Efficient Analysis of Experimental Panel Data: A Comparison of Categorical and Dynamic Models

Jeffrey A Mills
University of Cincinnati

Vikram K Suresh
University of Cincinnati

January 10, 2026

Abstract

This paper evaluates the finite-sample performance of categorical-time versus dynamic specifications for estimating treatment effect trajectories in short experimental panels. Using Monte Carlo simulations and STAR data, we show that autoregressive and log-trend models substantially improve precision and statistical power when treatment effects evolve smoothly, reducing path RMSE by 30-40 percent relative to categorical-time specifications. These efficiency gains arise from restricting the conditional mean rather than modeling error covariance. However, categorical-time models exhibit greater robustness to irregular treatment paths. The results highlight a fundamental bias-variance tradeoff, with specification choice depending on the plausibility of dynamic structure and inferential objectives.

Keywords:
JEL:

1 Introduction

Randomized experiments (RCTs) have become central to empirical economics, providing credible causal evidence upon interventions ranging from developmental to educational and labor market programs (Duflo, Glennerster, and Kremer, 2007; Banerjee and Duflo, 2009; Angrist and Pischke, 2009; Athey and Imbens, 2017). Increasingly, these experiments collect outcomes repeatedly over time, generating short panel datasets with modest number of units and multiple follow-up waves. Repeated measurements are intended to improve precision, document the evolution of treatment effects and distinguish short-run impacts from longer-run persistence or fade-out (McKenzie, 2012). Despite this, the standard empirical practice treats time categorically, estimating a separate treatment effect for each follow-up period.

The categorical time approach is flexible and transparent, but it is statistically costly in short panels. Estimating multiple treatment-time interactions consumes degrees of freedom, inflates variance and can produce erratic period-to-period estimates that may obscure true smooth treatment trajectories (Wooldridge, 2010). In field experiments where sample sizes are often constrained by cost and logistics (Muralidharan and Sundararaman, 2011; Bloom et al., 2013), outcomes such as achievement, earnings or health status exhibit strong temporal dependence. As a result, experiments that are well powered to detect endpoint effects may remain underpowered to characterize how treatment evolved over time.

When outcomes are observed repeatedly on a common scale, such as within wave standardized measures or change from baseline observations, this persistence can be modeled explicitly rather than treated implicitly or ignored. This allows researchers to replace high-dimensional set of period specific treatment effects with a small number of economically interpretable parameters governing outcome dynamics. Autoregressive specifications capture persistence and convergence while smooth trend models summarize nonlinear treatment trajectories consistent with learning curves or diminishing returns (Ben-Porath, 1967; Cunha and Heckman, 2007). Imposing such structure entails a bias vs. variance trade-off, while dynamic models restrict the space of possible trajectories, they can deliver efficiency gains when restrictions are appropriate.

This paper studies the potential efficiency of imposing parsimonious dynamic structure on treatment effect trajectories in randomized short panels. We focus on two classes of dynamic specifications commonly motivated by economic behavior, autoregressive models that capture outcome persistence and smooth log-trend models that summarize nonlinear treatment evolution. These are contrasted with the standard categorical-time specifications. The estimand of interest are the implied treatment effects on outcome trajectories defined as the difference in the expected outcomes between treatment and control at measurement wave, these are functions of underlying parameters rather than separate period-specific coefficients. The comparison is therefore one of representation and efficiency, not identification or estimand choice.

For this, we first conduct a Monte Carlo study calibrated to sample sizes, panel lengths and temporal dependence patterns observed in experimental conditions. The simulations are designed to assess how alternative representations of the same treatment effect trajectory perform in finite samples, both when dynamic restrictions are approximately current and when misspecified. We evaluate estimators using root mean squared error of the estimated treatment path and power for endpoint inference. Second, we provide an empirical subsampling exercise based on project STAR (Krueger, 1999), a canonical education experiment with four waves of achievement data. This exercise examines the finite-sample stability of competing specifications under resampling from an unknown data-generating process.

The results show that dynamic specifications can yield substantial efficiency gains relative to categorical time models when outcomes exhibit temporal structure. By pooling information across waves through economically motivated restrictions, dynamic models improve precision and power in short panels, enabling more informative inference about the timing and persistence of treatment effects. These gains are achieved without relying on large-sample approximations or complex covariance structures and are pronounced in sample sizes encountered in pilot studies and small-scale experiments. We provide systematic evidence on how alternative representations of treatment effect trajectories perform in finite samples under realistic experimental conditions. By isolating the bias–variance trade-off inherent in dynamic modeling, the paper offers practical guidance for applied researchers working with short experimental panels.

The remainder of the paper proceeds as follows. Section 2 presents the econometric framework, defining the estimands and outlining the assumptions underlying categorical-time and dynamic specifications. Section 3 describes the Monte Carlo simulation design, including the data-generating processes and estimation strategies. Section 4 provides an empirical illustration. Section 5 discusses implications for experimental practice and limitations, and Section 6 concludes.

2 Econometric Framework

2.1 Setup and Estimands

Consider a randomized controlled trial with N units indexed by $i = 1, \ldots, N$. Each unit is observed at T measurement waves indexed by $t = 1, \ldots, T$, with treatment assignment $D_i \in \{0, 1\}$ fixed over time. Let Y_{it} denote the observed outcome at wave t. We focus on settings in which outcomes are observed on a comparable scale across waves, either through within-wave standardization or normalization relative to the first observed measurement. The primary estimand of interest is the treatment effect on the outcome trajectory, defined at each wave t as,

$$\Delta_t \equiv \mathbb{E}\left[Y_{it}(1) - Y_{it}(0)\right]$$

Where, $Y_{it}(d)$ denotes the potential outcomes under treatment status d. In experimental applications, $\{\Delta_t\}_{t=1}^{T}$ are estimated directly as period specific coefficients with particular focus on endpoint treatment effect Δ_T. In dynamic models, $\{\Delta_t\}$ are implied functions of a small number of structural parameters governing outcome persistence or trend dynamics. Our focus is on comparing the finite-sample performance of these alternative representations of the same underlying estimand. Randomization ensures that in expectation treatment assignment is independent of potential outcome trajectories. Throughout, we abstract from missingness and attrition to isolate the efficacy of alternative mean specifications under ideal experimental conditions.

