

Beyond the Illusion of Power: Calibrating Quasi-Experiments in Observational IS

Spandan Ghose Chowdhury
Georgia Institute of Technology, GA, USA
spandan_gc@gatech.edu

Abstract

Information systems (IS) researchers increasingly use quasi-experimental methods such as difference-in-differences (DiD) and instrumental variables (IV) to recover causal effects from observational panel data. Power calculations that justify these designs assume i.i.d. errors, but the deeper problem is what even a cluster-robust calculator cannot see. We report a Monte Carlo study over 9,837 parameter conditions (≈ 9.8 million datasets) and decompose the planned-versus-achieved power gap. The serial-correlation component is recoverable by an AR(1)-aware calculator when ρ is known, and partially when ρ must be estimated from short pre-periods; but panel attrition, staggered-adoption bias, and parallel-trends pretesting are captured by no closed-form formula; exogenous attrition alone costs ≈ 8 – 11 percentage points at the few-hundred-to-thousand sample sizes IS studies use. Treatment-correlated, outcome-dependent attrition instead induces bias, not just power loss. For IV, holding first-stage F fixed, larger N neither raises power nor curbs exclusion bias—though with a fixed instrument more data does sharpen the first stage—so identification rests on instrument strength, not sample size. We provide a unified DiD+IV benchmark calibrated to IS effect sizes, a companion lookup tool, and design guidance for observational causal work in IS: <https://github.com/spandan1305/is-power-lookup>

Keywords: Power analysis, difference-in-differences, instrumental variables, Monte Carlo simulation, research methodology.

1. Introduction

Causal designs now occupy a central position in Information Systems (IS) empirical research (Tafti & Shmueli, 2025), particularly through difference-in-differences (DiD) and, where endogeneity concerns are stronger, instrumental variables (IV). Yet

ex ante design justification often remains anchored in power calculations derived from i.i.d. settings—poorly aligned with IS panels, where outcomes are serially correlated, attrition unbalances the panel, treatment timing is staggered, and effects are heterogeneous.

Our aim is not to re-establish that i.i.d. calculators are optimistic—that much is well known. We ask instead how much of the planned-versus-achieved power gap survives a competent, cluster-robust, AR(1)-aware calculator, and which frictions remain invisible to any closed-form formula. First, once the recoverable serial-correlation component is removed, how much power loss persists from panel attrition and staggered timing? Second, what is the inferential consequence of conditioning DiD analysis on pretrend-test non-rejection (Roth, 2022)? Third, how do weak-instrument and exclusion-restriction failures in IV differ as sample size increases?

Several methodological studies provide partial answers. Bertrand et al. (2004) document serial-correlation distortions in DiD inference; Roth (2022) formalizes the conditional-inference problem in pretrend testing; and Goodman-Bacon (2021), de Chaisemartin and D’Haultfœuille (2020), and Sun and Abraham (2021) characterize TWFE bias under staggered timing. What is missing is a *joint* quantification that treats panel frictions, estimator choice, conditional inference from pretesting, and IV identification failures as components of a single design problem, evaluated over an IS-calibrated parameter space with consistent metrics.

The individual mechanisms we study are established in econometrics—serial-correlation distortion (Bertrand et al., 2004), pretest selection (Roth, 2022), TWFE bias under staggered timing (Goodman-Bacon, 2021), and the weak-versus-exclusion IV trade-off. Our contribution is to quantify their *joint* magnitude in IS-typical panels and package that evidence for practice. First, we provide a unified Monte Carlo benchmark covering both DiD and IV under one IS-realistic data-generating process (DGP), reporting power, bias,

root-mean-squared error, and confidence interval coverage over 9,837 parameter conditions. Second, we calibrate the effect-size grid to documented IS effect magnitudes (Table 1), validate the simulation against canonical open datasets, and release a companion lookup tool over the full grid; the tool matches categorical inputs exactly, *snaps* numeric inputs to the nearest simulated grid value, and reports the matched design so users can judge how far the returned cell sits from their target. Third, we document an asymmetry with direct bearing on IV practice: weak-instrument bias attenuates with N , while exclusion bias is not cured by sample size, so identification rests on instrument strength rather than N . We translate these results into conditional design guidance rather than a universal checklist.

2. Background

In contemporary IS empirical work, two-way fixed effects (TWFE) DiD remains the default framework for policy rollouts, platform interventions, and staged feature launches. In the canonical model $Y_{it} = \alpha_i + \lambda_t + \delta D_{it} + \varepsilon_{it}$, identification of δ rests on parallel trends, typically assessed with event-study plots and joint tests on pre-treatment coefficients—diagnostics that are themselves finite-sample objects and noisy in observational panels.

Once timing is staggered and effects are heterogeneous, TWFE no longer has a straightforward ATT interpretation: it combines multiple 2x2 comparisons, including already-treated vs. not-yet-treated contrasts that can receive negative weight and induce bias (de Chaisemartin & D’Haultfœuille, 2020; Goodman-Bacon, 2021). We therefore benchmark TWFE against alternatives that re-center identification at the group-time level, notably Sun and Abraham (2021) and Callaway and Sant’Anna (2021); for a recent systematic comparison see Ulloa-Pérez et al. (2025).

Pretrend testing introduces a second inferential issue: Roth (2022) shows that conditioning interpretation on jointly insignificant pre-treatment coefficients induces selection bias, because samples that pass are those in which sampling noise masks true trend differences. Conditional error rates can then exceed unconditional ones, especially when pre-periods are short and noisy — conditions common in IS panels.

IV designs face a related dual-identification challenge: relevance and exclusion. In the just-identified case, $\hat{\beta}_{IV} = (Z'X)^{-1}Z'Y$, relevance is often summarized through first-stage diagnostics such as Stock and Yogo (2005), while exclusion requires

that Z affect Y only through X . IS instruments are often geographic, temporal, or peer-exposure based and plausibly vulnerable to contextual spillovers (Semadeni et al., 2014), making first-stage strength and exclusion plausibility distinct empirical concerns.

Because serially dependent, attrition-thinned panels reduce effective information relative to nominal row counts (Bertrand et al., 2004), our simulation treats serial dependence, attrition, and clustering as first-order design inputs rather than implementation details.

