Review of Economic Studies (2024) **00**, 1–33 doi:10.1093/restud/rdae007

Advance access publication 6 February 2024

Revisiting Event-Study Designs: Robust and Efficient Estimation

Kirill Borusyak
UC Berkeley, USA
Xavier Jaravel
LSE, UK
and
Jann Spiess
Stanford University, USA

First version received April 2022*; Editorial decision August* 2023*; Accepted January* 2024 *(Eds.)*

We develop a framework for difference-in-differences designs with staggered treatment adoption and heterogeneous causal effects. We show that conventional regression-based estimators fail to provide unbiased estimates of relevant estimands absent strong restrictions on treatment-effect homogeneity. We then derive the efficient estimator addressing this challenge, which takes an intuitive "imputation" form when treatment-effect heterogeneity is unrestricted. We characterize the asymptotic behaviour of the estimator, propose tools for inference, and develop tests for identifying assumptions. Our method applies with time-varying controls, in triple-difference designs, and with certain non-binary treatments. We show the practical relevance of our results in a simulation study and an application. Studying the consumption response to tax rebates in the U.S., we find that the notional marginal propensity to consume is between 8 and 11% in the first quarter—about half as large as benchmark estimates used to calibrate macroeconomic models—and predominantly occurs in the first month after the rebate.

Key words: Difference-in-differences, Efficiency, Marginal propensity to consume

JEL codes: C21, C23, E21

1. INTRODUCTION

Event studies are one of the most popular tools in applied economics and policy evaluation. An event study is a difference-in-differences (DiD) design in which a set of units in the panel receive treatment at different points in time. In this paper, we investigate the robustness and efficiency of estimators of causal effects in event studies, with a focus on the role of treatment effect heterogeneity. We first develop a simple econometric framework that delineates the identification assumptions from each other and from the estimation target, defined as some average of heterogeneous causal effects. We then apply this framework in three ways. First, we analyse the conventional practice of implementing event studies via two-way fixed-effect ordinary least squares (TWFE OLS) regressions and show how the implicit conflation of different assumptions

The editor in charge of this paper was Elias Papaioannou.

leads to biases. Second, leveraging event-study assumptions in an explicit and principled way allows us to derive the robust and efficient estimator, along with appropriate inference methods and tests. The estimator takes an intuitive "imputation" form when treatment-effect heterogeneity is unrestricted. Finally, we illustrate the practical relevance of our approach in an application estimating the marginal propensity for expenditure (MPX) out of tax rebates; our MPX estimates are lower than in prior work, implying that fiscal stimulus is less powerful than commonly thought.

Event studies are frequently used to estimate treatment effects when treatment is not randomized, but the researcher has panel data allowing them to compare outcome trajectories before and after the onset of treatment, as well as across units treated at different times. By analogy to conventional DiD designs without staggered rollout, event studies are commonly implemented by TWFE regressions, such as

$$Y_{it} = \alpha_i + \beta_t + \tau D_{it} + \varepsilon_{it}, \tag{1}$$

where outcome Y_{it} and binary treatment D_{it} are measured in periods t and for units i, α_i are unit fixed effects (FEs) that allow for different baseline outcomes across units, and β_t are period fixed effects that accommodate overall trends in the outcome. Specifications like (1) are meant to isolate a treatment effect τ from unit- and period-specific confounders. A commonly used dynamic version of this regression includes "lags" and "leads" of the indicator for the onset of treatment, to capture treatment effects for different "horizons" since the onset of treatment and test for the parallel trajectories of the pre-treatment outcomes.

To understand the problems with conventional TWFE estimators in event-study designs and provide a principled econometric approach to overcoming these issues, in Section 2 we develop a simple framework that makes the estimation targets and underlying assumptions explicit and clearly isolated. We suppose that the researcher chooses a particular weighted average (or weighted sum) of heterogeneous treatment effects they are interested in estimating. We make (and later test) two standard DiD identification assumptions: that potential outcomes without treatment are characterized by parallel trends and that there are no anticipatory effects. We also allow for—but do not require—an auxiliary assumption that the treatment effects themselves follow some model that restricts their heterogeneity for *a priori* specified economic reasons. This explicit approach is in contrast to regression specifications like (1), both static and dynamic, which implicitly conflate choices of estimation target and identification assumptions. Our framework covers a broad class of empirically relevant estimands beyond the standard average treatment-on-the-treated (ATT), including heterogeneous treatment effects by observed covariates and ATTs at different horizons that hold the composition of units fixed.

Through the lens of this framework, in Section 3 we uncover a set of challenges with conventional event-study estimation methods and trace them back to a mismatch between estimation target, identification assumptions, and the flexibility of the regression specification. First, we note that failing to rule out anticipation effects in "fully dynamic" specifications (with all leads and lags of the event included) leads to an under-identification problem when there are no never-treated units, such that the dynamic path of anticipation and treatment effects over time is not point-identified. We conclude that it is important to separate out testing the assumptions about pre-trends from the estimation of dynamic treatment effects under those assumptions. Second, implicit assumptions of homogeneous treatment effects embedded in static DiD regressions like (1) may lead to estimands that put negative weights on some long-run treatment effects. With staggered rollout, regression-based estimation leverages comparisons between groups that got treated over a period of time and reference groups which had been treated *earlier*. We label such cases "forbidden comparisons". Indeed, these comparisons are only valid when the homogeneity assumption is true; when it is violated, they can substantially distort the weights the

estimator places on treatment effects, or even make them negative. Third, in dynamic specifications, implicit assumptions about treatment-effect homogeneity across groups first treated at different times lead to the spurious identification of long-run treatment effects for which no DiD comparisons valid under heterogeneous treatment effects are available. The last two challenges highlight the danger of imposing implicit treatment effect homogeneity assumptions instead of allowing for heterogeneity and explicitly specifying the target estimand. We show that these challenges are not resolved by trimming the sample to a fixed window around the event date.

From the above discussion, the reader should not conclude that event-study designs are plagued by fundamental problems. On the contrary, these challenges only arise due to a mismatch between treatment-effect heterogeneity and specifications which restrict it. We therefore use our framework to circumvent these issues and derive robust and efficient estimators.

In Section 4, we first establish a simple characterization for the most efficient linear unbiased estimator of any pre-specified weighted sum of treatment effects, in the baseline case of spherical errors, that is, homoskedasticity with no serial correlation. This estimator explicitly incorporates the researcher's estimation goal and assumptions about parallel trends, anticipation effects, and restrictions on treatment-effect heterogeneity. It is constructed by estimating a flexible high-dimensional regression that differs from conventional event-study specifications, and aggregating its coefficients appropriately. While spherical errors are a natural starting point, the principled construction of this estimator more generally ensures unbiasedness and yields attractive efficiency properties, as we later confirm in simulations.

In our leading case where the heterogeneity of treatment effects is not restricted, the efficient robust estimator can be implemented using a transparent "imputation" procedure. First, the unit and period fixed effects $\hat{\alpha}_i$ and $\hat{\beta}_t$ are fitted by regressions using untreated observations only. Second, these fixed effects are used to impute the untreated potential outcomes and therefore obtain an estimated treatment effect $\hat{\tau}_{it} = Y_{it} - \hat{\alpha}_i - \hat{\beta}_t$ for each treated observation. Finally, a weighted sum of these treatment-effect estimates is taken, with weights corresponding to the estimation target.

To relate our efficient imputation estimator to other unbiased estimators that have been proposed in the literature, we derive two additional results showing the generality of the imputation structure. First, any other linear estimator that is unbiased in our framework with unrestricted causal effects can be represented as *an* imputation estimator, albeit with an inefficient way of imputing untreated potential outcomes. Second, even when assumptions that restrict treatment-effect heterogeneity are imposed, any unbiased estimator can still be understood as an imputation estimator for an adjusted estimand. Together, these two results allow us to characterize estimators of treatment effects in event studies as a combination of how they impute unobserved potential outcomes and which weights they put on treatment effects.

For the efficient estimator in our framework, we provide tools for valid inference. Specifically, we derive conditions under which the estimator is consistent and asymptotically normal and propose standard error estimates. Inference is challenging under arbitrary treatment-effect heterogeneity, because causal effects cannot be separated from the error terms. We instead show how asymptotically conservative standard errors can be derived, by attributing some variation in estimated treatment effects to the error terms.[1] Our inference results apply under mild conditions in short panels. Advancing the existing literature on DiD estimation with staggered adoption,

1. While the generality of our setting only allows for conservative inference (on any robust estimator, including ours), we obtain asymptotically exact standard errors in the special case that received the most attention in the literature: when units are randomly sampled from a population and the estimand consists of average treatment effects by period–cohort pairs.

we also provide conditions for consistency and inference that extend to panels where the number of time periods grows, as long as growth is not too fast. We also propose a leave-one-out modification to our conservative variance estimates with improved finite-sample performance.

Another important practical advantage of our approach is that it provides a principled way of testing the identifying assumptions of parallel trends and no-anticipation effects, based on OLS regressions with untreated observations only. Compared to conventional specifications with leads and lags of treatment that implicitly restrict treatment effects, this approach avoids the contamination of the tests by treatment-effect heterogeneity shown by Sun and Abraham (2021). Moreover, our strategy circumvents the inference problems after pre-testing that were pointed out by Roth (2022), under spherical errors. These attractive properties result from the clear separation of estimation and testing.

It is also useful to point out two limitations of our analysis. First, all event-study designs assume a restrictive parametric model for untreated outcomes. We do not evaluate when these assumptions may be applicable, and therefore when the event-study design are *ex ante*appropriate, as Roth and Sant'Anna (2023) do. We similarly do not consider estimation that is robust to violations of parallel-trend type assumptions, as Rambachan and Roth (2023) propose, although our framework allows relaxing those assumptions by including unit-specific trends and time-varying covariates. We instead take the standard assumptions of event-study designs as given and derive optimal estimators, valid inference, and practical tests to assess whether parallel-trend assumptions hold. Second, we also do not consider event studies as understood in the finance literature, based on high-frequency panel data, which typically do not use period fixed effects (MacKinlay, 1997).

In Section 5, we illustrate the practical relevance of our theoretical insights by revisiting the estimation of the marginal propensity to spend out of tax rebates in the event study of Broda and Parker (2014b). First, we show that the choice of a binned specification used by Broda and Parker (2014b) leads to a substantial upward bias in estimated MPXs. Indeed, we find that the binned specification puts a large weight on the effects happening in the first week after the rebate receipt, and negative weights on some longer-run effects, biasing the estimate upwards because the spending response quickly decays over time. Second, we highlight that, due to the implicit extrapolation of treatment effects in specifications restricting treatment-effect heterogeneity, some dynamic specifications could be mistakenly interpreted as evidence for a large and persistent increase in spending. Our imputation estimator eliminates unstable patterns found across such specifications. Finally, we illustrate the under-identification problem with the fully dynamic specification: the dynamic path of estimates is very sensitive to the choice of leads to drop.

Our findings deliver several insights for the macroeconomics literature. While commonly used estimates of the quarterly MPX covering all expenditures range from 50 to 90% and estimates of the quarterly MPX for non-durable expenditure range from 15 to 25%,[2] our estimates, when appropriately rescaled, are about half as large, at 25–37% for the MPX one quarter after tax rebate receipt for all expenditures and 8–11% for non-durables. Using the scaling methodology of Laibson *et al.* (2022), we estimate that the model-consistent, or "notional", MPC in the quarter following the tax rebate ranges between 7.8 and 11.4%, compared with 15.9–23.4% in the original estimation of Broda and Parker (2014b). Furthermore, our preferred estimates are much more short-lived than benchmark estimates, falling to a statistical zero beyond the first

2. Broda and Parker (2014b), Parker *et al.* (2013) and Johnson *et al.* (2006) estimate different versions of the MPX out of tax rebates. Laibson *et al.* (2022), Kaplan and Violante (2022) and Di Maggio *et al.* (2020) provide recent reviews of the literature on the estimation of the marginal propensity to spend and consume.

month after receiving the tax rebate. Thus, our new estimates imply that fiscal stimulus may be less potent than predicted by leading macroeconomic models targeting benchmark estimates.[3]

For convenient application of our results, we supply a Stata command, `did_imputation`, which implements the imputation estimator and inference for it in a computationally efficient way. Our command handles a variety of practicalities which are also covered by our theoretical results, such as time-varying covariates, triple-difference designs, and repeated cross-sections. We also provide a second command, `event_plot`, for producing "event-study plots" that visualize the estimates with both our estimator and the alternative ones.

Our paper contributes to a growing methodological literature on event studies. To the best of our knowledge, our paper is the first and only one to characterize the under-identification and spurious identification of long-run treatment effects that arise in conventional implementations of event-study designs. The negative weighting problem has received more attention. It was first shown by de Chaisemartin and D'Haultfœuille (2015, Supplement 1). The earlier manuscript of our paper (Borusyak and Jaravel, 2018) independently pointed it out and additionally explained how it arises because of forbidden comparisons and why it affects long-run effects in particular, which we now discuss in Section 3.3 below. The issue has since been further investigated by Goodman-Bacon (2021), Strezhnev (2018), and de Chaisemartin and D'Haultfœuille (2020), while Sun and Abraham (2021) have shown similar problems with dynamic specifications. Sun and Abraham (2021) and Roth (2022) have further uncovered problems with conventional pre-trend tests, and Schmidheiny and Siegloch (2023) have characterized the problems which arise from binning multiple lags and leads in dynamic specifications. Besides being the first to point out some of these issues, our paper provides a unifying econometric framework which explicitly relates these issues to the conflation of the target estimand and the underlying identification assumptions.

Several papers have proposed ways to address these problems, introducing estimators that remain valid when treatment effects can vary arbitrarily (Cengiz *et al.*, 2019; Marcus and Sant'Anna, 2020; Callaway and Sant'Anna, 2021; Sun and Abraham, 2021; de Chaisemartin and D'Haultfœuille, 2022). An important limitation of these robust estimators is that their efficiency properties are not known.[4] A key contribution of our paper is to derive a practical, robust, and finite-sample efficient estimator from first principles. We show that this estimator takes a particularly transparent form under unrestricted treatment-effect heterogeneity, while our construction also yields efficiency when some restrictions on treatment effects are imposed. By clearly separating the testing of underlying assumptions from the estimation step imposing these assumptions, we simultaneously increase estimation efficiency and avoid problems with inference after pre-testing under spherical errors. Our estimator uses all pre-treatment periods for

3. Orchard *et al.* (2023) apply our imputation estimator to the Parker *et al.* (2013) quarterly data, covering the full consumption basket, and also obtain estimates around half as large as in the original study. Our analysis complements their results since, thanks to the high-frequency data, it allows us to investigate the dynamics of the effect and explain the source of the bias of conventional approaches. See also Baker *et al.* (2022) for evidence that using robust event-study estimation methods matters in other empirical contexts.