2.2 Categorical Time Specification

The standard approach in experimental panel data with repeated measurements treats time categorically, allowing treatment effects to vary freely across measurement waves,

$$(1)\ Y_{it} = \lambda_t + \tau_t D_i + \epsilon_{it},\ t = 1,\ \ldots,\ T$$

Here, λ_t are period fixed effects, τ_t captures the treatment effect at wave t. This specification is highly flexible and transparent, imposing no restrictions on the shape of the treatment trajectory. However, when T is small and outcomes are serially correlated estimating each τ_t can result in high variance and unstable estimates across adjacent t (Wooldridge, 2010).

2.3 Dynamic Specifications

We contrast the categorical-time model with dynamic specifications that impose structure on outcome evolution. Keeping with the empirical contexts motivating this study, we omit period fixed effects in these models. This restriction is appropriate when outcomes are expressed as within-wave standardized measures or as changes relative to baseline measurement, so that common wave-level shifts are mechanically removed. Throughout the dynamic specifications, we omit period fixed effects λ_t because outcomes are normalized to be comparable across waves (e.g., within-wave standardization or change-from-baseline), so common shifts are absorbed by construction and do not contribute to identifying treatment trajectories.

2.3.1 Autoregressive Models

We first consider autoregressive specifications that model outcomes as persistent states. With T observed waves, the AR models are defined over $T - 1$ within-unit transitions, treating the first observation as an initial condition. For $t = 2, \ldots, T$,

$$(2)\ Y_{it} = (\phi + \kappa D_i)\, Y_{i,t-1} + \tau D_i + u_{it}$$

The parameter ϕ captures the outcome persistence in the control group, while κ allows treatment to modify persistence. The parameter τ represents a treatment-induced shift in the conditional outcome level. Stationarity is ensured if $|\phi| < 1$ and $|\phi + \kappa| < 1$. Although, τ and κ are not themselves period-specific treatment effects, they jointly determine the treatment effect on the outcome trajectory. Define the treatment effect on the levels at wave t as $\Delta_t \equiv \mathbb{E}\left(Y_{it}|D_i = 1\right) - \mathbb{E}\left(Y_{it}|D_i = 0\right)$. Under the AR specification and the normalization $\Delta_1 = 0$,

$$\Delta_t = (\phi + \kappa)\Delta_{t-1} + \tau,\ t = 2, \ldots,\ T$$

Thus, the implied treatment trajectory is a deterministic function of $\phi,\ \kappa, \tau$ with endpoint effect

$$\Delta_T = \tau \sum_{j=0}^{T-2} (\phi + \kappa)^j$$

This yields a closed form expression for the treatment effect at each wave. Thus, the AR model targets the same set of estimands $\{\Delta_t\}$ as the categorical time model, through a parsimonious dynamic structure. This specification formalizes a common intuition in experimental work, past outcomes are predictive of future outcomes, even when baseline measurements are imperfect or non-comparable across units (Duflo, Dupas, and Kremer, 2011; Andrabi et al., 2011).

To clarify the source of dynamic treatment effects, we also consider restricted AR variants that isolate (i) a level-only specification ($\kappa = 0$) and (ii) a moderation-only specification ($\tau = 0$). These restrictions correspond to whether treatment primarily shifts the conditional mean level or primarily alters persistence. This decomposition is used in the simulation study and mirrored in the empirical illustration.

2.3.2 Log-Trend Models

We also consider smooth trend models that summarize nonlinear treatment trajectories. Let the first observation serve as a common reference point and define outcome changes as $Y_{it} - Y_{i1}$. For $t = 2, \dots, T$,

$$(3)\ Y_{it} - Y_{i1} = \beta_0 \log(t) + \delta_0 D_i + \theta \left(D_i \cdot \log(t)\right) + v_{it}$$

The coefficient β_0 captures the average evolution of outcomes in the control group, while θ measures whether treatment accelerates or amplifies this trajectory. This functional form is motivated by learning and adjustment processes that exhibit early gains followed by diminishing returns (Ben-Porath, 1967; Cunha and Heckman, 2007). As with the AR model, the implied treatment effects $\{\Delta_t\}$ are smooth functions of small number of parameters rather than unrestricted period-specific coefficients.

$$\Delta_t = \delta_0 + \theta \log(t),\ t = 1, \dots, T$$

2.4 Error Structure and Identification

In all specifications, we allow for within-unit correlation in the error terms using standard covariance structures, including independence, compound symmetry, and autoregressive dependence (Wooldridge, 2010; Verbeke and Molenberghs, 2000). In short panels, correct specification of the conditional mean is often more consequential for efficiency than precise modeling of the residual covariance, particularly when outcomes exhibit strong persistence or smooth trajectories (McKenzie, 2012; Andrabi et al., 2011). Identification relies on random assignments and the assumption that attrition and measurement error do not become differential by treatment status over time. Under these conditions, common wave-level shocks affect precision but do not bias treatment comparisons (Imbens and Rubin, 2015). We interpret differences across specifications as reflecting trade-offs between flexibility and efficiency rather than changes in the underlying estimand.

Random assignment also distinguishes this setting from observational dynamic panel models that motivate instrumental-variable estimators such as Arellano and Bond (1991). In observational panels, lagged outcomes become endogenous after removing unit fixed effects, requiring deeper lags as instruments. In experimental panels, treatment is orthogonal to the shock process governing outcome dynamics by design, so the core identification concern addressed by difference-GMM does not arise for treatment effect estimation. Remaining issues are therefore primarily finite-sample and efficiency-related rather than identification-driven (Arellano and Bond, 1991; Wooldridge, 2010; Imbens and Rubin, 2015).

2.5 Existing Practices

Dynamic specifications formalize intuitions that are already implicit in much experimental work. Baseline outcomes are often viewed as informative but imperfect proxies for latent ability or state variables, and repeated measurements are collected precisely because outcomes are expected to evolve systematically over time (Duflo, Dupas, and Kremer, 2011; McKenzie, 2012). Our framework makes this structure explicit when repeated, comparable measurements are available, allowing treatment effects on trajectories to be estimated more efficiently.