Beyond Sun–Abraham and Callaway–Sant’Anna, the DiD literature has expanded to include imputation-based approaches (Borusyak et al., 2024), two-stage DiD (Gardner, 2022), panel counterfactual methods rooted in matrix completion and synthetic-control logic (Abadie, 2021; Athey et al., 2021), and causal machine learning (Chernozhukov et al., 2018; von Zahn et al., 2026; Wager & Athey, 2018). We do not implement these classes; the present grid is a benchmark against which they can be compared in follow-up work.

Finally, we extend simulation-based design diagnostics in management and IS (Semadeni et al., 2014) by jointly analyzing DiD and IV on one panel grid that models serial correlation and attrition and quantifies both estimator performance and inference calibration.

3. Simulation Design

3.1. Data-Generating Process

The DiD DGP generates $Y_{it} = \alpha_i + \lambda_t + \delta D_{it} + \varepsilon_{it}$ with $\varepsilon_{it} = \rho \varepsilon_{i,t-1} + u_{it}$, $u_{it} \sim \mathcal{N}(0, \sigma^2(1 - \rho^2))$, $\alpha_i \sim \mathcal{N}(0, 1)$, and $\lambda_t = 0.5t/T$. To span the design space most frequently encountered in IS panels, we vary nine inputs: sample size (`n_units` $\in \{50, 100, 200, 500, 1000, 5000\}$), panel length (`n_periods` $\in \{6, 10, 16, 24\}$), pre-period share (`pre_frac` $\in \{0.33, 0.50\}$), standardized effect magnitude (`effect_size` $\in \{0.1, 0.2, 0.5\}$ SD), serial persistence (`autocorr` $\rho \in \{0, 0.3, 0.5, 0.7\}$), adoption timing (`staggered` $\in \{\text{False}, \text{True}\}$), treatment-effect heterogeneity (`het_effects`, with early cohorts receiving 2δ and later cohorts δ), exogenous attrition (`churn_rate` $\in \{0, 0.3\}$), and covariance structure for inference (`cluster_type` $\in \{\text{unit}, \text{two-way}\}$). After removing infeasible cells, this yields 6,912 DiD conditions. (`het_effects` is defined only under staggered adoption, which removes one quarter of the 9,216 nominal cells.) We additionally run a 45-cell pretrend sub-study and a 2,880-cell IV grid that varies first-stage strength F , endogeneity ρ_{vu} ,

Effect (SD)	IS example
0.05–0.10	UX A/B test on conversion (button, copy)
0.10–0.20	Recommender change on click-through
0.20–0.35	Platform policy change (pricing, matching)
0.35–0.50	Major ERP/system implementation on firm productivity
> 0.50	Rarely observed in credible causal designs

Table 1. IS effect-size anchors: illustrative calibration points from a structured review plus domain judgment (see text), not meta-analytic estimates.

sample size, instrument count, and a direct $Z \rightarrow Y$ channel. Crucially, the grid fixes the *target* first-stage F rather than the first-stage coefficient, so the instrument strength is set to $\pi = \sqrt{F/(N+F)}$, which declines as $N^{-1/2}$; this normalization drives the IV scaling result below.

3.2. Estimators

For DiD, we estimate (i) TWFE under unit-clustered or two-way clustered covariance, (ii) the interaction-weighted event-study estimator of Sun and Abraham (2021), and (iii) cohort-specific ATT aggregation following Callaway and Sant’Anna (2021). Pretrend diagnostics use a joint Wald test over all pre-treatment event-time coefficients in a two-way FE event-study regression. For IV, we estimate just-identified and over-identified (three-instrument) 2SLS with heteroskedasticity-robust standard errors.

To calibrate the standardized effect-size grid to effects plausibly encountered in IS research, Table 1 maps each range to a representative study type. The anchors come from a structured (though not systematic) review of published IS quasi-experiments and large-scale digital experiments, combined with domain judgment: massive online field experiments report standardized effects at the low end (Aral & Walker, 2011), platform- and policy-scale interventions on short persistent panels fall in the middle ranges (Aziz et al., 2023; Huang et al., 2017; Seamans & Zhu, 2014), and effects above 0.5 SD are rare in credible causal designs (Gelman & Carlin, 2014). Researchers with better priors should query the lookup tool at those values instead.

3.3. Performance Metrics

For each design cell, we run $N_{\text{sim}} = 1,000$ independent Monte Carlo replications (500 in the dedicated pretrend exercise) and report empirical power

Test	Target	Achieved	Tol.
High-power recovery	≥ 0.90	0.94	± 0.03
Type I ($\beta = 0$)	≈ 0.05	0.047	± 0.01
95% CI coverage	≈ 0.95	0.955	± 0.01
Churn symmetry	$\text{diff} < 0.02$	0.008	—
AR(1) ρ recovery	$ \hat{\rho} - \rho_0 < 0.05$	0.023	—

Table 2. Pre-flight validation results against five implementation-fidelity targets. Each test uses 1,000 replications.

($\hat{P}(\text{reject } H_0 : \beta = 0)$), bias ($\mathbb{E}[\hat{\beta} - \beta]$), RMSE, and 95% confidence-interval coverage. In the pretrend module we additionally compute the unconditional versus conditional rejection gap emphasized by Roth (2022). The full experiment covers 9,837 parameter conditions (6,912 DiD + 45 pretrend + 2,880 IV cells) and approximately 9.8 million simulated datasets. Three supplementary sub-studies—a 48-cell attrition grid (Section 4.2), a 500-pilot-per-cell estimated- ρ analysis (Table 4), and a ten-cell reference-estimator validation (Section 4.1)—lie outside this count and are described where used.

3.4. Validation

Before launching the full grid, we validate implementation fidelity using five unit tests summarized in Table 2. All pass within their pre-registered tolerances.

4. DiD Results

4.1. Sample-Size Scaling Under Serial Correlation and Staggered Adoption

Two features of panel data drive a wedge between the power classical formulas predict and the power TWFE delivers: within-unit serial correlation, and the interaction between staggered timing and heterogeneous treatment effects. Panels (a) and (b) of Figure 1 summarize both.

Panel (a) reports TWFE power over N and residual autocorrelation ρ , holding `effect_size` = 0.2 SD, a balanced panel, non-staggered timing, and a 50% pre-period. At $\rho = 0$ — the i.i.d. benchmark underlying most closed-form power calculators — $N = 500$ reaches 92% power, consistent with the analytical formula. Raising ρ to 0.5, a value representative of weekly platform outcome series, reduces power at the same N to 76%; at $\rho = 0.7$ it falls to 70%, and $N = 200$ delivers only 34%. The decay is steep and non-linear, with the largest marginal losses between $\rho = 0$ and $\rho = 0.5$.

