Revisiting Event Study Designs,

with an Application to the Estimation of the Marginal Propensity to Consume*

Kirill Borusyak, Harvard University
Xavier Jaravel, London School of Economics

April 16, 2018

Abstract

A broad empirical literature uses "event study" research designs for treatment effect estimation, a setting in which all units in the panel receive treatment but at random times. We make four novel points about identification and estimation of causal effects in this setting and show their practical relevance. First, we show that in the presence of unit and time fixed effects, it is impossible to identify the linear component of the path of pre-trends and dynamic treatment effects. Second, we propose graphical and statistical tests for pre-trends. Third, we consider commonly-used "static" regressions, with a treatment dummy instead of a full set of leads and lags around the treatment event, and we show that OLS does not recover a reasonable weighted average of the treatment effects: long-run effects are weighted negatively, and we introduce different estimators robust to this issue. Fourth, we show that equivalent problems of under-identification and negative weighting arise in difference-in-differences settings when the control group is allowed to be on a different time trend or in the presence of unit-specific time trends.

We show the practical relevance of these issues in a series of examples from the existing literature. We focus on the estimation of the marginal propensity to consume out of tax rebates: according to our preferred specification, the marginal propensity to consume is much lower than (about half of) the main estimates in the literature.

The main message for practitioners is that because of identification issues and negative weighting in event study designs, results from common specifications are likely to seem non-robust. These problems can be alleviated in a principled way by using parametric and semi-parametric estimators and tests.

*We thank Alberto Abadie, Isaiah Andrews, Raj Chetty, Itzik Fadlon, Ed Glaeser, Peter Hull, Guido Imbens, Larry Katz, Jack Liebersohn and Jann Spiess for thoughtful conversations and comments. We are grateful to Jonathan Parker for his support in accessing and working with the data. The results in the empirical part of this paper are calculated based on data from The Nielsen Company (U.S.) LLC and provided by the Marketing Data Center and the University of Chicago Booth School of Business. The latest version of the paper is available here.

1 Introduction

A broad empirical literature in labor economics, public finance, finance and empirical macroeconomics uses event study research designs for treatment effect estimation, a setting in which all units in the panel receive treatment but at random times. We make a series of novel points about identification and estimation of causal effects in such a setting, which are closely related to the well-known age-cohort-time problem. We then establish the practical relevance of these points in light of the existing literature and in a specific application, estimating the impulse response function of consumption expenditures to tax rebate receipt.

We first show that in the presence of unit and time fixed effects, it is impossible to identify the linear component of the path of pre-trends and dynamic treatment effects. Identification of the dynamic causal effects only up to a linear trend is particularly problematic because researchers usually want to test for the absence of pre-trends prior to the event and to document changes in the outcome variable after treatment in a non-parametric way. Intuitively, the path of pre-trends and dynamic treatment effects is identified only up to a linear trend because, "within a treatment unit", one cannot disentangle the passing of absolute time from the passing of time relative to the treatment event. We show formally that the collinearity problem that arises in this setting is effectively the same as the age-cohort-time problem that others have studied.

We then propose two approaches to address this underidentification issue. Our first strategy is to restrict the pre-trends in the fully dynamic specification, while keeping unit fixed effects. Our second strategy consists in replacing unit fixed effects with unit random effects. We develop statistical tests for the validity of both approaches, as well as a graphical test for the first approach.

Next, we turn to estimation of the average treatment effect, with a particular focus on specifications that are meant to average all dynamic treatment effects post treatment. We show that the specification that is commonly used in the literature estimates an average of treatment effects that severely overweighs short-run effects and weighs long-run effects negatively. This issue can be serious, such that the estimated average treatment effect can be outside of the convex hull of the true dynamic treatment effects. To address the problem, we provide alternative parametric and semi-parametric estimation techniques, which always average dynamic effects in a convex way.

Intuitively, the short-run bias originates from a peculiar type of extrapolation implicitly performed by the OLS. Recall that a difference-in-differences estimator is meant to use the outcome gap between some units A and B in an early period s before they receive treatment for constructing counterfactuals for period t when B is treated but A is not. However, OLS regression conflates the situations where neither unit is treated at time s, *or both are*—A and B have the same treatment status in both cases, misleadingly opening room for comparison. Specifically, if causal effects exhibit dynamics, the difference between post-treatment outcomes of B versus A reflects the wedge between long- and short-run effects, since B was treated earlier. This difference is *subtracted* from an analogous one for period t, thereby creating a short-run bias.

We demonstrate that the problem is particularly severe in event studies with no control group. While a control group partially alleviates the problem, expanding the set of good comparisons, it has to be quite large for the problem to disappear, and the short-time bias is present even then.

Beyond these core results, we show that OLS suffers from similar weighting problems in two related but different designs. It happens in classical diff-in-diffs if the control group is allowed to be on a separate time trend—an approach that applied studies sometimes take when they are not certain of the quality of their control group. It also applies to regressions that include unit-specific time trends. In both cases, the problem arises even if the control group is very large.

Finally, we establish the empirical relevance of these various points by describing a series of recent and influential papers in the literature that run precisely the specifications whose undesirable properties we point out in this paper. Moreover, we use extended data following Broda and Parker (2014) to estimate the impulse response

function of consumption expenditures to tax rebate receipt, using the random timing of tax rebate receipt across households. In this application, we find that underidentification can lead to seemingly unstable results when it is not properly handled. We also find that the marginal propensity to consume estimated by the canonical regression has a strong short-run bias (upwards) due to the weighting issues. At the same time, semi-parametric specifications demonstrate that long-run effects are very difficult to identify in this type of natural experiment without undesirable extrapolation.

Although our paper is aimed primarily for the applied reader, it contributes to three micro-econometric literatures. Imbens (2015) notes that "possibly [...] because of lack of compelling evidence that simple methods (for example, ordinary least squares regression) are not adequate, few of the estimators proposed in the recent literature have been widely adopted". We view our paper as providing such evidence for difference-in-difference settings with differential event timing. Imbens's paper and Sloczynski (2016) are similar in spirit. The seminal paper by Angrist (1998), which argues in the other direction, provides a useful benchmark to highlight our points. We also speak to the literature that noted, more informally and for specific applications, some problems with event studies and unit-specific trends, such as Wolfers (2006), Fadlon and Nielsen (2015) and particularly Meer and West (2015). Finally, in developing our semi-parametric estimator, we contribute to the large literature including Hirano et al. (2003) and Abadie (2005).

The remainder of this paper is organized as follows. In Section 2, we describe our setting, the key specifications we study, and how they are consistent with a causal model. Section 3 presents the underidentification problem and our solutions, while Section 4 describes the negative weighting issue and how to address it. Section 5 considers a variety of extensions. Finally, Section 6 relates our points to the existing literature and presents the estimation of the marginal propensity to consume out of tax rebates as an application.

2 Setup

2.1 Data-generating Process

Consider a panel of $i = 1, \ldots, N$ units in which the outcome Y_{it} is observed for $t = 1, \ldots, T$ periods ("calendar time"), or possibly for a subset thereof. In our main setting, every unit receives treatment in some period E_i within the sample and stays treated forever.[1] Units with the same treatment period are referred to as a cohort. Let $K_{it} = t - E_i$ denote the "relative time"—the number of periods relative to the event. The indicator variable for being treated can therefore be written as $D_{it} = \mathbf{1}\{t \geq E_i\} = \mathbf{1}\{K_{it} \geq 0\}$.

Empirical papers using this event study setup often pursue some of the following three goals. They first estimate whether the treatment has an effect on average. Second, they test for pre-trends to lend credibility to the research design. Finally, they may study in more depth the dynamics of the causal effect. With these goals in mind, we choose a class of data-generating processes which is very flexible on the dynamics but abstracts away from a variety of other specification issues:[2]

$$Y_{it} = \tilde{\alpha}_i + \tilde{\beta}_t + \sum_{k=-\infty}^{\infty} \tilde{\gamma}_k \mathbf{1}\{K_{it} = k\} + \tilde{\varepsilon}_{it}. \tag{1}$$

Here $\{\tilde{\gamma}_k\}$ for $k < 0$ correspond to pre-trends, and for $k \geq 0$—to dynamic effects k periods relative to the event.[3] The average effect is $\sum_{k=0}^{\infty} \omega_k \tilde{\gamma}_k$ for some weighting scheme $\{\omega_k\}$, but researchers rarely specify it based on their

[1] We consider settings with a control group in Section 5.1.1.

[2] We allow for more general models in the Extensions section, such as heterogeneity of treatment effects across units. All of our results also directly extend to adding time-varying controls.

[3] By $k = \pm\infty$ we mean the largest number possible given the sample.

economic question and instead rely on the regression to do something reasonable. Tildes indicate the parameters of the "true model", reflecting the data generating process, and later on we express the estimands of commonly-used regression specification in terms of these parameters.[4] $\tilde{\alpha}_i$ and $\tilde{\beta}_t$ are unit and period fixed effects, respectively, and $\tilde{\varepsilon}_{it}$ is random noise. We call equation (1) the *fully dynamic specification*.

This formulation is consistent with a causal model in which each unit i for each period t has a set of potential outcomes $Y_{it}^{(k)}$ for each integer k, only one of which is realized. Treatment effects, expressed relative to one of them, e.g. $Y_{it}^{(-1)}$, are homogenous across units and calendar time periods, and depend only on k: $Y_{it}^{(k)} - Y_{it}^{(-1)} = \tilde{\gamma}_k$.[5] Furthermore, $Y_{it}^{(-1)} = \tilde{\alpha}_i + \tilde{\beta}_t + \tilde{\varepsilon}_{it}$, which is a standard assumption necessary for the validity of difference-in-difference approaches (Angrist and Pischke, 2008, p. 156). Together these assumptions deliver equation (1).

If the event is unpredictable, it is not known whether the current period corresponds to $K_{it} = -1, -2$, or any other negative number. As a consequence, $Y_{it}^{(-1)} = Y_{it}^{(-2)} = \ldots$, so $\tilde{\gamma}_k = 0$ for all $k < 0$. In that sense, random timing of the event implies that there cannot be any pre-trends.[6] Equation 1 reduces to the following specification, which we call *semi-dynamic* and take to be true if the empirical design is valid:

$$Y_{it} = \tilde{\alpha}_i + \tilde{\beta}_t + \sum_{k=0}^{\infty} \tilde{\gamma}_k \mathbf{1}\{K_{it} = k\} + \tilde{\varepsilon}_{it}. \tag{2}$$

2.2 Current Practice

In the current practice, it is prevalent to estimate models similar to (1) and (2) using OLS with two-way (unit and period) fixed effects. Different papers impose different restrictions on (1), but the following specification covers most of them:

$$Y_{it} = \alpha_i + \beta_t + \sum_{k=-A}^{B-1} \gamma_k \mathbf{1}\{K_{it} = k\} + \gamma_{B+} \mathbf{1}\{K_{it} \geq B\} + \varepsilon_{it}, \tag{3}$$

where $A \geq 0$ leads of treatment are included, together with $B \geq 0$ terms for specific short-run effects and a single last coefficient γ_{B+} for all longer-run effects.[7] Note the absence of tildes: unless $A = 0$ and $B = \infty$, this equation does not coincide with the true model (2). We will study where its coefficients converge to in large samples as functions of the true parameters. We will occasionally use hats to mark finite-sample objects.

The simplest and perhaps the most prevalent regression is (3) with $A = B = 0$, i.e.

$$Y_{it} = \alpha_i + \beta_t + \gamma_{0+} D_{it} + \varepsilon_{it}. \tag{4}$$

We refer to this specification as *static* or, following Allegretto et al. (2013), *canonical*, and will discuss it at great length later in the paper. Compared to the fully dynamic one, it imposes no pre-trends and constant treatment effects for all k. The other extreme is, of course, $A = B = \infty$, which is just the fully dynamic specification with no restrictions.

Often regressions are run using all available data, but sometimes the sample is balanced around the event time: only observations with $K_{it} \in [\underline{k}, \bar{k}]$, $\underline{k} < 0 \leq \bar{k}$, are included and only for units which are observed for all corresponding periods. We discuss pros and cons of this approach later on.

[4] The notation for these estimands does not use tildes.

[5] Note that one cannot hope to estimate treatment effects relative to the situation in which the event never happens, simply because this circumstance is not observed in the data. Picking $k = -1$ as the omitted category is an innocent and standard normalization.

[6] In some settings anticipation of the event is possible but limited to a fixed number of A periods. In that case $\tilde{\gamma}_k = 0$ for $k < -A$, and any $k < -A$ can be chosen as the omitted category.

[7] One term, e.g. $\gamma_{-1}\mathbf{1}\{k_{it} = -1\}$, can be omitted as a normalization.

3 Underidentification of the Fully Dynamic Specification

3.1 Problem

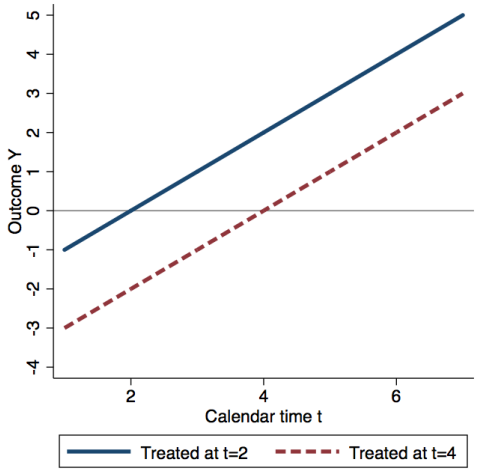
In this section, we show that the fully dynamic specification, given by equation (1), suffers from a fundamental underidentification problem. The goal of such a specification is to recover the dynamic path of causal effects $\{\tilde{\gamma}_k\}_{k=-\infty}^{\infty}$. We show that a linear trend in this path is not identified. One can start from any set of points estimates $\{\gamma_k\}_{k=-\infty}^{\infty}$, add a linear trend (in k) and adjust the sets of point estimates for the year fixed effects β_t and the unit fixed effects α_i to keep the same predicted value. Identification of the dynamic causal effects $\{\tilde{\gamma}_k\}_{k=-\infty}^{\infty}$ up to a linear trend is particularly problematic because researchers usually want to test for the absence of "pre-trends" prior to the event,[8] and more generally are hoping to document changes in the outcome variable after treatment in a non-parametric way.[9] In this section, we first illustrate the underidentification issue graphically to gain intuition. We then show mathematically where it stems from.