2.6 Within-Wave Normalization

Outcomes are standardized within wave prior to estimation for each period t,

$$Y_{it}^* = \frac{Y_{it} - \mu_{0t}}{\sigma_{0t}}$$

Where μ_{0t} and σ_{0t} are the mean and standard deviation of control group in period t. This normalization serves two purposes. First, it ensures treatment effects Δ_t in standard deviation units, ensuring measurement comparability across samples, grades, specifications and aligning effect magnitudes with standard practices in experimental and educational literatures. Second, within-wave absorbs common period effects by construction. Any aggregate shifts in the outcome levels due to test scaling, grade difficulty or general secular trends are removed mechanically. As a result, the outcome for the control group is zero for each period.

Explicit period fixed effects which usually control for these shifts are not required by construction in this normalized outcome space. All temporal evolution in the models reflects relative dynamics in treatment effects rather than aggregate time trends. In this setting, the autoregressive and smooth-trend restrictions characterize the persistence or attenuation of treatment effects over time, not changes in outcome scales. Normalization therefore simplifies interpretation of the treatment effect path $\{\Delta_t\}$ and ensures all estimands of interest are defined on a common comparable scale which is policy relevant.

3 Monte Carlo Simulation Design

We study finite-sample performance in randomized short panels by comparing the categorical-time representation of dynamic treatment effects to parsimonious dynamic specifications that impose structure on outcome evolution. The simulation design is calibrated to experimental settings in which researchers observe a balanced panel with some modest T (4, 8) and moderate-to-small N (50, 150). The scientific object to recover is the treatment effect trajectory and an endpoint effect. Our focus is explicitly on efficiency and bias-variance trade-offs induced by mean restrictions rather than identification which is assumed by construction in all designs.

3.1 Data and Estimand

For each Monte Carlo replication, we generate a balanced panel of $\{Y_{it}\}_{t=1}^{T}$ for units $i = 1, \ldots, N$ under randomized assignment $D_i \in \{0,1\}$, with $P(D_i = 1) = \frac{1}{2}$. Outcomes are simulated on a comparable scale across periods, consistent with within-wave normalization or baseline-reference changes, so that the treatment effect path is well defined and comparable across t. The estimand is the period-specific average treatment effect, $\Delta_t \equiv \mathbb{E}\left[Y_{it}(1) - Y_{it}(0)\right], \ t = 1, \ldots, T$.

3.2 Data Generating Processes

We consider a set of data-generating processes (DGPs) designed to span plausible forms of temporal dependence and treatment effect evolution in experimental panels. The DGPs differ along two dimensions, the structure of the conditional mean which determines treatment effect trajectory and the within-unit dependence in the error process which determines how information accumulates across time within a unit. The conceptual decomposition used in this econometric framework can differ either by imposing restrictions on the mean evolution or by modeling the residual covariance more flexibly.

The categorical time mean, we generate outcomes according to equation (1) with a monotone treatment path Δ_t that increases smoothly across periods so that Δ_T is approximately 0.2 standard deviations. This magnitude reflects policy relevant effect sizes commonly targeted in education and development experiments. The error term ϵ_{it} is generated under two dependence structures, independent and identically distributed and AR (1) dependence (geometrically decaying correlation within a unit over time).

The autoregressive mean DGP as defined in equation (2) generates outcomes that are persistent states and treatment may alter both level and persistence. In this DGP, the conditional mean depends on lagged outcomes and the treatment trajectory $\{\Delta_t\}$ is implied by the autoregressive recursion. This DGP is intended to represent settings in which outcomes exhibit genuine state dependence and interventions operate through persistence, which is a common empirical feature of repeated test scores, earnings and health indices.

Finally, for smooth trend mean, we generate treatment effects using the non-linear equation (3) motivated by trajectories with rapid gains followed by diminishing marginal effects. The mean structure is parameterized so that the implied endpoint effect is approximately 0.2 standard deviations.

3.3 Models Estimated

For each simulated dataset of sizes $N\ (50,\ 150) \times T(4,\ 8)$, we estimate a fixed set of competing specifications regardless of the true DGP. This 'fit all models to all DGPs' design is essential for interpreting the results as an evaluation of efficiency and robustness. The performance gains when a model's restrictions are approximately correct and losses when misspecified are revealed.

The specifications include categorical time estimated by ordinary least squares, categorical time model with AR (1) error structure (GLS), a log-trend model estimated on baseline-referenced changes, and autoregressive models estimated on within-unit transitions, including restricted variants that isolate level-only and persistence-moderated channels. Each specification produces an implied treatment effect path $\widehat{\Delta}_{1:T}$. For categorical time models, $\widehat{\Delta_t}$ is obtained directly as estimated time-treatment interaction per period. For dynamic models, $\widehat{\Delta_t}$ is computed as the model-implied function of estimated structural parameters (recursively for autoregressive specifications and closed form for the log-trend specification). All models are evaluated against the same estimand arrived through different specifications.

3.4 Performance Criteria

We evaluate finite sample performance using accuracy of the entire treatment-effect trajectory and accuracy of the endpoint effect. First, trajectory accuracy is summarized by the root mean squared error over the treatment path,

$$\mathrm{RMSE}_{\mathrm{path}} \equiv \left(\frac{1}{T}\sum_{t=1}^{T}(\widehat{\Delta}_t - \Delta_t)^2\right)^{1/2}$$

This criterion penalizes both systematic bias and sampling variance over the full trajectory and is therefore well suited to comparing flexible categorical models to parsimonious dynamic restrictions.

Second, endpoint accuracy is summarized by the absolute endpoint deviation,

$$\mathrm{AE}_T \equiv |\widehat{\Delta}_T - \Delta_T|$$

Because applied work often emphasizes a final follow-up impact, AE_T provides an interpretable complement to $\mathrm{RMSE}_{\mathrm{path}}$, distinguishing models that fit the overall path well from models that recover the endpoint particularly accurately.