Table 3 summarizes the nominal–achieved gap

across representative IS panel conditions. In the small-to-moderate- N regime where IS studies most commonly operate ($N \in [100, 1,000]$), the i.i.d.-versus-achieved gap ranges from ≈ 6 pp (at $N = 1,000$) to ≈ 33 pp (at $N = 500$ with attrition); the next paragraph decomposes how much of this is recoverable.

A natural objection is that this gap is measured against a *naive* i.i.d. calculator, whereas competent applied work already clusters and anticipates serial correlation. We therefore decompose it against a stronger baseline: an AR(1)-aware, cluster-robust calculator, operationalized as the same simulation under the cell's ρ but a balanced panel with correct unit clustering. This baseline is an *oracle*—it uses the true ρ and DGP—so the recoverability claims below are statements about that best-case upper bound. For balanced designs without attrition the baseline recovers essentially the entire gap: the serial-correlation loss is recoverable *in principle* by modeling the dependence. What survives is attrition: at `effect_size = 0.2 SD` and $\rho = 0.5$, an AR(1)-aware design still overstates achievable power by roughly 9 pp at $N = 200$ and 11 pp at $N = 500$ once 30% of unit-periods churn out (Table 3; Monte Carlo SE ≈ 2 pp). The staggered-timing bias and pretest selection documented below are likewise invisible to any closed-form formula. The headline: serial correlation is recoverable; attrition, staggered timing, and pretesting are not.

How much of this recoverability survives when ρ must be *estimated* rather than known? Table 4 quantifies the question for the same design cells. With the five pre-periods these designs afford, within-unit demeaning and the short estimation window jointly depress the per-unit lag-1 correlation well below its true value: the naive plug-in yields $\hat{\rho} \approx -0.12$ when the true ρ is 0.3, so a naive plug-in planner behaves almost like the i.i.d. calculator it was meant to replace, and the residual gap grows from 0–14 pp (oracle) to 7–35 pp. Adding the first-order Nickell-type correction $1/(T_{\text{pre}} - 1)$ moves $\hat{\rho}$ to ≈ 0.13 —about half of the way back—and correspondingly recovers roughly half of the shortfall (4–26 pp); the remainder reflects small-sample bias in the within-unit correlation itself, which the first-order term does not capture. Realizing the recovery on short pre-periods thus requires bias-aware estimation of ρ (or longer pre-periods); plugging a raw demeaned-residual autocorrelation into a calculator—or into our lookup tool's `autocorr` input—systematically overstates achievable power.

Panel (b) isolates the cost of staggered adoption with cohort-heterogeneous effects (early cohorts receive 2δ , later cohorts δ). Measured against the base effect

N	Condition	i.i.d.	AR(1)	Ach.	Resid.
100	$\rho=0.3$, no churn	0.29	0.20	0.20	0
200	$\rho=0.5$, churn 0.3	0.58	0.37	0.28	9
500	$\rho=0.5$, churn 0.3	0.92	0.70	0.59	11
500	$\rho=0.3$, no churn	0.94	0.80	0.80	0
1,000	$\rho=0.7$, no churn	1.00	0.94	0.94	0

Resid. = AR(1)-aware planned power minus achieved power, in percentage points: the part of the gap that survives a correctly specified AR(1)/cluster-robust calculation.

Table 3. TWFE power across representative IS panel conditions (`effect_size = 0.2 SD`, $T = 10$, 50% pre-period, unit clustering) under a naive i.i.d. calculator, an AR(1)-aware cluster-robust calculator (same ρ , balanced panel), and achieved (simulation). Serial correlation is recoverable (residual ≈ 0 absent attrition); the surviving loss is attrition. Monte Carlo SE ≈ 1.5 – 2 pp. Figure 1(a) averages over $T \in \{6, 10, 16, 24\}$, so its $\rho = 0.5$ value (0.76) differs from this table's $T = 10$ AR(1) column (0.70).

N	Condition	Oracle	Plug-in	Corr.	Ach.	Resid.
100	$\rho=0.3$, no churn	0.23	0.33	0.28	0.23	0 / 10 / 5
200	$\rho=0.5$, churn 0.3	0.35	0.56	0.46	0.27	8 / 29 / 19
500	$\rho=0.5$, churn 0.3	0.72	0.93	0.84	0.58	14 / 35 / 26
500	$\rho=0.3$, no churn	0.79	0.93	0.89	0.80	0 / 13 / 9
1,000	$\rho=0.7$, no churn	0.92	1.00	0.97	0.93	0 / 7 / 4

Resid. = planned minus achieved power (pp) under each baseline, reported as oracle / plug-in / corrected.

Table 4. Planned power when ρ is estimated from five pre-periods (500 pilot draws per cell; same cells as Table 3; Monte Carlo SE ≈ 2 pp). Oracle plans at the true ρ ; Plug-in at the raw demeaned-residual $\hat{\rho}$; Corr. adds the first-order $1/(T_{\text{pre}} - 1)$ demeaning-bias correction. Oracle and Ach. columns are re-simulated for this exercise and differ from Table 3 by Monte Carlo error.

δ , TWFE exhibits a persistent ≈ 0.042 SD deviation that does not attenuate with N . Only part of this is the Goodman-Bacon (2021) distortion from negatively weighted “forbidden” comparisons; the remainder is a benchmark mismatch that the reference-implementation check below isolates. Measured against the same base effect, our implementation of Callaway and Sant’Anna (2021) shows a smaller deviation (≈ 0.033 SD) and our implementation of Sun and Abraham (2021) a larger one (≈ 0.054 SD), stable across sample sizes (Monte Carlo SE ≈ 0.004 SD).