The intuition for why the fully dynamic specification is underidentified can easily be grasped with a simple example. We intentionally make it extreme, to isolate the fundamental problem from any additional effects or noise, which exist on top of it in real data; the mechanism is completely general. Consider Fig. 1, which plots the outcomes for a simulated dataset, which covers two cohorts. One is treated early at $t = 2$ (solid line), and the other one later at $t = 4$ (dashed line). Both groups are observed for all periods $t = 1, \ldots, 7$, and the outcomes exhibit linear growth with the same slope of one, albeit starting from different levels. There are two interpretations of what could cause such dynamics. On one hand, treatment could have no impact on the outcome, in which case the level difference corresponds to the cohort effect and trends are just a common feature of the environment, formalized by the time fixed effects. On the other hand, one could note that the outcome equals the number of periods since treated for both groups and all time periods: at the moment of treatment the outcome equals zero, it is negative before and positive after. So a possible interpretation is that the outcome is entirely driven by causal effects of treatment and anticipation of treatment. Note that in the former case, the solid line is a vertical shift up (a level shift) from the dashed line, while in the latter case it is a horizontal shift to the left that is due to the differential timing. With straight lines, these are observationally equivalent. One cannot hope to distinguish between dynamic causal effects and a combination of cohort effects with time trends, or more generally, of unit and time effects.

[8] In contrast, underidentification of the set of point estimates $\{\tilde{\gamma}_k\}_{k=-\infty}^{\infty}$ up to a constant, which is a well-known fact, is not problematic because it does not in itself prevent the study of dynamic effects and pre-trends.

[9] Alternatively, researchers could focus on testing for specific changes in the outcome variable after treatment that are invariant up to a linear transformation of the path of the dynamic causal effects $\{\gamma_k\}_{k=-\infty}^{\infty}$, i.e. they are identified. A change in the slope of the outcome variables or a sudden jump in the outcome variable after treatment are examples of parametric specifications that could be tested despite the underidentification problem.

Figure 1: Underidentification of Fully Dynamic Specification



We can now turn to a formal investigation of these issues. Observe that for any constant h, the following two representations are equivalent:

$$\begin{aligned} \hat{Y_{it}} &\equiv \alpha_i + \beta_t + \sum_{k=-\infty}^{\infty} \gamma_k \mathbf{1}\{K_{it} = k\} \\ &= (\alpha_i + h \cdot E_i) + (\beta_t - h \cdot t) + \sum_{k=-\infty}^{\infty} (\gamma_k + h \cdot k)\, \mathbf{1}\{K_{it} = k\} \end{aligned} \tag{5}$$

because by definition $t - E_i = K_{it}$. As a result, the dynamic causal effects $\{\gamma_k + h \cdot k\}$ fit the data just as well as the original $\{\gamma_k\}$ path, although these two sets of coefficients paint vastly different pictures about the causal effects.

To gain further intuition about the nature of the underidentification, we show that the empirical model above nests a specification with collinear terms. Specifically, replace the set of unit fixed effects $\{\alpha_i\}$ with a linear predictor in initial treatment period $\lambda + \alpha E_i$ (i.e. the outcomes of different "cohorts" of units experiencing the event at different times are allowed to differ in a linear way), the set of year fixed effects $\{\beta_t\}$ with a time trend βt, and the set of fully dynamic causal effects $\{\gamma_k\}$ with a trend in relative time γK_{it}. The fundamental underidentification problem described above can be seen immediately in the following regression:

$$Y_{it} = \lambda + \alpha E_i + \beta t + \gamma K_{it} + u_{it}$$

given that $t - E_i = K_{it}$. In other words, the presence of a linear term in the initial treatment period is necessary for the identification problem to arise. Unit fixed effects subsume such effects. In the presence of a control group, cohort fixed effects or unit fixed effects do not cause any identification problem because the control group pins down the year effects. The problem is effectively the same as the (well-known) age-cohort-time problem in the regression

$$Y_{it} = \underbrace{\alpha_{E_i}}_{\text{Cohort FE}} + \underbrace{\beta_t}_{\text{Time FE}} + \underbrace{\gamma_{t-E_i}}_{\text{Age FE}} + u_{it},$$

where E_i is the date of birth.

We want to stress that typically, but not always, only a linear component of the $\{\gamma_k\}$ path is not identified.[10]

[10] There are some exception to these rules, e.g. when treatment is staggered but happens at periodic intervals.

It is not possible to reproduce a *nonlinear* $\{\gamma_k\}$ path perfectly with unit and period fixed effects. The reason is that by definition, such a path is a nonlinear function of $\gamma(t - E_i)$, and it cannot be represented as a sum of any functions $\alpha(E_i)$ and $\beta(t)$.[11]

In sum, the linear trend in the dynamic schedule of causal effects $\{\tilde{\gamma}_k\}_{k=-\infty}^{\infty}$ is not identified, because one cannot disentangle the effects of passing of absolute time t and relative time k when there is no control group and in presence of unit fixed effects. More specifically, unit fixed effects create an identification problem because they subsume "linear cohort effects" (i.e. linear predictors of the form $\lambda + \alpha E_i$). The calendar year (t) is equal to the year in which the event happens for unit i (E_i) plus the "relative time" (K_{it}): there is a perfect linear relationship between these effects and it is therefore impossible to observe independent variation in these variables.

3.2 Solutions

3.2.1 Overview

As we explained above, calendar time, relative time, and a linear term in the initial treatment period cannot be included together in the regression. To avoid this, additional restrictions on the statistical model have to be imposed. Dropping unit fixed effects is an immediate fix, with the advantage of being very easy to implement, but it suffers from two important drawbacks: it requires strengthening the identification assumption, and it reduces power. Although dropping unit fixed effects may be a reasonable approach in some settings, we develop two other approaches that address underidentification and do not suffer from these drawbacks. Our first strategy is to restrict the pre-trends in the fully dynamic specification, while keeping unit fixed effects, and we show how to test this restriction. Our second strategy is to replace unit fixed effects with unit random effects, which is also testable.

Both of these strategies can be justified by reasonable assumptions about the nature of the variation in the timing of the event across units. The restriction of pre-trends is justified when the event is *unpredictable* conditional on unit characteristics, while the random effects model is warranted when the timing of the event is *randomly assigned* across units.[12] Consider two examples to clarify this distinction. When the effect of health shocks on income is of interest (e.g. Dobkin et al., 2015), it is plausible that low-income individuals may be less healthy and get hospitalized earlier on average. Yet, conditional on permanent income level, individuals may not be able to predict when the health shock will happen. In contrast, in the case of tax rebates studied by Parker et al. (2013), people could know when they are getting the rebate. However, the date was assigned based on the last two digits of the Social Security Number, so it was uncorrelated with any relevant individual characteristics.

The following subsections develop each of these approaches and discuss two related issues: the role of unit fixed effects in balanced samples, and the possibility of adding a control group as a solution to the identification problem.

3.2.2 Restricting Pre-Trends

We begin with the situation where event timing is supposed to be randomly assigned *conditionally* on the fixed effect $\tilde{\alpha}_i$, and unpredictable. The former assumption justifies the use of difference-in-differences type approaches. And latter one means that the outcome cannot be adjusted based on anticipation of the event, so there can be no pre-trends, $\tilde{\gamma}_k = 0$ for $k < 0$. This assumption can be tested statistically and graphically, and then imposed for efficient estimation of causal effects. We discuss these issues in order.

Under the no pre-trends null hypothesis, the true model is semi-dynamic, which has no identification issues. The alternative allows for the fully dynamic specification, which is only set identified. Despite this, the F-test works and can be implemented in the following straightforward way. Start from the fully dynamic regression and drop *any*

[11] To see why this is the case, imagine that t and E_i are continuous and take a cross-partial derivative $\partial^2\gamma(t - E_i)/\partial t \partial E_i = -\gamma''(t - E_i) \neq 0$, whenever γ is nonlinear. In contrast, $\alpha(E_i) + \beta(t)$ always has zero cross-partial.

[12] These two senses of "randomness" of the event timing appear to have been conflated in the existing literature.

two terms corresponding to $k_1, k_2 < 0$. This is the minimum number of restrictions for point identification, to pin down a constant and a linear term in K_{it}. Then use the F-test on the pre-trends remaining in the model.[13] The F-test compares the residual sums of squares under the restricted and unrestricted specifications, where the former is always semi-dynamic, and the latter is fully dynamic with two restrictions. Precisely due to underidentification, the fully dynamic specification with two restrictions is effectively unrestricted and its fit is identical for any k_1 and k_2, so the F-statistic will be invariant to k_1 and k_2 even in finite samples.

This test has power only against non-linear pre-trends. Indeed, nothing in the data can point to a linear pre-trend—that is the essence of underidentification. However, if the empirical design is actually flawed, i.e. event timing is correlated with unobservables, there is no reason for pre-trends to be exactly linear, and the test will detect them.

While we are not aware of any empirical papers implementing this F-test, a common way to check for pre-trends is to plot the path of $\hat{\gamma}_k$ before and after treatment. Sometimes this is called the event study approach. It originates from the standard difference-in-differences setup, where only an additive constant in $\tilde{\gamma}_k$ is not identified, and it is irrelevant for visual inspection; $\hat{\gamma}_{-1}$ is typically set to zero. In contrast, in the staggered design two restrictions have to be made. Different choices of the two restrictions, $\hat{\gamma}_{k_1} = \hat{\gamma}_{k_2} = 0$, will matter a lot for the whole picture: the whole path of estimated $\hat{\gamma}_k$ gets rotated by adding $\hat{h} \cdot k$ for some constant $\hat{h}$. If there are no pre-trends in the data-generating process, $\hat{h}$ asymptotically converges to zero for any k_1 and k_2 as the number of units grows. However, in finite samples difference may be large, particularly in longer panels, since $\hat{h}$ is multiplied by k.

So how does one pick the two omitted categories? While choosing $k_1 = -1$ and $k_2 = -2$ might seem natural, we propose setting the omitted categories far apart. Under the null hypothesis, this greatly reduces standard errors for most individual coefficients on the graph. To understand why, imagine that a line on a plane is drawn through two points with fixed x-coordinates $x_1 \neq x_2$, but stochastic y-coordinates, for simplicity with mean zero. The position of the line will be much more stable when x_1 and x_2 are far from each other. This is true both for the slope of the line (the analog of $\hat{h}$) and its value at a typical x (the analog of the $\hat{\gamma}_k$). The fully-dynamic specification with two restriction effectively draws a line and evaluates all dynamic effects relative to it. When k_1 is far from k_2, e.g. $k_1 = -1$ and k_2 close to the most negative value of K in the sample, it will be much less likely that a linear pre-trend (although perhaps statistically insignificant) will be visible. Remember that linear pre-trends are never possible to detect in the data, so choosing $k_2 = -2$ would only reduce the usefulness of the graph, distracting attention from nonlinearities in the pre-trends.

Even if the two restrictions are chosen well, this graph should only be used to evaluate pre-trends; it does not estimate the treatment effects efficiently. Once the researcher is comfortable with the assumption of no pre-trends, all γ_k, $k < 0$, should be set to zero. The semi-dynamic specification should be estimated and its coefficients plotted to provide a graphical illustration of the dynamics of causal effects.[14]

3.2.3 Unit Random Effects

We now consider a second sense in which the timing of the event is random: the treatment period E_i is independent of the relevant unit characteristics—in our model (1), the time-invariant unit intercept $\tilde{\alpha}_i$. In such a setting, the estimation can be carried out without unit fixed effects, which are no longer necessary for the research design. Dropping unit fixed effects immediately addresses the underidentification problem.[15] However, doing so reduces efficiency: instead, we propose to carry out estimation in a random effects model. In addition to increasing efficiency, another advantage of using a random effects model is that we can test the hypothesis that the treatment period is

[13] Both the restricted and unrestricted specifications are identified now, so standard results about the F-test behavior apply.

[14] Note that in case there is truly a linear trend in the set of $\{\tilde{\gamma}_k\}_{-\infty}^{\infty}$, then the results from the fully dynamic specification can be interpreted as a test for any change relative to this linear trend around the time of the event.

[15] Recall that unit fixed effects create an identification problem because they subsume "linear cohort effects"

independent of the unit fixed effects. As in the case of the no pre-trends assumption, the random effects assumption can be tested against some, although not all, alternatives, and then imposed to regain identification.

When we discussed underidentification of the fully dynamic specification, we emphasized that for any path of $\{\gamma_k\}$, identical fit of the model can be produced with the path $\{\gamma_k + h \cdot k\}$. But the same result holds for the unit fixed effects: $\{\alpha_i - h \cdot E_i\}$ and $\{\alpha_i\}$ fit the data equally well, as long as other coefficients in the model are adjusted appropriately (see equation (5)). As a consequence, it is impossible to test whether $\tilde{\alpha}_i$ is *uncorrelated* with E_i—the estimates can always be made uncorrelated by choosing h. [16]

Yet, *independence* is testable; consider the following simple F-test. The restricted specification allows for arbitrary time and causal effects, but no cohort (or unit) effects. The unrestricted one is

$$Y_{it} = \sum_e \alpha_e \mathbf{1}\{E_i = e\} + \beta_t + \sum_{k=-\infty}^{\infty} \gamma_k \mathbf{1}\{K_{it} = k\} + \text{noise}.$$

It is almost equivalent to the fully dynamic specification, except that cohort effects are included instead of unit effects. As in the F-test of Section 3.2.2, two normalizations are required, but this time two arbitrary cohort terms $\alpha_e \mathbf{1}\{E_i = e\}$ should be dropped.[17] A joint restriction $\alpha_e \equiv 0$ is then tested, but because of underidentification, the test is only powerful against nonlinear functions $\mathbb{E}[\alpha_i \mid E_i]$.

When the researcher is comfortable to impose the independence assumption, they should use the random effects estimator as their preferred one for the full path of $\{\tilde{\gamma}_k\}$. Remember that in general, the setup does not imply there are no pre-trends. If the units' outcomes can adjust to the randomized, yet known in advance event timing, the pre-trends are part of the treatment effect.

