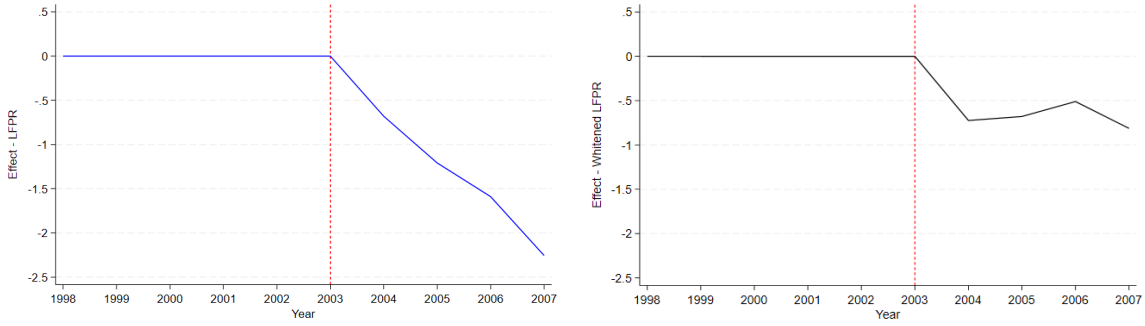
(i) LFPR: Replication of K-L 2024's Figure 5

(ii) Whitened LFPR: Modified SCM



(iii) LFPR: Effect (Replication)

(iv) Whitened LFPR: Effect (Modified SCM)

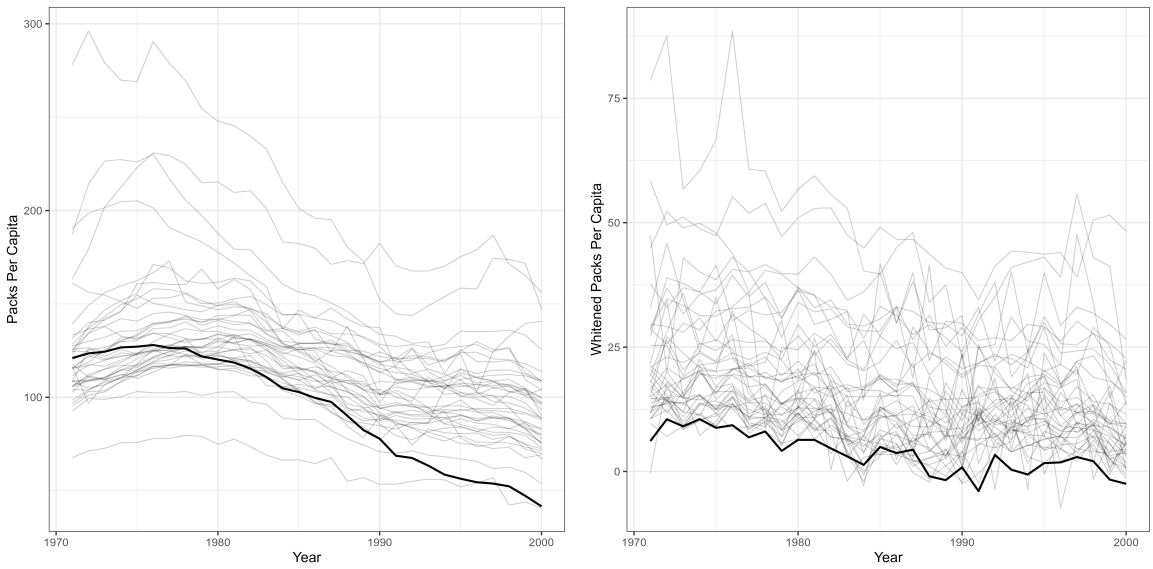


Note: The left panels in this figure replicates Figure 5 from Kumar and Liang (2024) (i) and the effect (iii) of the 2003 policy change on Texas labor force participation (compared to "synthetic" Texas constructed using synthetic control method). The right panels (ii and iv) modify the SCM by pre-whitening the time-series of each state before determining the weights and constructing "synthetic Texas." The results from panels (ii) and (iv) show a smaller but more stable effect over the post-treatment period, consistent with what we would expect to recover once we purge AR(1) dependence from the series.