4. There are three notable exceptions. Marcus and Sant'Anna (2020) consider a two-stage generalized method of moments (GMM) estimator and establish its semi-parametric efficiency under heteroskedasticity in a large-sample framework with a fixed number of periods. However, they find this estimator to be impractical, as it involves many moments, for example, almost as many as the number of observations in the application they consider. Second, Roth and Sant'Anna (2023) characterize the efficient DiD estimator which leverages random timing of the treatment, rather than a more conventional parallel-trend assumption, as we do. Finally, Harmon (2022) builds on our framework to characterize the efficiency properties of DiD estimators when error terms follow a random walk—the opposite case from our benchmark analysis of efficiency which imposes no serial correlation of errors. In Supplementary Appendix A.5, we generalize our results to intermediate cases, allowing for models of heteroskedasticity and serial correlation.

imputation, as appropriate under the standard DiD assumptions, while alternative estimators use more limited information.[5]

In the MPX application, we find large gains of our imputation estimator: the confidence interval is about 50% longer for each week relative to the rebate for de Chaisemartin and D'Haultfœuille (2022), and 2–3.5 times longer for Sun and Abraham (2021) (which without extra controls are equivalent to the two versions of the Callaway and Sant'Anna (2021) estimator). We confirm these gains in a simulation study, finding that the standard deviations of alternative robust estimators are 1.3–3.6 times higher with spherical errors, and that these gains are generally preserved under heteroskedasticity and serial correlation of errors.

Finally, our paper is related to a nascent literature that develops robust estimators similar to the imputation estimator. To the best of our knowledge, this idea has been first proposed for factor models (Gobillon and Magnac, 2016; Xu, 2017). Athey *et al.* (2021) consider a general class of "matrix-completion" estimators for panel data that first impute untreated potential outcomes by regularized factor- and fixed-effects models and then average over the implied treatment-effect estimates. The imputation idea has been explicitly applied to fixed-effect estimators in event studies by Liu *et al.* (2022), Gardner (2021), Thakral and Tô (2022), and Gardner *et al.* (2023). Specifically, the counterfactual estimator of Liu *et al.* (2022), the two-stage estimator of Gardner (2021), Thakral and Tô (2022), and Gardner *et al.* (2023), and a version of the matrix-completion estimator from Athey *et al.* (2021) without factors or regularization coincide with the imputation estimator in our model for the specific class of estimands their papers consider. Relative to these papers, we make four contributions: we derive a general imputation estimator from first principles, show its efficiency, provide tools for valid asymptotic inference when unit fixed effects are included, and show its robustness to pre-testing. Subsequently to our work, Wooldridge (2021) derives a two-way Mundlak estimator, which is also equivalent to the imputation estimator for a restricted class of estimands in complete panels with controls that are not allowed to change over time (but that may have time-varying effects). The robustness and efficiency properties of our estimator are not limited to those situations.

2. SETTING

We consider estimation of causal effects of a binary treatment D_{it} on an outcome Y_{it} in a panel of units i and periods t. We focus on "staggered rollout" designs in which being treated is an absorbing state. For each unit there is an event date E_i when D_{it} switches from 0 to 1 forever: $D_{it} = \mathbf{1}[K_{it} \geq 0]$, where $K_{it} = t - E_i$ is the number of periods since the event date ("horizon"). Some units may never be treated, denoted by $E_i = \infty$. Units with the same event date are referred to as a cohort.

We do not make any random sampling assumptions and work with a set of observations $it \in \Omega$ of total size N, which may or may not form a complete panel. We similarly view the event date for each unit, and therefore all treatment indicators, as fixed. We define the set of treated observations by $\Omega_1 = \{it \in \Omega\colon\ D_{it} = 1\}$ of size N_1 and the set of untreated (*i.e.* never-treated and not-yet-treated) observations by $\Omega_0 = \{it \in \Omega\colon\ D_{it} = 0\}$ of size N_0.[6]

5. This efficiency gain relative to de Chaisemartin and D'Haultfœuille (2022) and Sun and Abraham (2021) is obtained without stronger assumptions. The Callaway and Sant'Anna (2021) assumptions are also equivalent to ours when there is only one period before any unit is treated and there are no covariates (see Marcus and Sant'Anna, 2020).

6. Viewing the set of observations and event times as non-stochastic is not essential. In Supplementary Appendix A.1, we show how this framework can be derived from one in which both are stochastic, by appropriate conditioning. Our conditional framework avoids random sampling assumptions made in other work on DiD designs (*e.g.* de Chaisemartin and D'Haultfœuille, 2020; Callaway and Sant'Anna, 2021; Sun and Abraham, 2021).

We denote by $Y_{it}(0)$ the period-t stochastic potential outcome of unit i if it is never treated. Causal effects on the treated observations $it \in \Omega_1$ are denoted by $\tau_{it} = \mathbb{E}[Y_{it} - Y_{it}(0)]$. We suppose a researcher is interested in a statistic which sums or averages treatment effects $\tau = (\tau_{it})_{it\in\Omega_1}$ over the set of treated observations with pre-specified non-stochastic weights $w_1 = (w_{it})_{it\in\Omega_1}$ that can depend on treatment assignment and timing, but not on realized outcomes:

Estimation target. $\tau_w = \sum_{it\in\Omega_1} w_{it}\tau_{it} \equiv w_1'\tau$.

For notation brevity, we consider scalar estimands.

Different weights are appropriate for different research questions. The researcher may be interested in the overall ATT, formalized by $w_{it} = 1/N_1$ for all $it \in \Omega_1$. In event-study analyses, a common estimand is the average effect h periods since treatment for a given horizon $h \geq 0$: $w_{it} = \mathbf{1}[K_{it} = h]/|\Omega_{1,h}|$ for $\Omega_{1,h} = \{it: \ K_{it} = h\}$. Our approach also allows researchers to specify target estimands that place unequal weights on units within the same cohort-by-horizon cell. For example, one may be interested in weighting units by their size, or in estimating a "balanced" version of horizon-average effects: the ATT at horizon h computed only for the subset of units also observed at horizon h', such that the gap between two or more estimates is not confounded by compositional differences. Finally, we do not require the w_{it} to add up to one; for example, a researcher may be interested in the difference between average treatment effects at different horizons or across some groups of units (*e.g.* women and men), corresponding to $\sum_{it\in\Omega_1} w_{it} = 0$.[7]

To identify τ_w, we consider three assumptions. We start with the parallel-trends assumption, which imposes a TWFE model on the untreated potential outcomes.

Assumption 1 (Parallel trends). *There exist non-stochastic* α_i *and* β_t *such that* $\mathbb{E}[Y_{it}(0)] = \alpha_i + \beta_t$ *for all* $it \in \Omega$.[8]

An equivalent formulation requires $\mathbb{E}[Y_{it}(0) - Y_{it'}(0)]$ to be the same across units i for all periods t and t' (whenever it and it' are observed).

Parallel-trend assumptions are standard in DiD designs, but their details may vary. First, we impose the TWFE model on the entire sample. Although weaker assumptions can be sufficient for identification of τ_w (*e.g.* Callaway *et al.*, 2021), those alternative restrictions depend on the realized treatment timing. Since parallel trends is an assumption on *potential* outcomes, we prefer its stronger version which can be made *a priori*.[9] Moreover, Assumption 1 can be tested by using pre-treatment data, while minimal assumptions cannot. Second, we impose Assumption 1 at the unit level, while sometimes it is imposed on cohort-level averages. Our approach is in line with the practice of including unit, rather than cohort, FEs in DiD analyses and allows us to avoid

7. More broadly, the choice of weights allows for estimation of treatment-effect heterogeneity by observed characteristics R_{it}. Indeed, the slope of the linear projection of τ_{it} on some observable R_{it} (which may or may not be time-varying) is a weighted sum of treatment effects, $\sum_{it\in\Omega_1} w_{it}\tau_{it}$ for $w_{it} = (R_{it} - \bar{R})/\sum_{js\in\Omega_1}(R_{js} - \bar{R})^2$ and $\bar{R} = \frac{1}{|\Omega_1|}\sum_{js\in\Omega_1} R_{js}$. The same logic generalizes when R_{it} is a vector, via the Frisch–Waugh–Lowell theorem. This approach also allows for tests of restrictions on treatment-effect heterogeneity, for example, to assess whether ATTs vary across time horizons.

8. In estimation, we will set the fixed effect of either one unit or one period to zero, such as $\beta_1 = 0$. This is without loss of generality, since the TWFE model is otherwise over-parameterized.

9. Specifically, Assumption 4 in Callaway and Sant'Anna (2021) requires that the TWFE model only holds for all treated observations ($D_{it} = 1$), observations directly preceding the treatment onset ($K_{it} = -1$), and in all periods for never-treated units. Similarly, Goodman-Bacon (2021) proposes to impose parallel trends on a "variance-weighted" average of units, as the weakest assumption under which static specifications we discuss in Section 3 identify some average of causal effects. While technically weaker, this assumption may be hard to justify *ex ante* without imposing parallel trends on all units as it is unlikely that non-parallel trends will cancel out by averaging.

biases in incomplete panels where the composition of units changes over time. Moreover, we show in Supplementary Appendix A.2 that, under random sampling and without compositional changes, assumptions on cohort-level averages imply Assumption 1.

Our framework extends immediately to richer models of $Y_{it}(0)$:

Assumption 1′ (General model of $Y(0)$). *For all $it \in \Omega$, $\mathbb{E}[Y_{it}(0)] = A'_{it}\lambda_i + X'_{it}\delta$, where λ_i is a vector of unit-specific nuisance parameters, δ is a vector of nuisance parameters associated with common covariates, and A_{it} and X_{it} are known non-stochastic vectors.*

The first term in this model of $Y_{it}(0)$ nests unit FEs, but also allows to interact them with some observed covariates unaffected by the treatment status, for example, to include unit-specific trends. This term looks similar to a factor model, but differs in that regressors A_{it} are observed. The second term nests period FEs but additionally allows any time-varying covariates, that is, $X'_{it}\delta = \beta_t + \tilde{X}'_{it}\tilde{\delta}$. In Supplementary Appendix A.1, we clarify that X_{it} have to be unaffected by treatment and strictly exogenous to be included in the specification.

We next rule out anticipation effects, that is, the causal effects of being treated in the future on current outcomes (*e.g.* Abbring and Van den Berg, 2003):

Assumption 2 (No-anticipation effects). *$Y_{it} = Y_{it}(0)$ for all $it \in \Omega_0$.*

Assumptions 1 and 2 together imply that the observed outcomes Y_{it} for untreated observations follow the TWFE model. It is straightforward to weaken this assumption, for example, by allowing anticipation for some k periods before treatment: this simply requires redefining event dates to earlier ones. However, some form of this assumption is necessary for DiD identification, as there would be no reference periods for treated units otherwise.

Finally, researchers sometimes impose restrictions on causal effects, explicitly or implicitly. For instance, τ_{it} may be assumed to be homogeneous for all units and periods, or only depend on the number of periods since treatment (but be otherwise homogeneous across units and calendar periods). We will consider such restrictions as a possible auxiliary assumption:

Assumption 3 (Restricted causal effects). *$B\tau = 0$ for a known $M \times N_1$ matrix B of full row rank.*

It will be more convenient for us to work with an equivalent formulation of Assumption 3, based on $N_1 - M$ free parameters driving treatment effects rather than M restrictions on them:

Assumption 3′ (Model of causal effects). *$\tau = \Gamma\theta$, where θ is a $(N_1 - M) \times 1$ vector of unknown parameters and Γ is a known $N_1 \times (N_1 - M)$ matrix of full column rank.*

Assumption 3′ imposes a parametric model of treatment effects. For example, the assumption that treatment effects all be the same, $\tau_{it} \equiv \theta_1$, corresponds to $N_1 - M = 1$ and $\Gamma = (1, \ldots, 1)'$. Conversely, a "null model" $\tau_{it} \equiv \theta_{it}$ that imposes no restrictions is captured by $M = 0$ and $\Gamma = \mathbb{I}_{N_1}$.

If restrictions on the treatment effects are implied by economic theory, imposing them will increase estimation power. Often, however, such restrictions are implicitly imposed without an *ex ante* justification, but just because they yield a simple model for the outcome. We will show in Section 3 how estimators that rely on this assumption can fail to estimate reasonable averages of treatment effects, let alone the specific estimand τ_w, when the assumption is violated.[10]

10. We view the null Assumption 3 as a conservative default. We note, however, that this makes the assumptions inherently asymmetric in that they impose restrictive models on potential control outcomes $Y_{it}(0)$ (Assumption 1), but not on treatment effects τ_{it}. This asymmetry reflects the standard practice in staggered rollout DiD designs and is natural when the structure of treatment effects is *ex ante* unknown, while our framework also accommodates the case

While we formulated our setting for staggered-adoption DiD designs with binary treatments in panel data, our framework applies without change in many related research designs. In *repeated cross-sections*, a different random sample of units i (*e.g.* individuals) from the same groups $g(i)$ (*e.g.* regions) is observed in each period. Unit FEs are not possible to include but can be replaced with group FEs in Assumption 1′: $\mathbb{E}[Y_{it}(0)] = \alpha_{g(i)} + \beta_t$. In *triple-differences designs*, the data have two dimensions in addition to periods, for example, i corresponds to a pair of region $j(i)$ and demographic group $g(i)$. Assumption 1′ can be specified as $\mathbb{E}[Y_{it}(0)] = \alpha_{j(i)g(i)} + \alpha_{j(i)t} + \alpha_{g(i)t}$.[11] With *non-binary treatment intensity*, our setting applies if each unit is observed untreated before E_i and treated with heterogenous intensity $R_{it} \neq 0$ (that may or may not vary over time) from period E_i. Assumptions 1 and 2 can apply, and the researcher can consider estimands such as the "ATT per unit of intensity", $\frac{1}{|\Omega_1|}\mathbb{E}\left[\sum_{it\in\Omega_1}(Y_{it} - Y_{it}(0))/R_{it}\right]$, by setting w_{it} proportionally to $1/R_{it}$. The challenges we describe in Section 3 for standard staggered DiDs and the solutions of Section 4 directly apply in all of these cases.[12]

3. CHALLENGES PERTAINING TO CONVENTIONAL PRACTICE

In this section, we first introduce the common TWFE regressions with restricted treatment-effect heterogeneity that have traditionally been used in DiD designs. We then discuss several estimation challenges that pertain to these specifications, including under-identification in certain dynamic specifications, negative weighting, and spurious identification of long-run causal effects. We conclude the section by discussing how our framework also relates to other problems that have been pointed out by Roth (2022) and Sun and Abraham (2021).