For each (N, T, DGP) design cell, we report distributional summaries of these metrics across replications (e.g., medians and upper quantiles). We additionally summarize comparative performance via the frequency with which a model attains the lowest out-of-sample path error across design cells, which provides a compact representation of which restrictions dominate across plausible experimental environments.

4 Project STAR – Empirical Experiments

The empirical analysis based on Project STAR is organized into two experiments. They share a common motivation to understand finite sample performance of alternative representations in canonical short panels. Treating these designs separately is to ensure calibrated power is different from empirical stability.

4.1 Classical Power

In the first experiment, we construct a classical power analysis by using STAR to calibrate a pseudo-data generating process and then evaluating rejection probabilities under a controlled sequence of alternatives. The STAR panel provides two objects required for this, an empirical estimate of baseline treatment effect trajectory shape across the four grade level measurements ($T_{STAR} = 4$), and an empirical distribution of within-student error vectors that preserves the observed time-series dependence.

Formally, let $\widehat{s}_t$ denote an estimated treatment effect shape with $\widehat{s}_1 = 0$ and $\widehat{s}_4 \neq 0$, obtained from a baseline categorical time estimated on the balanced STAR panel ($N_{STAR} = 2668$ *unique students over* $T_{STAR} = 4$ *measurement waves*). For any targeted endpoint effect magnitude δ_4 (in standard deviation units), define a scaled treatment path, $\Delta_t(\delta_4) = \left(\frac{\delta_4}{\widehat{s}_4}\right)\widehat{s}_t,\ t = 1,\ \ldots,\ 4$, so that $\Delta_4(\delta_4) = \delta_4$ by construction. Pseudo panels of size N (50, 150, 250, 400, 600) are then generated as follows. Treatment is assigned as $D_i \sim Bernoulli(0.5)$, we then draw a residual vector $e^*_{it} = (\epsilon_{i1}, \epsilon_{i2}, \epsilon_{i3}, \epsilon_{i4})$ by sampling with replacement from the empirical distribution of STAR residual vectors at the student level. Outcomes are then constructed as $Y^*_{it} = \Delta_t(\delta_4) D_i + e^*_{it},\ t = 1, \ldots, 4$. This preserves the empirical time dependence in the error process while imposing a controlled treatment trajectory with endpoint δ_4. We then estimate each competing specification on the pseudo-panel and test the endpoint null $H_0 : \Delta_4 = 0$ using Wald standard errors. Power at (N, δ_4) is computed as the Monte Carlo rejection probability at a two-sided 5% level. Holding the error structure close to the empirical STAR environment, the experiment is designed to identify which specification delivers highest probability of detecting an endpoint effect as signal increases (δ_4 *is increased in steps*, $0.0 : 0.05 : 0.2\ SD$).

4.2 STAR Subsampling Experiment

The second experiment uses STAR directly to re-estimate treatment effects using the competing specifications through repeated subsample draws from the balanced panel ($N_{STAR} = 2668$) of size N (50, 150, 250, 400, 600). For each subsample and each model, we compute the model's endpoint estimate $\widehat{\Delta}_t$ and its associated test statistic for $H_0 :\ \Delta_4 = 0$. Aggregating across repeated subsamples yields a rejection frequency. Since the true data-generating process for STAR is unknown and is not manipulated, this rejection frequency is not a calibrated power curve. It is a measure of finite-sample stability and sensitivity. It reflects how each specification behaves under realistic data conditions and sampling variation. A specification may perform well in pseudo-experimental conditions when its restrictions are approximately satisfied but exhibit weaker stability under subsampling if the true trajectory or mean dynamics are not well captured by those restrictions.

The subsampling experiment complements the STAR-calibrated power curves by shifting the question to empirical robustness; how stable is inference when DGP is unknown? Taken together, the experiments separate efficiency gains achievable under plausible dynamic structure from the costs of imposing structure when the empirical environment only partially conforms to the restrictions.

5 Results

5.1 Recovery of Treatment Effect Trajectories

Across all design cells, autoregressive specifications outperform other specifications in terms of path accuracy. Using median RMSE over the treatment-effect trajectory as the primary performance metric, autoregressive models achieve the lowest error in many design environments (Figure 1). AR-level specifications attain the lowest median RMSE in over half of all N, T, DGP cells, with AR-parent models also performing strongly. Smooth log-trend models outperform categorical-time specifications in several settings, but do not match the consistency of autoregressive models. Table 1, summarizes RMSE over treatment effect trajectory across specifications for the Monte Carlo simulations.

Table 1: Monte Carlo performance across dynamic specifications

Model	Median RMSE (Path)	90th pct. RMSE (Path)	Median $\lvert\widehat{\Delta}_T - \Delta_T\rvert$	90th pct. $\lvert\widehat{\Delta}_T - \Delta_T\rvert$
TE OLS	0.198	0.341	0.176	0.458
TE GLS AR(1)	0.198	0.341	0.176	0.458
Log OLS	0.164	0.293	0.153	0.401
AR Level	0.132	0.262	0.141	0.387
AR Parent	0.139	0.289	0.168	0.457

Relative to the categorical-time OLS benchmark, dynamic mean restrictions reduce median path RMSE by approximately 30–40 percent in environments with outcome persistence or smooth treatment evolution. These reductions are comparable to the improvement one would obtain from a substantial increase in sample size, holding the specification fixed. In contrast, modeling serial correlation in the error term alone via GLS applied to a categorical time mean delivers negligible improvements over OLS, indicating that efficiency gains arise primarily from restricting the mean structure, not from refining the covariance estimator.