3.2.4 Related Issues

Using a Balanced Sample: As previously discussed, without unit fixed effects there is no underidentification problem, but in some settings the research design requires the inclusion of unit fixed effects, for instance if treatment is unpredictable only conditional on some time-invariant characteristics. In such settings, a potential easy fix for the underidentification problem would be to drop fixed effects and balance the sample around the initial treatment period (i.e. restrict the sample such that each unit appears for the same number of periods before and after the initial treatment period). Indeed, there is a view that unit fixed effects are essentially irrelevant in event studies on a balanced panel.

The intuition underlying this view is that balancing the sample addresses one key issue that is handled by unit fixed effect in unbalanced panels: the changing composition of the sample. Omitting unit fixed effects when working with an unbalanced panel may be a big assumption because of selection into treatment. For instance, units that are treated earlier in the sample may be different from other units and because the sample is unbalanced they spend a bigger share of the sample under treated status. Therefore, in the absence of unit fixed effect one would worry that the estimated coefficient on the treatment dummy may partly reflect selection.

In fact, balancing the sample does *not* help address such selection effects. We show in Appendix D that omitting unit fixed effects when working with *balanced* panels is just as worrisome as when working with *unbalanced* panels (for the purpose of dealing with selection effects). A balanced panel appears to solve the "selection issue" that is salient in the case of unbalanced panels because each unit gets treated for the same number of years during the panel. However, in practice year fixed effects absorb all of the variation for years at the beginning (where all units are untreated) or end (where all units are treated) of the sample. For this reason, the point estimate we obtain for

[16] An example of a conceptually different moment restriction which could be tested but would not provide much intuition is $Cov(\alpha_i^2, E_i) = 0$.

[17] The unrestricted specification can be estimated by OLS or random effects—the results will be identical.

the treatment effect by running a regression in a balanced sample is exactly the same as the one obtained by running the same regression in this balanced sample further restricted to years in the middle of the sample (which means we are effectively running a regression on an *unbalanced* panel). Appendix D discusses this more formally and shows that endogenous selection effects in the data generating process affects the consistency of the point estimate in the same way under balanced and unbalanced panels.

Adding a Control Group: A number of empirical papers using random timing research designs start with a sample which includes units that are never treated, or could conceivably collect such a sample. Therefore, it would be possible to include a control group of units that never experience treatment in the estimation sample. In many instances, it is difficult to construct a control group that plausibly provides valid counterfactuals for the treatment group. But assuming that such a control group is available, then including it in the estimation sample solves the underidentification problem because the control group can be used to estimate the year effects independently of the causal effect of treatment. However, the strategy of adding a control group has two important limitations. First, if the control group is small relative to the treatment group, important finite-sample issues can arise. Second, one cannot allow the control group to be on its own time trend, otherwise the underidentification problem is left intact. We discuss both of these issues in greater depth in Section 5.

4 Negative Weighting in Canonical Regression

4.1 Problem

In this section we show that fundamental underidentification discussed above creates problems for more restricted specifications. Our flagship case will be the canonical regression (4), but the argument extends to all specifications with two-way fixed effects, which do not allow for flexible dynamic treatment effects. We show that these regressions estimate an average of treatment effects that severely overweighs short-run effects and weighs long-run effects negatively. That is, if programs P1 and P2 have similar short-run effects but P1 is uniformly more efficient in the long-run, the canonical regression will show that P1 has *lower* average effect.

Assume that the design is valid in the sense that there are no pre-trends, so the true model is semi-dynamic (2). People often summarize treatment effects by the γ coefficient from the canonical regression,

$$Y_{it} = \alpha_i + \beta_t + \gamma D_{it} + \varepsilon_{it}.$$

This specification is valid under a restriction that $\tilde{\gamma}_k$ are equal for all $k \geq 0$, i.e. that treatment leads to an immediate and permanent jump in the outcome variable and no further effects. This restriction should not hold in most applications, in our view. Quite often treatment effects are growing or decaying over time, may kick in with a delay, etc. However, there is a perception that γ should estimate average treatment effects with some reasonable weights. While true in some other contexts (e.g. Angrist 1998), we show that this logic does not apply to the canonical regression estimand-–the weights are not even necessarily positive.[18]

The first step in the argument is that γ is *some* weighted average of $\{\tilde{\gamma}_k\}$ with weights which can be easily estimated from the data and solely depend on the *grid*—the distribution of treatment periods E_i and sample selection.

Lemma 1. *The canonical regression OLS estimand can be expressed as a weighted average of dynamic treatment*

[18] Imbens (2015) is one of the few papers that shows negative weighting by OLS, in a different context.

effects,

$$\gamma = \sum_{k=0}^{\infty} \omega_k \tilde{\gamma}_k$$

with weights ω_k that sum up to one and equal the coefficients for D_{it} from in the following regressions:

$$\mathbf{1}\{K_{it} = k\} = FE_i + FE_t + \omega_k D_{it} + noise, \qquad k \geq 0. \tag{6}$$

where FE_i denotes unit fixed effects and FE_t time fixed effects.

To gain intuition for this lemma, note the following: by linearity of OLS, one can recover the coefficient for D_{it} in the canonical regression with Y_{it} as the outcome by instead running two canonical regressions with subcomponents of Y_{it} as the outcome variable and then summing up the coefficients on D_{it} obtained in each of these regressions.[19] Consider first running a canonical regression with $\left(\tilde{\alpha}_i + \tilde{\beta}_t + \tilde{\varepsilon}_{it}\right)$ as the outcome (one could do this in theory if the true parameters of the model were known), and then another canonical regression with $\sum_{k=0}^{\infty} \tilde{\gamma}_k \mathbf{1}\{K_{it} = k\}$ as the outcome, and finally sum up the coefficients for D_{it} obtained in each of these two regressions to recover γ. The first regression will load on the fixed effects and not on D_{it}, so γ comes solely from the second regression. By the same logic, one further notes that γ can be recovered by running a series of canonical regressions with $\tilde{\gamma}_k \mathbf{1}\{K_{it} = k\}$ as the outcomes (repeating for all $k \geq 0$) and summing up the coefficients. Since $\tilde{\gamma}_k$ is a multiplicative constant in $\tilde{\gamma}_k \mathbf{1}\{K_{it} = k\}$, each of these canonical regressions generates coefficients for D_{it} that can be written $\tilde{\gamma}_k \cdot \omega_k$, for ω_k determined by specification (6). Importantly, the only variables required by (6) are the unit identifier, calendar time, and relative time—what we call the grid.[20]

Now the question is whether these ω_k weights are "reasonable" in some sense, and our answer is strongly negative. The canonical regression estimand suffers from a severe short-run bias, and weights long-run effects negatively. Although a general characterization of ω_k does not seem feasible, we demonstrate our result and intuitions for it in four ways. First, we show that the short-run bias originates from a peculiar type of extrapolation performed by the OLS when treatment timing is heterogenous across units. Second, we connect negative weighting to the propensity score regression to show why long-run effects are more likely to be negatively weighted. Third, we solve for an approximation of the weights in the general case, where the short-run bias always arises. Finally, we get a striking closed-form solution for ω_k in a simplified special case.

4.1.1 Intuition: Forbidden Extrapolations

In a nutshell, the intuition for the short-run bias is related to "forbidden extrapolations" performed by OLS, where post-treatment periods are used to provide counterfactuals for the earlier ones, rather than the other way round.[21]

As a benchmark with no bias, we first remind the reader how classical diff-in-diff estimators perform extrapolations using double-differences. Consider the simplest setting which involves two groups of homogenous units, A

[19] To be explicit, if $Y = A + B$, then the specifications

$$Y = \beta^Y X + \epsilon$$
$$A = \beta^A X + \eta$$
$$B = \beta^B X + \zeta$$

yield

$$\beta^Y = \frac{Cov(Y, X)}{Var(X)} = \frac{Cov(A, X) + Cov(B, X)}{Var(X)} = \beta^A + \beta^B.$$

The result holds for multivariate X.

[20] To show that weights always add up to one, imagine that $\tilde{\gamma}_k = \tilde{\gamma} \neq 0$ for all $k \geq 0$. The canonical regression is then correctly specified and provides consistent estimate $\gamma = \tilde{\gamma}$. But $\gamma = \sum_{k \geq 0} \omega_k \tilde{\gamma}$, so $\sum_{k \geq 0} \omega_k = 1$. Since weights do not depend on the true outcomes, this result holds generally.

[21] This intuition derives from Meer and West (2015, Sec. 2.1). We are very grateful to Jonathan Meer for pointing us at their insightful paper.

Table 1: Treatment Status in the Minimal Examples

Panel A

Period \ Group	A	B
0	0	0
1	0	1

Panel B

Period \ Group	A	B
0	0	0
1	0	1
2	1	1

(control) and B (treatment), and two time periods 0 and 1. Units in A are never treated, and units in B receive treatment at $t = 1$. Table 1A illustrates how the treatment status depends on the group and period. Then in the canonical regression

$$y_{it} = \alpha_i + \beta_t + \gamma D_{it} + \text{noise},$$

the estimate of the treatment effect, γ, is determined by a double difference

$$\gamma_{classic} = (y_{B1} - y_{A1}) - (y_{B0} - y_{A0}).$$

The difference between period-0 outcomes is extrapolated to period 1 to serve as a no-treatment counterfactual. It is permitted by the conventional parallel trends assumption, so we call it an "admissible extrapolation".

We now proceed to the case of a staggered treatment diff-in-diff, where a different type of a double-difference also provides information on treatment effects. Extend the previous example by adding period 2 when unit A also gets treated (see Table 1B). Then under the canonical specification, the difference $y_{B2} - y_{A2}$ identifies the permanent level difference between A and B, $\alpha_B - \alpha_A$, in the same way as $y_{B0} - y_{A0}$ does. As a result,

$$\gamma_{staggered} = (y_{B1} - y_{A1}) - (y_{B2} - y_{A2})$$

is also a valid estimator of the treatment effects, provided they are homogenous.

This double-difference becomes problematic when short- and long-run effects differ. Indeed, y_{B1} and y_{A2} both reflect the immediate causal effect, denote it γ_S, whereas y_{B2} reflects the long-run effect γ_L, and the remaining outcome observations are pre-treatment. It is then evident that $\gamma_{classic} = \gamma_S$ and $\gamma_{staggered} = 2\gamma_S - \gamma_L$. Instead of estimating the average, the second number puts double weight on the short-run effect and a negative weight on the long-run one (in earlier notation, $\omega_0 = 2$ and $\omega_1 = -1$). We call that a forbidden extrapolation.

A typical panel, like the one from our three-period example, will lend itself for both admissible and forbidden extrapolations, and both types will be used by OLS for maximal efficiency under treatment effect homogeneity. But robustness to treatment effect dynamics is, in our view, more important for empiricists than efficiency, so forbidden extrapolations should not be used.

4.1.2 Formal Results

We now turn to a more formal discussion of how weights ω_k are determined and when they can be negative. We show in Appendix C that the weighting scheme implicitly used by OLS is proportionate to the residuals in the linear propensity score regression. That is, suppose treatment indicator D_{it} is regressed on all other right-hand side variables—unit and period fixed effects in our case. Then observations it which have fitted values above D_{it} (and hence negative residuals) will have a negative weight in the original canonical regression—larger Y_{it} produces smaller γ. Although D_{it} is a dummy variable, the linear probability model can easily generate fitted values $\hat{D}_{it} > D_{it} = 1$ for some post-treatment observations. The outcome variable Y_{it} contains treatment effects for these observations, and those will be weighted negatively.

Which observations could suffer from this problem? As once-treated units stay treated forever, the probability of being treated increases over time, so time fixed effects in the probability score regression should be increasing in t. Similarly, units that are treated earlier (with small E_i) are treated for a larger fraction of the periods. Therefore, fitted values will be particularly large for observations corresponding to early treated units at the end of the sample—precisely those which identify long-run treatment effects. They will be negatively weighted, or at least underweighted, by the canonical regression.

This argument also implies that negative weighting will not happen in models without individual fixed effects, such as

$$Y_{it} = FE_t + \gamma D_{it} + \text{noise}.$$

Fitted values from the simple propensity score regression of D_{it} on all calendar time dummies are fractions of treated units in each year, which always lie between zero and one. This result is consistent with Angrist (1998) who show that when the regression is *saturated*, i.e. solely includes dummies for all levels of a single categorical variable, OLS weights treatment effects by the variance of treatment conditional on the controls (see also Angrist and Pischke, 2008, sec. 3.3.1). Such variance is of course non-negative.

Appendix B.2 provides another treatment of the problem. Instead of solving for ω_k, we look at the weighting scheme implied by a simplified specification that is nested in the canonical specification. More precisely, unit and time fixed effects are replaced by corresponding linear terms (we used this trick in discussing the underidentification problem as well). We show in Proposition 2 that values of k that are larger than average (among treated observations) are always underweighted relative to their sample weights.

4.1.3 Example of Negative Weighting

Now we consider a very simple grid that may approximate the structures of the data in some applications.

Proposition 1. *Suppose there are $T \geq 2$ time periods, E_i is distributed uniformly among them, and for each unit i, the outcome is observed for all periods. Then,*

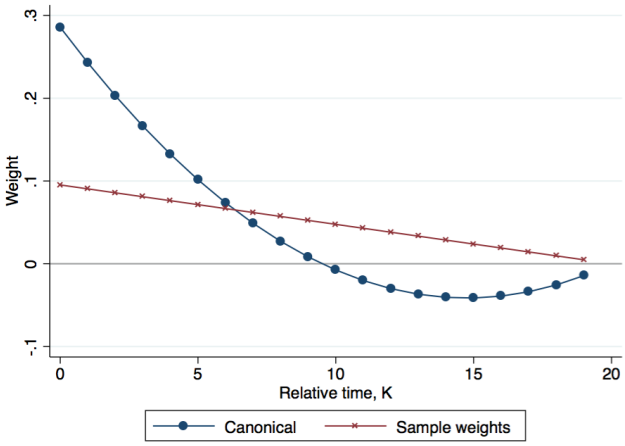
$$\omega_k = \frac{(T-k)(T-2k-1)}{T(T^2-1)/6}, \quad k = 0, \ldots, T-1.$$

Strikingly, $\omega_k < 0$ for $k > (T-1)/2$.[22]

Figure 2 illustrates this proposition by plotting the weights ω_k for $T = 20$. For comparison, it also shows the fraction s_k of observations with each k in the post-treatment sample, which could be a possible definition of a reasonable weighting scheme. It is clear that short-run effects are severely overweighted, whereas long-run effects enter negatively. That means that a program that produces uniformly larger effects may look worse in the canonical regression.