We caution strongly against reading this ordering as an estimator ranking; two robustness checks show it is an artifact. First, re-analyzing the full grid, the apparent SA excess declines monotonically with pre-period length (from 0.080 SD at two pre-periods to 0.031 SD at twelve) but persists everywhere—pointing to something systematic. Second, we validated our simplified implementations on a ten-cell subset (200 replications per cell) against `fixest::sunab` (via its `pyfixest` port) and the differences

implementation of Callaway–Sant’Anna. Under this DGP the two reference estimators are *numerically identical* and approximately unbiased for the realized cohort-share ATT—which under heterogeneous effects is 0.23–0.30 SD, not the base $\delta = 0.2$. Against that estimand, TWFE shows a modest ≈ -0.02 SD attenuation (the Goodman–Bacon distortion proper), simplified SA tracks the reference, and simplified CS *understates* the ATT by ≈ 0.03 SD via its equal-weight aggregation. The panel-(b) ordering thus stacks a benchmark mismatch (deviations from δ rather than the estimand) on simplified aggregation weights: with reference implementations and the correct estimand, the cohort-aware estimators are unbiased and effectively tied, while TWFE retains a real but modest distortion. We retain the simplified implementations only for the power and coverage columns of the main grid, where they track the references at short panels (per-replication correlations 0.70–0.83 at $T \leq 10$, falling to ≈ 0.47 in the two $T = 24$ validation cells; $T = 16$ was not validated).

4.2. Panel Attrition and Effective Sample Size

Panel (c) of Figure 1 isolates the effect of exogenous attrition on TWFE power at $\rho = 0.3$ (effect_size = 0.2 SD, non-staggered, averaged over T and pre-period share). Moving from churn_rate = 0 to 0.3 — 30% of unit-period observations missing, drawn exogenously from both arms — reduces power from 0.822 to 0.742 at $N = 500$, and from 0.491 to 0.411 at $N = 200$. In Table 3’s $\rho = 0.5$, $T = 10$ cells the contrast costs 9–11 pp. The loss exceeds what a naive “effective sample size” correction predicts: surviving observations remain within-unit correlated, so usable signal contracts faster than the observation count. The penalty is most pronounced at the sample sizes IS studies most commonly operate in (roughly $N \in [200, 1000]$) and shrinks as power saturates.

The main grid draws attrition identically from both arms, and the 8–11 pp penalty is specific to that mechanism. Because the more common real-world failure mode is attrition that differs by treatment status—adopters who churn after a disappointing experience, controls who exit a comparison platform—we ran a dedicated 48-cell sub-grid ($N \in \{200, 500\}$, $\rho \in \{0.3, 0.5\}$, staggered and non-staggered, 1,000 replications per cell) varying the mechanism while holding total attrition near 30%. Two findings qualify the headline. First, attrition that differs by treatment status but remains *exogenous in timing* (40% of treated units versus 20% of controls, or the mirror image) stays

essentially unbiased ($|\text{bias}| < 0.01$ SD) and costs only a few additional percentage points of power ($0.72 \rightarrow 0.66$ at $N = 500$, $\rho = 0.3$): it remains a power problem, and the lookup tool’s single churn_rate input approximates it serviceably. Second, *outcome-dependent* attrition—elevated exit hazard after adverse outcome shocks—changes the nature of the problem. Symmetric across arms, the selection cancels in the DiD contrast; concentrated in the treated arm, the estimate acquires an attenuation bias of ≈ -0.05 SD, coverage drops to 0.92–0.94, and power falls to 0.44 at $N = 500$, $\rho = 0.3$. No power calculation repairs this case: it is a bias problem requiring design or estimation remedies (tracking exiters, selection modeling, bounds); the cohort-specific Callaway and Sant’Anna (2021) estimator is largely robust to it in our staggered cells because its short cohort-window comparisons limit exposure to post-exit selection. Attrition claims elsewhere in the paper—including the abstract’s 8–11 pp figure—refer to the exogenous case; outcome-dependent attrition is strictly worse and differently shaped.

4.3. Parallel-Trends Pretesting and Conditional Inference

We next examine pretesting for parallel trends. Figure 2(a) reports the rejection rate of the joint Wald test on pre-period event-time coefficients: under a true null it rejects at 0.028–0.037 across all sample sizes (approximately correct size), and its power rises monotonically in the violation magnitude and in N (at $N = 1,000$, violations of 0.05 SD/period are detected in 63% of replications).

The substantive concern, as Roth (2022) emphasizes, is the inferential consequence of *conditioning* on having passed the pretest. Figure 2(b) uses a DGP with a *true null* treatment effect ($\beta = 0$) but non-zero pre-trend violations, so any rejection of $H_0 : \beta = 0$ is a false positive. At a violation of 0.10 SD/period, unconditional rejection is 0.717, while among replications that pass the pretest, conditional rejection remains 0.249—nearly five times the nominal 0.05 level. The mechanism is selection: pretest-passing replications are those in which sampling noise attenuates the estimated pre-trend, and the same noise propagates into a biased post-treatment estimate. A binary “pretest-gate” rule therefore confers far less Type I error protection than its nominal size suggests, and the gap widens for the small violations most likely to escape detection.

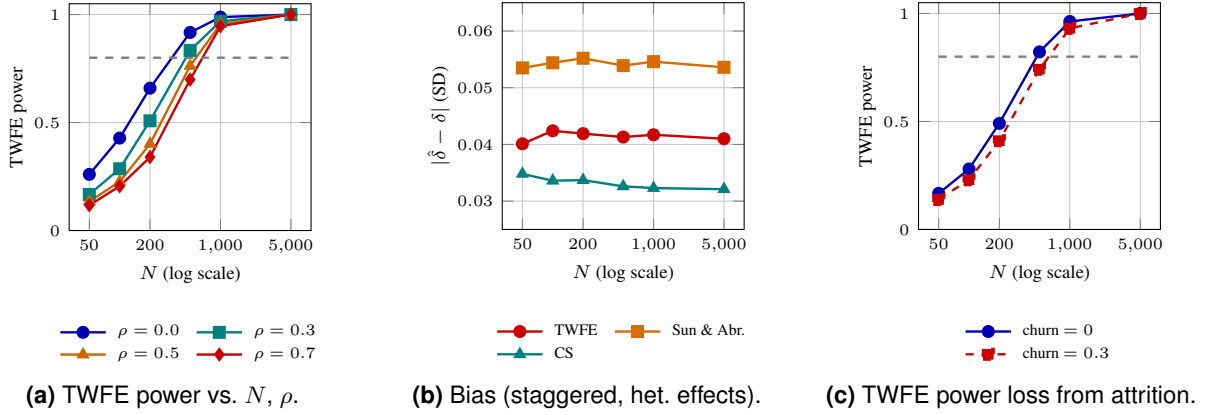


Figure 1. DiD precision under panel frictions. (a) TWFE power decays steeply with residual autocorrelation (effect size = 0.2 SD, non-staggered, 50% pre-period). (b) Deviation from the base effect δ (not the estimand) under staggered, heterogeneous effects for TWFE and our simplified Sun–Abraham and Callaway–Sant’Anna implementations ($\rho = 0.3$). The ordering shown is an artifact of measuring against δ rather than the cohort-share ATT; see text. (c) Exogenous attrition (30% missing unit-periods) further erodes power at moderate N ($\rho = 0.3$). Dashed lines mark the 80% target; 1,000 replications per cell.