Using the median absolute deviation of the endpoint effect $|\widehat{\Delta}_T - \Delta_T|$ as a metric, dynamic specifications again outperform categorical-time models in most environments. Autoregressive models achieve the lowest endpoint error in roughly half of the design cells, with log-trend models performing well when the true

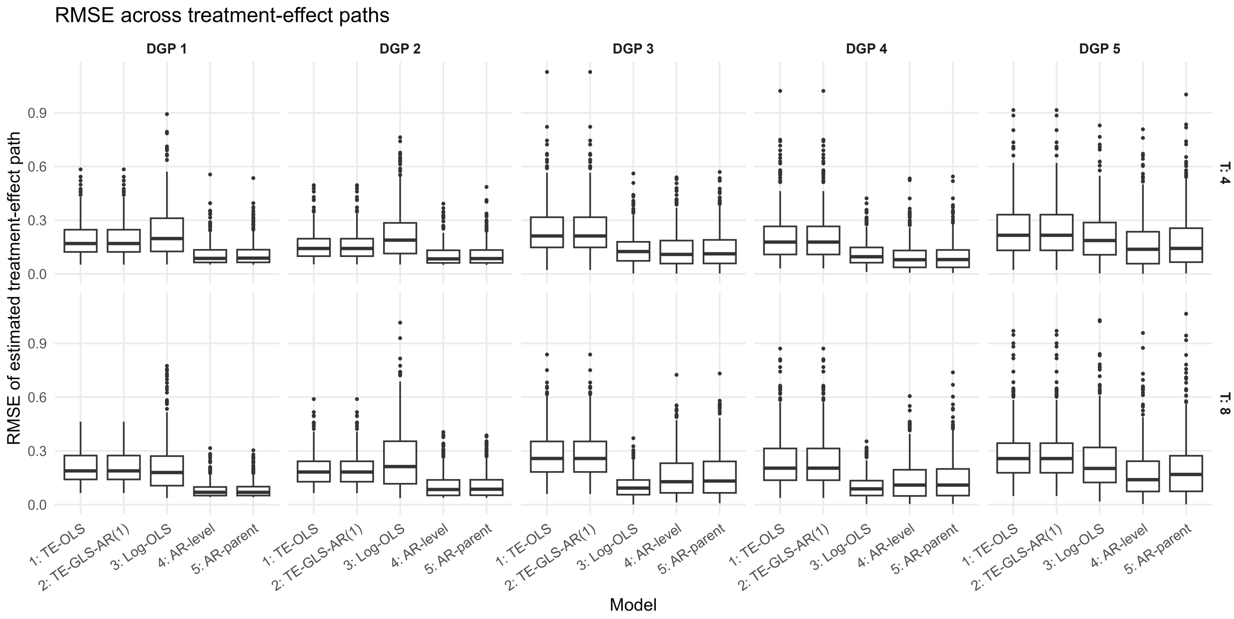

Figure 1: RMSE Boxplot for treatment effect trajectory across specifications and DGPs

treatment path is smooth. Categorical time models rarely minimize endpoint error, despite their flexibility, reflecting the variance costs of estimating multiple period-specific parameters in short panels.

Performance patterns are stable across sample sizes and panel lengths. Increasing Nreduces RMSE for all specifications, as expected, but does not alter the relative ranking of models. Similarly, increasing T improves overall precision but amplifies the advantage of dynamic restrictions by allowing models to exploit temporal dependence more effectively. Importantly, no single specification dominates uniformly, when the true treatment path departs sharply from the imposed dynamic structure, categorical time models become competitive. This reflects a bias–variance trade-off rather than estimator failure. Because all specifications are evaluated using Wald tests based on their own variance estimators, rejection probabilities incorporate the joint effects of finite-sample bias and standard-error accuracy. In short panels ($T = 4, 8$), standard bias corrections are unstable and rarely used in practice, and any residual bias would manifest as size or power distortions, which we do not observe.

5.2 Power for Endpoint Inference

Figure 2 and 3 report rejection probabilities for tests of the null hypothesis $H_0 : \Delta_T = 0$ under STAR-calibrated pseudo data, where the treatment effect trajectory is rescaled to deliver a known endpoint effect Δ_4 while preserving the empirical error structure of the STAR experiment. This design isolates the power consequences of alternative mean specifications under empirically realistic noise.

Two findings emerge clearly. First, autoregressive mean specifications dominate in terms of power for endpoint inference across all sample sizes and effect magnitudes. Both AR-level and AR-parent models exhibit substantially higher rejection rates than categorical-time specifications at moderate signal strengths, with differences that widen as N increases. For example, at $\Delta_4 = 0.15$ and $N = 400$, autoregressive models achieve rejection rates exceeding 60 percent, compared to approximately 40 percent for log-trend models and roughly 35 percent for categorical-time OLS and GLS.

Second, modeling serial correlation in the error term alone does not materially improve power. GLS estimation with AR (1) disturbances closely tracks OLS with categorical time effects across all designs, reinforcing the conclusion from the Monte Carlo RMSE results that efficiency gains arise from restricting the conditional mean rather than from covariance modeling.

The ordering of specifications is stable across sample sizes and converges monotonically as N grows. At small sample sizes, power is uniformly low for all models, but autoregressive specifications maintain