[22] All proofs are given in Appendix B.1.

Figure 2: Weights in Canonical Regressions



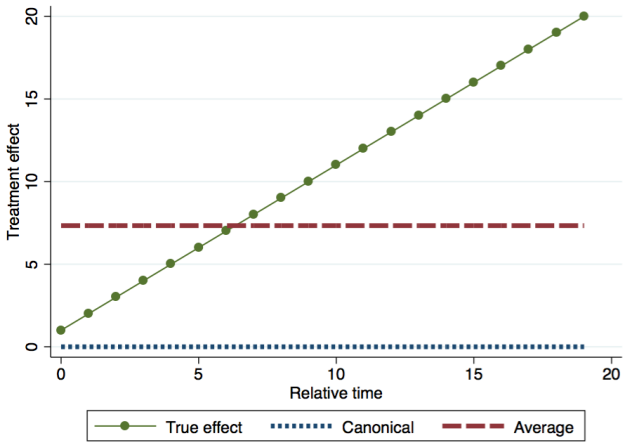
When treatment effects have strong dynamics, there will be a wide discrepancy between the sample size-weighted average treatment effects and the canonical regression estimand. Figure 3 shows two examples of this. Panel A corresponds to the case treatment permanently changes the *slope* (growth rate) rather than the level of the outcome, i.e. $\tilde{\gamma}_k = k+1$ for $k \geq 0$. Canonical regression completely misses the effects, estimating γ to be zero! The following corollary formalizes the result:

Corollary 1. *Suppose the assumptions of Proposition 1 hold. Then, when treatment changes the slope of the outcome's growth, i.e.* $\tilde{\gamma}_k = \varphi(k+1)$ *for* $k \geq 0$ *and some constant* φ*, then the canonical regression OLS estimand* $\gamma = 0$*, regardless of* φ *and the panel length* T*.*

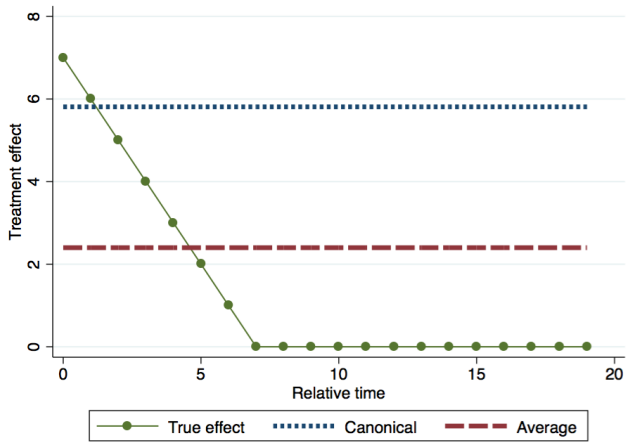
The canonical regression coefficient in this case lies outside of the convex hull of the effects at each time horizon. We show later that this is not just a theoretical possibility but happens in applications. Although negative weights in Figure 2 are not very large, they are multiplied by the large treatment effects at longer horizons in this example.

Figure 3: Biases of Canonical Regressions

A: Treatment Affects the Slope



B: Temporary Effects



Panel B of Figure 3 considers a situation when treatment effects are temporary, and the outcome gradually (in seven periods) reverts back to the original trajectory. Canonical regression will produce a large coefficient close to the largest, very short-run effect, and does not characterize the average well.

The weighting problem extends to more flexible specifications with two-way fixed effects. Consider for example, a "capped" specification

$$Y_{it} = \alpha_i + \beta_t + \sum_{k=0}^{B-1} \gamma_k \mathbf{1}\{K_{it} = k\} + \gamma_{B+} \mathbf{1}\{K_{it} \geq B\} + \varepsilon_{it}.$$

If treatment effects after $K_{it} = B$ are not constant, as assumed by this specification, this will make the estimate of γ_{B+} unreliable. But also, through the wrong choice of individual and time fixed effects, short-run effects will be biased. Figure 8 provides an illustration for $B = 6$ when the true effects are growing linearly, as in Figure 3A. The long-run effect is outside of the convex hull of the true effects for $k \geq 6$, and short-run effects are downward biased.

4.2 Solutions

Unlike with the underidentification issue, the solution for the weighting problem is quite straightforward. Researchers should never run regressions which impose any restrictions on the dynamics of treatment effects post-treatment. They should estimate them flexibly and average the coefficients manually with some weights (e.g. proportionately to the sample size). When pre-trends can be assumed away, that amounts to fitting the semi-dynamic regression (2). If some anticipation effects are possible, a more general specification (3) should be employed with the number of leads A reflecting the horizon of anticipation. Here $A < \infty$ is required for identification, and $B = \infty$ is necessary to avoid weighting problems.

Another solution is to find a valid control group that never experiences treatment, yet faces the same time effects $\tilde{\beta}_t$. The control group helps identify the time effects, alleviating the problem. Importantly, the control group should not be allowed to be on a separate time trend (see Section 5.1.1).

Note, however, that if the control group is small *relative to the treatment group*, it would not help. For the problem to disappear, time effects should be identified *solely* from the control group. Figure 9 shows, in the context of Proposition 1, how weights ω_k vary with the fraction of units (and, equivalently, observations) in the control group. Having a 10% control group does almost nothing, and even with equally sized groups, the weighting scheme is still quite far from s_k. Running a flexible specification and averaging the effects manually seems worthwhile even in presence of a control group.[23]

Although this approach solves the problem when the only type of treatment effect heterogeneity is across the time horizon, averaging of the effects across individuals or calendar times, that is implied by the semi-dynamic specification, may not be perfect. We return to this problem in Extensions, proposing a matching procedure that works more generally.

5 Extensions

It should by now be clear to the reader that both problems—underidentification and negative weighting—stem from the ability of two-way fixed effects to recover the relative time variable K_{it} in our staggered design without the control group. Here we demonstrate that equivalent problems may arise in a variety of different empirical designs.

We also extend our setting to the case with more general treatment effect heterogeneity and show that OLS for the semi-dynamic specification has undesirable extrapolation properties. We propose a matching scheme robust to this type of heterogeneity.

[23] A lazy man's alternative is to "boost" the control group: use weighted OLS where all control group observations are weighted by a very large number, e.g. 1,000 times more than the treatment group. This may increase standard errors but from the identification point of view, it is equivalent to having a very large control group. This weighted OLS can be viewed as a two-stage procedure: time effects are first identified from a regression $Y_{it} = \alpha_i + \beta_t +$ noise on the control group only, and then used in the canonical regression for the treatment group.

5.1 Related Problems in Other Empirical Designs

5.1.1 Difference-in-differences with Group-specific Time Trend

Consider first the standard difference-in-difference setup where all units in the treatment group ($G_i = 1$) are treated in same period E, whereas control group units ($G_i = 0$) are never treated. The treatment indicator is $D_{it} = Post_t \cdot G_i$, where $Post_t = \mathbf{1}\{t \geq E\}$. While the plain vanilla event study specification

$$Y_{it} = \alpha_i + \beta_t + \sum_{k=-\infty}^{\infty} \gamma_k \mathbf{1}\{t - E = k\} \cdot G_i + \text{noise}$$

does not suffer from any identification problems, sometimes researchers are not confident in the quality of their control group and would like to include a group-specific trend,

$$Y_{it} = \alpha_i + \beta_t + \mu t \cdot G_i + \sum_{k=-\infty}^{\infty} \gamma_k \mathbf{1}\{t - E = k\} \cdot G_i + \text{noise}.$$

Because E is the same for all units, the group-specific time trend is collinear with $(t - E)\, G_i$ in presence of individual fixed effects.

As a consequence, restricted specifications, such as

$$Y_{it} = \alpha_i + \beta_t + \mu t \cdot G_i + \gamma D_{it} + \text{noise}, \tag{7}$$

suffer from negative weighting of dynamic effects, *regardless* of the relative size of the control group. Figure 10 illustrates this point for the case when there are $t = 1, \ldots, 20$ periods, and treatment happens in the middle, $E = 11$.

One can replace individual fixed effects with group-specific intercepts in (7),

$$Y_{it} = \beta_t + (\alpha + \mu t) \cdot G_i + \gamma D_{it} + \text{noise},$$

or include unit-specific trends instead,

$$Y_{it} = \alpha_i + \beta_t + \mu_i \cdot t + \gamma D_{it} + \text{noise}.$$

All of these specifications are affected by the same problem.[24]

As we mentioned in the previous sections, the issues equally arise in *staggered* designs with a control group if either group- or unit-specific time trend is included. Indeed, the relative time in the treatment group, $(t - E_i)\, G_i$, can be recovered from $\alpha_i + \mu t \cdot G_i$.

5.1.2 Unit-Specific Trends without Control Group

Return now to the staggered design without a control group. When the empirical strategy is not perfectly convincing, it is sometimes recommended to check robustness to including unit-specific (e.g., state-specific) time trends:

$$Y_{it} = \alpha_i + \beta_t + \mu_i \cdot t + \sum_{k=-\infty}^{\infty} \gamma_k \mathbf{1}\{K_{it} = k\} + \text{noise}.$$

[24] In all of these cases, semi-dynamic specifications, e.g. $Y_{it} = \alpha_i + \beta_t + \mu t \cdot G_i + \sum_{k=0}^{\infty} \gamma_k \mathbf{1}\{t - E = k\} \cdot G_i + \text{noise}$, are fine. In this regression μ is estimated only using non-treated observations, while in restricted specifications treatment effects influence the estimate of μ, biasing γ.

Without the control group, this creates an additional problem. Not only a linear term K_{it} can now be recovered by fixed effects, but also a quadratic term $(K_{it})^2$. Indeed,

$$(K_{it})^2 = (E_i)^2 + t^2 - 2E_i t.$$

These three components are nested by α_i, β_t, and $\mu_i \cdot t$, respectively. As a result, the fully dynamic path of treatment effects is identified only up to a quadratic polynomial. Three restrictions instead of two must be imposed to regain identification, and the F-test described above is only powerful against pre-trends which have more complicated shape.

Correspondingly, weighting in the canonical-type regression

$$Y_{it} = \alpha_i + \beta_t + \mu_i \cdot t + \gamma D_{it} + \text{noise}$$

becomes even worse than before.[25]

Given our results in the previous subsection and here, we do not recommend including unit-specific time trends in any difference-in-difference or event study specifications (except for the case discussed in footnote 24).

5.2 Treatment Effect Heterogeneity

Throughout the paper, we imposed a strong assumption that the fully dynamic specification characterizes the true data generating process. Treatment effects are required to depend only on the time relative to treatment, K_{it}, but otherwise are homogenous across units and calendar time periods. Formally, we define treatment effects in terms of potential outcomes as $\tau_{itk} = Y_{it}^{(k)} - Y_{it}^{(-1)}$, and the fully dynamic specification requires $\tau_{itk} \equiv \gamma_k$. Such homogeneity, particularly across time periods, is very restrictive and difficult to reconcile with economic models—we provide an example related to the optimal choice of consumption and saving under the permanent income hypothesis in Appendix E.

What is the estimand of the semi-dynamic regression when heterogeneity is allowed for? To understand this, we again invoke the result on OLS as a weighting estimator, which weights each observation by the residual from the linear propensity score regression (see Appendix C). The propensity score regression behind γ_k, $k \geq 0$, in the semi-dynamic specification is the following one:

$$\mathbf{1}\{K_{it} = k\} = FE_i + FE_t + \sum_{\substack{l=0 \\ l \neq k}}^{\infty} \rho_{kl} \mathbf{1}\{K_{it} = l\} + \text{noise}. \tag{8}$$

An observation is weighted negatively in two cases: if $K_{it} = k$ and the fitted value in (8) is above one, or if $K_{it} \neq 0$ and the fitted value is any positive number. While we did not observe the former situation in simulations, the latter was quite prevalent among treated observations. The semi-dynamic estimand γ_k is a positively-weighted average of observations with $K_{it} = k$, minus a weighted average of control observations,, plus additional terms for treated observations with $K_{it} = l \neq k$ which have both positive and negative weights summing up to zero.

The intuition for having these additional terms is that the semi-dynamic specification imposes $\tau_{itk} \equiv \gamma_k$, allowing for substantial extrapolation. For instance, γ_1 can be estimated by comparing units treated at $E_i = 1$ and 2, both observed at periods $t = 1$ and 3. Alternatively, it can be estimated from units treated at $E_i = 5$ and 6 observed at $t = 5$ and 7. The difference between the resulting estimates is consistent for *zero* when the semi-dynamic specification is true, so an estimator for γ_0 can add or subtract this difference multiplied by any constant. Such extrapolation can improve efficiency, but also makes OLS non-robust to heterogeneity.

[25] Illustrative simulations are available from the authors upon request.

Severe extrapolation is inevitable in some cases. Suppose for example that treatment happens to all units between periods t_F and t_F+3, but we want to estimate γ_7—the effect of being treated seven periods ago compared to not having been treated yet. There is no direct difference-in-differences quasi-experiment that would help for this task, because in calendar periods where some unit has been treated for 7 periods, all units have already been treated.

However, for k smaller than the range of treatment periods, $\tilde{\gamma}_k$ can be estimated without extrapolation. As long as individual heterogeneity is captured by unit fixed effects, as we have always assumed, one can find periods $t' < t$, as well as units i and j treated at $E_i < E_j$ and observed in these periods, which satisfy $t' < E_i < t < E_j$ and $t = E_i + k$. That is, unit i has been treated at t for k periods, but not treated yet at t', whereas unit j has not been treated in either period. When there are no pre-trends, j provides a valid counterfactual to i.

