FROM THE RESEARCH BENCH

Staggered difference-in-differences needs a cohort-time estimand

When treatment starts at different times, name the cohort-time effect and aggregation target before interpreting a familiar fixed-effects coefficient.

Sinan Isoglu, MBA (Quantic)

Commercial growth leader, lecturer and doctoral researcher

PUBLISHED
1 September 2026

READING TIME
8 min

WORDS
1,760

https://isoglu.com/insights/journal/staggered-difference-in-differences-needs-a-cohort-time-estimand/

CONTENTS

MANAGEMENT SUMMARY

Staggered difference-in-differences is not one coefficient with a new label. When cohorts receive a treatment at different times, the analyst must define the group, calendar time, event time, comparison set, outcome, horizon, and aggregation target. Callaway and Sant'Anna provide group-time effects and aggregation schemes. Goodman-Bacon shows how a two-way fixed-effects coefficient can combine weighted two-by-two comparisons, including timing comparisons that become difficult under heterogeneous effects. Roth and colleagues separate estimands, pre-trend tests, sensitivity, and inference. Wooldridge supplies a bounded nonlinear extension. This article turns those boundaries into a cohort-time release card. It does not estimate a current effect, rank software, or use the blocked Sun and Abraham version.

Keywords: Difference-in-differences · Causal inference · Treatment-effect heterogeneity · Cohort-time estimand

When treatment starts at different times, the first task is not to choose a familiar regression. It is to name the effect being estimated.

The short answer is precise: a staggered difference-in-differences design needs a cohort-time estimand, an explicit comparison set, and a declared aggregation rule before a coefficient can be given a causal meaning. A two-way fixed-effects coefficient may be useful in some settings, but under staggered timing it can combine several two-by-two comparisons with weights that do not match the effect a decision-maker thinks they are reading.

Callaway and Sant'Anna define group-time average treatment effects for settings with multiple periods and variation in treatment timing. Goodman-Bacon decomposes the conventional two-way fixed-effects estimator into weighted two-by-two comparisons. Roth, Sant'Anna, Bilinski, and Poe organize recent work around heterogeneity, parallel trends, and inference. Wooldridge develops a nonlinear extension for a narrower panel-data setting. Together, the sources support a design discipline, not one universal estimator.

Why must causal inference name the cohort-time estimand before running regressions?

Suppose two customer cohorts receive the same commercial intervention. Cohort A receives it in January. Cohort B receives it in April. A third cohort has not received it by the end of the observation window. The phrase "the treatment effect" is incomplete. Which cohort? Which month? Compared with which units? For how long after treatment? Weighted how?

The minimum object description is:

Table 1 Why must causal inference name the cohort-time estimand before running regressions?

Field	Question	Why it matters
Cohort	Which units first receive treatment in this group?	Treatment timing is part of the object.
Calendar time	In which period is the outcome observed?	A cohort can have different effects over time.
Event time	How far before or after treatment is the period?	Dynamic effects are not the same as one post-period average.
Comparison	Which untreated or not-yet-treated units provide the contrast?	Already-treated units may not be valid controls for a later cohort.
Outcome	What is measured, in which unit, and at what horizon?	A coefficient cannot repair an unclear outcome.
Aggregation	Which cohort-time effects are combined, and with what weights?	The overall number depends on the aggregation target.
Inference	Where was treatment assigned, and where should uncertainty be clustered?	Precision is part of the design, not an afterthought.

Table from this essay. Sources and interpretation are given in the article.

These are design fields, not software settings. A package can return a number while leaving the causal object under-specified.

Figure 1 The staggered DiD estimand release gate

Release field	Required input	Permitted statement	Stop signal
Cohort and time	First-treatment group, calendar period, event time	“This is the effect for cohort g at time t.”	The article says only “the treatment effect.”
Comparison	Never-treated or not-yet-treated set, with conditions	“This comparison supplies the stated contrast.”	Already-treated units silently serve as controls.
Outcome and horizon	Outcome unit, measurement window, post-treatment horizon	“The estimate concerns this outcome over this horizon.”	The outcome changes between sections.
Aggregation	Cohorts, periods, and weights used for the summary	“This overall result answers this weighted question.”	The summary is treated as a natural ATT.
Inference and sensitivity	Assignment level, uncertainty method, trend and heterogeneity checks	“Uncertainty and sensitivity were reviewed at this boundary.”	A pre-trend test is treated as proof.

Author's release framework grounded in Callaway and Sant'Anna (2021), Goodman-Bacon (2021), Roth et al. (2023), and Wooldridge (2023). The prompts are synthetic and do not contain an effect estimate.

Why are group-time average treatment effects the necessary baseline in staggered rollouts?

Callaway and Sant’Anna’s contribution is to define group-time average treatment effects in a setting with multiple periods and variation in treatment timing. The group identifies when units first receive treatment. Time identifies the period in which the outcome is evaluated. The effect is therefore a set of objects, not necessarily one naturally occurring scalar.

That structure makes the decision question clearer. A team may want the effect for early adopters in the first post-treatment month. It may want the effect for late adopters after a full quarter. It may want an overall average across cohorts. Those are not interchangeable requests.

The identification conditions also belong to the object. Callaway and Sant’Anna allow parallel trends after conditioning on observed covariates in the relevant formulation. Their estimation strategies include outcome regression, inverse-probability weighting, and doubly robust procedures. The method choice does not remove the need to explain what is assumed about untreated potential outcomes.

The aggregation schemes are equally important. If effects vary by cohort or event time, a summary depends on which groups and periods receive weight. An overall value can be useful if it answers a declared policy question. It is not automatically the average a reader imagines when a table reports one treatment coefficient.

Why does standard two-way fixed effects introduce negative weighting in staggered designs?