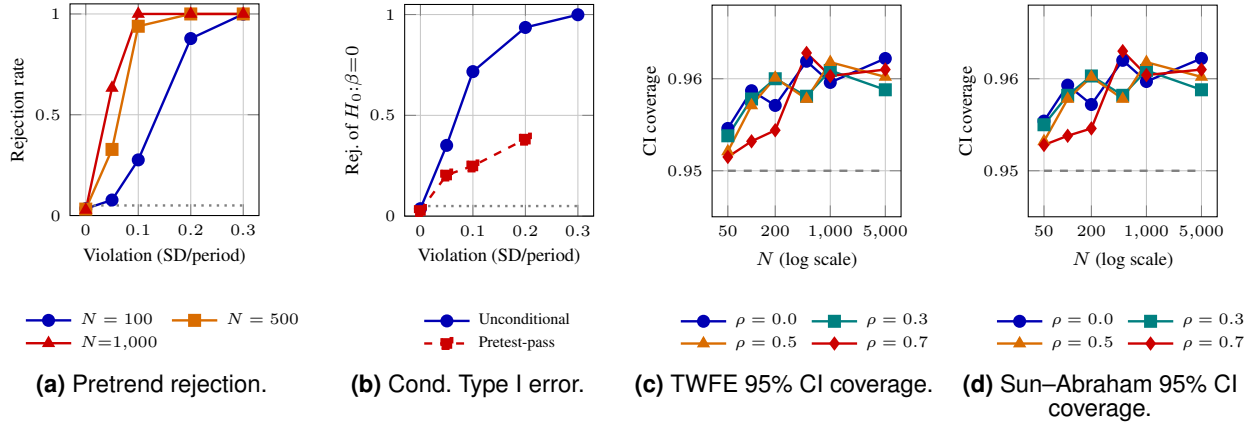


Figure 2. Pretrend behavior and inferential calibration. (a) The pretrend test has approximately nominal size and gains power with N . (b) Conditioning on pretest passage inflates Type I error ($\beta = 0$ DGP; the conditional curve ends at 0.20 because the pretest rejects everywhere at larger violations). (c)–(d) 95% CI coverage for TWFE and Sun–Abraham (non-staggered, homogeneous, unit clustering).

4.4. Confidence-Interval Coverage

Panels (c) and (d) of Figure 2 report empirical 95% CI coverage for TWFE and Sun–Abraham across N and ρ under the non-staggered, homogeneous benchmark. Coverage remains near nominal (0.951–0.963) with no systematic drift, indicating that the power losses above reflect effective-sample-size contraction rather than under-coverage. For Callaway–Sant’Anna, whose simplified across-cohort standard error is undefined in this single-cohort benchmark, a unit-level bootstrap yields coverage of 0.91–0.97 (mean ≈ 0.94) across $N \in \{200, 500\}$ and $\rho \in \{0.3, 0.5\}$ —near-nominal and comparable to the other estimators.

The picture is different in the *staggered* cells, where

the simplified C&S analytic interval—the across-cohort dispersion of per-cohort ATTs—undercovers systematically. Across all 2,304 staggered homogeneous-effects cells its coverage averages 0.906, never exceeds 0.955, and one third of cells fall below 0.90. The undercoverage is flat in N and improves only with panel length (0.856 at $T = 6$ rising to 0.934 at $T = 24$): the signature of a standard error estimated from a handful of cohort-level ATTs, whose count grows with T but not N . The supplement’s worked example (0.876, a staggered heterogeneous cell with attrition) is part of this pattern, not an outlier. By comparison, TWFE coverage under unit clustering in the same cells is nominal (0.957) and Sun–Abraham close to it (0.937); a separate tail arises under *two-way*

clustering, where the truncated CGM variance estimator undercovers (mean 0.865, worsening with T)—a distinct issue we flag but do not pursue. Practically: treat the simplified C&S analytic interval in staggered designs as anti-conservative by roughly 4–9 coverage points, and prefer the reference multiplier-bootstrap inference of the `did` package whenever its intervals matter for a conclusion.

4.5. Supplementary DiD Robustness Modules

Beyond the core grid, a treatment-process module allows non-absorbing and continuous-dose exposure. Binary TWFE degrades under continuous treatment (bias ≈ 0.101 , RMSE ≈ 0.143 , power ≈ 0.491) while absorbing and non-absorbing designs stay near-unbiased with coverage ≈ 0.95 – 0.96 : treatment-process misspecification can be first-order. We restrict supplementary estimators to specifications we can implement faithfully, leaving imputation, matrix-completion, and causal-ML comparisons to their reference implementations in follow-up work.

4.6. Empirical Anchor: Classic DiD Design

To anchor the simulation evidence in observed data, we replicate a classic open DiD panel design using the firm-level job-training dataset (54 firms, 3 years; grant adoption in 1988–1989; Wooldridge, 2010). The estimated TWFE effect is -0.279 (SE 0.198). Treating this magnitude as an *ex ante* design anchor—a posited target effect rather than observed power, which is an uninformative transform of the p -value (Hoening & Heisey, 2001)—the design-based i.i.d. power for this panel is 0.291, the calculation most frequently reported in applied IS design sections. Mapping the design to the nearest cells in the simulation grid (staggered adoption, unit clustering, nearest N , T , pre-period share, autocorrelation, and effect size) yields materially lower achieved power: 0.067 for TWFE, 0.097 for Sun–Abraham, and 0.134 for Callaway–Sant’Anna. Because the anchor is mapped to the *nearest* grid cell rather than simulated at its exact parameters, part of this gap reflects cell coarseness, and here that part cannot be tightly bounded: the panel’s three periods lie below the grid’s minimum $T = 6$, so the mapped cell is a coarse approximation. For the IV anchor below, an exact-parameter replication bounds the corresponding contribution at about one percentage point. The nominal-achieved gap for TWFE is about 22 percentage points, consistent with the main result that classical power heuristics overstate inferential strength in short, persistent panels.