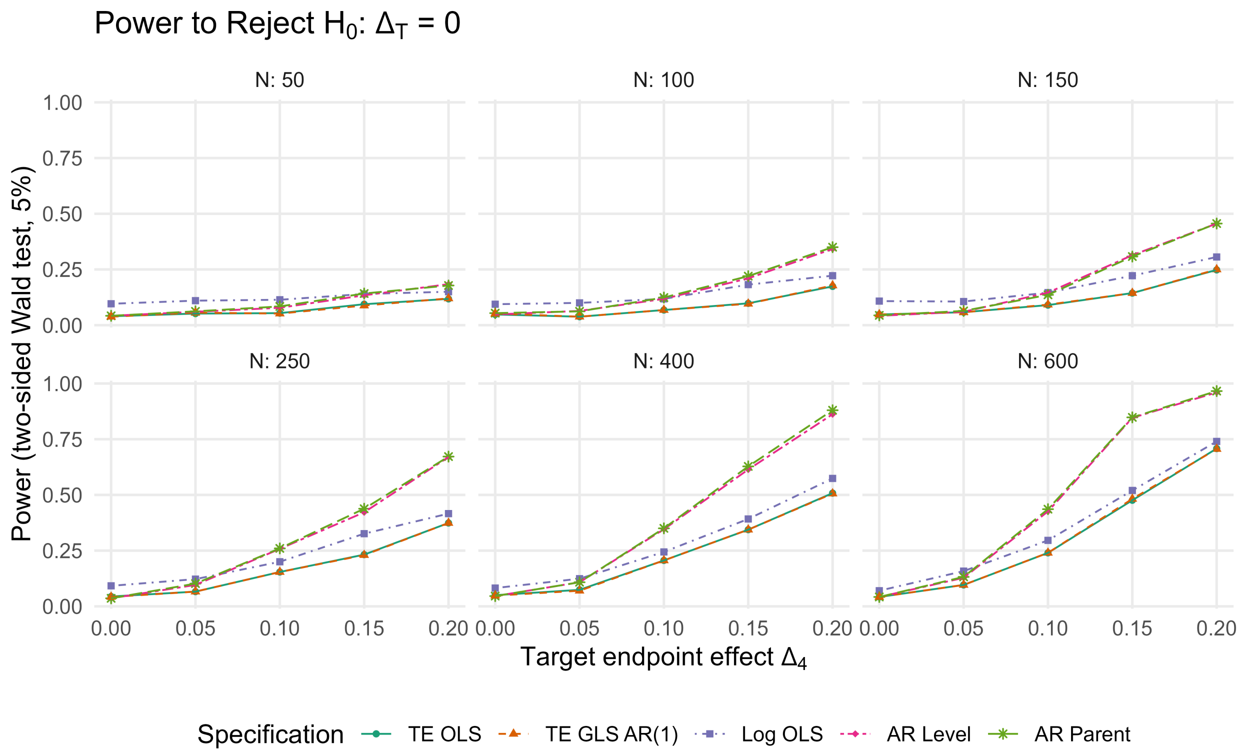

Figure 2: Power curves for endpoint inference over various sample sizes and signal strengths calibrated from estimates using the STAR dataset.

a consistent advantage even in these settings. As N increases, the gap between dynamic and categorical specifications becomes more pronounced, indicating that dynamic mean restrictions exploit information that categorical time models fail to aggregate efficiently.

5.3 Empirical Subsampling in STAR

We complement the STAR-calibrated power analysis with two empirical subsampling exercises conducted directly on the original STAR data. Unlike the pseudo data experiments above, these exercises do not impose a known alternative and therefore do not deliver power in the formal sense. Instead, they assess the finite-sample stability of inference under repeated random sampling when the true treatment path is unknown.

Figure 5 reports rejection frequencies for tests based on the average treatment effect over periods 2–4, while Figure 4 reports rejection frequencies for tests of the endpoint effect. In both cases, subsamples of size $N \in \{100, 150, 250, 400, 600\}$ are drawn repeatedly from the original balanced STAR panel ($N_{STAR} = 2668$, $T_{STAR} = 4$), and inference is conducted using the same specifications considered above.

Two patterns stand out. First, categorical time specifications exhibit higher rejection rates than dynamic mean models, particularly when inference targets the average treatment effect over the trajectory. For example, when testing the mean effect over periods 2–4, categorical time OLS rejects at rates exceeding 70 percent in larger subsamples, while autoregressive models reject at rates below 20 percent. A similar, though attenuated, pattern appears for endpoint inference.

Second, the relative ordering across specifications differs sharply from that observed in the STAR-calibrated power experiments. In contrast to the pseudo data results where autoregressive models dominate uniformly dynamic specifications appear conservative in the empirical subsampling exercises, while categorical time models are substantially more responsive to sampling variation in the observed data.

The subsampling exercises do not condition on a fixed alternative and therefore conflate statistical power

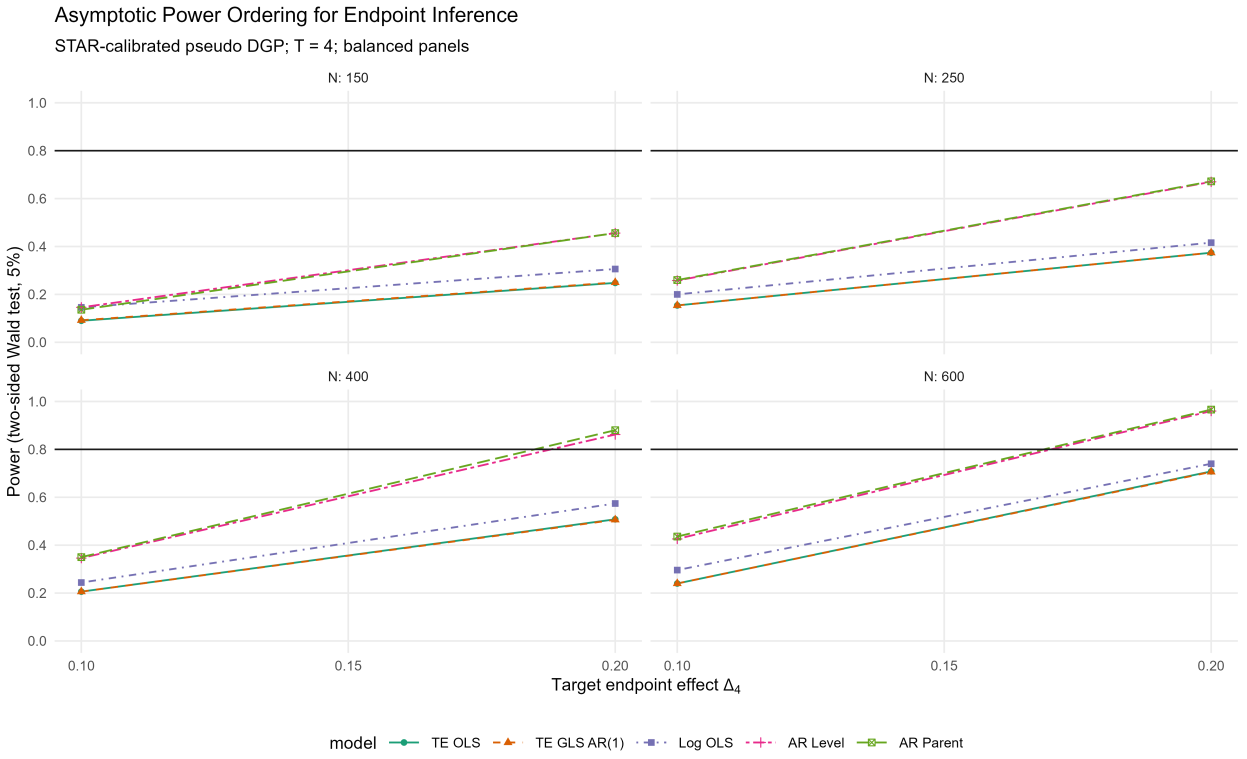

Figure 3: Power ordering for inference when treatment effects are 0.1, 0.15 and 0.2 SD at endpoint. AR models dominate other specifications.