The simplest way to estimate treatment effects without extrapolation is to run the semi-dynamic regression

$$Y_{it} = \alpha_i + \beta_t + \gamma_k D_{it} + \text{noise} \tag{9}$$

for each $k \geq 0$ separately on the subsample of the data which includes observations with $K_{it} = k$, as well as the entire control group $K_{it} < 0$. This regression cannot suffer from extrapolation of treated observations because the only type of treatment effects the is present in the estimation sample here is γ_k. We will use this strategy in the application in Section 6.2.

Another approach is matching. For each treated observation (i,t) there are multiple t' and potentially many units j that i can be matched with, subject to the rules discussed above. For efficiency reasons, all of them should be used with some weights $w_{it,jt'}$, and the resulting matching estimator can be written as

$$\hat{\gamma}_k = \frac{\sum_{i,t} \left(Y_{it} - Y_{it}^{CF}\right) \mathbf{1}\left\{K_{it} = k\right\}}{\sum_{i,t} \mathbf{1}\left\{K_{it} = k\right\}}$$

where $Y_{it}^{CF} = \sum_{j,t'} w_{it,jt'} \cdot (Y_{it'} - Y_{jt'} + Y_{jt})$ is the counterfactual outcome and $\sum_{j,t'} w_{it,jt'} = 1$.

Importantly, $\hat{\gamma}_k$ constructed in this way weights treated observations positively, in fact equally, to obtain the average treatment effect on the treated. This matching estimator can be rewritten as a weighting estimator, which extends Hirano et al. (2003) (on regressions with controls) and Abadie (2005) (on difference-in-differences with a single pre-period) to our event study design.

The only paper known to us which uses a similar estimator is Fadlon and Nielsen (2015), except that they require $E_j = E_i + 5$ and $t' = E_i - 2$. That reduces efficiency without relaxing any assumptions.[26]

6 Empirical Relevance

6.1 A Common Issue

A variety of papers suffer from the identification problems described above. Indeed, these papers use specifications analogous to those described in Sections 3, 4 and 5 as their main specifications. Prominent examples include Di Maggio et al. (2014), Hoynes and Schanzenbach (2012), Hoynes et al. (2016) and Duggan et al. (2016).

Di Maggio et al. (2014) investigate how indebted households' consumption and saving decisions are affected by anticipated changes in monthly interest payments. They study borrowers with adjustable rate mortgages featuring an automatic reset of the interest rate after five years, which is used for identification because the reset occurs at

[26] Fadlon and Nielsen (2015) test the identification assumptions by comparing 5-year pre-trends of treated observations and their counterfactuals. In our notation this is a test for $\tilde{\gamma}_{-1} - \tilde{\gamma}_{-6} = \cdots = \tilde{\gamma}_{-5} - \tilde{\gamma}_{-10}$, a hypothesis weaker than the linearity of pre-trends discussed in Section 3.2.2.

different points in time across households. They run the following specification:

$$Y_{i,t,d,\tau} = \sum_{\theta=-4}^{4} \beta_\tau \mathbf{1}\{\tau = \theta\} + \beta_5 \mathbf{1}\{\tau \geq 5\} + \lambda_i + \eta_{g,t} + \Gamma X_{i,t} + \epsilon_{i,t,\tau}$$

where i denotes the household, g the county, t the month or quarter and τ the quarter since the interest rate adjustment. Therefore, Di Maggio et al. (2014) regain identification by imposing that all causal effects β_θ should be equal for $\tau \geq 5$, and that all causal effects should be equal to 0 for $\tau \leq -4$. These linear restrictions do solve the underidentification problem, but their validity is not tested and our proposal is to instead run a semi-dynamic specification after testing that the linear restriction on the pretrend is consistent with the data.

Hoynes and Schanzenbach (2012) study the work incentive effects of the Food Stamp Program, using staggered introduction of the program across counties in the 1960s and 1970s to estimate the impact of the program on the extensive and intensive margins of labor supply, earnings, and family cash income. They estimate the following model:

$$y_{ict} = \alpha + \delta FSP_{ct} + X_{it}\beta + \sigma CB60_c \cdot t + \gamma REIS_{ct} + \eta_c + \lambda_t + \mu_{st} + \epsilon_{ict}$$

where i indexes family, c the county, t the year, and FSP_{ct} is an indicator variable equal to one if there is a Food Stamp Program in county c at time t, X_{it} are time-varying family characteristics, $CB60_c$ are 1960 county characteristics (interacted with linear time), $REIS_{ct}$ are county-level per-capita income transfer, η_c and λ_t are county- and year-fixed effects and μ_{st} are state-specific linear time trends. This exactly corresponds to the canonical regression we described in Section 2 and it therefore suffers from the issues discussed in Section 4. Using the same research design, Hoynes et al. (2016) study the effect of the Food Stamp Program on adult health and economic outcomes for individuals who where exposed to the introduction of the program in utero or during childhood. They run almost exactly the same specification as described above, which suffers from the issues that are inherent to the canonical regression.

Duggan et al. (2016) exploit variation in the timing of patent decisions to estimate the causal effect of a patent on the price of a molecule as well as on quantities sold and on the number of pharmaceutical firms operating in the market. In various specifications, the authors use a control group of molecules but allow this control group to be on a different time trend compared to the control group, which we showed in Section 5 creates identification and weighting problems that are identical to those in the setting without a control group. For instance, they run:

$$\begin{aligned} Y_{mt} \quad = \quad & \alpha_m + \delta_t + \lambda \cdot t \cdot I\{EverPatent\}_m + \delta_t \cdot I\{Post95Launch\}_m \\ & + \eta_1 \cdot I\{HasPatent\}_m + \eta_2 \cdot I\{HasPatent\}_{mt} \cdot I\{Post95Launch\}_m + \epsilon_{mt} \end{aligned}$$

with obvious notation.

More broadly, our points are relevant for a large set of papers that use some of the specifications we caution against in some of their sections, either as main specifications or as robustness checks. Examples include Grogger (1995), Glaeser and Mare (2001), Bertrand and Mullainathan (2003), Reber (2005), Papaioannou and Siourounis (2008), and Dobkin et al. (2015). Our points are also relevant for the literature using the AKM model, because this model requires to separately identify age effects, cohort effects and time effects (see for instance Card et al. 2016).

In the rest of this section, we carry out analyses in the spirit of Parker et al. (2013), extending their data and estimating the impulse response function of consumption expenditures to tax rebate receipt. The results illustrate the quantitative relevance of our points about underidentification in the fully dynamic specification and negative weighting in estimation of the average treatment effect in the canonical regression. In ongoing work, we examine the importance of our points for the literatures on the impact of unilateral divorce laws and compulsory school

reforms.

6.2 Estimating the Impulse Response Function of Nondurable Consumption Spending to Tax Rebate Receipt

The marginal propensity to consume out of tax rebates is a crucial parameter for economic policy. The Economic Stimulus Act of 2008 consisted primarily of a 100 billion dollar program that sent tax rebates to approximately 130 million US tax filers. Whether this stimulus was effective at boosting the economy and counteracting the Great Recession depends on the extent to which these tax cuts directly changed household spending, as well as on any subsequent multiplier or price effects affecting aggregate demand.

Using event study designs, a recent literature provides estimates of the marginal propensity to consume out of tax rebates. In particular, Parker et al. (2013) and Broda and Parker (2014) estimate the change in household spending on nondurables using a natural experiment provided by the structure of the 2008 tax cut. The tax cuts varied across households in amount, method of disbursement, and timing. Typically, single individuals received between $300 and $600, while couples received between $600 and $1,200; moreover, households received $300 per child who qualified for the child tax credit. Within each disbursement method, the timing of tax rebate receipt was determined by the final two digits of the recipient's Social Security number (SSN), and these digits are effectively randomly assigned across households. The causal effect of the receipt of the payments on household spending is estimated by comparing the spending of households that received payments in a given period to the spending of households that received payments in other periods. This kind of random variation requires precisely the type of estimators studied in this paper. Note that such variation can only be used to estimate the change in household spending on nondurable directly caused by the receipt of the tax rebates and inherently ignores general equilibrium effects (e.g. Keynesian multipliers and price effects).

We estimate the performance of various estimators at estimating the impulse response function of nondurable consumption to tax rebate receipt using the same data as Broda and Parker (2014). Both Parker et al. (2013) and Broda and Parker (2014) examine the response of nondurable consumption to tax rebate receipt. While Parker et al. (2013) estimate the causal effect of tax rebate receipt on nondurable consumption using quarterly consumption data from the Consumer Expenditure Survey, Broda and Parker (2014) use more detailed data from the Nielsen Homescan Consumer Panel. The Nielsen dataset has the advantage of tracking transactions at a much higher (daily) frequency, which is why we choose it as our main dataset. The Nielsen data covers approximately 15% of overall households expenditures and includes the following product categories: food, alcohol, beauty and health products, household supplies, and general merchandise. The analysis using the same CEX data as in Parker et al. (2013) is available from the authors upon request.

This setting is ideal for our purposes for two reasons: first, the source of variation — the final two digits of the recipient's Social Security Number — is well understood and is truly as good as random; second, precise estimation of the average marginal propensity to consume out of tax rebates as well as its dynamics is crucial for fiscal policy.

In the remainder of this section, we use the variation in the week in which households received tax rebates to estimate the impulse response function of consumption to tax rebate receipt. We run specifications analogous to those described in Section 2, with weekly nondurable consumption as the outcome and households as the unit of treatment. We use a balanced sample of households and observe 16 weeks prior to the tax rebate and 22 weeks after. We run the fully dynamic specification and semi-dynamic specifications using this data, with household fixed effects.[27]

The analysis delivers four conclusions that show the relevance of the points developed in the previous sections:

[27] We have also checked that we are able to replicate the results of Broda and Parker (2014), using their specifications. The results are available upon request.

1. underidentification can lead to seemingly unstable and confusing results when it is not properly handled; 2. our proposed solutions to address underidentification work well (namely: using a semi-dynamic specification after running a F-test for non-linear pretrends, and/or using a semi-parametric specification); 3. canonical regressions yield misleading results and are another source of seeming instability and inconsistency of results; 4. semi-parametric specifications that leave no room for extrapolation show that long-run effects are very difficult to identify in this type of natural experiment.

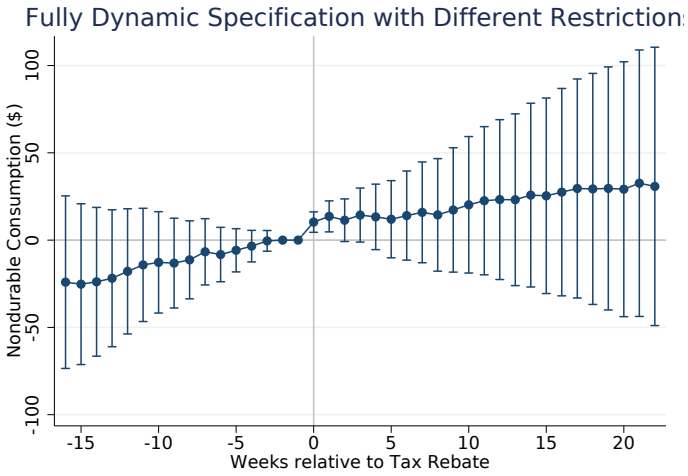
Overall, our main message for practitioners is twofold. First, we show that because of identification and negative weighting issues in event study designs, researchers are likely to produce results that appear to be unstable and/or inconsistent with each other. These issues can be avoided in a principled way by using the estimators and tests introduced in this paper. Second, our application illustrates that long-run effects are likely to be difficult to identify in event study research designs. Researchers should be mindful of estimators introducing extrapolation, which are likely to be unstable. We recommend that they emphasize the types of effects that can be identified in the semi-parametric estimators we introduce, which are likely to be the short-run effects, as in the application below.

The first finding from the analysis of the consumption response to tax rebate receipt, using the data from Broda and Parker (2014), is that the underidentification problem can lead to completely different estimated paths for pre-trends and treatment effects, in line with the discussion in Section 3.1. Panel A of Figure 4 shows the dynamic path of pre-trends and treatment effects, plotting the regression coefficients on all leads and lags in specification (1), where the parameters are estimated after dropping leads 1 and 2. The estimated coefficients exhibit a strong upward trend. Intuitively, dropping leads that are close to each other is likely to generate strong trends because of noise in finite sample.

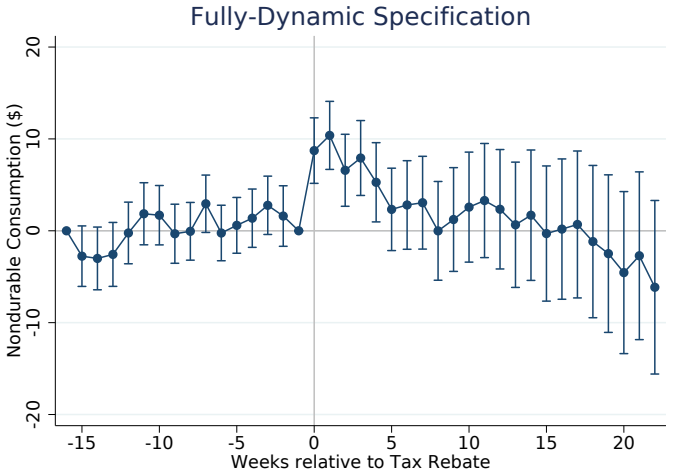
In contrast, Panel B of Figure 4 shows that when dropping the very first and very last leads, the pattern of consumption during the event study window is much more stable. There is no discernible pre-trend, and weekly consumption jumps exactly at the time of tax rebate receipt for about five weeks, before reverting to its usual level. Note that the set of coefficients shown in Panels A and B of Figure 4 are identical up to a rotation (or linear trend). The additional noise introduced by the omission of leads 1 and 2 in Panel A is reflected in larger standard errors. Researchers that are not attentive to under-identification run the risk of obtaining seemingly unstable results across specifications, depending on which set of leads (or lags) of the treatment indicator they (or their statistical software) drops. In a balanced sample, dropping the very first and very last leads of the treatment indicator is a sensible approach because it will reduce noise. Next, we discuss a principled approach to addressing underidentification: first testing for statistical significance of non-linear pretrends with a F-test, then running the semi-dynamic specification.