5. IV Results

IV identification rests on two conceptually independent requirements — relevance and exclusion — often evaluated through the single lens of first-stage diagnostics. Our IV grid separates them, varying 2SLS performance over first-stage strength, sample size, and direct $Z \rightarrow Y$ violations of exclusion, with the true structural effect at $\beta = 0.1$ (panels (a)–(b) of Figure 3) or $\beta = 0$ (panels (c)–(d)).

5.1. Power as a Function of First-Stage Strength and Sample Size

Panel (a) of Figure 3 plots IV power over the target first-stage F and N , conditional on a true structural effect of $\beta = 0.1$ (a “small” effect by the Table 1 anchors, and the range most common in IS platform and recommender studies). For this effect size, 80% power is infeasible within our grid: power never exceeds ≈ 0.19 , even at $F = 100$ and $N = 10,000$. At $F = 10$ — the canonical Stock and Yogo (2005) reference point — power is ≈ 0.08 – 0.10 regardless of N , and in the $F = 10$ – 20 range typical of IS instruments (geographic variation, peer exposure) even $N = 1,000$ delivers power below 0.11. The ceiling reflects the fixed- F normalization: holding $F \approx 100$ pins the concentration parameter near 100 regardless of N , so $\text{SE}(\hat{\beta})$ stays ≈ 0.10 and $|\hat{\beta}|/\text{SE} \approx 1$ for $\beta = 0.1$, far short of conventional power. This is a statement about fixed F , not data collection: with a *fixed* instrument (constant first-stage coefficient), F grows with N , $\text{SE}(\hat{\beta})$ falls as $N^{-1/2}$, and power rises in the usual way. The grid holds F fixed precisely to isolate its role, so the lever it identifies is instrument strength, not sample size—and the $F > 10$ convention, while informative about size, is an inadequate proxy for power when the target effect is small.

5.2. Two Failure Modes: Weak Instruments versus Exclusion Violations

The two canonical IV assumption failures — a weak first stage and a non-zero direct $Z \rightarrow Y$ channel — produce bias signatures that, under our fixed- F normalization, scale with N in opposite directions. Panel (b) of Figure 3 plots the absolute finite-sample bias of 2SLS as F ranges from 5 to 100. At $F \geq 50$ bias is in the 10^{-3} range across all sample sizes; at $F = 5$ it is visible but attenuates as N grows, consistent with weak-instrument bias being $O(N^{-1})$ conditional on identification (Angrist & Pischke, 2009).

Panel (c) of Figure 3 fixes F at 100 and varies

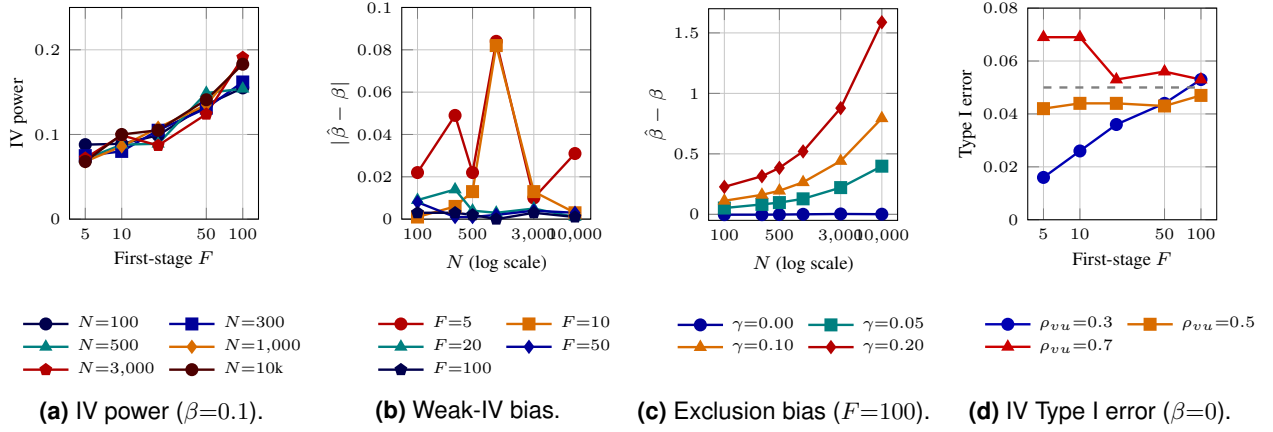


Figure 3. IV diagnostics across first-stage strength, sample size, and assumption failures. (a) Power never exceeds ≈ 0.19 —far below the 0.80 target—even at $N = 10,000$ and $F = 100$ ($\beta = 0.1$, $\rho_{vu} = 0.5$, single instrument). (b) Weak-instrument bias attenuates with F and N ; fluctuations at $F \in \{5, 10\}$, $N = 1,000$ are Monte Carlo noise. (c) With $F = 100$ but a direct $Z \rightarrow Y$ channel γ , bias scales with N (true $\beta = 0$, $\rho_{vu} = 0.5$). (d) Heteroskedasticity-robust 2SLS holds size near 0.05 (dashed) even at $F = 5$; distortion rises only at $\rho_{vu} = 0.7$.

the direct-effect parameter γ governing the $Z \rightarrow Y$ channel outside the structural equation. Because the design holds the target F fixed, $\pi = \sqrt{F/(N+F)} \propto N^{-1/2}$, and the just-identified 2SLS plim bias from a direct channel is γ/π , so it grows as $\sqrt{N}$: for $\gamma = 0.05$ it rises from 0.052 at $N = 100$ to 0.398 at $N = 10,000$ ($7.7\times$ observed versus $7.1\times$ predicted). Under a fixed instrument coefficient instead, the plim bias γ/π is constant in N while sampling variance shrinks—the estimator becomes confidently wrong. The message shared across both normalizations is narrow but firm: sample size does not cure exclusion bias. The asymmetry with panel (b) has direct diagnostic implications: the standard toolkit (first-stage F , Hansen J) cannot separate the two regimes with a single-instrument design, so over-identification tests provide limited leverage in exactly the settings where the bias grows fastest.