3.1. *Conventional restrictive specifications in staggered-adoption DiD*

Causal effects in staggered-adoption DiD designs have traditionally been estimated via OLS regressions with TWFEs, using specifications that implicitly restrict treatment effect heterogeneity across units. While details may vary, the following specification covers many studies:

$$Y_{it} = \tilde{\alpha}_i + \tilde{\beta}_t + \sum_{\substack{h=-a \\ h\neq -1}}^{b-1} \tau_h \mathbf{1}[K_{it} = h] + \tau_{b+}\mathbf{1}[K_{it} \geq b] + \varepsilon_{it}. \tag{2}$$

Here, $\tilde{\alpha}_i$ and $\tilde{\beta}_t$ are the unit and period ("two-way") fixed effects, $a \geq 0$ and $b \geq 0$ are the numbers of included "leads" and "lags" of the event indicator, respectively, and ε_{it} is the error term. The first lead, $\mathbf{1}[K_{it} = -1]$, is often excluded as a normalization, while the coefficients on the other leads (if present) are interpreted as measures of "pre-trends", and the hypothesis that $\tau_{-a} = \cdots = \tau_{-2} = 0$ is tested visually or statistically. Conditionally on this test passing, the

where the researcher is willing to impose structure. Restrictions on treatment effects, when appropriate, are also useful for external validity: unless some structure is imposed on treatment effects, one cannot use estimates from past data to inform future policy, for instance extending a given treatment to currently untreated units. However, one can use our framework without restrictions to learn *about the structure of treatment effects, for example, whether they vary across cohorts for each horizon*.

11. Another variation is when the outcome is measured in a single period but across *two cross-sectional dimensions*, such as regions i and birth cohorts g, with the treatment implemented in a set of regions for the cohorts born after some cut-off period E_i (*e.g.* Hoynes *et al.*, 2016). Then one may write $\mathbb{E}\left[Y_{ig}(0)\right] = \alpha_i + \beta_g$.

12. This is also the case of *non-staggered DiD* designs, in which units receive treatment in a single period or never. Our insights in Sections 3 and 4 are still relevant if continuous covariates or unit-specific trends are included (see Sant'Anna and Zhao, 2020; Wolfers, 2006 for related ideas).

coefficients on the lags are interpreted as a dynamic path of causal effects: at $h = 0, \ldots, b-1$ periods after treatment and, in the case of τ_{b+}, at longer horizons binned together. We will refer to this specification as "*dynamic*" (as long as $a + b > 0$) and, more specifically, "*fully dynamic*" if it includes all available leads and lags except $h = -1$, or "*semi-dynamic*" if it includes all lags but no leads.

Viewed through the lens of the Section 2 framework, these specifications make implicit assumptions on untreated potential outcomes, anticipation and treatment effects, and the estimand of interest. First, they make Assumption 1 but, for $a > 0$, do not fully impose Assumption 2, allowing for anticipation effects for a periods before treatment.[13] Typically this is done as a means to *test* Assumption 2 rather than to *relax* it, but the resulting specification is the same. Second, equation (2) imposes strong restrictions on causal effect heterogeneity (Assumption 3), with treatment (and anticipation) effects assumed to only vary by horizon h and not across units and periods otherwise. Most often, this is done without an *a priori* justification. If the lags are binned into the term with τ_{b+}, the effects are further assumed to be time-invariant once b periods have elapsed since the event. Finally, dynamic specifications do not explicitly define the estimands τ_h as particular averages of heterogeneous causal effects, even though researchers often consider that effects may vary across observations, as evidenced by a literature on the interpretation of OLS estimands going back to at least Angrist (1998) and Humphreys (2009).

Besides dynamic specifications, equation (2) also nests a very common specification used when a researcher is interested in a single parameter summarizing all causal effects. With $a = b = 0$, we have the "*static*" specification in which a single treatment indicator is included:

$$Y_{it} = \tilde{\alpha}_i + \tilde{\beta}_t + \tau^{\text{static}} D_{it} + \varepsilon_{it}. \tag{3}$$

In line with our Section 2 setting, the static equation imposes the parallel trends and no-anticipation Assumptions 1 and 2. However, it also makes a particularly strong version of Assumption 3—that all treatment effects are the same. Moreover, the target estimand is again not written out as an explicit average of potentially heterogeneous causal effects.

In the rest of this section, we turn to the challenges associated with OLS estimation of equations (2) and (3). We explain how these issues result from the conflation of the target estimand, Assumptions 2 and 3, providing a new and unified perspective on the problems of static and dynamic specifications with restricted treatment-effect heterogeneity.

3.2. *Under-identification of the fully dynamic specification*

The first problem pertains to fully dynamic specifications and arises because a strong enough Assumption 2 is not imposed. We show that those specifications are under-identified if there is no never-treated group:

Proposition 1. *If there are no never-treated units, the path of $\{\tau_h\}_{h \neq -1}$ coefficients is not point-identified in the fully dynamic specification. In particular, for any $\kappa \in \mathbb{R}$, the path $\{\tau_h + \kappa(h+1)\}$ fits the data equally well, with the fixed-effect coefficients appropriately modified.*

Proof. All proofs are given in Supplementary Appendix B. □

13. One can alternatively view this specification as imposing Assumption 2 but making a weaker Assumption 1 which includes some pre-trends into $Y_{it}(0)$. This difference in interpretation is immaterial for our results.

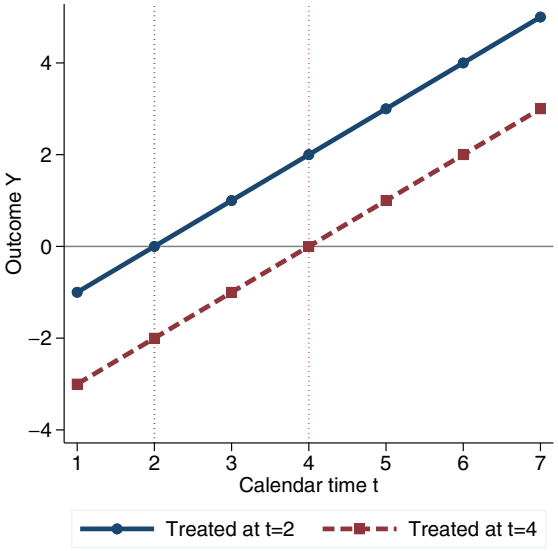

FIGURE 1
Under-identification of fully dynamic specification

Notes: This figure shows the evolution of outcomes over seven periods for two units (or cohorts), for the illustrative example of Section 3.2. Vertical lines mark the periods in which the two units are first treated.

To illustrate this result with a simple example, Figure 1 plots the outcomes for a simulated dataset with two units (or equal-sized cohorts), one treated at $t = 2$ and the other at $t = 4$. Both units exhibit linear growth in the outcome, starting from different levels. There are two interpretations of these dynamics. First, treatment could have no impact on the outcome, in which case the level difference corresponds to the unit FEs, while trends are just a common feature of the environment, through period FEs. Alternatively, note that the outcome equals the number of periods since the event for both groups and all time periods: it is zero at the moment of treatment, negative before, and positive after. A possible interpretation is that the outcome is entirely driven by causal effects and anticipation of treatment. Thus, one cannot hope to distinguish between unrestricted dynamic causal effects and a combination of unit effects and time trends.[14]

The problem may be important in practice, as statistical packages may resolve this collinearity by dropping an arbitrary unit or period indicator. Some estimates of $\{\tau_h\}$ would then be produced, but because of an arbitrary trend in the coefficients they may suggest a violation of parallel trends even when the specification is in fact correct, that is, Assumptions 1 and 2 hold and there is no heterogeneity of treatment effects for each horizon (Assumption 3).

To break the collinearity problem, stronger restrictions on anticipation effects, and thus on Y_{it} for untreated observations, have to be introduced. One could consider imposing minimal restrictions on the specification that would make it identified. In typical cases, only a linear trend in $\{\tau_h\}$ is not identified in the fully dynamic specification, while non-linear paths cannot

14. Formally, the problem arises because a linear time trend t and a linear term in the cohort E_i (subsumed by the unit FEs) can perfectly reproduce a linear term in horizon $K_{it} = t - E_i$. Therefore, a complete set of treatment leads and lags, which is equivalent to the horizon FEs, is collinear with the unit and period FEs.

TABLE 1
Two-unit, three-period example

$\mathbb{E}[Y_{it}]$	$i = A$	$i = B$
$t = 1$	α_A	α_B
$t = 2$	$\alpha_A + \beta_2 + \tau_{A2}$	$\alpha_B + \beta_2$
$t = 3$	$\alpha_A + \beta_3 + \tau_{A3}$	$\alpha_B + \beta_3 + \tau_{B3}$
Event date	$E_i = 2$	$E_i = 3$

Notes: This table shows the evolution of expected outcomes over three periods for two units (or cohorts), for the illustrative example of Proposition 3. Without loss of generality, we normalize $\beta_1 = 0$.

be reproduced with unit and period fixed effects. Therefore, just one additional normalization, for example, $\tau_{-a} = 0$ in addition to $\tau_{-1} = 0$, breaks multi-collinearity.[15]

However, minimal identified models rely on *ad hoc* identification assumptions which are *a priori* unattractive. For instance, just imposing $\tau_{-a} = \tau_{-1} = 0$ means that anticipation effects are assumed away 1 and a periods before treatment, but not in other pre-periods. This assumption therefore depends on the realized event times. Instead, a systematic approach is to impose the assumptions—some forms of no-anticipation effects and parallel trends—that the researcher has an *a priori* argument for and which motivated the use of DiD. Such assumptions also give much stronger identification power.[16]

3.3. *Negative weighting in the static regression*

We now show how, by imposing Assumption 3 instead of specifying the estimation target, the static TWFE specification does not identify a reasonably weighted average of heterogeneous treatment effects: the underlying weights may be negative, particularly for the long-run causal effects. The issues we discuss here also arise in dynamic specifications that bin multiple lags together.

First, we note that, if the parallel-trends and no-anticipation assumptions hold, the static specification identifies *some* weighted average of treatment effects.[17]

Proposition 2. *If Assumptions 1 and 2 hold, then the estimand of the static specification in equation (3) satisfies* $\tau^{static} = \sum_{it \in \Omega_1} w_{it}^{static} \tau_{it}$ *for some weights* w_{it}^{static} *that do not depend on the outcome realizations and add up to one,* $\sum_{it \in \Omega_1} w_{it}^{static} = 1$.

The underlying weights w_{it}^{static} can be computed from the data using the Frisch–Waugh–Lovell theorem (see equation (17) in the proof of Proposition 2) and only depend on the timing of treatment for each unit and the set of observed units and periods. The static specification's estimand, however, cannot be interpreted as a *proper* weighted average, as some weights can be negative, which we illustrate with a simple example.

Proposition 3. *Suppose Assumptions 1 and 2 hold and the data consist of two units (or equal-sized cohorts), A and B, treated in periods 2 and 3, respectively, both observed in periods* $t = 1, 2, 3$ *(as shown in Table 1). Then the estimand of the static specification (3) can be expressed as* $\tau^{static} = \tau_{A2} + \frac{1}{2}\tau_{B3} - \frac{1}{2}\tau_{A3}$.

15. Additional collinearity arises, for example, when treatment is staggered but happens at periodic intervals.
16. Our suggestion to impose identification assumptions at the estimation stage does not mean that those assumptions should not also be tested; we discuss testing in detail in Section 4.4.
17. This result was previously stated in Theorem 1 of de Chaisemartin and D'Haultfœuille (2020) for general designs, and later in Appendix C of Borusyak and Jaravel (2018) for staggered-adoption designs.

This example illustrates the severe short-run bias of the static specification: the long-run causal effect, corresponding to the early treated unit A and the late period 3, enters with a negative weight $(-1/2)$. Thus, larger long-run effects make the coefficient smaller.

This problem results from what we call "forbidden comparisons" performed by the static specification. Recall that the original idea of DiD estimation is to compare the evolution of outcomes over some time interval for the units which got treated during that interval relative to a reference group of units which didn't, identifying the period FEs. In the Proposition 3 example, such an "admissible" comparison is between units A and B in periods 2 and 1, $(Y_{A2} - Y_{A1}) - (Y_{B2} - Y_{B1})$. However, panels with staggered treatment timing also lend themselves to a second type of comparisons—which we label "forbidden"—in which the reference group has been treated throughout the relevant period. For units in this group, the treatment indicator D_{it} does not change over the relevant period, and so the restrictive specification uses them to identify period FEs, too. The comparison between units B and A in periods 3 and 2, $(Y_{B3} - Y_{B2}) - (Y_{A3} - Y_{A2})$, in Proposition 3 is a case in point. While a comparison like this is appropriate and increases efficiency when treatment effects are homogeneous (which the static specification was designed for), forbidden comparisons are problematic under treatment-effect heterogeneity. For instance, subtracting $(Y_{A3} - Y_{A2})$ not only removes the gap in period FEs, $\beta_3 - \beta_2$, but also deducts the evolution of treatment effects $\tau_{A3} - \tau_{A2}$, placing a negative weight on τ_{A3}. The restrictive specification leverages comparisons of both types and estimates the treatment effect by $\hat{\tau}^{static} = (Y_{B2} - Y_{A2}) - \frac{1}{2}(Y_{B1} - Y_{A1}) - \frac{1}{2}(Y_{B3} - Y_{A3})$.[18]

Fundamentally, this problem arises because the specification imposes very strong restrictions on treatment-effect homogeneity, that is, Assumption 3, instead of acknowledging the heterogeneity and specifying a particular target estimand (or perhaps a class of estimands that the researcher is indifferent between).

With a large number of never-treated units or a large number of periods before any unit is treated (relative to other units and periods), our setting becomes closer to a classical non-staggered DiD design, and therefore negative weights disappear, as our next result illustrates:

Proposition 4. *Suppose all units are observed for all periods $t = 1, \dots, T$ and the earliest treatment happens at $E_{\text{first}} > 1$. Let N_1^* be the number of observations for never-treated units before period E_{first} and N_0^* be the number of untreated observations for ever-treated units since E_{first}. Then there is no negative weighting, that is, $\min_{it \in \Omega_1} w_{it}^{\text{static}} \geq 0$, if and only if $N_1^* \geq N_0^*$.*[19]

Even when weights are non-negative, they may remain highly unequal and diverge from the estimands that the researcher is interested in. Our preferred strategy is therefore to commit to the estimation target and explicitly allow for treatment-effect heterogeneity, except when some form of Assumption 3 is *ex ante* appropriate.

18. The proof of Proposition 2 shows why long-run effects in particular are subject to the negative weights problem. In general, negative weights arise for the treated observations, for which the residual from an auxiliary regression of D_{it} on the two-way FEs is negative. de Chaisemartin and D'Haultfœuille (2020) show that, in complete panels, the unit FEs are higher for early treated units (which are observed treated for a larger shares of periods) and period FEs are higher for later periods (in which a larger shares of units are treated). The early treated units observed in later periods correspond to the long-run effects.

19. N_1^* and N_0^* respectively correspond to the numbers of admissible and forbidden 2×2 DiD comparisons available for the earliest-treated units in the latest period T. The gap between them drives negative weights with complete panels, as in Strezhnev (2018, Proposition 1).