with sensitivity to unmodeled features of the realized treatment path. Categorical time models remain agnostic about the temporal structure of treatment effects and can accommodate irregular or non-monotonic patterns that are present in the STAR data but suppressed by dynamic mean restrictions. As a result, they exhibit higher empirical rejection rates when the signal shape is uncertain, even though they are less efficient when the underlying trajectory is smooth and known up to scale.

Dynamic mean restrictions deliver substantial efficiency gains and higher true power when the treatment path is well approximated by a low-dimensional structure, as demonstrated in the STAR-calibrated pseudo data. However, categorical time specifications offer greater robustness to unknown or irregular signal shapes, which manifests as higher rejection rates under empirical subsampling. These findings underscore that differences across specifications reflect distinct bias–variance trade-offs rather than violations of asymptotic theory.

6 Discussion and Conclusion

This paper examines how alternative specifications for treatment effect dynamics perform in short experimental panels when inference targets both entire trajectories and endpoint effects. We study a constrained set of widely used specifications under controlled Monte Carlo designs and empirically grounded experiments calibrated to the STAR data.

First, imposing parsimonious dynamic structure on the conditional mean delivers substantial efficiency gains when treatment effects evolve smoothly over time. Across Monte Carlo designs and STAR-calibrated pseudo–data, autoregressive specifications dominate categorical-time models in terms of path recovery and statistical power for endpoint inference. These gains arise from restricting the mean structure rather than from modeling serial correlation in the error term, a distinction that is often blurred in applied work.

Second, the advantage of dynamic specifications is not universal. Empirical subsampling in STAR where

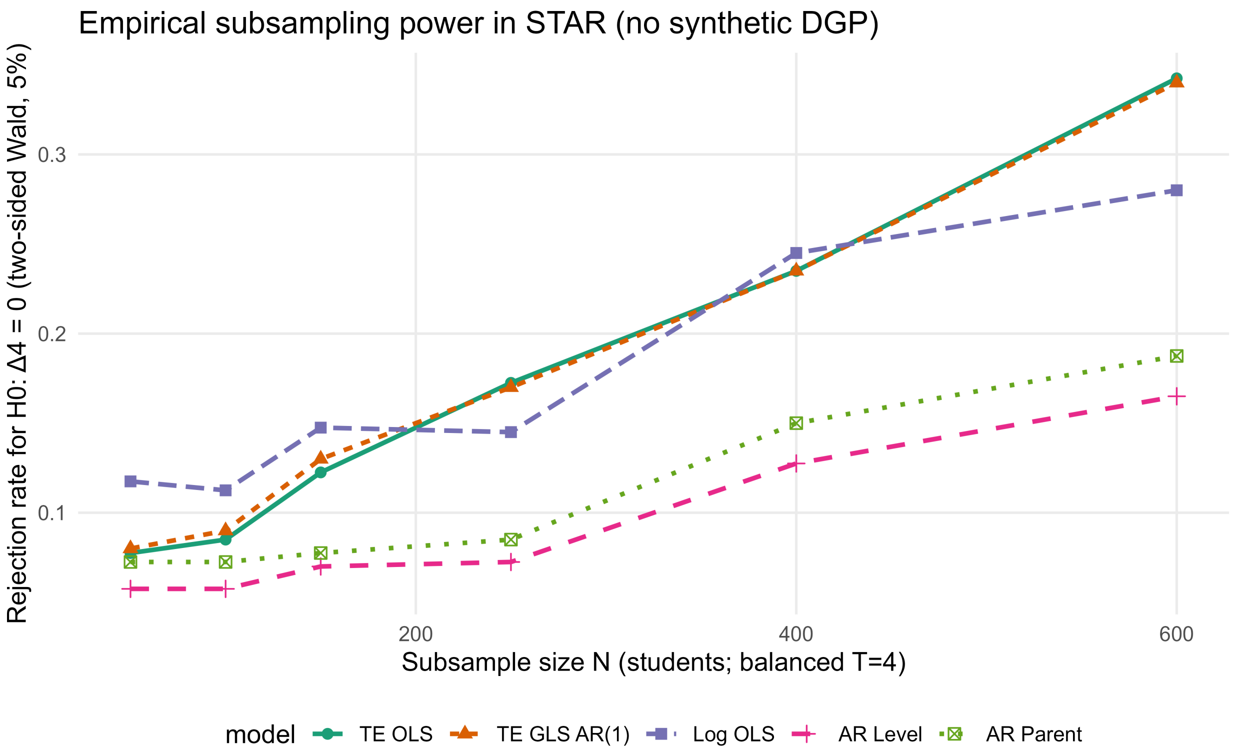

Figure 4: Empirical robustness of treatment effect recovery in random subsamples drawn from the STAR balanced panel.

the true treatment path is unknown and potentially irregular reveals that categorical-time models exhibit greater sensitivity to realized variation in the data. This reflects a familiar bias–variance trade-off, dynamic restrictions improve efficiency when approximately correct but suppress local features of the signal when misspecified. Importantly, this divergence highlights the distinct inferential objectives captured by power under a known alternative versus stability under unknown signal shape.

Third, these results clarify the practical implications of specification choice in short panels. When the goal is precise estimation or powerful testing of a smooth treatment trajectory, dynamic mean restrictions are highly effective. When the objective is robustness to unknown or non-smooth dynamics, categorical-time specifications may be preferable despite their higher variance. No single specification dominates across all environments, and the appropriate choice depends on the inferential target and the plausibility of dynamic structure.