5.3. Size Properties of Robust 2SLS Under Weak Instruments

Panel (d) of Figure 3 sets $\beta = 0$ and reports the Type I error of heteroskedasticity-robust 2SLS. Even at $F = 5$, size never exceeds 0.069, and it lies above 0.05 only for strong endogeneity ($\rho_{vu} = 0.7$); for $F \geq 20$ it is indistinguishable from nominal. Size distortion is thus more limited than classical Stock and Yogo (2005) warnings imply, consistent with Angrist and Pischke (2009) and Semadeni et al. (2014) on robust standard errors (see Lonschien, 2025 for a recent overview). Weak instruments in our DGP damage designs primarily through low power and first-stage bias

rather than inflated false-positive rates.

5.4. Empirical Anchor: Classic IV Design

We also replicate a standard open IV design using the Card returns-to-schooling dataset (instrument: `nearc4`, $N = 3,010$; Card, 1995). The estimated 2SLS coefficient on schooling is 0.119 (SE 0.053), with first-stage $F = 14.49$. Treating 0.119 as an *ex ante* design anchor (Hoenig & Heisey, 2001), the design-based i.i.d. power is 0.614. Mapping this design to the nearest IV-grid cell ($N = 3,000$, $F = 20$, $\rho_{vu} = 0.3$, $\beta = 0.1$) gives achieved power of 0.060, a nominal-achieved gap of roughly 55 percentage points. The gap is not an artifact of nearest-cell coarseness: a targeted replication at Card’s exact parameters ($\beta = 0.119$, $F = 14.49$, $N = 3,010$) shifts achieved power by under one percentage point, bounding the cell-coarseness contribution for this anchor at about one point. The remainder reflects a weak first stage interacting with a small true effect: clearing conventional relevance heuristics does not guarantee adequate power at Card’s *realized* N and F . Because `nearc4` is a fixed instrument, a larger- N version of the same design would have proportionally higher F and materially higher power; the lesson concerns instrument strength at the available sample, not the futility of collecting more data.

6. Discussion

6.1. Implications for Study Design

We offer these as evidence-based guidance, not prescriptions. (1) Residual autocorrelation is a first-order but *recoverable* input to DiD power (Table 3). Where pre-period data exist, estimate ρ from unit-demeaned residuals—with the bias correction of Table 4 on short pre-periods—and propagate it through a calibrated simulation; otherwise report a power curve over $\rho \in [0, 0.7]$. (2) What is *not* recoverable—attrition and staggered timing—must be built into the design directly, and is better read off the companion lookup tool’s nine-dimensional cell than 2D slices. (3) Estimator choice matters, and so do implementation fidelity and the estimand: reference implementations of Sun–Abraham and Callaway–Sant’Anna are approximately unbiased for the cohort-share ATT while TWFE attenuates it by ≈ 0.02 SD, but orderings among *simplified* reimplementations can invert that conclusion (Section 4). Report TWFE alongside a cohort-aware estimator *from a reference package* and state the estimand explicitly. (4) Binary pretesting inflates conditional Type I error (Roth, 2022); prefer reporting the unconditional estimate with a sensitivity analysis (Rambachan & Roth, 2023) over a pass/fail gate. (5) The $F > 10$ convention speaks to size, not power (at $F = 10$, $N = 500$, IV power < 0.11); report achieved power explicitly. (6) Exclusion bias is not cured by sample size; defend exclusion on substantive or design grounds, not via larger panels. (7) CI coverage is a useful second-order diagnostic: deviations from 0.95 signal mis-specification such as inadequate clustering. (8) Advanced estimators (imputation, matrix completion, causal ML) rely on their reference inference machinery; validate coverage on a study-like DGP before trusting their default intervals.

6.2. Limitations

Our DGP uses AR(1) residuals, unit and time effects, and a simple trend; multi-way error dependence in the DGP (we vary only the clustering used for inference), unit-specific trends, and seasonality are out of scope, so the irreducible gap we isolate is a conservative lower bound for messier panels. The attrition sub-grid covers treatment-differential and outcome-dependent exit, but not richer dynamic selection (anticipation, selection on effect size). Our simplified Callaway–Sant’Anna analytic SE undercovers across the staggered grid (Section 4.4); porting the reference `did` multiplier-bootstrap SE across the full grid is future work. We do not evaluate

parallel-trends sensitivity analyses (Rambachan & Roth, 2023); the main DiD grid uses `effect_size > 0` (DiD Type I error is taken from the pretest and IV modules); Monte Carlo precision is 1,000 replications per cell; and, as noted at the point of each claim, the AR(1)-aware baseline is an oracle (Table 4 quantifies the estimated- ρ case) and each empirical anchor is mapped to the nearest grid cell (coarseness bounded for IV by the exact-parameter Card replication; unbounded for the DiD anchor, whose T lies outside the grid).

6.3. Next Steps: Toward IS Research Methodology for the 2030s

The grid—an empirically anchored, validated simulation released as a queryable artifact—is a prototype for how *ex ante* design evidence in IS could be produced and consumed. Four next steps: richer failure modes (informative missingness, dynamic selection, anticipation—mapping rather than assuming the power-versus-bias boundary); estimator fidelity as infrastructure, extending reference Sun–Abraham, Callaway–Sant’Anna (multiplier-bootstrap inference), imputation (Borusyak et al., 2024; Gardner, 2022), matrix-completion (Athey et al., 2021), and causal-ML (Chernozhukov et al., 2018; Wager & Athey, 2018) estimators across the full grid so the community compares estimators on equal footing rather than through per-paper reimplementations; design tooling that folds estimated-parameter uncertainty and parallel-trends sensitivity (Rambachan & Roth, 2023) into the lookup tool for registration-ready workflows; and a living, versioned benchmark open to community-contributed cells, as shared benchmarks did for machine learning. As IS panels grow larger and messier, we expect empirically calibrated simulation evidence to displace closed-form power formulas as the default justification standard.

6.4. Reproducibility and Computational Cost

The pipeline is implemented in Python with per-cell hash-derived seeds, so any cell is independently reproducible. The full grids (~ 9.8 million datasets) required roughly four days of wall-clock time on a 14-core consumer laptop—no cluster access was used or needed. This cost matters mainly for full replication: a single design cell reproduces in minutes and a targeted sub-grid within an hour on commodity hardware, and the lookup tool removes the need to re-simulate at all for designs inside the released grid.