Goodman-Bacon shows that, with variation in treatment timing, the two-way fixed-effects difference-in- differences estimator can be represented as a weighted average of two-by-two comparisons. The comparisons include treated versus never-treated units, early-treated versus later-treated units, and later-treated versus early-treated units.

The problem is not that every two-way fixed-effects design is mechanically unusable. The problem is that the coefficient's weights and comparison objects may differ from the target effect. An early-treated group can serve as a control for a later-treated group. Once the early group is already affected, that comparison is no longer a simple untreated contrast.

Under homogeneous effects, the mixture may still have a straightforward interpretation. Under heterogeneous effects, the mixture can be difficult to interpret and can receive weights that are negative or otherwise problematic. The correct response is to expose the timing and heterogeneity conditions, then choose an estimand and estimator that answer the intended question. "Two-way fixed effects is always wrong" is as imprecise as "the coefficient is the ATT."

This is why a regression table should not be the first exhibit. The first exhibit should show who is treated, when treatment begins, who is available as a comparison, and which effects are being averaged.

Why are parallel pre-trends contextual evidence rather than proof of unconfoundedness?

Roth and colleagues place pre-trend assessment inside a wider methods agenda. Their review emphasizes multiple periods and staggered timing, violations of parallel trends, heterogeneity, and inference. It also warns that pre-trend tests can be underpowered and that using them as a pretest can introduce bias.

A failed pre-trend test can raise a serious concern. A non-failed test does not prove parallel trends. A test may lack power to detect a relevant departure, or the selected pre-period may not represent the untreated path that matters after treatment. The release sentence should therefore state the test, its period, its power or uncertainty where available, and the sensitivity of the result to plausible trend differences.

The same separation applies to inference. Roth and colleagues recommend explicit estimand and estimator choices, heterogeneity-robust methods, sensitivity analysis, and clustering at the treatment-assignment level when appropriate. Clustering cannot rescue a comparison set that has already been contaminated. It answers a different question about uncertainty conditional on the design.

How do nonlinear binary and count outcomes complicate staggered difference-in-differences?

Many commercial outcomes are not continuous. They may be binary adoption, a count of orders, a bounded rate, or a positive but skewed amount. Wooldridge develops simple nonlinear difference-in-differences strategies for panel data with staggered interventions and optional covariates.

That extension is useful because it makes the outcome model visible. It is not a license to apply a nonlinear link to any staggered design and carry over the same interpretation. The estimand, scale, covariates, timing, and infer-

ence remain model-specific. A probability difference, a log-scale effect, and a ratio are different objects even when they use the same panel.

The practical rule is simple: if the outcome scale changes, rewrite the release card. Do not treat “nonlinear” as a technical footnote that leaves the causal sentence unchanged.

What runnable econometric workflow ensures unbiased staggered rollout evaluations?

Use the following sequence before a result enters a commercial decision:

1. Write the treatment definition and the first-treatment date for each cohort.
2. Name the outcome, unit, observation window, and post-treatment horizon.
3. Define the comparison set for each cohort-time cell and state when already-treated units are excluded.
4. State the untreated-trend condition, including any covariates or conditioning set.
5. Choose the estimator family after the estimand, not before it.
6. Choose the aggregation weights and explain why they answer the decision question.
7. State the inference cluster and review pre-trends, heterogeneity, and sensitivity separately.
8. If the outcome is nonlinear, declare the model-specific scale and interpretation.
9. Write the permitted sentence and the stronger sentence that remains outside the evidence.

This sequence is deliberately slower than copying a regression specification. It is faster than debating an unexplained coefficient after the decision has already been made.

Where are the econometric boundaries of modern difference-in-differences estimators?

The four sources do not establish a current treatment effect for a private pipeline, customer cohort, or employer. They do not rank every software implementation. They do not say that one estimator dominates under all timing, outcome, or heterogeneity conditions. They provide the ingredients for an estimand-first design and a warning about automatic interpretation.

The stopping rule is therefore concrete. Do not release a staggered difference-in-differences result until the cohort-time target, comparison set, aggregation rule, inference choice, and sensitivity boundary are written in the same place. If one is missing, the result may still be calculable. It is not yet a defensible answer to a decision question.

The estimand boundary connects to the counterfactual attribution model and the uplift claim that needs a specification.

Callaway, B., and P. H. C. Sant'Anna. (2021). Difference-in-Differences with Multiple Time Periods. Journal of Econometrics, 225(2), 200-230. DOI

Goodman-Bacon, A. (2021). Difference-in-Differences with Variation in Treatment Timing. Journal of Econometrics, 225(2), 254-277. DOI

Roth, J., P. H. C. 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. DOI

Wooldridge, J. M. (2023). Simple Approaches to Nonlinear Difference-in-Differences with Panel Data. The Econometrics Journal, 26, C31-C66. DOI

END OF ESSAY

HOW TO CITE THIS ESSAY

Isoglu, S. (2026, September 1). Staggered difference-in-differences needs a cohort-time estimand. Sinan Isoglu. https://isoglu.com/insights/journal/staggered-difference-in-differences-needs-a-cohort-time-estimand/

KEEP READING

From the research bench
What are parallel trends? The assumption behind difference-in-differences
https://isoglu.com/insights/journal/what-are-parallel-trends/

From the research bench
What is external validity? A result needs a transfer boundary
https://isoglu.com/insights/journal/what-is-external-validity/

From the research bench
An uplift claim needs a specification before it needs a number
https://isoglu.com/insights/journal/an-uplift-claim-needs-a-specification/

Sinan Isoglu, MBA (Quantic)
Commercial growth leader, lecturer and doctoral researcher

Online version: https://isoglu.com/insights/journal/staggered-difference-in-differences-needs-a-cohort-time-estimand/