3.4. *Spurious identification of long-run effects in dynamic specifications*

Another consequence of inappropriately imposing Assumption 3 concerns estimation of long-run causal effects. Conventional dynamic specifications (except those subject to the under-identification problem) yield *some* estimates for all τ_h coefficients. Yet, for large enough h, no averages of treatment effects are identified under Assumptions 1 and 2 with unrestricted treatment-effect heterogeneity. Therefore, estimates from restrictive specifications are fully driven by unwarranted extrapolation of treatment effects across observations and may not be reliable, unless strong *ex ante* reasons for Assumption 3 exist.

This issue is well illustrated in the example of Proposition 3. To identify the long-run effect τ_{A3} under Assumptions 1 and 2, one needs to form an admissible DiD comparison, of the outcome growth over some period between unit A and another unit not yet treated in period 3. However, by period 3 both units have been treated. Mechanically, this problem arises because the period fixed effect β_3 is not identified separately from the treatment effects τ_{A3} and τ_{B3} in this example, absent restrictions on treatment effects. Yet, the semi-dynamic specification

$$Y_{it} = \tilde{\alpha}_i + \tilde{\beta}_t + \tau_0 \mathbf{1}[K_{it} = 0] + \tau_1 \mathbf{1}[K_{it} = 1] + \tilde{\varepsilon}_{it}$$

will produce an estimate $\hat{\tau}_1$ via extrapolation. Specifically, two different parameters, $\tau_{A3} - \tau_{B3}$ and τ_{A2}, are identified by comparing the two units in periods 2 or 3, respectively, with period 1. Therefore, when imposing homogeneity of short-run effects across units, $\tau_{A2} = \tau_{B3} \equiv \tau_0$, we estimate the long-run effect $\tau_{A3} \equiv \tau_1$ as the sum of $\tau_1 - \tau_0$ and τ_0:

$$\hat{\tau}_1 = [(Y_{A3} - Y_{A1}) - (Y_{B3} - Y_{B1})] + [(Y_{A2} - Y_{A1}) - (Y_{B2} - Y_{B1})].$$

However, when $\tau_{A2} \neq \tau_{B3}$, this estimator is biased.

In general, the gap between the earliest and the latest event times observed in the data provides an upper bound on the number of dynamic coefficients that can be identified without extrapolation of treatment effects. This result, which follows by the same logic of non-identification of the later period effects, is formalized by our next proposition:

Proposition 5. *Suppose there are no never-treated units and let* $\bar{H} = \max_i E_i - \min_i E_i$. *Then, for any non-negative weights* w_{it} *defined over the set of observations with* $K_{it} \geq \bar{H}$ *(that are not identically zero), the weighted sum of causal effects* $\sum_{it:\, K_{it} \geq \bar{H}} w_{it} \tau_{it}$ *is not identified by Assumptions 1 and 2.*[20]

Robust estimators, including the one we characterize in Section 4, can only be computed for identified estimands, never resulting in spurious estimates.

We finally note that the challenges described in Section 3 apply even if the sample is "trimmed" to a fixed window around the event time; see Supplementary Appendix A.3.

4. IMPUTATION-BASED ESTIMATION AND TESTING

To overcome the challenges affecting conventional practice, we now derive the robust and efficient estimator and show that it takes a particularly transparent "imputation" form when no

20. The requirement that the weights are non-negative rules out some estimands on the *gaps* between treatment effects for $K_{it} \geq \bar{H}$ which are in fact identified. For instance, adding period $t = 4$ to Table 1 example, the difference $\tau_{A4} - \tau_{B4}$ would be identified (by $(Y_{A4} - Y_{B4}) - (Y_{A1} - Y_{B1})$), even though neither τ_{A4} nor τ_{B4} is identified.

restrictions on treatment-effect heterogeneity are imposed. We then perform asymptotic analysis, establishing the conditions for the estimator to be consistent and asymptotically normal, derive conservative standard error estimates for it, and discuss appropriate pre-trend tests.

Throughout, we continue to suppose that the researcher chose the estimation target τ_w and assumed a model of $Y_{it}(0)$ (Assumption 1′) and no anticipation. Some model of treatment effects (Assumption 3) may also be assumed, although our main focus is on the null model, under which treatment-effect heterogeneity is unrestricted. Letting $\varepsilon_{it} = Y_{it} - \mathbb{E}[Y_{it}]$ for $it \in \Omega$, we thus have under Assumptions 1′, 2, and 3′:

$$Y_{it} = A_{it}'\lambda_i + X_{it}'\delta + D_{it}\Gamma_{it}'\theta + \varepsilon_{it}. \tag{4}$$

We assume throughout that $\tau_w = w_1'\Gamma\theta$ is identified. Supplementary Proposition A1 provides conditions for identification. First, we provide a general rank condition on the matrices of unit-specific and other covariates that requires that the covariate space of treated observations is spanned by that of the untreated ones. This assumption allows us to estimate from the untreated observations those nuisance parameters that are necessary to impute control outcomes of the treated observations, thus providing identification. Second, we derive specific conditions for the case where the parameter δ represents time fixed effects and A_{it} may vary over time but not across units (as with unit FEs and unit-specific linear trends). In this case, we show that there is identification if (1) the A_{it} are not collinear for any relevant unit and (2) there is at least one untreated unit at the end of the time period of interest.

4.1. *Efficient estimation*

For our efficiency result, we impose an additional assumption on the error variances:

Assumption 4 (Spherical errors). *Error terms ε_{it} are spherical, that is, homoskedastic and mutually uncorrelated across all $it \in \Omega$: $\mathbb{E}\left[\varepsilon\varepsilon'\right] = \sigma^2\mathbb{I}_N$.*

While this assumption is strong, our efficiency results also apply without change under dependence that is due to unit random effects, that is, if $\varepsilon_{it} = \eta_i + \tilde{\varepsilon}_{it}$ for $\tilde{\varepsilon}_{it}$ that satisfy Assumption 4 and for some η_i. Moreover, these results are straightforward to relax to any known form of heteroskedasticity or mutual dependence.[21] Under Assumption 4 and allowing for restrictions on causal effects, we have:

Theorem 1 (Efficient estimator). *Suppose Assumptions 1′, 2, 3′, and 4 hold. Then among the linear unbiased estimators of τ_w, the (unique) efficient estimator $\hat{\tau}_w^*$ can be obtained with the following steps:*

1. *Estimate θ by the OLS solution $\hat{\theta}^*$ from the regression (4) (where we assume that θ is identified).*
2. *Estimate the vector of treatment effects τ by $\hat{\tau}^* = \Gamma\hat{\theta}^*$.*
3. *Estimate the target τ_w by $\hat{\tau}_w^* = w_1'\hat{\tau}^*$.*

Moreover, this estimator $\hat{\tau}_w^*$ is unbiased for τ_w under Assumptions 1′, 2, and 3′ alone, even when error terms are not spherical.

21. For instance, if error terms are uncorrelated and have known variances σ_{it}^2 (up to a scaling factor), efficiency requires the estimation step (Step 1) of Theorems 1 and 2 to be performed with weights proportional to σ_{it}^{-2}. One example for this is when the data are aggregated from n_{it} individuals randomly drawn from group i in period t and spherical individual-level errors, in which case efficiency is obtained with weights proportional to n_{it}.

Under Assumptions 1′, 2, and 3′, regression (4) is correctly specified. Thus, this estimator for θ is unbiased by construction, and efficiency under spherical error terms is a direct consequence of the Gauss–Markov theorem. Moreover, OLS yields the most efficient estimator for any linear combination of θ, including $\tau_w = w_1'\Gamma\theta$. While assuming spherical errors may be unrealistic in practice, we think of this assumption as a natural conceptual benchmark to decide between the many unbiased estimators of τ_w.[22]

In the important special case of unrestricted treatment-effect heterogeneity, $\hat{\tau}_w^*$ has a useful "imputation" representation. The idea is to estimate the model of $Y_{it}(0)$ using the untreated observations $it \in \Omega_0$ and leverage it to impute $Y_{it}(0)$ for treated observations $it \in \Omega_1$. Then, observation-specific causal effect estimates can be averaged appropriately. Perhaps surprisingly, the estimation and imputation steps are identical regardless of the target estimand. Applying any weights to the imputed causal effects yields the efficient estimator for the corresponding estimand. We have:

Theorem 2 (Imputation representation for the efficient estimator). *With a null Assumption 3′ (that is, if $\Gamma = \mathbb{I}_{N_1}$), the unique efficient linear unbiased estimator $\hat{\tau}_w^*$ of τ_w from Theorem 1 can be obtained via an imputation procedure:*

1. *Within the untreated observations only ($it \in \Omega_0$), estimate the λ_i and δ (by $\hat{\lambda}_i^*, \hat{\delta}^*$) by OLS in*
$$Y_{it} = A_{it}'\lambda_i + X_{it}'\delta + \varepsilon_{it}. \tag{5}$$
2. *For each treated observation ($it \in \Omega_1$) with $w_{it} \neq 0$, set $\hat{Y}_{it}(0) = A_{it}'\hat{\lambda}_i^* + X_{it}'\hat{\delta}^*$ and $\hat{\tau}_{it}^* = Y_{it} - \hat{Y}_{it}(0)$ to obtain the estimate of τ_{it}.*
3. *Estimate the target τ_w by a weighted sum $\hat{\tau}_w^* = \sum_{it\in\Omega_1} w_{it}\hat{\tau}_{it}^*$.*

The imputation representation offers computational and conceptual benefits. First, it is computationally efficient as it only requires estimating a simple TWFE model, for which fast algorithms are available (Guimarães and Portugal, 2010; Correia, 2017). This is in contrast to the OLS estimator from Theorem 1, as equation (4) has regressors $\Gamma_{it}D_{it}$ in addition to the fixed effects, which are high-dimensional unless a low-dimensional model of treatment-effect heterogeneity is imposed.

Second, the imputation approach is intuitive and transparently links the parallel trends and no-anticipation assumptions to the estimator. Indeed, Imbens and Rubin (2015) write: "*At some level, all methods for causal inference can be viewed as imputation methods, although some more explicitly than others*" (p. 141). We formalize this statement in the next proposition, which shows that *any* estimator unbiased for τ_w can be represented in the imputation way, but the way of imputing the $Y_{it}(0)$ may be less explicit and no longer efficient.

Proposition 6 (Imputation representation for all unbiased estimators). *Under Assumptions 1′ and 2, any linear estimator $\hat{\tau}_w$ of τ_w that is unbiased under arbitrary treatment-effect heterogeneity (that is, a null Assumption 3) can be obtained via imputation:*

1. *For every treated observation, estimate expected untreated potential outcomes $A_{it}'\lambda_i + X_{it}'\delta$ by some unbiased linear estimator $\hat{Y}_{it}(0)$ using data from the untreated observations only.*
2. *For each treated observation, set $\hat{\tau}_{it} = Y_{it} - \hat{Y}_{it}(0)$.*

22. This benchmark appears natural as it parallels the Gauss–Markov theorem which also relies on spherical errors. In Monte Carlo simulations (Supplementary Appendix A.11), the estimator performs well even under deviations from spherical errors. In Supplementary Appendix A.5, we generalize the results to parametric models of heteroskedasticity and serial correlation, in the spirit of generalized least squares and relating to Wooldridge (2021).

3. *Estimate the target by a weighted sum* $\hat{\tau}_w = \sum_{it\in\Omega_1} w_{it}\hat{\tau}_{it}$.

This result establishes an imputation representation when treatment effects can vary arbitrarily. Proposition A2 in the Supplementary Appendix establishes that the imputation structure applies even when restrictions $\tau = \Gamma\theta$ are imposed, albeit with an additional step in which the weights w_1 defining the estimand are adjusted in a way that does not change τ_w under the imposed model.[23] In this sense, unbiased causal inference is equivalent to imputation in our framework.

4.2. *Asymptotic properties*

Having derived the linear unbiased estimator $\hat{\tau}^*_w$ for τ_w in Theorem 1 that is also efficient under spherical error terms, we now consider its asymptotic properties without imposing that assumption. We study convergence along a sequence of panels indexed by the sample size N, where randomness stems from the error terms ε_{it} only, as in Section 2. Our approach applies to asymptotic sequences where both the number of units and the number of time periods may grow, but the assumptions are least restrictive when the number of time periods remains constant or grows slowly, as in short panels.

Instead of assuming that error terms are spherical, we now assume that error terms are clustered by units i.

Assumption 5 (Clustered error terms). *Error terms ε_{it} are independent across units i and have bounded variance, $Var[\varepsilon_{it}] \leq \bar{\sigma}^2$ for $it \in \Omega$ uniformly.*

The key role in our results is played by the weights that the Theorem 1 estimator places on each observation. Since the estimator is linear in the observed outcomes Y_{it}, we can write it as $\hat{\tau}^*_w = \sum_{it\in\Omega} v^*_{it} Y_{it}$ with non-stochastic weights v^*_{it}, derived in Proposition A3 in the Supplementary Appendix.

We now formulate high-level conditions on the sequence of weight vectors that ensure consistency, asymptotic normality, and will later allow us to provide valid inference. These results apply to any unbiased linear estimator $\hat{\tau}_w = \sum_{it\in\Omega} v_{it} Y_{it}$ of τ_w, not just the efficient estimator $\hat{\tau}^*_w$ from Theorem 1—that is, if the respective conditions are fulfilled for the weights v_{it}, then consistency, asymptotic normality, and valid inference follow as stated. For the specific estimator $\hat{\tau}^*_w$ introduced above, we then provide sufficient low-level conditions for short panels.

First, we obtain consistency of $\hat{\tau}_w$ under a Herfindahl condition on the weights v that takes the clustering structure of error terms into account.

Assumption 6 (Herfindahl condition). *Along the asymptotic sequence, $\|v\|^2_{\mathrm{H}} \equiv \sum_i (\sum_{t;it\in\Omega} |v_{it}|)^2 \to 0$, for weights v_{it} in the unbiased linear estimator $\hat{\tau}_w = \sum_{it\in\Omega} v_{it} Y_{it}$.*

The condition on the clustered Herfindahl index $\|v\|^2_{\mathrm{H}}$ states that the sum of squared weights vanishes, where weights are aggregated by units. One can think of the inverse of the sum of squared weights, $n_H = \|v\|^{-2}_{\mathrm{H}}$, as a measure of effective sample size, which Assumption 6

23. As a special case, we can still write the efficient estimator from Theorem 1 as an imputation estimator from Theorem 2 with alternative weights v^*_1 on the imputed treatment effects. Supplementary Proposition A4 shows that these adjusted weights v^*_1 solve a quadratic variance-minimization problem with a linear constraint that preserves unbiasedness under Assumption 3. We also provide an explicit formula for the resulting weights in Supplementary Proposition A3.

requires to grow large along the asymptotic sequence. If it is satisfied, and variances are uniformly bounded, we obtain consistency of $\hat{\tau}_w$.[24]

Proposition 7 (Consistency of $\hat{\tau}_w$). *Under Assumptions 1′, 2, 3′, 5, and 6, $\hat{\tau}_w - \tau_w \xrightarrow{\mathcal{L}_2} 0$ for an unbiased linear estimator $\hat{\tau}_w$ of τ_w, such as $\hat{\tau}_w^*$ in Theorem 1.*

We note that the large number of unit-specific parameters $\{\lambda_i\}_i$ cannot generally be estimated consistently in panels with a small number of time periods, raising a potential incidental-parameters problem. However, consistency of $\hat{\tau}_w$ does not rely on consistency for unit-specific parameters, since our estimator averages over many units.