Figure 9: State-level Cigarette Packs Per Capita, Raw versus Whitened (1970-2000)

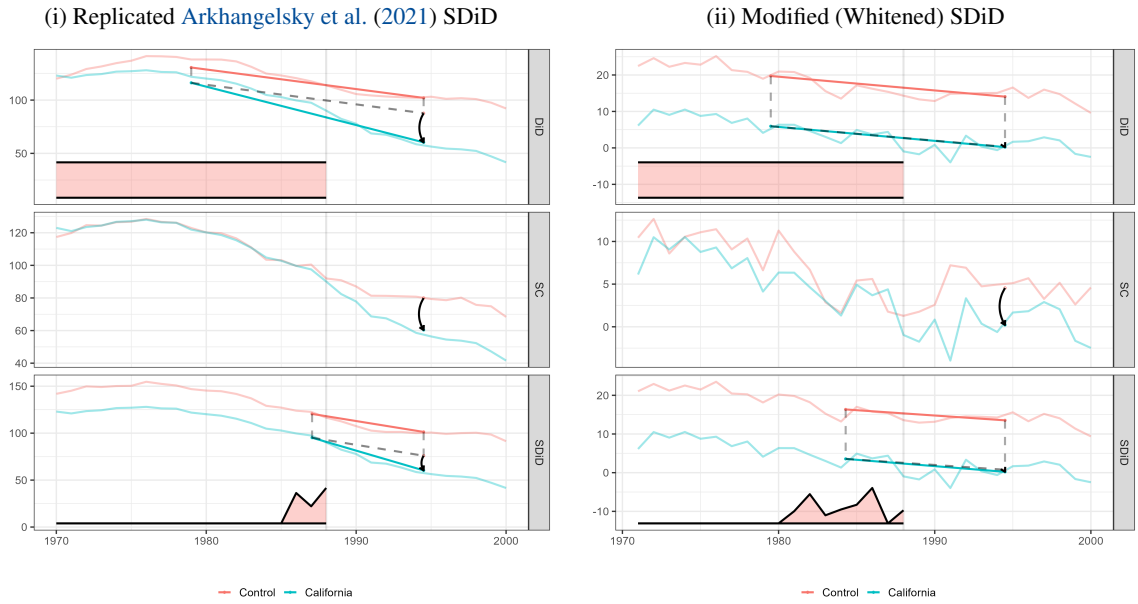
(i) Packs Per Capita Over Time

(ii) Whitened Packs Per Capita Over Time



Note: This figure depicts the time-series of 39 individual states used in Abadie et al. (2010) and Arkhangelsky et al. (2021)'s study of a 1989 policy change on cigarette taxes in California. Panel (i) replicates the series of raw data (cigarette consumption quantified as packs per capita) used in these studies from 1970 to 2000. Panel (ii) depicts the corresponding whitened time-series, which is the variation in the outcome independent of each individual state's AR(1) process. California is in the foreground in black, while all other states are represented in the background in gray.

Figure 10: Revisiting the Effect of California Proposition 99 - SDiD versus Modified (Whitened) SDiD



Note: This figure compares a replication of Arkhangelsky et al. (2021)'s analysis of California's cigarette tax policy change in 1989 and a modification that pre-whitens the time-series of each state. Panel (i) shows the original replication, which depicts the results from methods across three smaller panels: DiD (top), SCM (middle), and SDiD (bottom). Panel (ii) modifies this replication by whitening the time-series of each state to purge AR(1) dependence from the outcome. The resulting comparison shows that the whitened series is parallel before and after the policy change, recovering no significant time-invariant structural treatment effect. The results suggest the policy effect we observe in the data is largely driven by temporal dependence.

References

Abadie, A. (2005). Semiparametric difference-in-differences estimators. *Review of Economic Studies 72*(1), 1–19.

Abadie, A., A. Diamond, and J. Hainmueller (2010). Synthetic control methods for comparative case studies: Estimating the effect of california's tobacco control program. *Journal of the American statistical Association 105*(490), 493–505.

Abdallah, C. S. and W. D. Lastrapes (2012). Home equity lending and retail spending: Evidence from a natural experiment in texas. *American Economic Journal: Macroeconomics 4*(4), 94–125.

Angrist, J. D. and J.-S. Pischke (2009). *Mostly harmless econometrics: An empiricist's companion*. Princeton University Press.

Arellano, M. and S. Bond (1991). Some tests of specification for panel data: Monte carlo evidence and an application to employment equations. *The Review of Economic Studies 58*(2), 277–297.