Acknowledgment

LLMs were used strictly for language polishing (grammar and readability); we reviewed and verified all content and assume full responsibility for its accuracy and integrity.

References

- Abadie, A. (2021). Using synthetic controls: Feasibility, data requirements, and methodological aspects. *Journal of Economic Literature*, 59(2), 391–425. <https://doi.org/10.1257/jel.20191450>
- Angrist, J. D., & Pischke, J.-S. (2009). *Mostly harmless econometrics: An empiricist's companion*. Princeton University Press. <https://doi.org/10.2307/j.ctvc4j72>
- Aral, S., & Walker, D. (2011). Creating social contagion through viral product design: A randomized trial of peer influence in networks. *Management Science*, 57(9), 1623–1639. <https://doi.org/10.1287/mnsc.1110.1421>
- Athey, S., Bayati, M., Doudchenko, N., Imbens, G., & Khosravi, K. (2021). Matrix completion methods for causal panel data models. *Journal of the American Statistical Association*, 116(536), 1716–1730. <https://doi.org/10.1080/01621459.2021.1891924>
- Aziz, A., Li, H., & Telang, R. (2023). The consequences of rating inflation on platforms: Evidence from a quasi-experiment. *Information Systems Research*, 34(2), 590–608. <https://doi.org/10.1287/isre.2022.1134>
- Bertrand, M., Duflo, E., & Mullainathan, S. (2004). How much should we trust differences-in-differences estimates? *Quarterly Journal of Economics*, 119(1), 249–275. <https://doi.org/10.1162/003355304772839588>
- Borusyak, K., Jaravel, X., & Spiess, J. (2024). Revisiting event-study designs: Robust and efficient estimation. *Review of Economic Studies*, 91(6), 3253–3285. <https://doi.org/10.1093/restud/rdae007>
- Callaway, B., & Sant'Anna, P. H. C. (2021). Difference-in-differences with multiple time periods. *Journal of Econometrics*, 225(2), 200–230. <https://doi.org/10.1016/j.jeconom.2020.12.001>
- Card, D. (1995). Using geographic variation in college proximity to estimate the return to schooling. In *Aspects of labour market behaviour: Essays in honour of John Vanderkamp* (pp. 201–222). University of Toronto Press.
- Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1), C1–C68. <https://doi.org/10.1111/ectj.12097>
- de Chaisemartin, C., & D'Haultfoeuille, X. (2020). Two-way fixed effects estimators with heterogeneous treatment effects. *American Economic Review*, 110(9), 2964–2996. <https://doi.org/10.1257/aer.20181169>
- Gardner, J. (2022). Two-stage differences in differences. *arXiv preprint*. <https://doi.org/10.48550/arXiv.2207.05943>
- Gelman, A., & Carlin, J. (2014). Beyond power calculations: Assessing Type S (Sign) and Type M (Magnitude) errors. *Perspectives on Psychological Science*, 9(6), 641–651. <https://doi.org/10.1177/1745691614551642>
- Goodman-Bacon, A. (2021). Difference-in-differences with variation in treatment timing. *Journal of Econometrics*, 225(2), 254–277. <https://doi.org/10.1016/j.jeconom.2021.03.014>
- Hoenig, J. M., & Heisey, D. M. (2001). The abuse of power: The pervasive fallacy of power calculations for data analysis. *The American Statistician*, 55(1), 19–24. <https://doi.org/10.1198/000313001300339897>
- Huang, N., Hong, Y., & Burtch, G. (2017). Social network integration and user content generation: Evidence from natural experiments. *MIS Quarterly*, 41(4), 1035–1058. <https://doi.org/10.25300/MISQ/2017/41.4.02>
- Londschien, M. (2025). *A statistician's guide to weak-instrument-robust inference in instrumental variables regression with illustrations in Python*. arXiv: 2508.12474. <https://arxiv.org/abs/2508.12474>
- Rambachan, A., & Roth, J. (2023). A more credible approach to parallel trends. *Review of Economic Studies*, 90(5), 2555–2591. <https://doi.org/10.1093/restud/rdad018>
- Roth, J. (2022). Pretest with caution: Event-study estimates after testing for parallel trends. *American Economic Review: Insights*, 4(3), 305–322. <https://doi.org/10.1257/aeri.20210236>
- Seamans, R., & Zhu, F. (2014). Responses to entry in multi-sided markets: The impact of craigslist on local newspapers. *Management Science*, 60(2), 476–493. <https://doi.org/10.1287/mnsc.2013.1785>
- Semadeni, M., Withers, M. C., & Certo, S. T. (2014). The perils of endogeneity and instrumental variables in strategy research: Understanding through simulations. *Strategic Management Journal*, 35(7), 1070–1079. <https://doi.org/10.1002/smj.2136>
- Stock, J. H., & Yogo, M. (2005). Testing for weak instruments in linear IV regression. In D. W. K. Andrews & J. H. Stock (Eds.), *Identification and inference for econometric models: Essays in honor of Thomas Rothenberg* (pp. 80–108). Cambridge University Press.
- Sun, L., & Abraham, S. (2021). Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *Journal of Econometrics*, 225(2), 175–199. <https://doi.org/10.1016/j.jeconom.2020.09.006>
- Tafti, A., & Shmueli, G. (2025). Causal inference grounded in causal diagrams: Benefits, limitations, and opportunities for the Information Systems field. *MIS Quarterly*, 49(2), 409–428. <https://doi.org/10.25300/MISQ/2024/18422>
- Ulloa-Pérez, E., Bair, E. F., Navathe, A. S., & Linn, K. A. (2025). *Comparative evaluation of difference in differences methods for staggered adoption interventions*. arXiv: 2508.14365. <https://arxiv.org/abs/2508.14365>
- von Zahn, M., Güler, A., Pfeiffer, J., Reijers, H. A., & Hinz, O. (2026). Causal machine learning in information systems research. *Business & Information Systems Engineering*, 68, 235–242. <https://doi.org/10.1007/s12599-026-00999-x>
- Wager, S., & Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523), 1228–1242. <https://doi.org/10.1080/01621459.2017.1319839>
- Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data* (2nd ed.). MIT Press.