We next consider the asymptotic distribution of the estimator around the estimand.

Proposition 8 (Asymptotic normality). *If the assumptions of Proposition 7 hold, there exists $\kappa > 0$ such that $\mathbb{E}[|\varepsilon_{it}|^{2+\kappa}]$ is uniformly bounded, the weights are not too concentrated in the sense that $\sum_i (\sqrt{n_H} \sum_{t;it\in\Omega} |v_{it}|)^{2+\kappa} \to 0$, and the variance does not vanish, $\liminf n_H \sigma_w^2 > 0$ for $\sigma_w^2 = Var\left[\hat{\tau}_w\right]$, then we have that $\sigma_w^{-1}(\hat{\tau}_w - \tau_w) \xrightarrow{d} \mathcal{N}(0,1)$.*

This result establishes conditions under which the difference between estimator and estimand is asymptotically normal. Besides regularity, this proposition requires that the estimator variance σ_w^2 does not decline faster than $1/n_H$. It is violated if the clustered Herfindahl formula is too conservative: for instance, if the number of periods is growing along the asymptotic sequence while the within-unit over-time correlation of error terms remains small. Alternative sufficient conditions for asymptotic normality can be established in such cases, for example, along the lines of Footnote 24.

So far, we have formulated high-level conditions on the weights v_{it} of any linear unbiased estimator of τ_w. Supplementary Appendix A.7 presents low-level sufficient conditions for consistency and asymptotic normality of the imputation estimator $\hat{\tau}_w^*$ for the benchmark case of a panel with unit and period FEs, a fixed or slowly growing number of periods, and no restrictions on treatment effects. Unlike Propositions 7 and 8, these conditions are imposed directly on the weights w_1 chosen by the researcher, and not on the the implied weights v_{it}^*, such that the researcher can assess more directly whether the asymptotic approximation is likely to be precise. In particular, the estimator achieves consistency and asymptotic normality in the common case where the number of time periods is fixed, the size of all cohorts increases, the weights on treatment effects do not vary within the same period and cohort, and the sum of (absolute) weights is bounded. In addition, the sufficient conditions are also fulfilled when the number of periods grows slowly and when weights differ across observations within the same cohort and period, but not by too much. With covariates other than unit and period FEs, for example, with unit-specific linear trends, the general weight conditions in Assumption 6 and Proposition 8 can also be used to verify consistency and asymptotic normality. In those cases, the sufficient conditions are typically fulfilled for convex combinations of cohort-average treatment effects whenever the

24. The Herfindahl condition can be restrictive since it allows for a worst-case correlation of error terms within units. When such correlations are limited, other sufficient conditions may be more appropriate instead, such as $R(\sum_{it\in\Omega} v_{it}^2) \to 0$ with $R = \max_i(\text{largest eigenvalue of } \Sigma_i)/\bar{\sigma}^2$, where $\Sigma_i = (Cov\left[\varepsilon_{it}, \varepsilon_{is}\right])_{t,s}$. Here, R is a measure of the maximal joint covariation of all observations for one unit. If error terms are uncorrelated, then $R \leq 1$, since the maximal eigenvalue of Σ_i corresponds to the maximal variance of an error term ε_{it} in this case, which is bounded by $\bar{\sigma}^2$. An upper bound for R is the maximal number of periods for which we observe a unit, since the maximal eigenvalue of Σ_i is bounded by the sum of the variances on its diagonal.

size of cohorts grows sufficiently fast relatively to the number of periods (see Supplementary Appendix A.7).

4.3. *Conservative inference*

We next estimate the variance of $\hat{\tau}_w = \sum_{it\in\Omega} v_{it} Y_{it}$, which equals $\sigma_w^2 = \mathbb{E}\left[\sum_i (\sum_{t;it\in\Omega} v_{it}\varepsilon_{it})^2\right]$ with clustered error terms (Assumption 5). We start with the case where treatment effect heterogeneity is unrestricted (*i.e.* $\Gamma = \mathbb{I}$). As in de Chaisemartin and D'Haultfœuille (2020), exact inference becomes infeasible when treatment effects are heterogeneous, but conservative inference is possible. Following Section 4.2, the inference tools we propose apply to a generic linear unbiased estimator but we use them for the efficient estimator $\hat{\tau}_w^*$. Our strategy is to estimate individual error terms by some $\tilde{\varepsilon}_{it}$ and then use a plug-in estimator,

$$\hat{\sigma}_w^2 = \sum_i \left(\sum_{t;it\in\Omega} v_{it}\tilde{\varepsilon}_{it} \right)^2. \tag{6}$$

Estimating the error terms presents two challenges, which become apparent when we consider the benchmark choice $\tilde{\varepsilon}_{it} = \hat{\varepsilon}_{it}$ based on the regression residuals $\hat{\varepsilon}_{it} = Y_{it} - A_{it}'\hat{\lambda}_i^* - X_{it}'\hat{\delta}^* - D_{it}\hat{\tau}_{it}^*$ in the regression (4). The first challenge is the incidental-parameters problem in estimating λ_i. However, by using cluster-robust variance estimates, our inference does not suffer from this problem since the variance estimator $\hat{\sigma}_w^2$ does not rely on the consistent estimation of λ_i any more, similar to the insight of Stock and Watson (2008).

A second challenge arises from unrestricted treatment-effect heterogeneity. In Theorem 2, treatment effects are estimated by fitting the corresponding outcomes Y_{it} perfectly, with residuals $\hat{\varepsilon}_{it} \equiv 0$ for all treated observations. This issue is not specific to our estimation procedure: one generally cannot distinguish between τ_{it} and ε_{it} from observations of $Y_{it} = A_i'\lambda_i + X_{it}'\delta + \tau_{it} + \varepsilon_{it}$ for treated observations, making it impossible to produce unbiased estimates of σ_w^2 (see Lemma 1 in Kline *et al.* (2020) for a similar impossibility result).

While unbiased estimation of σ_w^2 is not possible, we show that this variance can be estimated conservatively. Our variance estimator is based on an auxiliary parsimonious model of treatment effects. We do not require this model to be correct, in the sense that inference is weakly asymptotically conservative under mis-specification. However, auxiliary models which better approximate τ_{it} will make confidence intervals tighter and closer to asymptotically exact. In the computation of $\hat{\sigma}_w^2$ we set $\tilde{\varepsilon}_{it}$ for the treated observations equal to the residuals of the auxiliary model. We require the model to be parsimonious, such that it does not overfit and the residuals include ε_{it}. When the model is incorrect, $\tilde{\varepsilon}_{it}$ also include a component due to the mis-specification of τ_{it}, leading to conservative inference.

We formalize the auxiliary model by considering estimators $\tilde{\tau}_{it}$ for each $it \in \Omega_1$ which satisfy two properties: (1) $\tilde{\tau}_{it}$ converges to *some* non-stochastic limit $\bar{\tau}_{it}$ and (2) if the auxiliary model is correct, $\bar{\tau}_{it} = \tau_{it}$. The following theorem presents conditions under which our construction yields asymptotically conservative inference:

Theorem 3 (Conservative clustered standard error estimates). *Assume that the assumptions of Proposition 7 hold, that the model of treatment effects is trivial ($\Gamma = \mathbb{I}$), that the estimates $\tilde{\tau}_{it}$ converge to some non-random $\bar{\tau}_{it}$ in the sense that $\|v\|_{\mathrm{H}}^{-2}\sum_i(\sum_{t;it\in\Omega_1} v_{it}(\tilde{\tau}_{it} - \bar{\tau}_{it}))^2 \xrightarrow{p} 0$, that $\hat{\delta}^*$ from Theorem 1 is sufficiently close to δ in the sense that $\|v\|_{\mathrm{H}}^{-2}\sum_i(\sum_{t;it\in\Omega} v_{it}X_{it}'(\hat{\delta}^* - \delta))^2 \xrightarrow{p} 0$, and that $|\tau_{it}|$, $|\bar{\tau}_{it}|$ and $\mathbb{E}\left[\varepsilon_{it}^4\right]$ are uniformly bounded and the weights are not too*

concentrated in the sense that $\sum_i (\frac{\sum_{t;it\in\Omega}|v_{it}|}{\|v\|_H})^4 \to 0$. *Then the variance estimate*

$$\hat{\sigma}_w^2 = \sum_i \left(\sum_{t;it\in\Omega} v_{it}\tilde{\varepsilon}_{it} \right)^2, \quad \tilde{\varepsilon}_{it} = Y_{it} - A_{it}'\hat{\lambda}_i^* - X_{it}'\hat{\delta}^* - D_{it}\tilde{\tau}_{it} \tag{7}$$

is asymptotically conservative: $\|v\|_H^{-2}(\hat{\sigma}_w^2 - \sigma_w^2 - \sigma_\tau^2) \xrightarrow{p} 0$, *where* $\sigma_\tau^2 = \sum_i (\sum_{t;D_{it}=1} v_{it}(\tau_{it} - \bar{\tau}_{it}))^2 \geq 0$. *If* $\bar{\tau}_{it} = \tau_{it}$ *for all* $it \in \Omega_1$, $\sigma_\tau^2 = 0$, *meaning that the variance estimate is asymptotically exact.*

The theorem shows that the proposed variance estimate addresses the two challenges laid out above. First, the estimates remain valid even though we may not be able to estimate the unit-specific parameters λ_i consistently. This is because unit-specific parameter estimates drop out when summing over all observations of one unit in equation (7), as shown in the proof. Second, by using estimates $\tilde{\tau}_{it}$ that fulfil the convergence condition of the theorem, we avoid the issue of obtaining trivial residuals for the treated observations. The resulting variance estimates are asymptotically conservative. From these estimates we can also obtain conservative confidence intervals (that asymptotically have coverage that is at least nominal) if the estimator is also asymptotically normal, such as under the sufficient conditions of Proposition 8.

It remains to choose the estimates $\tilde{\tau}_{it}$. We focus on auxiliary models that impose the equality of treatment effects across large groups of treated observations: for a partition $\Omega_1 = \bigcup_g G_g$, $\tau_{it} \equiv \tau_g$ for all $it \in G_g$. The τ_g can then be estimated by some weighted average of $\hat{\tau}_{it}^*$ among $it \in G_g$. Specifically, we propose averages of the form

$$\tilde{\tau}_g = \frac{\sum_i \left(\sum_{t;it\in G_g} v_{it}\right)\left(\sum_{t;it\in G_g} v_{it}\hat{\tau}_{it}^*\right)}{\sum_i \left(\sum_{t;it\in G_g} v_{it}\right)^2}. \tag{8}$$

In Supplementary Appendix A.8, we show that this choice of weights leads to minimal excess variance σ_τ^2 in the case where there is only a single group g, corresponding to a conservative auxiliary model which requires all treatment effects to be the same. The choice of the partition aims to maintain a balance between avoiding overly conservative variance estimates and ensuring consistency. If the sample is large enough, one may want to partition Ω_1 into multiple groups of observations such that treatment-effect heterogeneity is expected to be smaller within them than across. For instance, with many units, a group may consist of observations corresponding to the same horizon relative to treatment onset. If cohorts are large, one can further partition observations into groups defined by cohort and period, which we use as the default in our Stata command.

While sufficiently large groups in equation (8) avoid overfitting asymptotically (under appropriate conditions), in finite samples these $\tilde{\tau}_{it}$ still use $\hat{\tau}_{it}^*$ and thus partially overfit to ε_{it}. In Supplementary Appendix A.9, we therefore also consider leave-out versions of these $\tilde{\tau}_{it}$.

We make four final remarks on Theorem 3. First, our strategy for estimating the variance extends directly to conservative estimation of variance–covariance matrices for vector-valued estimands, for example, for average treatment effects at multiple horizons h. Second, the result applies in short panels under the low-level conditions of Supplementary Appendix A.7 (see Supplementary Proposition A7). Third, while we have focused here on the case of unrestricted

heterogeneity ($\Gamma = \mathbb{I}$), Theorem 3 can be extended to the case with a non-trivial treatment-effect model imposed in Assumption 3.[25] Finally, computation of $\hat{\sigma}^2_w$ for the estimator $\hat{\tau}^*_w$ from Theorem 1 involves the implied weights v^*_{it}, which becomes computationally challenging with multiple sets of high-dimensional FEs. In Supplementary Appendix A.10, we develop a computationally efficient algorithm for computing v^*_{it} based on the iterative least squares algorithm for conventional regression coefficients (Guimarães and Portugal, 2010).

4.4. *Testing for parallel trends*

In this section, we discuss testing the (generalized) parallel-trend and no-anticipation assumptions Assumptions 1′ and 2. We propose a testing procedure based on OLS regressions with untreated observations only, departing from both traditional regression-based tests and more recent placebo tests. This procedure is robust to treatment-effect heterogeneity and, under spherical errors, has attractive power properties and avoids the problem of inference after pre-testing explained by Roth (2022). We propose:

Test 1 (Robust OLS-based pre-trend test).

1. *Choose an alternative model for Y_{it} for untreated observations $it \in \Omega_0$ that is richer than that imposed by Assumptions 1′ and 2: for an observable vector W_{it} (which we consider non-stochastic, like A_{it} and X_{it}),*

$$Y_{it} = A'_{it}\lambda_i + X'_{it}\delta + W'_{it}\gamma + \varepsilon_{it}. \tag{9}$$

2. *Estimate γ by $\hat{\gamma}$ in equation (9) using OLS on untreated observations only.*
3. *Test $\gamma = 0$ using the heteroskedasticity- and cluster-robust Wald test.*

This test is valid because equation (9) is implied by Assumptions 1′ and 2 if the null $\gamma = 0$ holds.[26]

The test requires choosing W_{it} to parametrize the possible violation of Assumptions 1′ and 2. A natural choice for W_{it}, which parallels conventional pre-trend tests, is a set of indicators for observations $1, \dots, k$ periods before the onset of treatment for some k, with periods before $E_i - k$ serving as the reference group.[27] This choice is appropriate, for instance, if the researcher's main worry is the possible effects of treatment anticipation, that is, violations of Assumption 2. This choice of W_{it} also lends itself to making "event-study plots", which combine the ATT estimates by horizon $h \geq 0$ with a series of pre-trend coefficients; we supply the `event_plot` Stata command for this goal. Alternatively, the researcher may focus on possible violations of

25. By Supplementary Proposition A2, the general efficient estimator can be represented as an imputation estimator for a modified estimand, that is, by changing w_1 to some v_1. Theorem 3 then yields a conservative variance estimate for it. We note that under sufficiently strong restrictions on treatment effects, asymptotically exact inference may be possible, as the residuals $\hat{\varepsilon}_{it}$ in equation (4) may be estimated consistently even for treated observations (except for the inconsequential noise in $\hat{\lambda}_i$), alleviating the need for an additional auxiliary model.