Arkhangelsky, D., S. Athey, D. A. Hirshberg, G. W. Imbens, and S. Wager (2021). Synthetic difference-in-differences. *American Economic Review 111*(12), 4088–4118.

Arkhangelsky, D. and G. Imbens (2024). Causal models for longitudinal and panel data: A survey. *The Econometrics Journal 27*(3), C1–C61.

Athey, S. and G. W. Imbens (2006). Identification and inference in nonlinear difference-in-differences models. *Econometrica 74*(2), 431–497.

Bahadir, B., I. Gumus, and M. Schaffer (2024). Does relaxing household credit constraints hurt small business lending? evidence from a policy change in texas. *Journal of Corporate Finance 89*, 102682.

Baker, A., B. Callaway, S. Cunningham, A. Goodman-Bacon, and P. H. Sant'Anna (2025). Difference-in-differences designs: A practitioner's guide. *Journal of Economic Literature, Forthcoming*.

Baker, A. C., D. F. Larcker, and C. C. Wang (2022). How much should we trust staggered difference-in-differences estimates? *Journal of Financial Economics 144*(2), 370–395.

Bertrand, M., E. Duflo, and S. Mullainathan (2004). How much should we trust differences-in-differences estimates? *Quarterly Journal of Economics 119*(1), 249–275.

Bilinski, A. and L. A. Hatfield (2018). Nothing to see here? non-inferiority approaches to parallel trends and other model assumptions. *arXiv preprint arXiv:1805.03273*.

Born, B. and J. Breitung (2016). Testing for serial correlation in fixed-effects panel data models. *Econometric Reviews 35*(7), 1290–1316.

Brockwell, P. J. and R. A. Davis (2009). *Time series: theory and methods*. Springer science & business media.

Burlig, F., L. Preonas, and M. Woerman (2020). Panel data and experimental design. *Journal of Development Economics 144*, 102458.

Caetano, C. and B. Callaway (2024). Difference-in-differences when parallel trends holds conditional on covariates. *arXiv preprint arXiv:2406.15288*.

Caetano, C., B. Callaway, S. Payne, and H. S. Rodrigues (2022). Difference in differences with time-varying covariates. *arXiv preprint arXiv:2202.02903*.

Callaway, B. (2023). Difference-in-differences for policy evaluation. *Handbook of labor, human resources and population economics*, 1–61.

Callaway, B. and P. H. Sant'Anna (2021). Difference-in-differences with multiple time periods. *Journal of Econometrics 225*(2), 200–230.

Cameron, A. C. and D. L. Miller (2015). A practitioner's guide to cluster-robust inference. *Journal of Human Resources 50*(2), 317–372.

Card, D. and A. B. Krueger (1994). Minimum wages and employment: A case study of the fast-food industry in new jersey and pennsylvania. *American Economic Review 84*(4), 772–793.

Cefalu, M., B. G. Vegetabile, M. Dworsky, C. Eibner, and F. Girosi (2020). Reducing bias in difference-in-differences models using entropy balancing. *arXiv preprint arXiv:2011.04826*.

Cunningham, S. (2021). *Causal inference: The mixtape*. Yale University Press.

Currie, J., H. Kleven, and E. Zwiers (2020). Technology and big data are changing economics: Mining text to track methods. In *AEA Papers and Proceedings*, Volume 110, pp. 42–48. American Economic Association 2014 Broadway, Suite 305, Nashville, TN 37203.

De Chaisemartin, C. and X. d'Haultfoeuille (2023). Two-way fixed effects and differences-in-differences with heterogeneous treatment effects: A survey. *The Econometrics Journal 26*(3), C1–C30.

Dube, A., D. Girardi, O. Jorda, and A. M. Taylor (2025). A local projections approach to difference-in-differences. *Journal of Applied Econometrics*.

Ferman, B. (2023). Inference in difference-in-differences: How much should we trust in independent clusters? *Journal of Applied Econometrics 38*(3), 358–369.

Goldsmith-Pinkham, P., P. Hull, and M. Kolesár (2024). Contamination bias in linear regressions. *American Economic Review 114*(12), 4015–4051.

Goodman-Bacon, A. (2021). Difference-in-differences with variation in treatment timing. *Journal of Econometrics 225*(2), 254–277. Themed Issue: Treatment Effect 1.