Figure 4: Under-identification in Fully Dynamic Specifications

Panel A: Dropping Leads 1 and 2



Panel B: Dropping Leads 1 and 16



The second finding is that our proposed solution to handle underidentification problem works well in this application. We proceed in two steps. To begin with, we show that we cannot reject that there is truly no (nonlinear) pre-trend before the tax receipt: the p-value for the F-test of the hypothesis that all lags of the treatment indicators are equal to zero is 0.13. We also note that Panel B of Figure 4 provides a graphical test for (nonlinear) pretrends: all leads are statistically indistinguishable from zero, are very stable and relatively precisely estimated zeroes.[28] This makes us confident that the variation used to estimate the causal effect of tax rebate receipt on consumption is as good as random and provides a justification for running the semi-dynamic specifications, which sets all leads equal to zero.

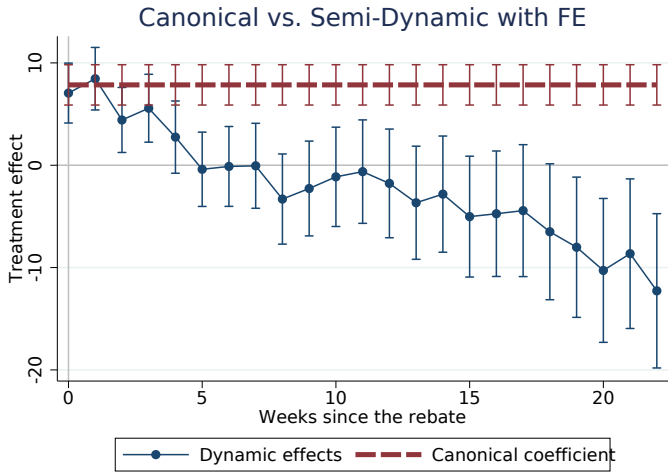
Our preferred estimates of the path of dynamic causal effects are obtained after imposing the restriction that all leads of the treatment indicator are equal to zero. The results are shown in Panel A of Figure 5 and show that the receipt of a tax rebate leads to an increase in nondurable consumption immediately after the tax rebate receipt.

[28] The value-added of this graphical test relative to the F-test described is that we can show that the results are not driven by noise: we are dealing with precisely estimated zeroes.

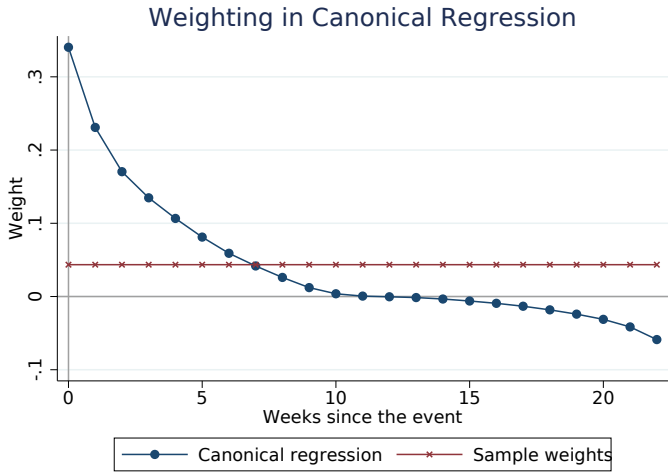
The standard errors are smaller in Panel A of Figure 5 compared with Panel B of Figure 4, because the formed imposes the restriction that all lags of the treatment indicator, which increases efficiency when this restriction is true. Note that the point estimates for the lags of the treatment indicators appear to exhibit a downward trend in both Panel A of Figure 5 and Panel B of Figure 4, which seems surprising since one would expect the treatment to go to zero a few weeks or months after the receipt of the tax rebate (instead of becoming negative). We discuss below how this pattern results from extrapolation and can be addressed using the semi-parametric specifciations introduced in Section 5.2.

Figure 5: Semi-Dynamic and Canonical Specifications

Panel A: Results for Semi-Dynamics and Canonical Specifications



Panel B: Weights in Canonical Regression



Third, we find that summarizing the treatment effect with a canonical regression would be very misleading, because of the negative weighting problems described in Section 4. Panel B of Figure 5 shows the weights given to the various dynamic causal effects by the canonical regression, many of which are negative. Moreover, the first lags of the treatment indicator get very large weights in the canonical regression. The causal effect of tax rebate receipt is a short-run increase in consumption, with no effect on weekly spending after five to eight weeks, as shown in Panel B of Figure 4 and in Panel A of Figure 5.

Given that the treatment effect is declining over time, the weights used in the canonical regression yield an upward biased estimate of the "average" (with sample weights) treatment effect. The point estimate for the average treatment effect in the canonical regression is \$7.85 (with a standard error of 1.004). Interpreting this coefficient as the average treatment effect over 20 weeks would suggest that, during this period, households spent about 25% of their rebate amount on nondurable consumption alone (as covered in the Nielsen data, which account for about 15% of overall consumption). Assuming that the consumption response is similar in other consumption categories, this estimate would imply a marginal propensity to consume out of tax rebates of over 100% over 20 weeks, which is extremely large. In fact, as shown in Panel A of Figure 5, the point estimate from the canonical regression is almost outside of the convex hull of the dynamic treatment effects in the semi-dynamic specification. These patterns are in line with the theory presented in Section **4.1** and should not cause worry about the validity of the research design—but they show that the results from the canonical regression must be interpreted with caution.

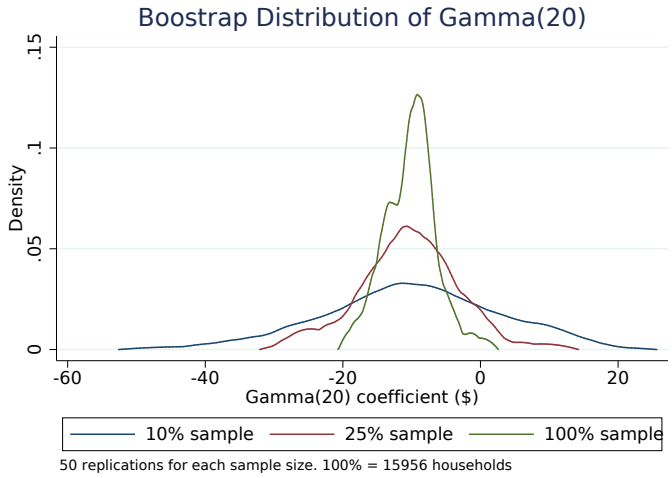
Fourth, we find that long-run effects are very difficult to identify in this setting. Panel B of Figure 4 and Panel A of Figure 5 exhibit downward trend in the point estimates on the lags of the treatment indicator. Such trend can be due to extrapolation in parametric specifications and is more likely to occur in smaller samples. As first evidence for this, Panel A of Figure 6 provides bootstrapped distributions of the estimate of $\hat{\gamma}_{20}$ by the semi-dynamic specification, using subsamples of various sizes. The distribution is extremely noisy for the samples that are 10% of the actual data, but also with the actual sample size. This large volatility is due to the spurious trends in coefficients. To see that directly, we estimate parameter μ from the regression

$$Y_{it} = \alpha_i + \beta_t + \sum_{k=0}^{10} \gamma_k \mathbf{1}\left\{K_{it} = k\right\} + \mu\left(K_{it} - 10\right) \cdot \mathbf{1}\left\{K_{it} > 10\right\}.$$

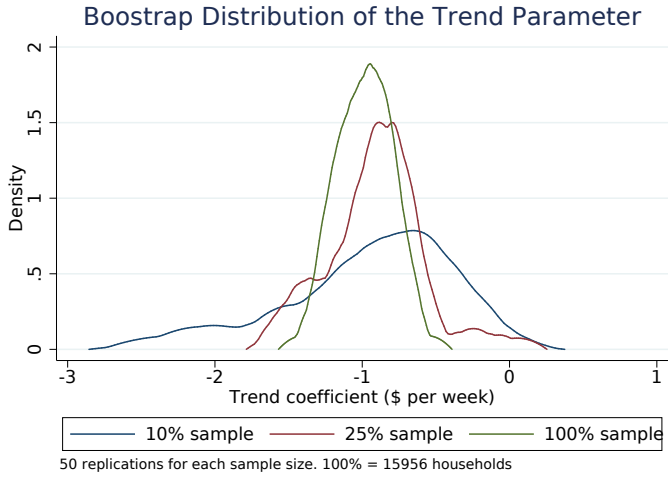
This parameter measures the slope of the impulse response function after 10 weeks since treatment. Economic intuition suggests that behavior should stabilize there, so $\mu \approx 0$. Panel B of Figure 6 shows bootstrapped distributions of $\hat{\mu}$, and a typical value is around -\$1; that is, each week after the tenth one the causal effect is estimated to fall by \$1. This is implausible given that the effect is close to zero after several weeks, and it cannot be substantially negative, based on the economic intuition. With the 10% sample size, there is also a large tail where $\hat{\mu}$ is positive, so that the marginal propensity to consumer steadily grows after ten weeks since the rebate.

Figure 6: Bootstrap Results

Panel A: Bootstrapped Distribution of $\hat{\gamma}_{20}$



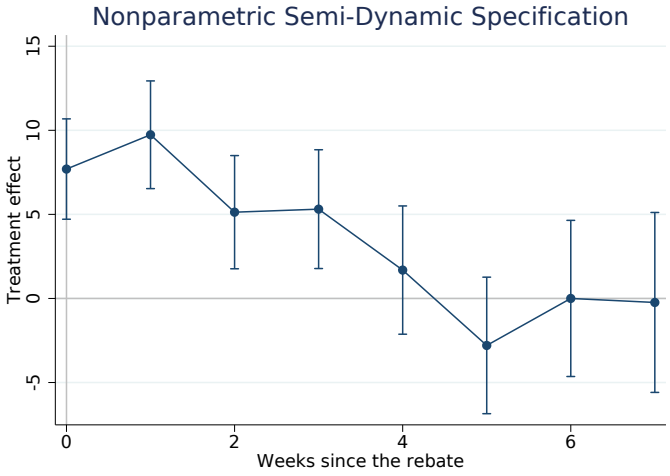
Panel B: Bootstrapped Distribution of the Trend Parameter



A more direct test for our claim that extrapolation is behind the spurious trend is to use the semi-parametric estimator (9) introduced in Section 5.2. Recall that these estimators do not leave room for extrapolation within the treatment group. Figure 7 presents the results. In contrast with the parametric specifications discussed above, there is no longer a downward trend in the point estimates of the lags of the treatment indicator. Instead, consumption increases for only three weeks after the receipt of the tax rebate, and then reverts to its usual value.

In the semi-parametric specification, it is not possible to study causal effects beyond seven weeks after the receipt of the tax rebate. This application illustrates the more general point that long-run effects are likely to be difficult to identify in event study research designs with a limited range of treatment dates and no control group, and that researchers should be mindful of estimators introducing extrapolation, which are likely to be unstable. We recommend to emphasize short-run effects that can be identified by the semi-parametric estimator.

Figure 7: Results from Semi-Parametric Specification



7 Conclusion

Difference-in-difference estimators are some of the most prevalent tools in economics for estimating causal effects of treatments. While the classical design compares two groups of units before and after a single date when one of them gets treated, researchers also frequently work with panels where different units are treated at different moments. For instance, in broadly used event studies all units in a panel eventually get treated, but some earlier than others. This paper showed that because of underidentification issues and negative weighting in event study designs, results from common specifications are likely to seem non-robust. These problems can be alleviated in a principled way by using parametric and semi-parametric estimators and tests.

Two approaches were introduced to address the underidentification issue. The first is to restrict the pre-trends in the fully dynamic specification, while keeping unit fixed effects. The second consists in replacing unit fixed effects with unit random effects. Statistical and graphical tests were developed to prove the validity of these approaches. The paper also showed that the canonical estimator does not provide a reasonable average of treatment effects along an important dimension---the time since first treatment---although other methods can. Specifically, the canonical estimator has a severe short-term bias and in many cases weights long-term effects negatively. To address the problem, alternative parametric and semi-parametric estimation techniques were introduced, which always average dynamic effects in a convex way. The parametric approach amounts to estimating the dynamic regression that includes dummies for all lags of treatment and then manually averaging its coefficients, whereas the semi-parametric one is a new weighting estimator in the spirit of Hirano et al. (2003).

The practical relevance of this problem was established in light of the existing literature and in a specific application about the estimation of the marginal propensity to consume out of tax rebates.

References

Abadie, Alberto, "Semiparametric Difference-in-Difference Estimators," *Review of Economic Studies*, 2005, *72*, 1–19.

_ , Alexis Diamond, and Jens Hainmueller, "Comparative Politics and the Synthetic Control Method," *American Journal of Political Science*, 2015, *59* (2), 495–510.

Allegretto, Sylvia, Arindrajit Dube, Michael Reich, and Ben Zipperer, "Credible Research Designs for Minimum Wage Studies," *IZA Discussion Papers*, 2013, *7638* (7638).

Angrist, Joshua, "Estimating the labor market impact of voluntary military service using social security data on military applicants," *Econometrica*, 1998, *66* (2), 249–288.

Angrist, Joshua D. and JS Pischke, *Mostly harmless econometrics: An empiricist's companion*, Princeton University Press, 2008.

Bertrand, Marianne and Sendhil Mullainathan, "Enjoying the Quiet Life? Corporate Governance and Managerial Preferences," *Journal of Political Economy*, 2003, *111* (5), 1043–1075.

Broda, Christian and Jonathan A. Parker, "The economic stimulus payments of 2008 and the aggregate demand for consumption," *Journal of Monetary Economics*, 2014, *68* (S), S20–S36.

Card, David, Ana Rute Cardoso, and Patrick Kline, "Bargaining, Sorting, and the Gender Wage Gap: Quantifying the Impact of Firms on the Relative Pay of Women," *The Quarterly Journal of Economics*, 2016, *131* (2), 633–686.

Di Maggio, Marco, Amir Kermani, and Rodney Ramcharan, "Monetary Policy Pass-Through: Household Consumption and Voluntary Deleveraging," *Working Paper*, 2014.