26. There is a natural alternative test of the null $\gamma = 0$ in the model (9), namely the Hausman test based on the difference between the imputation estimator $\hat{\tau}^W_w$ based on the model in equation (9) and the efficient imputation estimator $\hat{\tau}^*_w$ that is only valid when $\gamma = 0$. Like our test, this test only uses untreated observations and avoids the Roth (2022) pre-testing problem for spherical errors. The Hausman approach has the advantage of quantifying the magnitude of bias from omitting W_{it}, while Test 1 has the advantage that it is also informative about violations that cancel out in the Hausman test. The two tests are equivalent for scalar W_{it}.

27. The optimal choice of k is a challenging question. As usual with Wald tests, choosing a k that is too large can lead to low power against many alternatives, in particular those that generate large biases in treatment-effect estimates that impose invalid Assumption 1′.

Assumption 1′. For instance, with data spanning many years one could test for the presence of a structural break in unit FEs.

Test 1 can be contrasted with two existing strategies to test parallel trends. Traditionally, researchers estimated a dynamic specification including lags and leads of treatment onset, and tested—visually or statistically—that the coefficients on leads are equal to zero. More recent papers (*e.g.* de Chaisemartin and D'Haultfœuille, 2020; Liu *et al.*, 2022) replace it with a placebo strategy: pretend that treatment happened k periods earlier for all eventually treated units, and estimate the average effects $h = 0, \ldots, k-1$ periods after the placebo treatment using the same estimator as for actual estimation.

Both of these alternatives strategies have drawbacks. Because the traditional regression-based test uses the full sample, including treated observations, and imposes restrictions on treatment effects (which are assumed homogeneous within each horizon), it is *not* a test for Assumptions 1′ and 2 only. Rather, it is a joint test that is sensitive to violations of the implicit Assumption 3 (Sun and Abraham, 2021). Even if a researcher has reasons to impose a non-trivial Assumption 3 in estimation, a robust test for parallel trends and no-anticipation *per se* should avoid those restrictions on treatment-effect heterogeneity. With a null Assumption 3, treated observations are not useful for testing, and our test only uses the untreated ones.[28]

Tests based on placebo estimates appropriately use untreated observations only and may have intuitive appeal. However, mimicking the estimator does not generally correspond to an efficient test of a class of plausible alternatives. In contrast, Test 1 possesses well-known asymptotic efficiency properties when W_{it} is correctly specified. For example, when ε_{it} are spherical and normal, it is asymptotically equivalent to the homoskedastic F-test, which is a uniformly most powerful invariant test (Lehmann and Romano, 2006, ch. 7.6).

Finally, we show an additional advantage of Test 1: if the researcher conditions on the test passing (*i.e.* does not report the results otherwise), inference on $\hat{\tau}^*_w$ is still asymptotically valid under the null of no violations of Assumptions 1′ and 2 and under spherical errors. This avoids the issue pointed out by Roth (2022, Proposition 4) in the context of restrictive dynamic event study regressions: that variance estimates which do not take pre-testing into account are inflated, leading to unnecessarily conservative inference.[29]

Proposition 9 (Pre-test robustness). *Suppose the model in equation (9) and Assumption 4 hold. Then $\hat{\tau}^*_w$ constructed as in Theorem 1 is uncorrelated with any vector $\hat{\gamma}$ constructed as in Test 1. If the error terms are also normally distributed,[30] then $\hat{\tau}^*_w$ and $\hat{\gamma}$ are independent, and inference on τ_w based on $\hat{\tau}^*_w$ is unaffected by pre-tests based on $\hat{\gamma}$.*[31]

28. Wooldridge (2021) shows that as long as treatment effects are allowed to vary flexibly, tests based on specifications estimated on the full sample do not use treated observations. Therefore, such tests are also not contaminated by treatment-effect heterogeneity.

29. Roth (2022) points out another issue, which we also avoid under the assumptions of Proposition 9: when pre-trend and treatment-effect estimators are correlated, the bias arising from violations of Assumptions 1′ and 2 is affected by pre-testing; it is exacerbated in specific cases (Roth, 2022, Proposition 2).

30. The normality assumption is not essential; in the proof, we show that a similar, asymptotic result holds generally under regularity conditions.

31. An early version of Roth (2022) shows how to construct an adjustment that removes the dependence when it exists, provided the covariance matrix between $\hat{\tau}^*$ and $\hat{\gamma}$ can be estimated. By Proposition 9, this adjustment is not needed for the Theorem 1 estimator under spherical errors.

5. APPLICATION

Having derived the attractive theoretical properties of the imputation estimator, we now illustrate their practical relevance by revisiting the estimation of the marginal propensity to spend in the event study of Broda and Parker (2014b). We also use this empirical setting to verify the properties of the imputation estimator in a simulation study.

5.1. *Setting*

The marginal propensity to spend out of tax rebates is a crucial parameter for economic policy. In the U.S., the Economic Stimulus Act of 2008 consisted primarily of a 100 billion dollar programme that sent tax rebates to approximately 130 million tax filers. Parker *et al.* (2013) and Broda and Parker (2014b) estimate the marginal propensity for expenditure (MPX) out of the 2008 tax rebates. The rebate was disbursed using two methods: either via direct deposit to a bank account, if known by the IRS, or with a mailed paper check. For each method, the week in which the funds were disbursed depended on the second-to-last digit of the taxpayer's social security number (SSN). This number provides a source of quasi-experimental variation because the last four digits of a SSN are assigned sequentially to applicants within geographic areas.

Broda and Parker (2014b, henceforth BP) use an event-study design to examine the response of non-durable spending to tax rebate receipt, leveraging the quasi-experimental variation in the *timing* of the receipt. The quasi-random assignment of the last digits of the SSN makes the parallel-trends assumption for expenditures *a priori* plausible.[32] The no-anticipation assumption may also be expected to hold: although the disbursement schedule was known in advance, households were directly notified by mail only several days before disbursement.

We estimate the performance of various estimators at estimating the impulse response function of non-durable spending to tax rebate receipt using the same data as BP.[33] While earlier work by Parker *et al.* (2013) estimates the impulse responses using quarterly spending data from the Consumer Expenditure Survey, BP leverage more detailed data from the NielsenIQ Homescan Consumer Panel. The NielsenIQ dataset tracks transactions at a much higher (in principle, daily) frequency, which is why we choose it for our analysis. The NielsenIQ data cover expenditures on consumer packaged goods (food, beverages, beauty and health products, household supplies, and general merchandise), representing around 15% of total household expenditures. Our dataset, identical to that of BP, is a complete panel of 21,760 households (including 21,690 with non-missing disbursement method information) observed over 52 weeks of year 2008.

5.2. *Comparison between robust and conventional estimates*

We show how BP's estimates of the MPX suffer from an upward bias in the short-run due to the choice of a binned specification (Section 5.2.1) and how they may be spurious in the long-run (Section 5.2.2). In Section 5.2.3, we present our preferred robust estimates and discuss implications for the macroeconomics literature.

32. Thakral and Tô (2022) point out that for the paper check group pre-rebate household characteristics (in *levels*) are not balanced with respect to the timing of the receipt. While this is problematic for randomization-based approaches to DiD (*e.g.* Arkhangelsky and Imbens, 2022; Roth and Sant'Anna, 2023), parallel *trends* in expenditures may still hold. Indeed, we fail to reject them with pre-trend tests below.

33. Specifically, we combine the NielsenIQ data provided by the Kilts Center for Marketing at the University of Chicago (Kilts Center, University of Chicago, n.d.) with the NielsenIQ supplemental survey providing information on 2008 Stimulus payments, which was developed by BP and is available from them by request.

5.2.1. Negative weighting and upward bias with binning. We replicate BP's estimates, focusing on the first 3 months since the receipt, while leaving longer-run effects to Section 5.2.2. BP estimate conventional dynamic specifications of the form:

$$Y_{it} = \alpha_i + \beta_t + \sum_{h=-a}^{b} \tau_h \mathbf{1}[K_{it} = h] + \varepsilon_{it}, \tag{10}$$

where Y_{it} is the dollar amount of spending in calendar week t for household i, α_i are household FEs, and β_t are week FEs. In some specifications, week FEs are interacted with the disbursement method $m(i)$ (*i.e.* $\beta_{m(i)t}$ is included instead of β_t in equation (10)) to leverage the variation in timing only within each disbursement method; we refer to those specifications as "with disbursement method FEs". The set of $\mathbf{1}[K_{it} = h]$ are the lead/lag indicator variables tracking the number of weeks $K_{it} = t - E_i$ since the week of the tax rebate receipt for the household, E_i; b is chosen such that all possible lags in the sample are covered; a varies as discussed below. MPXs for each horizon, as well as pre-trend coefficients, are captured by τ_h. Regressions are weighted by the NielsenIQ projection weights.

BP's preferred specification is a binned version of equation (10) which constrains τ_h to be constant across 4-week periods—"months"—around the event, starting with the week of tax rebate receipt: for example, $\tau_0 = \cdots = \tau_3$. This specification also includes one monthly pre-trend coefficient, that is, $a = 4$ with $\tau_{-1} = \cdots = \tau_{-4}$. These estimates, without and with disbursement method FEs, are replicated in Table 2, columns 1 and 2, suggesting that tax rebate receipt led to an increase in spending in the contemporaneous month of \$42.6 (s.e. 7.2) in column 1 to \$47.6 (s.e. 9.2) in column 2, and a cumulative increase over 3 months of \$60.5 (s.e. 25.7) in column 1 to \$94.4 (s.e. 33.5) in column 2. As we will discuss in Section 5.2.3, extrapolating these estimates from NielsenIQ products to all consumption implies very large total MPX.

Next, we show that the MPX estimates are much smaller without binning. In columns 3 and 4 of Table 2, we report the estimates from the conventional specification (10) without binning and with one weekly lead ($a = 1$), as in BP's Table 3. We report the coefficients aggregated to the monthly level. Compared to columns 1 and 2, there is a large fall in the cumulative 3-month MPX, from \$60.5 (s.e. 25.7) to \$26.8 (s.e. 21.4) without disbursement method FEs and from \$94.4 (s.e. 33.5) to \$9.6 (s.e. 34.4) with these FEs. In columns 5 and 6, we use the robust and efficient imputation estimator to estimate weekly average responses and aggregate them to the monthly level. The point estimates are similar to columns 3 and 4 for the contemporaneous month response, while for the quarterly MPX they are in between the results obtained with a binned specification and without binning.[34]

Could the difference between binned and other estimates indicate a violation of the DiD assumptions? Figure 2 provides evidence against this possibility, showing that there is no sign

34. Supplementary Table A1 reports the differences between the estimates from OLS with no binning or imputation and the binned OLS specification, with standard errors and p-values. All binned estimates are significantly different from those without binning at the 10% (5%) significance level without (with) disbursement method FEs. The difference between the binned and imputation estimates is only significant at the 10% level for the contemporaneous month with disbursement method FEs. We explain below that the difference in estimates is due to the difference in estimands, rather than statistical noise.

TABLE 2
Estimates of the monthly and quarterly MPX out of tax rebates

	Dollars spent after tax rebate receipt					
	OLS monthly binned		OLS no binning		Imputation estimator	
	(1)	(2)	(3)	(4)	(5)	(6)
Contemporaneous month	42.59	47.57	35.02	27.88	38.13	30.54
	(7.19)	(9.15)	(5.75)	(7.75)	(5.68)	(9.08)
First month after	9.31	26.26	−2.28	−4.48	−2.47	7.43
	(9.00)	(11.95)	(7.59)	(12.48)	(7.81)	(16.17)
Second month after	8.63	20.52	−5.96	−13.82	13.08	4.01
	(11.17)	(14.57)	(10.06)	(16.38)	(22.51)	(29.89)
Third month total	60.53	94.35	26.79	9.58	48.75	41.97
	(25.73)	(33.54)	(21.43)	(34.42)	(30.97)	(46.56)
Disbursement method FE	No	Yes	No	Yes	No	Yes
N observations	1,131,520	1,127,880	1,131,520	1,127,880	631,040	536,553
N households	21,760	21,690	21,760	21,690	21,760	21,690

Notes: Columns 1 and 2 estimate the binned version of specification (10) with $a = 4$ and imposing that the coefficients are the same in each month, that is, 4 weeks since the rebate receipt. Columns 3 and 4 estimate the same specification without binning, with $a = 1$. These specifications are identical to Broda and Parker (2014b, Tables 3 and 4, columns 1 and 4). Columns 5 and 6 report the efficient imputation estimator. All columns aggregate coefficients by month for the first 3 months after the rebate receipt and suppress the other coefficients. Columns 1, 3, and 5 use household and week FEs, while columns 2, 4, and 6 additionally interact week FEs with disbursement method dummies. The estimates in column 6 exclude the last week of the quarter ($h = 11$) due to insufficient sample size. All estimates use projection weights from the NielsenIQ Consumer Panel, and standard errors are clustered by household.

of pre-trends.[35] The Wald test confirms this finding: the *p*-value for the null of no pre-trends is 0.185 (0.403) without (with) disbursement method FEs.

We find instead that the higher estimates from binned specifications are explained by the estimand they implicitly choose. Specifically, this estimand places a very large weight on the first weeks after the rebate, when the effects are the largest, and negative weights on other weeks. Figure 2 shows that the increase in spending after the receipt is concentrated in the first weeks since the rebate. Figure 3 in turn shows the weights with which the quarterly MPX estimated from the monthly binned specification of Table 2 aggregates the MPXs at each weekly horizon. These weights show how the estimand of the binned specification diverges from the true quarterly MPX, which is a simple sum of the effects at each horizon $h = 0, \ldots, 11$ weeks, that is, with constant weights of one on each week.[36] In the specification without disbursement method FEs, the weight placed on the first-week response is three times larger than it would be with an equally weighted sum; it is five times larger with the FEs. Furthermore, within each month the weights become negative for the last weeks of the month. Applying the weights of the binned specification across weeks from Figure 2 to the estimates without binning (underlying column 3 of Table 2), we obtain a point estimate of $42.6 for the contemporaneous

35. This figure reports the imputation estimates for 8 weeks since the rebate along with the pre-trend coefficients from the Section 4.4 test that allows for 8 weeks of anticipation effects. The figure also reports conventional specifications without binning augmented to include $a = 8$ weeks of pre-trends, dropping observations more than 8 weeks since the rebate.

36. The binned specification's estimand also diverges from the true MPX in how it weights different households for the same weekly horizon (similar to the issues studied theoretically by Sun and Abraham (2021) for dynamic specifications without binning). We focus on the variation across horizons here because MPXs have a very strong dynamic pattern.