Ho, D. E., K. Imai, G. King, and E. A. Stuart (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political analysis 15*(3), 199–236.

Hsiao, C. and Q. Zhou (2024). Panel treatment effects measurement: Factor or linear projection modelling? *Journal of Applied Econometrics 39*(7), 1332–1358.

Khan, S. and E. Tamer (2010). Irregular identification, support conditions, and inverse weight estimation. *Econometrica 78*(6), 2021–2042.

Kumar, A. (2018). Do restrictions on home equity extraction contribute to lower mortgage defaults? evidence from a policy discontinuity at the texas border. *American Economic Journal: Economic Policy 10*(1), 268–297.

Kumar, A. and C.-Y. Liang (2019). Credit constraints and gdp growth: Evidence from a natural experiment. *Economics Letters 181*, 190–194.

Kumar, A. and C.-Y. Liang (2024). Labor market effects of credit constraints: Evidence from a natural experiment. *American Economic Journal: Economic Policy 16*(3), 1–26.

Lastrapes, W. D., I. Schmutte, and T. Watson (2022). Home equity lending, credit constraints and small business in the us. *Economic Inquiry 60*(1), 43–63.

Marcus, M. and P. H. Sant'Anna (2021). The role of parallel trends in event study settings: An application to environmental economics. *Journal of the Association of Environmental and Resource Economists 8*(2), 235–275.

Moulton, B. R. (1990). An illustration of a pitfall in estimating the effects of aggregate variables on micro units. *The Review of Economics and Statistics*, 334–338.

Nickell, S. (1981). Biases in dynamic models with fixed effects. *Econometrica*, 1417–1426.

Piccininni, M., E. J. T. Tchetgen, and M. J. Stensrud (2025). Refining the notion of no anticipation in difference-in-differences studies. *arXiv preprint arXiv:2507.12891*.

Piger, J. and T. Stockwell (2025). Differences from differencing: Should local projections with observed shocks be estimated in levels or differences? *Journal of Applied Econometrics*.

Rambachan, A. and J. Roth (2023). A more credible approach to parallel trends. *Review of Economic Studies 90*(5), 2555–2591.

Roth, J. (2022). Pretest with caution: Event-study estimates after testing for parallel trends. *American Economic Review: Insights 4*(3), 305–322.

Roth, J. and P. H. Sant'Anna (2023). When is parallel trends sensitive to functional form? *Econometrica 91*(2), 737–747.

Roth, J., P. H. Sant'Anna, A. Bilinski, and J. Poe (2023). What's trending in difference-in-differences? a synthesis of the recent econometrics literature. *Journal of Econometrics 235*(2), 2218–2244.

Sant'Anna, P. H. and Q. Xu (2023). Difference-in-differences with compositional changes. *arXiv preprint arXiv:2304.13925*.

Sant'Anna, P. H. and J. Zhao (2020). Doubly robust difference-in-differences estimators. *Journal of econometrics 219*(1), 101–122.

Słoczyński, T. (2022). Interpreting ols estimands when treatment effects are heterogeneous: Smaller groups get larger weights. *Review of Economics and Statistics 104*(3), 501–509.

Wooldridge, J. M. (2023). Simple approaches to nonlinear difference-in-differences with panel data. *The Econometrics Journal 26*(3), C31–C66.

Ye, T., L. Keele, R. Hasegawa, and D. S. Small (2024). A negative correlation strategy for bracketing in difference-in-differences. *Journal of the American Statistical Association 119*(547), 2256–2268.

Zevelev, A. A. (2021). Does collateral value affect asset prices? evidence from a natural experiment in texas. *The Review of Financial Studies 34*(9), 4373–4411.

9 Appendix

9.1 Appendix Tables and Figures

Table 7: Replication of Kumar and Liang's (2019) Table 1 with DR-PT Corrections

Panel A - Replication of K-L (2019)			
Dependent Var.: real GDP Growth	T1-1	T1-2	T1-3
Texas X post 1997	-0.604	-0.937	1.234
	(0.391)	(0.448)	(0.798)
Observations	800	800	192
Adjusted R-Squared	0.405	0.424	0.399
AIC	3125	3106	805
BIC	3130	3148	834
Panel B - Replication Using Dynamically Robust Parallel Trends			
Dependent Var.: real GDP Growth	T1-1	T1-2	T1-3
Texas X post 1997	-0.451	-0.751	1.160
	(0.275)	(0.329)	(0.919)
GDPGrt-1 ρ_1	0.243	0.210	0.016
	(0.058)	(0.051)	(0.103)
Texas X GDPGrt-1 ρ_2	0.059	0.106	0.474
	(0.075)	(0.071)	(0.067)
Observations	800	800	192
Adjusted R-Squared	0.439	0.450	0.431
AIC	3079	3071	796
BIC	3093	3123	732
Additional Controls by Model			
State FE	Yes	Yes	Yes
Year FE	Yes	Yes	Yes
Demographic Controls	No	Yes	Yes