Dobkin, Carlos, Amy Finkelstein, Raymond Kluender, and Matthew J Notowidigdo, "The Economic Consequences of Hospital Admissions," 2015.

Duggan, Mark, Craig Garthwaite, and Aparajita Goyal, "The Market Impacts of Pharmaceutical Product Patents in Developing Countries: Evidence from India," *American Economic Review*, 2016, *106* (1), 99–135.

Fadlon, Itzik and Torben Heien Nielsen, "Household Responses to Severe Health Shocks and the Design of Social Insurance," 2015.

Glaeser, Edward L. and David C. Mare, "Cities and Skills," *Journal of Labor Economics*, 2001, *19* (2), 316–342.

Grogger, Jeffrey, "The Effect of Arrests on the Employment and Earnings of Young Men," *The Quarterly Journal of Economics*, 1995, *110* (1), 51–71.

Hirano, Keisuke, Guido W. Imbens, and Geert Ridder, "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score," *Econometrica*, 2003, *71* (4), 1161–1189.

Hoynes, Hilary W, Diane Whitmore Schanzenbach, and Douglas Almond, "Long Run Impacts of Childhood Access to the Safety Net," *American Economic Review*, 2016, *106* (4), 903–934.

Hoynes, Hilary Williamson and Diane Whitmore Schanzenbach, "Work incentives and the Food Stamp Program," *Journal of Public Economics*, 2012, *96* (1-2), 151–162.

Imbens, Guido W., "Matching methods in practice: Three examples," *Journal of Human Resources*, 2015, *50* (2), 373–419.

Meer, Jonathan and Jeremy West, "Effects of the Minimum Wage on Employment Dynamics," *Journal of Human Resources*, 2015.

Papaioannou, Elias and Gregorios Siourounis, "Democratisation and growth," *Economic Journal*, 2008, *118* (532), 1520–1551.

Parker, Jonathan A., Nicholas S. Souleles, David S. Johnson, and Robert McClelland, "Consumer Spending and the Economic Stimulus Payments of 2008," *American Economic Review*, 2013, *103* (6), 2530–2553.

Reber, Sarah J, "Court-Ordered Desegregation : Successes and Failures Integrating American Schools since Brown versus Board of Education," *The Journal of Human Resources*, 2005, *40* (3), 559–590.

Sloczynski, Tymon, "A General Weighted Average Representation of the Ordinary and Two-Stage Least Squares Estimands," 2016.

Wolfers, Justin, "Did Unilateral Divorce Laws Raise Divorce Rates? A Reconciliation and New Results," *American Economic Review*, 2006, pp. 1802–1820.

A Additional Figures and Tables

Figure 8: Biases in the Capped Regression

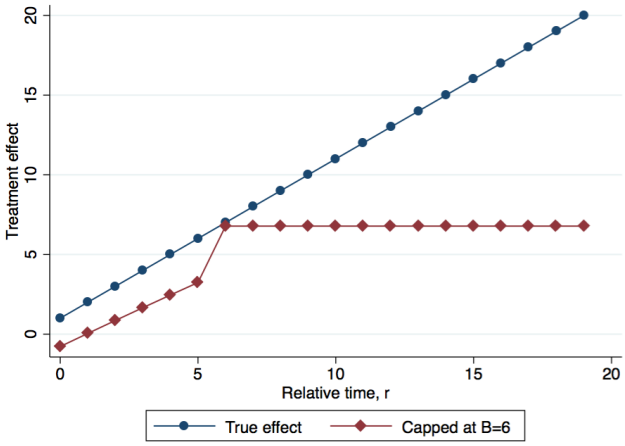


Figure 9: Canonical Weights with Control Group

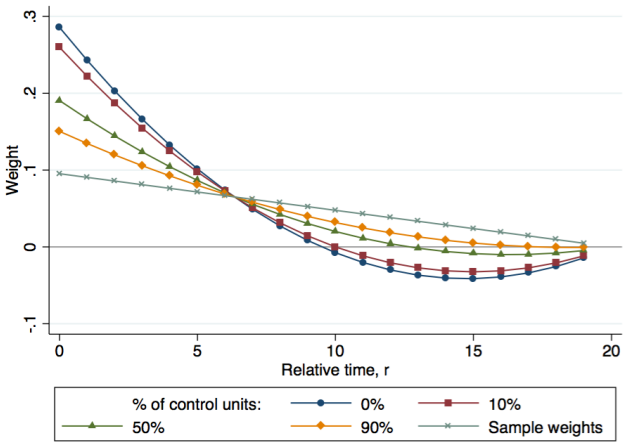
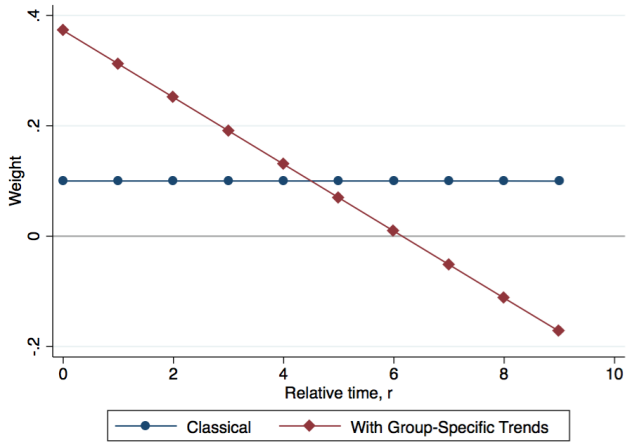


Figure 10: Weights in Diff-in-Diff Specifications



Notes: "Classical" refers to specification $Y_{it} = \alpha_i + \beta_t + \gamma D_{it}$. The specification with group-specific trends is given by (7).

B Additional Results and Proofs

B.1 Proofs

Proof of Proposition 1. In the case where E_i takes all values $1, \ldots, T$ with equal probabilities, the OLS estimand is the same as in the case where there is exactly one observation per each t and E_i, and there no noise, $\tilde{\varepsilon}_{it} = 0$. We consider such dataset with T^2 observations and index units such that $E_i = i$.

To compute ω_k, we apply the general result from Appendix C. OLS estimation of the canonical regression produces the estimate

$$\gamma = \sum_{i,t} \omega_{it} y_{it}$$

where ω_{it} is proportionate to the residual in the regression of D_{it} on unit and time fixed effects. The square structure of the data allows us to compute this residual in a very simple way:

$$\omega_{it} \propto D_{it} - \frac{1}{T}\sum_{i'} D_{i't} - \frac{1}{T}\sum_{t'} D_{it'} + \frac{1}{T^2}\sum_{i',t'} D_{i't'}.$$

For a treated observation with $t - i = k \geq 0$,

$$\begin{aligned}\omega_{it} &\propto 1 - \frac{t}{T} - \frac{T-(i-1)}{T} + \frac{T(T+1)/2}{T^2} \\ &= \frac{T-2k-1}{2T} \\ &\propto T - 2k - 1.\end{aligned}$$

Since ω_k is the coefficient in the regression of $\mathbf{1}\{K_{it} = k\}$ on D_{it} with two-way fixed effects,

$$\begin{aligned}\omega_k &= \sum_{i,t} \omega_{it} \mathbf{1}\{K_{it} = k\} \\ &\propto (T-k)(T-2k-1)\end{aligned}$$

because there are $T-k$ observations satisfying $i-t=k$. Given $\sum_{k=0}^{T-1}\omega_k=1$, this implies

$$\omega_k=\frac{(T-k)(T-2k-1)}{\sum_{k=0}^{T-1}(T-k)(T-2k-1)}.$$

It is a matter of simple algebra to simplify the denominator. Denote $a_k=(T-k)(T-2k-1)$. Then,

$$\begin{aligned}a_k+a_{T-1-k}&=(T-k)(T-2k-1)-(k-1)(T-2k-1)\\&=(T-2k-1)^2.\end{aligned}$$

Therefore,

$$\begin{aligned}\sum_{k=0}^{T-1}a_k&=\frac{1}{2}\left(\sum_{k=0}^{T-1}a_k+\sum_{k=0}^{T-1}a_{T-1-k}\right)\\&=\frac{1}{2}\sum_{k=0}^{T-1}(T-2k-1)^2.\end{aligned}$$

If T is odd,

$$\begin{aligned}\sum_{k=0}^{T-1}a_k&=\frac{1}{2}\cdot 2\cdot\sum_{k=0}^{(T-1)/2}(2k)^2\\&=4\cdot\frac{\frac{T-1}{2}\cdot\frac{T+1}{2}\cdot T}{6}\\&=\frac{(T-1)T(T+1)}{6}.\end{aligned}$$

If T is even,

$$\begin{aligned}\sum_{k=0}^{T-1}a_k&=\frac{1}{2}\cdot 2\cdot\left(\sum_{k=0}^{T}k^2-\sum_{k=0}^{T/2}(2k)^2\right)\\&=\frac{T(T+1)(2T+1)}{6}-4\cdot\frac{\frac{T}{2}\cdot\frac{T+2}{2}\cdot(T+1)}{6}\\&=\frac{(T-1)T(T+1)}{6},\end{aligned}$$

so the same expression holds all T.

Proof of Corollary 1. It is straightforward to verify that weights from Proposition 1 satisfy

$$(k+1)\omega_k=-(l+1)\omega_l$$

whenever $k+l=T-1$. Using this fact and Proposition 1, we can write

$$\gamma=\sum_{k=0}^{T-1}(k+1)\omega_k=\frac{1}{2}\left(\sum_{k=0}^{T-1}(k+1)\omega_k+\sum_{l=0}^{T-1}(l+1)\omega_l\right)=0.$$

B.2 Weights in a Simplified Specification

Recall that two-way fixed effects can reproduce a linear trend in the relative time K_{it}, which is at the core of the underidentification problem. Consider the following regression that is nested within the canonical one:[29]

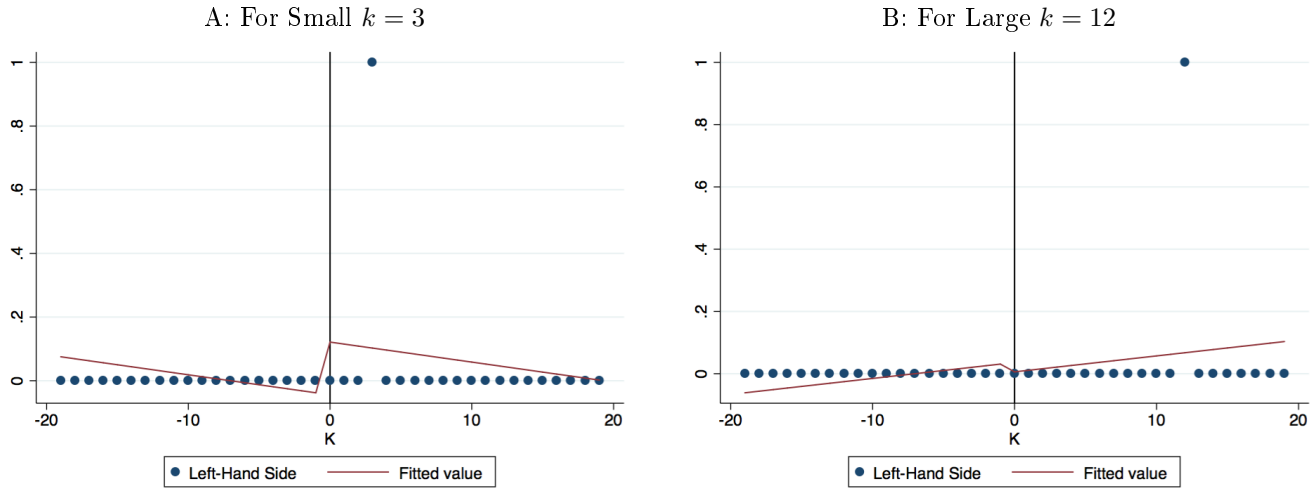
$$Y_{it} = \mu + hK_{it} + \gamma^R D_{it} + \varepsilon_{it}. \tag{10}$$

As in Lemma 1, $\gamma^R = \sum_{k\geq 0} \omega_k^R \tilde{\gamma}_k$ with weights that can be estimated from a series of regressions

$$\mathbf{1}\{K_{it} = k\} = \mu + h_k^R K_{it} + \omega_k^R \mathbf{1}\{K_{it} \geq 0\} + \text{noise}, \quad k \geq 0. \tag{11}$$

This regression (11) is defined solely in terms of a single variable K_{it}. Figure 11 illustrates it in the context of Proposition 1 for two values of k: small ($k = 3 \ll T = 20$) and large ($k = 12$). Specification (11) fits the left-hand side variable with two parallel lines of any location and slope, and ω_k^R is the shift between them. When k is large, this shift can be negative (see panel B). Moreover, short-run effects are always overweighted, as formalized by the following proposition:[30]

Figure 11: Estimation of Specification (11)



Proposition 2. *The weight schedule implied by regression (10) is always short-run biased. That is, $\omega_k^{Restricted}/s_k$ is decreasing in k. It is greater than 1 if and only if k is smaller than the average K_{it} post-treatment, i.e. $k < \sum_{l=0}^{\infty} l s_l$.*

Proof. Our approach is similar to the one from the proof of Proposition 1. Consider regression

$$\mathbf{1}\{K \geq 0\} = \mu_1 + h_1 K \tag{12}$$

with weights m_K, measuring sample shares of each K. Because $\mathbf{1}\{K > 0\}$ and K are co-monotone, the slope is positive, $\hat{h}_1 > 0$. Moreover, the intercept satisfies $\hat{\mu}_1 \leq 1$. Indeed, if $\hat{\mu}_1 > 1$, estimated residuals are non-positive for all $K \geq 0$. But in that case a marginal rotation of the regression line that makes it flatter while keeping the x-intercept constant will reduce squared errors for all K, and hence the total as well.

The general result of Appendix C implies that each treated observation in (10) is weighted in proportion to the

[29] To get it, plug in $\alpha_i = -hE_i$ and $\beta_t = \mu + ht$.