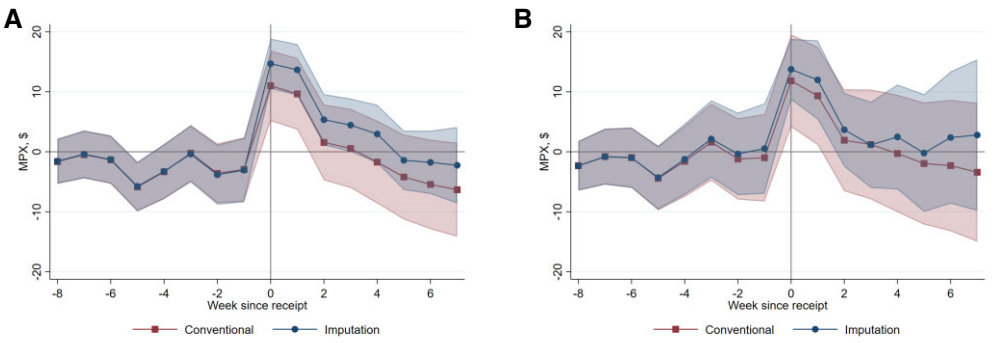

FIGURE 2
Dynamic specifications and pre-trends: (A) without disbursement method FEs and (B) with disbursement method FEs

Notes: Panel A reports estimates of the response of spending to tax rebate receipts and pre-trend coefficients, using specification (10) with $a = 8$ and without binning ("Conventional") and with the efficient imputation estimator and the pre-trend test from Section 4.4 ("Imputation"). Panel B additionally interacts week FEs with the disbursement method. Observations 8 or more weeks since the rebate receipt are excluded. Estimation is weighted by the projection weights from the NielsenIQ Consumer Panel. 95% confidence bands are shown, using standard errors clustered by household.

month—indistinguishable from column 1, instead of \$35.0 in column 3. Similarly, we get \$60.4 for the quarter, nearly identical to \$60.5 in column 1, instead of \$26.8 in column 3. Thus, the short-run biased weighting scheme due to binning explains nearly all the difference between columns 1 and 3 of Table 2.[37]

5.2.2. Spurious identification of long-run causal effects. We now examine the long-run dynamics of MPXs obtained with conventional specifications and the imputation estimator. The timing of the tax rebate is such that we simultaneously observe treated and untreated households for at most 13 weeks.[38] Per Proposition 5, without restrictive assumptions on treatment-effect heterogeneity it is not possible to estimate causal effects beyond 12 weeks. Yet conventional dynamic specifications produce estimates for longer horizons via extrapolation. We examine whether the estimates obtained in this way could paint a misleading picture of the long-run dynamics of MPXs.

In Figure 4, we use the same specifications as in Table 2 but we report the full set of dynamic estimates for the treatment effects. Panel A reports the estimates from the binned specification. With disbursement method FEs, the point estimates are large and positive for all 9 months following the receipt of the tax rebate. Thus, due to the extrapolation resulting from binning, this specification could be mistakenly interpreted as evidence for a very large and persistent increase in spending. Without these FEs, the estimates tend to hover around zero.

In Panel B, we show estimates with the conventional dynamic specification without binning. Both specifications with and without disbursement method FEs yield point estimates that are almost all negative in the long run. Taken at face value, these estimates could misleadingly suggest that households intertemporally substitute consumption by making purchases at the time

37. Short-run biased weighting also explains the majority, although not all, of the difference between the specification with disbursement method FEs in columns 2 and 4 of Table 2. Applying the binned specification weights to the specifications without binning, we get an estimate of \$40.0 for the contemporaneous month and \$69.0 for the quarter, thus reducing the discrepancy between columns 2 and 4 by 62% and 70% for the month and quarter, respectively.

38. The first treated households received the rebate during week 17 of 2008 (week ending 26 April), while the last treated households received it during week 30 (week ending 26 July).

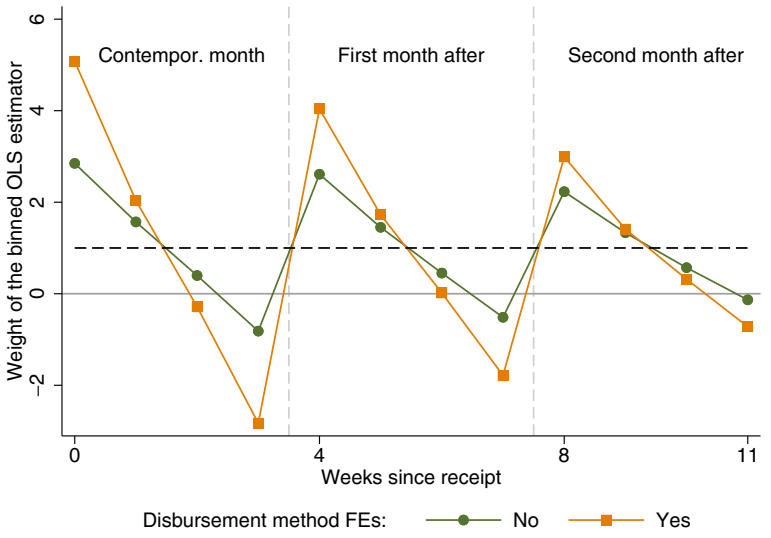

FIGURE 3
Short-term bias in weights for binned specifications

Notes: This figure reports the cumulative weight that the monthly binned OLS estimator of the quarterly MPX from Table 2, with or without disbursement methods FEs, places on the true effects at each horizon $h = 0, \ldots, 11$ weeks since the rebate receipt. These weights are computed using the Frisch–Waugh–Lovell theorem, analogously to equation (17), and aggregated across the first 3 months since the rebate receipt. The black dashed line indicates the weight corresponding to the true quarterly MPX, that is, a simple sum of the effects at each horizon.

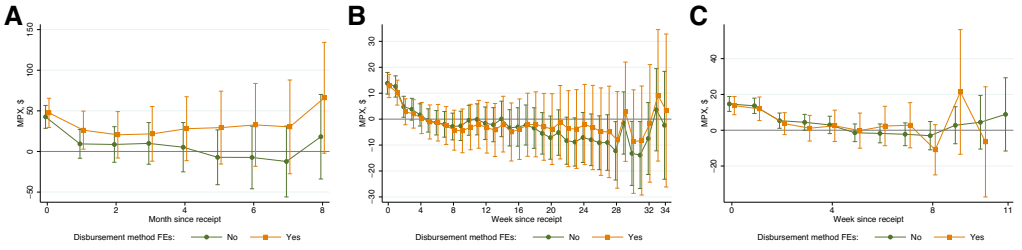

FIGURE 4
Long-run MPXs, conventional specifications versus robust imputation estimator: (A) conventional binned estimates, (B) conventional dynamic estimates, and (C) robust imputation estimates

Notes: Panels A–C plot MPX coefficients and 95% confidence bands using the same specifications as in Table 2. Coefficients on the leads of treatment are not shown. The last horizon in Panel B ($h = 35$ weeks) and Panel C ($h = 12$ weeks without disbursement methods FEs or $h = 11$ with FEs) are suppressed because of the very large standard errors, due to a limited sample size. Standard errors are clustered by household.

of tax rebate that they would have made 20–30 weeks later. As in Panel A, these point estimates are noisy but could lend themselves to some economic interpretation.

In contrast, Panel C describes the results from the robust imputation estimator, which does not allow extrapolation in the absence of an explicit control group. This panel shows that, for the horizons for which imputation is possible, there is no evidence of any impact on spending beyond 2–4 weeks after tax rebate receipt. The patterns are the same both with and without disbursement FEs. These results highlight the practical relevance of the insights from Section 3.4: the imputation estimators avoid extrapolation, thus eliminating seemingly unstable patterns found across conventional specifications.

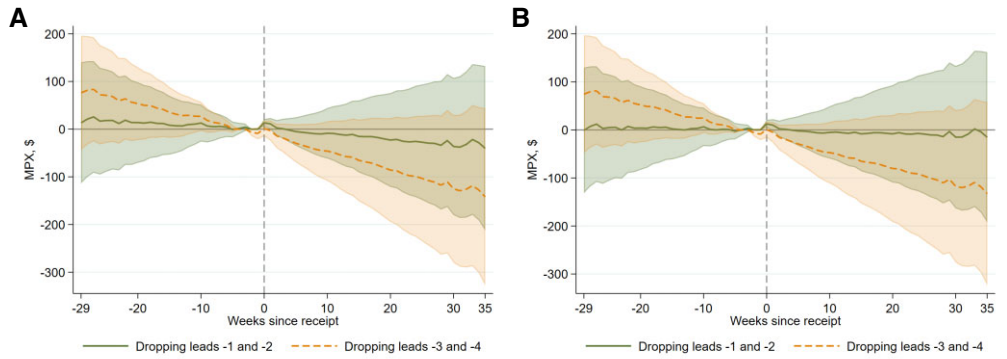

FIGURE 5
Under-identification of the fully dynamic specification: (A) without disbursement method FEs and (B) with disbursement method FEs

Notes: This figure reports MPX and pre-trend estimates and 95% confidence bands for specification (10) with all leads and lags of the tax receipt included, except for two chosen as indicated. Standard errors are clustered by household.

Finally, in Figure 5 we illustrate the importance of the insights on the under-identification of fully dynamic specifications from Section 3.2. Unlike earlier specifications, which only included a small number of treatment leads, here we run the specification (10) with a full set of weekly leads and lags around tax rebate receipt. We drop two leads since the set of lead and lag coefficients is only identified up to a linear trend, as discussed in Section 3.2. We find that the fully dynamic estimates change drastically depending on which two leads are dropped. We illustrate this by comparing the MPXs when dropping leads -1 and -2 or -3 and -4. This shows another source of instability in conventional practice, which the imputation estimator directly avoids.

5.2.3. Preferred robust estimates and macroeconomic implications. We now discuss the implications of our findings for the macroeconomics literature. We proceed in two steps: selecting our preferred MPX estimate from Section 5.2.1 for the NielsenIQ products and then extrapolating it to broader consumption baskets, following the strategy of BP.

Our preferred estimate for the average cumulative MPX out of the tax rebate for the NielsenIQ products is \$30.5, corresponding to the imputation estimator with disbursement method FEs (Table 2, column 6) in the first month since the rebate. This constitutes 3.4% of the average rebate amount. We choose the specification with disbursement method FEs because the variation in timing is more plausibly exogenous within disbursement methods. We focus on the first (*i.e.* contemporaneous) month and impose zero effects for the following months based on the evidence from Figures 2 and 4 that the MPXs rapidly decay to zero, while estimation noise increases.[39] Finally, we choose the imputation estimator over conventional specifications for its robustness properties. In contrast to columns 1 and 2 of Table 2, it avoids the short-term bias due to binning. Moreover, in contrast to column 3 and 4 it avoids extrapolation of long-run effects (the estimates are similar for the contemporaneous month). Robustness to treatment-effect heterogeneity is gained without an efficiency loss in this application: the standard errors are similar across columns of Table 2.

39. Our preferred estimate is robust to the choice of the time window: the cumulative MPX would have been similar (at \$25.7 instead of \$30.5) if we focused on the first 2 weeks only.

To obtain MPX estimates covering the full consumption basket, BP propose to rescale the estimates obtained with the NielsenIQ data. This scaling is done in three different ways: (1) by the ratio of spending per capita in the National Income and Product Account and NielsenIQ data; (2) by the ratio of the self-reported change in spending on all goods after the rebate relative to that on NielsenIQ goods alone; and (3) by a factor based on the relative shares of spending and relative responsiveness to the rebate across subcategories of goods as measured in Consumer Expenditure Survey (CE). Using these three approaches and BP's preferred MPX estimate (reproduced in our Table 2, column 1), they estimate that the tax rebate raised the annualized expenditure growth rate by 1.3–1.9 percentage points (p.p.) in 2008Q2 and by 0.6–0.9 p.p. in 2008Q3, depending on the choice of rescaling.

Applying the same scaling methods to our preferred MPX estimate for the contemporaneous month and assuming zero response in the following months paints a very different picture, with an increase in annualized expenditure growth of only 0.8–1.1p.p. in 2008Q2 and 0.15–0.22p.p. in 2008Q3. Our estimate implies a 40% smaller response of consumption expenditures in 2008Q2, and 75% smaller in 2008Q3. Correspondingly, while BP conclude that the propensity to spend at the individual level from a tax rebate over 3 months since the rebate is between 51 and 75%, our preferred estimates are half as large, between 25 and 37%.[40]

In Table 3, we summarize the MPX estimates for the first quarter after tax rebate obtained with BP's and our preferred specification. The first row reports the observed marginal propensity to spend on products included in the NielsenIQ sample during that quarter, as a fraction of the average rebate amount. The next rows rescale these estimates to extrapolate the marginal propensity to spend to broader samples, that is, the full consumption basket (second row) and non-durables (third row). For the full consumption basket, we implement the three scaling procedures from BP and report the lower and upper bounds; for non-durables we leverage the scaling method of Laibson *et al.* (2022). The fourth row reports the model-consistent, or "notional", marginal propensity to consume (MPC) that can be used as a target for macroeconomic models, also following the methodology of Laibson *et al.* (2022).[41] The estimates based on BP in column 1 are closely in line with the literature: typical estimates of the quarterly MPX for all expenditures range from 50 to 90%, while estimates of the quarterly MPX for non-durable expenditure range from 15 to 25%.[42] In contrast, the imputation estimator in column 2 of Table 3 delivers estimates that are about half as large in all rows. These smaller MPC estimates imply a lower effectiveness of fiscal stimulus.

Thus, our new estimates for the impact of the 2008 fiscal stimulus on the U.S. economy yield two lessons for the calibration of macroeconomic models: (1) that the targeted MPC should be significantly smaller—about half as large—and (2) that it is best to calibrate the model using weekly level estimates of the MPC, as we report in Figure 2, rather than monthly or, especially, quarterly MPC estimates, which are much noisier. Indeed, models should reflect that most of the spending response occurs in the very short run, in the first 2–4 weeks after tax rebate receipt.

40. We obtain these estimates by replicating the first row of Panel A of BP's Table 5 and using our preferred estimates.

41. Standard macroeconomic models assume a notional consumption flow that does not distinguish between non-durable and durable consumption. Prior to Laibson *et al.* (2022) showing that the notional MPC should be the relevant target, state-of-the-art macroeconomic models targeted non-durable MPX estimates. For instance, Kaplan and Violante (2014) targeted the estimates from Johnson *et al.* (2006), which are quantitatively similar to those from BP when rescaled as in Table 3, despite using more aggregated data and a different rebate episode.

42. Laibson *et al.* (2022) provide a recent review of the literature. Kaplan and Violante (2022) review non-durable MPX, and Di Maggio *et al.* (2020) review total MPX.