Note: This table replicates the main TWFE results from Kumar and Liang (2019). Panel A replicates Table 1 in their paper, estimating the effect of a Texas policy changing credit constrains on real GDP. Panel B uses the same specifications as Panel A but incorporates additional AR(1) terms (a lagged dependent variable - real GDP growth rate in the prior period - and its interaction with the treatment group variable, Texas) as an DR-PT correction. Robust standard errors (clustered by state) are in ( ).

Table 8: Replication of Kumar and Liang's (2019) in logged Real GDP (in levels)

Panel A - Replication of K-L (2019)			
Dependent Var.: ln(real GDP)	T1-1	T1-2	T1-3
Texas X post 1997	0.081	0.053	0.049
	(0.013)	(0.020)	(0.036)
Observations	800	800	192
Adjusted R-Squared	0.997	0.998	0.997
AIC	-2459	-2583	-597
BIC	-2455	-2540	-568
Panel B - Replication Using Dynamically Robust Parallel Trends			
Dependent Var.: ln(real GDP)	T1-1	T1-2	T1-3
Texas X post 1997	-0.028	-0.028	-0.006
	(0.004)	(0.004)	(0.008)
ln(RGDP)-1 ρ_1	0.890	0.897	0.867
	(0.015)	(0.014)	(0.038)
Texas X ln(RGDP)-1 ρ_2	0.092	0.085	0.053
	(0.010)	(0.015)	(0.018)
Observations	800	800	192
Adjusted R-Squared	1.000	1.000	0.999
AIC	-4099	-4106	-899
BIC	-4085	-4055	-863
Additional Controls by Model			
State FE	Yes	Yes	Yes
Year FE	Yes	Yes	Yes
Demographic Controls	No	Yes	Yes

Note: This table replicates a variation of the main TWFE results from Kumar and Liang (2019). Panel A re-estimates Table 1 in their paper using logged real GDP (in levels) as the dependent variable instead of growth rates. Panel B uses the same specifications as Panel A but incorporates additional AR(1) terms (a lagged dependent variable - ln(real GDP) in the prior period - and its interaction with the treatment group variable, Texas) as an DR-PT correction. Robust standard errors (clustered by state) are in ( ).

Table 9: Replication of Kumar and Liang's (2024) Table 3 Panel A with DR-PT Corrections

Panel A - Replication of K-L (2024) Table 3, Panel A. All States Sample.

Dependent Var.: LFPR	T3-1a	T3-2a	T3-3a	T3-4a	T3-5a	T3-6a
Texas X 1998-2003	-1.08	-0.717	-0.085	-0.501	-0.501	-0.884
	(0.140)	(0.628)	(0.667)	(0.677)	(0.423)	(0.363)
Texas X post 2003	-2.069	-2.625	-1.474	-0.935	-1.27	-1.901
	(0.212)	(0.290)	(0.526)	(0.451)	(0.664)	(0.643)
Observations	800	800	800	800	797	597
Adjusted R-squared	0.943	0.951	0.961	0.959	0.975	0.98
AIC	1757	1488	1158	1196	788	409
BIC	1767	1497	1167	1205	849	470