The broader contribution of this paper is to disentangle these trade-offs in a setting that mirrors canonical experimental panels. By combining Monte Carlo designs with STAR-calibrated pseudo data and empirical subsampling, we provide a unified framework for understanding how specification choices map into efficiency, power, and robustness. While our analysis is intentionally focused, the results highlight for short panels, restrictions on the treatment-effect path are first-order determinants of finite-sample performance and should be chosen with explicit reference to the underlying empirical context.

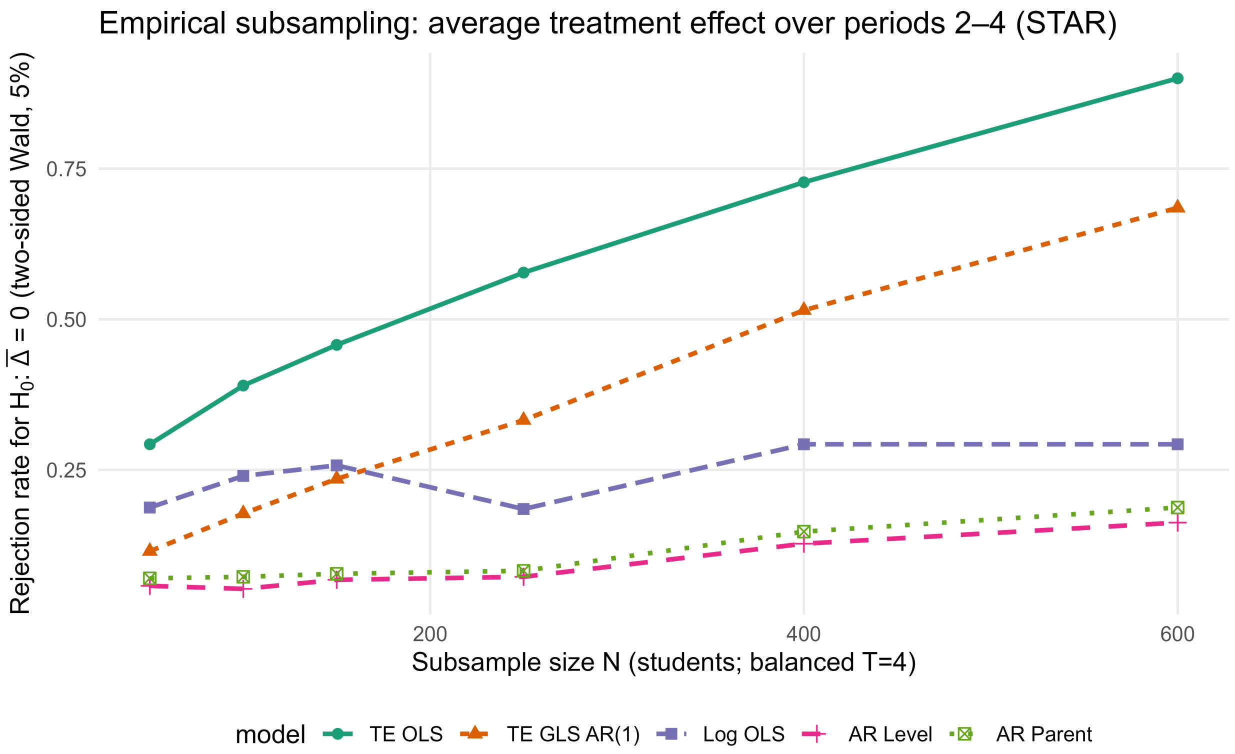

Figure 5: Empirical robustness of average trajectory treatment effect recovery in random subsamples drawn from the STAR balanced panel.

References

Andrabi, Tahir, Jishnu Das, Asim Ijaz Khwaja, and Tristan Zajonc. 2011. "Do value-added estimates add value? Accounting for learning dynamics." *American Economic Journal: Applied Economics* 3 (3):29–54.

Angrist, Joshua D and Jörn-Steffen Pischke. 2009. *Mostly harmless econometrics: An empiricist's companion*. Princeton University Press.

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

Athey, Susan and Guido W Imbens. 2017. "The econometrics of randomized experiments." In *Handbook of economic field experiments*, vol. 1. North-Holland, 73–140.

Banerjee, Abhijit V and Esther Duflo. 2009. "The experimental approach to development economics." *Annual Review of Economics* 1 (1):151–178.

Ben-Porath, Yoram. 1967. "The production of human capital and the life cycle of earnings." *Journal of Political Economy* 75 (4, Part 1):352–365.

Bloom, Nicholas, Benn Eifert, Aprajit Mahajan, David McKenzie, and John Roberts. 2013. "Does management matter? Evidence from India." *Quarterly Journal of Economics* 128 (1):1–51.

Cunha, Flavio and James Heckman. 2007. "The technology of skill formation." *American Economic Review* 97 (2):31–47.

Duflo, Esther, Pascaline Dupas, and Michael Kremer. 2011. "Peer effects, teacher incentives, and the impact of tracking: Evidence from a randomized evaluation in Kenya." *American Economic Review* 101 (5):1739–1774.

Duflo, Esther, Rachel Glennerster, and Michael Kremer. 2007. "Using randomization in development economics research: A toolkit." In *Handbook of Development Economics*, vol. 4. Elsevier, 3895–3962.

Imbens, Guido W and Donald B Rubin. 2015. *Causal inference for statistics, social, and biomedical sciences: An introduction*. Cambridge University Press.

Krueger, Alan B. 1999. "Experimental estimates of education production functions." *The Quarterly Journal of Economics* 114 (2):497–532.

McKenzie, David. 2012. "Beyond baseline and follow-up: The case for more T in experiments." *Journal of Development Economics* 99 (2):210–221.

Muralidharan, Karthik and Venkatesh Sundararaman. 2011. "Teacher performance pay: Experimental evidence from India." *Journal of Political Economy* 119 (1):39–77.

Verbeke, Geert and Geert Molenberghs. 2000. *Linear mixed models for longitudinal data*. Springer.

Wooldridge, Jeffrey M. 2010. *Econometric analysis of cross section and panel data*. MIT Press, 2nd ed.