[30] It should be primarily viewed as providing intuition because $\omega_k \neq \omega_k^R$ in general.

residual from (12), i.e. $1-\hat{\mu}_1-\hat{h}_1 K$, and the total weight

$$\omega_k^R = \frac{m_k\left(1-\hat{\mu}_1-\hat{h}_1 k\right)}{\sum_{k\geq 0} m_k\left(1-\hat{\mu}_1-\hat{h}_1 k\right)}.$$

Since $s_k = m_k/m_+$ for $m_+ = \sum_{k\geq 0} m_k$, we can rewrite

$$\frac{\omega_k^R}{s_k} = \frac{1-\hat{\mu}_1-\hat{h}_1 k}{\sum_{l\geq 0} s_l\left(1-\hat{\mu}_1-\hat{h}_1 l\right)}.$$

The denominator, which is the average residual from (12) corresponding to $K \geq 0$, must be positive. It is then clear that ω_k^R/s_k is falling in k and is greater than one if and only if

$$1-\hat{\mu}_1-\hat{h}_1 k > \sum_{l\geq 0} s_l\left(1-\hat{\mu}_1-\hat{h}_1 l\right),$$

i.e. $k < \sum_{l=0}^{\infty} l s_l$.

C Negative Weighting

The derivation below illustrates why in regressions with non-saturated controls OLS may assign "negative weights" to treated observations, although one wishes to have positive weights for treated observations and negative weights for control observations. We start from a simple setup with treatment indicator D and control variables X.

$$Y = \beta D + \alpha X + \varepsilon$$

By the Frisch-Waugh-Lovell theorem,

$$Y - f'X = \beta\left(D - \pi' X\right) + \varepsilon^*$$

where $f'X$ and $\pi'X$ are linear projections of Y and D on X, respectively. The latter is the propensity score estimated using the linear probability model. Then,

$$\beta = \frac{Cov(Y - f'X, D - \pi'X)}{Var(D - \pi'X)}$$

We can write this as a sum of Y_is multiplied by weights, or using the population notation,

$$\beta = \mathbb{E}\left[(Y - f'X)(D - \pi'X)\right]\frac{1}{Var(D-\pi'X)} = \mathbb{E}\left[Y \cdot \omega(D,X)\right], \qquad \text{for } \omega(D,X) = \frac{D-\pi'X}{Var(D-\pi'X)},$$

where the second equality uses the fact that $f'X$ is a linear function of X, hence must be uncorrelated with $(D - \pi'X)$. It is clear that the OLS-implied weights are proportionate to the residual in the propensity score regression.

It is straightforward to show that $\mathbb{E}\left[\omega \cdot D\right] = 1$ and $\mathbb{E}\left[\omega \cdot (1-D)\right] = -1$. That is, β is an average of the outcomes in the treatment group *minus* an average of outcomes in the control group, with weights adding up to one in both cases. In the standard Rubin causal model, one can write $Y = Y_0 + \tau D$, where Y_0 is the no-treatment potential

outcome, and τ is the (heterogenous) treatment effect. Therefore, OLS estimates

$$\beta = \mathbb{E}\left[Y_0 \cdot \omega\right] + \mathbb{E}\left[\tau \cdot (\omega \cdot D)\right].$$

The first term (with the weights adding up to zero) represents selection bias, and the second one—the average of treatment effects. Even if selection is not a problem, the second term is problematic when $\omega < 0$, i.e. $\pi' X > D = 1$. This can never happen in a saturated regression, but is very likely for at least some observations in regressions with continuous controls or multiple sets of fixed effects.

As simple as the result is, we are unaware of any other paper showing it. Abadie et al. (2015) note for the setting with one treated and many untreated observations that the untreated ones can be weighted negatively. However, they do not connect this to the propensity score regression or, since they do not allow for multiple treated observations, to the averaging of treatment effects.

In the setting discussed in Section 4, X is a set of individual and time dummies. Then, people who are treated earlier in the sample have more observations with treated status $D = 1$. Also, there are also more treated observations in the later periods because treatment status doesn't revert to 0. Therefore, the largest treated values are for the long-run treatment effects (high K_{it}) and we can get negative weights. These weights just depend on the "grid"—the distribution of calendar time t and initial treatment periods E_i in the sample—because there is no other variable in the propensity score regression.

This result is particularly worrisome if the treatment effect is dynamic because large treatment effects in the long run get assigned a weight of the wrong sign. Because of this, the γ estimated in the canonical regression could be outside of the convex hull of the true treatment effects $\tilde{\gamma}_k$. For instance if the treatment effect is positive and growing over time, the estimated γ could be *negative* although we were hoping it would be a weighted average of $\tilde{\gamma}_k$ (we show in Section 6 that this in fact happens in several important empirical applications).

D Unbalanced vs. Balanced Panels and Individual Fixed Effects

As discussed in Section 3.2.3, one approach to regain identification is to do away with unit fixed effects and hope that this does not pose a threat to identification. In this section, we clarify the nature of the potential threats to identification when excluding unit fixed effects for both unbalanced and balanced panels. For this section, assume throughout that there are no year effects to simplify the analysis (also note that since we are considering specifications without individual fixed effects, the other concerns with the canonical regression discussed in 4 do not apply). We believe there is a conventional wisdom in applied work that omitting unit fixed effects when working with *unbalanced* panels is a big assumption because of selection into treatment, while omitting unit fixed effects in the case of *balanced* panels is much less problematic. We discuss below why the two settings in fact pose similar issues (as in the rest of this note, the discussion considers setting without a control group of units that never experience treatment).

Unbalanced panels. It is well understood that when panels are unbalanced, if there is a correlation between the time of treatment and the unit fixed effects (in our notation, $Cov(\alpha_i, E_i) \neq 0$), then including individual fixed effects is key. Such a correlation could result from intuitive patterns of endogenous selection into treatment. Consider for instance a setting where i) treatment has a positive and constant effect, ii) unit fixed effects α_i reflect the bargaining power of the unit, which allows for better outcomes Y_{it} in general and also for earlier selection into treatment (e.g. $Cov(\alpha_i, E_i) < 0$). In this setting, running a regression of the form $Y_{it} = \beta_t + \gamma T_{it} + u_{it}$ (without unit fixed effects) yields an upwardly biased estimate of the true constant treatment effect. Intuitively, T_{it} conveys information about the "type" of the unit: units with higher individual fixed effects are treated earlier in the sample,

i.e. for a longer period of time in the sample, and the estimated treatment effect coefficient partly captures these higher fixed effects. In simulated data, this can be checked by running a regression of the form $T_{it} = \lambda\alpha_i + u_{it}$,which yields $\lambda < 0$. Balanced panel may at first glance appear to be impervious to this issue.

Balanced panels. By construction, in a balanced sample each unit gets treated for the same number of periods of the observed sample. Therefore, in simulated data with the same data generating process as discussed for unbalanced panels, running a regression of the form $T_{it} = \lambda\alpha_i + u_{it}$ yields $\lambda \approx 0$. Is it sufficient to restrict the sample to a balanced panel to address the concerns resulting from endogenous selection into treatment discussed in the case of unbalanced panels? And if not, why not and what is the link with unbalanced panels? We have verified in simulations that balancing the sample does not solve the problem, and here we provide intuition for why selection into treatment correlated with individual fixed effects in fact poses exactly the same problem in balanced and unbalanced panels. The intuition can be best seen based on the following figure:

Figure 12: Share of Treated Units and Unit Fixed Effects over Time in Balanced Sample

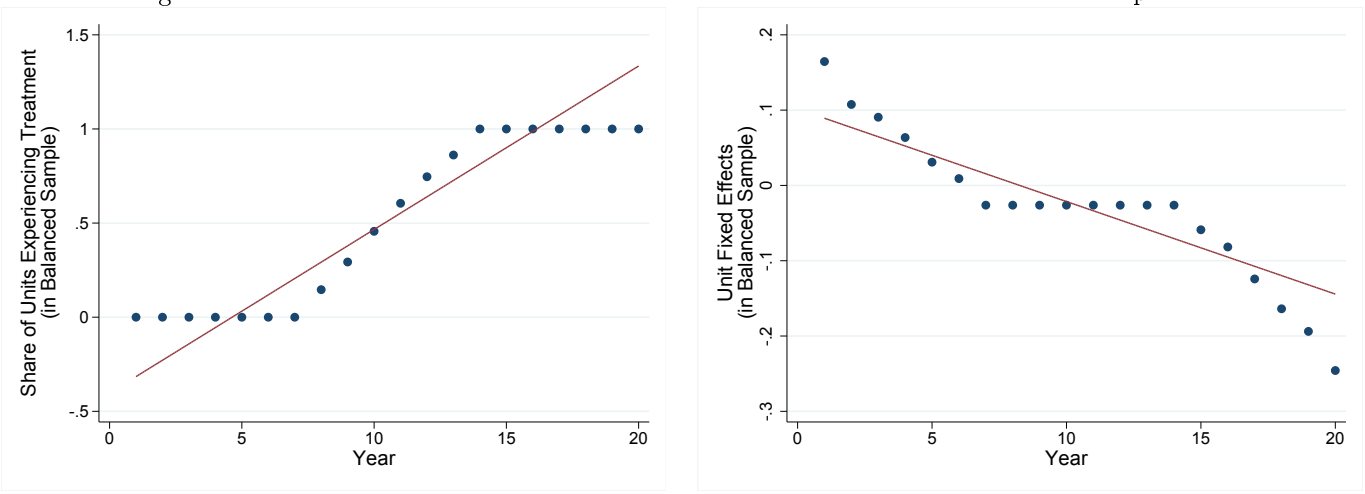


Figure 12 is based on a simulation with $Cov(\alpha_i, E_i) < 0$ ("better" units get treated earlier in calendar time), where the sample used for regressions is restricted such that each unit is observed for k years before the first year of treatment, as well as during the year of treatment and for $k-1$ years after. In other words, each unit is observed for k years under treated status and for k years under untreated status (in our simulation, $k = 7$). Figure 12 illustrates two points. First, in the first k years of the balanced sample none of the units are treated, and in the last k years of the sample all units are treated. This means that the year fixed effects for years that are early and late in the sample will apply to a homogeneous group of units, which are all either treated or untreated - in other words, they will absorb the treatment effect in those years. This means that identification of the treatment effect coefficient will come entirely from observations in the interval of time between those cutoff dates (i.e. with t such that $k < t < \bar{T}-k$, which in our simulation amounts to 6 years, with a total number of years given by $\bar{T} = 20$). Second, there are composition effects over the years in terms of the unit fixed effects: units that show up earlier in the sample tend to have higher fixed effects (because $Cov(\alpha_i, E_i) < 0$). The composition of unit fixed effects is stable in a balanced sample for years t such that $k \leq t \leq \bar{T} - k$, which in our simulation amount to 8 years. Intuitively, a balanced sample appears to solve the "selection issue" discussed in the setting of unbalanced samples because each unit gets treated for the same number of years during the sample: however, in practice year fixed effects absorb all of the variation for years t that are not such that $k < t < \bar{T} - k$ (i.e. that are at the beginning or end of the sample). For this reason, the point estimate we obtain for the treatment effect by running $Y_{it} = \beta_t + \gamma T_{it} + u_{it}$ in the balanced sample is exactly the same as the one we obtain by running the same specification in the balanced sample further

restricted to t such that $k < t < \bar{T} - k$ (the middle of the sample).[31] In this restricted sample (which is *unbalanced*!) we have $Cov(\alpha_i, E_i) < 0$.[32] Intuitively, once we eliminate the observations that are at the very beginning and the very end of the sample, we are left with a sample where units with higher fixed effects spend a bigger share of the sample under treated status.[33] This shows that endogenous selection effects in the data generating process affects the consistency of the point estimate in the same way under balanced and unbalanced panels.

Another way to summarize the intuition is as follows: the idea of irrelevance of unit fixed effects in balanced panel is coming from the observation that T_{it} is orthogonal to the unit dummies in the balanced sample. This holds unconditionally but fails conditionally on the time dummies, which are always included in these regressions. This failure is obvious: if ones fixes t, the dependence between T_{it} (equivalently, E_i) and the true unit fixed effects is precisely the problem we were hoping to address - and it exists even restricting the sample to individuals observed at t in the balanced sample.

E An Economic Model: the Permanent Income Hypothesis

To see how restrictive the assumption of homogenous treatment effects across cohorts and time is, we consider the simplest setting where event timing is fully random and unpredictable. All units (for simplicity, let us think of them as people for now) have the same prior belief about when the event may happen, and they update it over time using the Bayes rule. If a person does not receive treatment at the beginning of period t, this may be a minor shock if she thinks she is likely to get it next period, or a large shock if the expected date of the event moves far into the future, and behavior will respond accordingly. Only under very specific prior distributions can the treatment effect be independent of t.

For a specific example, consider an agent who lives for $t = 1, \ldots, T$ periods and is expecting to get a permanent raise of R from her baseline wage normalized to zero. The raise will happen at the beginning of a random period drawn from some distribution, and she does not get any information until the date of the raise. Her consumption follows the permanent income hypothesis, and there is no discounting. What is the effect of getting a raise on consumption at impact, τ_{it0}? If she gets the raise at time t, she is certain that the permanent income per period is R. If the event does not happen at t, it equals $R \cdot \mathbb{E}\left[T - E_i + 1 \mid E_i > t\right] / \left(T - t + 1\right)$. The treatment effect equals the difference between the two:

$$\tau_{ito} = R \cdot \frac{\mathbb{E}\left[E_i - t \mid E_i > t\right]}{T - t + 1}.$$

It is independent of t only if E_i is uniformly distributed across periods.

[31] We have verified in simulation that the point estimate and standard errors for the treatment effect are indeed exactly the same in the balanced sample and the balanced sample with the further restriction that $k < t < \bar{T} - k$.

[32] We have verified this in the simulation by running $T_{it} = \lambda \alpha_i + u_{it}$ in the balanced sample restricted to t such that $k < t < T - k$, which indeed yields $\lambda < 0$

[33] For instance, the units that were treated at the earliest possible time, i.e. with $T_i = k + 1$, are now treated in 100% of observations. Before further restricting the balanced sample, they were (by definition!) treated in 50% of observations. Conversely, the units that were treated at the latest possible time, i.e. with $T_i = \bar{T} - k - 1$, appear as treated in $\frac{1}{\bar{T} - 2k}$% of observation, or 16.66% of observations in our simulation.