Panel B - Replication Using Dynamically Robust Parallel Trends

Dependent Var.: LFPR	T3-1a	T3-2a	T3-3a	T3-4a	T3-5a	T3-6a
Texas X 1998-2003	-0.301	-0.385	-0.069	-0.465	-0.329	-0.722
	(0.063)	(0.215)	(0.103)	(0.294)	(0.084)	(0.194)
Texas X post 2003	-1.024	-1.866	-1.361	-1.495	-1.193	-1.863
	(0.115)	(0.321)	(0.328)	(0.221)	(0.374)	(0.613)
LFPRt-1 (ρ_1)	0.756	0.761	0.666	0.673	0.501	0.372
	(0.025)	(0.031)	(0.032)	(0.030)	(0.037)	(0.040)
LFPRt-1 X Texas (ρ_2)	-0.07	-0.358	-0.375	-0.621	-0.268	0.250
	(0.055)	(0.122)	(0.171)	(0.262)	(0.158)	(0.473)
Observations	800	800	800	800	797	597
Adjusted R-squared	0.976	0.979	0.978	0.979	0.983	0.984
AIC	1073	818	689	679	479	278
BIC	1091	837	707	697	550	349
Additional Controls by Model						
State FE	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Div. x Year Effects	No	Yes	Yes	Yes	Yes	Yes
State x Linear Trend	No	No	Yes	No	Yes	Yes
Oil price x State FE	No	No	No	Yes	No	No
Other controls	No	No	No	No	Yes	Yes
Bank branching	No	No	No	No	No	Yes

Note: This table replicates the main TWFE results from Kumar and Liang (2024) for their sample using all 50 states, estimating the effect of Texas policy changing credit constraints on labor force participation rates. Panel A replicates Table 3, Panel A in their paper. Panel B uses the same specifications as Panel A but incorporates additional AR(1) terms (a lagged dependent variable - labor force participation rate (LFPR) in the prior period - and its interaction with the treatment group variable, Texas) as an DR-PT correction. Robust standard errors (clustered by state) are in ( ).

Table 10: Replication of Kumar and Liang's (2024) Table 3 Panel B with DR-PT Corrections

Panel A - Replication of K-L (2024) Table 3, Panel B. Energy States Sample

Dependent Var.: LFPR	T3-1b	T3-2b	T3-3a	T3-4b	T3-5b	T3-6a
Texas X 1998-2003	-1.152	-1.31	-0.714	-1.115	-0.833	-0.983
	(0.147)	(0.305)	(0.104)	(0.34)	(0.292)	(0.608)
Texas X post 2003	-2.357	-2.888	-1.796	-1.36	-1.573	-1.954
	(0.28)	(0.117)	(0.677)	(0.044)	(0.682)	(1.097)
Observations	192	128	192	128	128	144
Adjusted R-squared	0.979	0.973	0.977	0.974	0.976	0.975
AIC	275	112	98	81	62	5
BIC	282	118	105	86	81	26

Panel B - Replication Using Dynamically Robust Parallel Trends

Dependent Var.: LFPR	T3-1a	T3-2a	T3-3a	T3-4a	T3-5a	T3-6a
Texas X 1998-2003	-0.447	-0.703	-0.219	-0.842	-0.274	-0.378
	(0.10)	(0.12)	(0.16)	(0.10)	(0.32)	(0.60)
Texas X post 2003	-1.509	-2.018	-1.255	-1.467	-1.301	-1.712
	(0.24)	(0.43)	(0.56)	(0.25)	(0.55)	(1.10)
LFPRt-1 ρ_1	0.733	0.588	0.569	0.53	0.541	0.521
	(0.07)	(0.06)	(0.05)	(0.10)	(0.06)	(0.20)
LFPRt-1 X Texas ρ_2	-0.23	-0.236	-0.385	-0.7	-0.193	-0.363
	(0.08)	(0.14)	(0.19)	(0.29)	(0.22)	(0.84)
Observations	192	128	192	128	128	144
Adjusted R-squared	0.989	0.982	0.984	0.981	0.981	0.978
AIC	149	56	27	41	23	-24
BIC	162	64	37	50	43	-3
Additional Controls by Model						
State FE	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Div. x Year Effects	No	Yes	Yes	Yes	Yes	Yes
State x Linear Trend	No	No	Yes	No	Yes	Yes
Oil price x State FE	No	No	No	Yes	No	No
Other controls	No	No	No	No	Yes	Yes
Bank branching	No	No	No	No	No	Yes

Note: This table replicates TWFE results from Kumar and Liang (2024) for their sample using energy-intensive states, estimating the effect of a Texas policy changing credit constraints on labor force participation rates. Panel A here replicates Table 3, Panel B in their paper. Panel B uses the same specifications as Panel A but incorporates additional AR(1) terms (a lagged dependent variable - labor force participation rate (LFPR) in the prior period - and its interaction with the treatment group variable, Texas) as an DR-PT correction. Robust standard errors (clustered by state) are in ( ).