TABLE 3
First-quarter MPX and MPC estimates for calibration of macroeconomic models

Statistic	Replication of Broda and Parker (2014b) (1)	Imputation estimator (2)
NielsenIQ MPX	6.7%	3.4%
Total MPX	50.8–74.8%	24.8–36.6%
Non-durable MPX	14.1–20.8%	6.9–10.2%
Notional MPC	15.9–23.4%	7.8–11.4%

Notes: This table reports the first-quarter MPX and MPC using the preferred binned specification of Broda and Parker (2014b) and our preferred specification based on the imputation estimator. The first row reports the marginal propensity to spend on products included in the NielsenIQ sample, as a fraction of the average rebate amount. The second row rescales these estimates to extrapolate them to the marginal propensity to spend on all goods using the three rescaling methods from Broda and Parker (2014b). The ranges correspond to the lowest and highest values among the three rescaling methods. To obtain the estimate for the non-durables MPX in the third row, we use the scaling factor of Laibson *et al.* (2022), who show that the total MPX is equal to 3.6 times the non-durables MPX. The fourth row also follows the methodology of Laibson *et al.* (2022) and reports the model-consistent ("notional") MPC that can be used as a target for macroeconomic models, equal to the total MPX divided by 3.2.

5.3. *Efficiency gains relative to alternative robust estimators*

Finally, we compare the efficiency of the imputation estimator to the alternative robust estimators of de Chaisemartin and D'Haultfœuille (2022) and Sun and Abraham (2021), abbreviated dCDH and SA. We document the in-sample efficiency gains in Figure 6 by showing the point estimates and confidence intervals for weekly average MPXs based on the imputation estimator and the two alternatives in Panel A. We use the specification without disbursement method FEs.[43] The point estimates are very similar for dCDH and the imputation estimator, but they differ from those of SA, because this estimator uses a much smaller control group (only the households who received the rebate in the latest possible week) and is therefore much noisier.

Panel B zooms in on the efficiency comparison by reporting the lengths of the confidence intervals for SA and dCDH relative to that of the imputation estimator. The differences are large: the confidence interval from dCDH is about 50% longer for all periods, and 2–3.5 times longer for SA.

In Supplementary Appendix A.11, we confirm these efficiency gains, obtained from a single sample, in a Monte Carlo study based on the BP data, for several data-generating processes. We find that the imputation estimator has sizable efficiency advantages over alternative robust estimators not only with spherical errors but also in presence of heteroskedasticity, serial correlation, or both. Moreover, these gains do not come at a cost of systematically higher sensitivity to parallel-trend violations. We also confirm that our analytical standard errors have correct coverage.

6. CONCLUSION

In this paper, we provided a unified framework that formalizes an explicit set of goals and assumptions underlying event-study designs, reveals and explains challenges with conventional

43. We implement the dCDH method using the `csdid` Stata command developed for the Callaway and Sant'Anna (2021) estimator: the two estimators are identical absent additional controls, and `csdid` allows for projection weights.

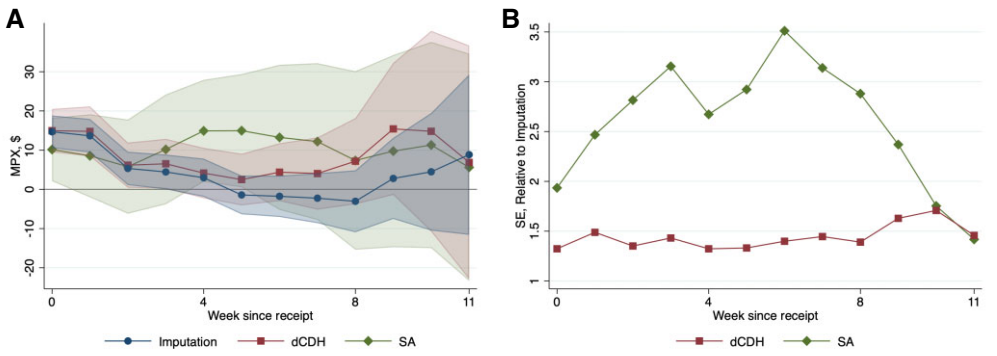

FIGURE 6
Alternative robust MPX estimates and in-sample efficiency: (A) point estimates and confidence intervals and (B) confidence interval lengths, relative to the imputation estimator

Notes: Panel A shows the estimates and 95% confidence bands for the average MPXs by week since rebate using three robust estimators: the imputation estimator, de Chaisemartin and D'Haultfœuille (2022) (dCDH), and Sun and Abraham (2021) (SA). The specifications do not include disbursement method fixed effects. Panel B reports the ratios of the lengths of confidence intervals for dCDH and SA relative to the imputation estimator. Standard errors are clustered by household.

practice, and yields an efficient estimator. In a benchmark case where treatment-effect heterogeneity remains unrestricted, this robust and efficient estimator takes a particularly simple "imputation" form that estimates fixed effects among the untreated observations only, imputes untreated outcomes for treated observations, and then forms treatment-effect estimates as weighted averages over the differences between actual and imputed outcomes. We developed results for asymptotic inference and testing and compared our approach to other estimators. We also highlighted the importance of separating testing of identification assumptions from estimation, which increases estimation efficiency and helps address inference biases due to pre-testing. We demonstrated the practical relevance of these insights in an empirical application documenting that the notional MPC is between 8 and 11% in the first quarter, about half as large as benchmark estimates.

Acknowledgments. This draft supersedes our 2018 manuscript, "Revisiting Event Study Designs, with an Application to the Estimation of the Marginal Propensity to Consume". We thank Alberto Abadie, Isaiah Andrews, Raj Chetty, Itzik Fadlon, Ed Glaeser, Peter Hull, Guido Imbens, Larry Katz, Jack Liebersohn, Benjamin Moll, Jonathan Roth, Pedro Sant'Anna, Amanda Weiss, and three anonymous referees for thoughtful conversations and comments. We are particularly grateful to Jonathan Parker for his support in accessing and working with the data and code from Broda and Parker (2014b). Two accompanying Stata commands are available from the SSC repository: `did_imputation` for treatment-effect estimation with our imputation estimator and pre-trend testing, and `event_plot` for making dynamic event-study plots. Our own analyses are calculated (or derived) based in part on data from Nielsen Consumer LLC and marketing databases provided through the NielsenIQ Datasets at the Kilts Center for Marketing Data Center at The University of Chicago Booth School of Business. The conclusions drawn from the NielsenIQ data are ours and do not reflect the views of NielsenIQ. NielsenIQ is not responsible for, had no role in, and was not involved in analysing and preparing the results reported herein.

Supplementary Data

Supplementary data are available at *Review of Economic Studies* online.

Data Availability

The data underlying this article are publicly available on Zenodo at doi.org/10.5281/zenodo.10037585.

REFERENCES

ABBRING, J. H. and VAN DEN BERG, G. J. (2003), "The Nonparametric Identification of Treatment Effects in Duration Models", *Econometrica*, **71**, 1491–1517.

ANGRIST, J. (1998), "Estimating the Labor Market Impact of Voluntary Military Service Using Social Security Data on Military Applicants", *Econometrica*, **66**, 249–288.

ARKHANGELSKY, D. and IMBENS, G. W. (2022), "Doubly Robust Identification for Causal Panel Data Models", *The Econometrics Journal*, **25**, 649–674.

ATHEY, S., BAYATI, M., DOUDCHENKO, N., *et al.* (2021), "Matrix Completion Methods for Causal Panel Data Models", *Journal of the American Statistical Association*, **116**, 1716–1730.

BAKER, A. C., LARCKER, D. F. and WANG, C. C. Y. (2022), "How Much Should We Trust Staggered Difference-In-Differences Estimates?", *Journal of Financial Economics*, **144**, 370–395.

BORUSYAK, K. and JARAVEL, X. (2018), "Revisiting Event Study Designs" (Working Paper, Harvard University). http://web.archive.org/web/20210201163348/https://scholar.harvard.edu/files/borusyak/files/borusyak_jaravel_event_studies.pdf.

BRODA, C. and PARKER, J. A. (2014a), "Replication Data for: The Economic Stimulus Payments of 2008 and the Aggregate Demand for Consumption", *ScienceDirect*.

—— and —— (2014b), "The Economic Stimulus Payments of 2008 and the Aggregate Demand for Consumption", *Journal of Monetary Economics*, **68**, S20–S36.

CALLAWAY, B., GOODMAN-BACON, A. and SANT'ANNA, P. H. (2021), "Difference-in-Differences with a Continuous Treatment", arXiv.

CALLAWAY, B. and SANT'ANNA, P. H. (2021), "Difference-in-Differences with Multiple Time Periods and an Application on the Minimum Wage and Employment", *Journal of Econometrics*, **225**, 200–230.

CENGIZ, D., DUBE, A. and LINDNER, A. (2019), "The Effect of Minimum Wages on Low-Wage Jobs", *The Quarterly Journal of Economics*, **134**, 1405–1454.

CORREIA, S. (2017), "Linear Models with High-Dimensional Fixed Effects: An Efficient and Feasible Estimator" (Working Paper, Duke University).

DE CHAISEMARTIN, C. and D'HAULTFŒUILLE, X. (2015), "Fuzzy Differences-in-Differences", arXiv.

—— and —— (2020), "Two-way Fixed Effects Estimators with Heterogeneous Treatment Effects", *American Economic Review*, **110**, 2964–2996.

—— and —— (2022), "Difference-in-Differences Estimators of Intertemporal Treatment Effects" (Working Paper, University of Copenhagen).

DI MAGGIO, M., KERMANI, A. and MAJLESI, K. (2020), "Stock Market Returns and Consumption", *Journal of Finance*, **75**, 3175–3219.

GARDNER, J. (2021), "Two-Stage Differences in Differences", arXiv, preprint: not peer reviewed.

GARDNER, J., THAKRAL, N., TÔ, L. T., *et al.* (2023), "Two-Stage Differences in Differences" (Working Paper, University of Mississippi). https://linh.to/files/papers/tsdd.pdf.

GOBILLON, L. and MAGNAC, T. (2016), "Regional Policy Evaluation: Interactive Fixed Effects and Synthetic Controls", *Review of Economics and Statistics*, **98**, 535–551.

GOODMAN-BACON, A. (2021), "Difference-in-differences with Variation in Treatment Timing", *Journal of Econometrics*, **225**, 254–277.

GUIMARÃES, P. and PORTUGAL, P. (2010), "A Simple Feasible Procedure to fit Models with High-Dimensional Fixed Effects", *Stata Journal*, **10**, 628–649.

HARMON, N. A. (2022), "Difference-in-Differences and Efficient Estimation of Treatment Effects" (Working Paper, University of Copenhagen).

HOYNES, H. W., SCHANZENBACH, D. W. and ALMOND, D. (2016), "Long Run Impacts of Childhood Access to the Safety Net", *American Economic Review*, **106**, 903–934.

HUMPHREYS, M. (2009), "Bounds on Least Squares Estimates of Causal Effects in the Presence of Heterogeneous Assignment Probabilities" (Working Paper, Columbia University).

IMBENS, G. W. and RUBIN, D. B. (2015), *Causal Inference in Statistics, Social, and Biomedical Sciences* (Cambridge: Cambridge University Press).

JOHNSON, D. S., PARKER, J. A. and SOULELES, N. S. (2006), "Household Expenditure and the Income tax Rebates of 2001", *American Economic Review*, **96**, 1589–1610.

KAPLAN, G. and VIOLANTE, G. L. (2014), "A Model of the Consumption Response to Fiscal Stimulus Payments", *Econometrica*, **82**, 1199–1239.

—— and —— (2022), "The Marginal Propensity to Consume in Heterogeneous Agent Models", *Annual Review of Economics*, **14**, 747–775.

Kilts Center, University of Chicago (n.d.), "Nielseniq Consumer Panel Data".

KLINE, P., SAGGIO, R. and SOLVSTEN, M. (2020), "Leave-Out Estimation of Variance Components", *Econometrica*, **88**, 1859–1898.

LAIBSON, D., MAXTED, P. and MOLL, B. (2022), "A Simple Mapping from MPCs to MPXs" (Working Paper).

LEHMANN, E. L. and ROMANO, J. P. (2006), *Testing Statistical Hypotheses* (New York: Springer Science & Business Media).

LIU, L., WANG, Y. and XU, Y. (2022), "A Practical Guide to Counterfactual Estimators for Causal Inference with Time-Series Cross-Sectional Data", *American Journal of Political Science*, **68**, 160–176.

MACKINLAY, A. C. (1997), "Even Studies in Economics and Finance", *Journal of Economic Literature*, **XXXV**, 13–39.
MARCUS, M. and SANT'ANNA, P. H. (2020), "The Role of Parallel Trends in Event Study Settings: An Application to Environmental Economics", *Journal of the Association of Environmental and Resource Economists*, **8**, 235–275.
ORCHARD, J., RAMEY, V. A. and WIELAND, J. (2023), "Micro MPCs and Macro Counterfactuals: The Case of the 2008 Rebates" (NBER Working Paper 31584).
PARKER, J. A., SOULELES, N. S., JOHNSON, D. S., *et al.* (2013), "Consumer Spending and the Economic Stimulus Payments of 2008", *American Economic Review*, **103**, 2530–2553.
RAMBACHAN, A. and ROTH, J. (2023), "A More Credible Approach to Parallel Trends", *The Review of Economic Studies*, **90**, 2555–2591.
ROTH, J. (2022), "Pretest with Caution: Event-Study Estimates after Testing for Parallel Trends", *American Economic Review: Insights*, **4**, 305–322.
ROTH, J. and SANT'ANNA, P. H. (2023), "Efficient Estimation for Staggered Rollout Designs", *Journal of Political Economy Microeconomics*, **1**, 669–709.
—— and —— (2023), "When Is Parallel Trends Sensitive to Functional Form?", *Econometrica*, **91**, 737–747.
SANT'ANNA, P. H. and ZHAO, J. (2020), "Doubly Robust Difference-In-differences Estimators", *Journal of Econometrics*, **219**, 101–122.
SCHMIDHEINY, K. and SIEGLOCH, S. (2023), "On Event Studies and Distributed-Lags in Two-Way Fixed Effects Models: Identification, Equivalence, and Generalization", *Journal of Applied Econometrics*, **38**, 695–713.
STOCK, J. H. and WATSON, M. W. (2008), "Heteroskedasticity-robust Standard Errors for Fixed Effects Panel Data Regression", *Econometrica*, **76**, 155–174.
STREZHNEV, A. (2018), "Semiparametric Weighting Estimators for Multi-Period Difference-in-Differences Designs" (Working Paper, University of Pennsylvania).
SUN, L. and ABRAHAM, S. (2021), "Estimating Dynamic Treatment Effects in Event Studies with Heterogeneous Treatment Effects", *Journal of Econometrics*, **225**, 175–199.
THAKRAL, N. and TÔ, L. T. (2022), "Anticipation and Consumption" (Working Paper).
WOLFERS, J. (2006), "Did Unilateral Divorce Laws Raise Divorce Rates? A Reconciliation and New Results", *American Economic Review*, **96**, 1802–1820.
WOOLDRIDGE, J. M. (2021), "Two-Way Fixed Effects, the Two-Way Mundlak Regression, and Event Study Estimators" (Working Paper).
XU, Y. (2017), "Generalized Synthetic Control Method: Causal Inference with Interactive Fixed Effects Models", *Political Analysis*, **25**, 57–76.