FEDERAL RESERVE BANK *of* ATLANTA WORKING PAPER SERIES

Modeling Event Studies with Heterogeneous Treatment Effects

Laura M. Argys, Thomas A. Mroz, and M. Melinda Pitts

Working Paper 2023-11a
September 2023 (Revised August 2025)

Abstract: In this paper, we develop a new event-study approach that has the ability to capture model misspecifications in the presence of heterogeneous treatment effects. It is less prone to false rejections of the null hypothesis of no event-time variations than an event study that assumes a single treatment effect in the presence of heterogenous effects. We also introduce an alternative approach to calculating the standard errors for this event-study model to better assess the underlying trends in the model specification. These features contribute to a more precise understanding of the impacts of policy changes on outcomes and behaviors.

JEL classification: C22, C23

Key words: event studies, heterogeneous treatment effects

https://doi.org/10.29338/wp2023-11

The views expressed here are those of the authors and not necessarily those of the Federal Reserve Bank of Atlanta or the Federal Reserve System. Any remaining errors are the authors' responsibility. The authors report there are no competing interests to declare.

Please address questions regarding content to Laura M. Argys, College of Liberal Arts and Sciences, CB 181, University of Colorado, Denver, PO Box 173364, Denver, CO 80217-3364, laura.argys@ucdenver.edu; Thomas A. Mroz, Georgia State University, Department of Economics, PO Box 3992, Atlanta, GA 30302-3992, tmroz@gsu.edu; or M. Melinda Pitts, Federal Reserve Bank of Atlanta, Research Department, 1000 Peachtree Street NE, Atlanta, GA 30309-4470, 404-498-7009, melinda.pitts@atl.frb.org.

I. Introduction

Researchers commonly use a difference-in-difference methodology to assess the impact of a change in policy on behavior and outcomes. In recent years, event study analyses frequently accompany these difference-in-difference analyses (Currie, Kleven, and Zwiers (2020)). Event studies typically analyze the impact of a single policy event, with the primary focus on identifying pre-trends and a baseline comparison for post-treatment effects. However, in the presence of heterogeneous impacts of a treatment across treated units, as is likely in many economic program evaluations, the diagnostic information obtained from a simple, standard event-study analysis can be inadequate or misleading. In situations with only a small number of unique (and treatment-unit assignable) effects, one could introduce several new groups of event-time effects, one for each unique treatment effect, into the analysis to provide unbiased information, as suggested by Miller (2023). With a large number of group-specific event-time effects, however, the estimates of trends in the outcomes could be imprecise and thus, uninformative. In situations where treatment effects are functions of continuous covariates, an approach that allows for a different "event study" for each level of the treatment effect would not be feasible.

In this paper, we develop an approach to overcome the shortcomings of using a standard, single-treatment effect event study to assess the ability of an empirical model to measure heterogenous treatment effects. In addition, we consider an alternative approach to calculating the standard errors for this event-study model to better assess the underlying trends in the model specification. This approach we suggest for evaluating event-time effects directly carries over to generalized difference-in-differences (GDiD) models where treatment effects are not restricted to be single-valued and constant over time. In these GDiD models, one allows for the realistic possibility that the impact of a treatment could depend on measured variations in individual and location characteristics over time. For example, the impact of a job training program on employment could depend on one's age and education, as well as local labor market conditions and the availability of public transportation.

Once one introduces such realistic, heterogeneous, and time-varying responses to a treatment, there is no longer a "single treatment effect" in the post-treatment period. Any attempt to construct an event-study graph displaying a simple shift in the outcome immediately following the initiation of the treatment would not capture the treatment effect heterogeneity. Specifically, for example, should the "treatment effect" singled out in the simple event study be the impact of a change in school attendance requirements on a seventeen-year-old white youth from a poorly educated family in a high-unemployment period, on a similar Hispanic youth during a low-unemployment regime, or a somewhat arbitrary "average" effect estimated from the statistical model?

The fact that there is no unchanging single effect, however, does not imply that event-study analyses cannot provide important information. Instead of using a simple event-time graph to display the levels of "treatment effects" as a function of the time since treatment initiation (i.e., the event time), it can be informative to focus instead on another important purpose of using an event-study analysis, namely, uncovering any persistent patterns or trends related to the timing of the treatment in the outcome of interest that are not captured by the empirical model.

In the next section, the shortcoming of using a single event study in the presence of heterogeneous treatment effects is explored. The first step is to precisely define an event-study formulation with a simple, conventional single-treatment effect specification. We then present a slightly revised formulation of the event study that estimates "different" event study time effects but is otherwise a statistically and substantively identical model specification. In this second model, the reformulation is akin to a change in the "base category" for a set of mutually exclusive dummy variables; differences arise only because we focus on post-treatment deviations from the "treatment effect" estimated for the first post-treatment time period.[1] This reformulation yields standard errors that allow us to focus the analysis to our main objective: studying policy impacts in the presence of heterogenous treatment effects. We then apply these event-study approaches to artificial data sets where there are heterogeneous treatment effects that depend upon exogenous characteristics. We close with a discussion of the benefits arising from explicitly modeling sets of heterogenous treatment effects.

II. A "Single Treatment Effect" Estimate in a Heterogeneous Effect World

It is important to recognize the limitations of estimating a single treatment effect in the presence of heterogeneous treatment effects. Suppose there are two groups, g taking values 1 and 2, with treatment effects $\beta_1$and β_2, respectively, where the true model is

$$y_{g,i} = \beta_0 + \alpha \cdot 1(g = 1) + \beta_1 T_{g,i} \cdot 1(g = 1) + \beta_2 T_{g,i} \cdot 1(g = 2) + \varepsilon_{g,i} \tag{1}$$

with $E\left(\varepsilon_{g,i} | T_{g,i}, 1(g = 1), 1(g = 2)\right) = 0$. Under this conditional mean assumption, OLS estimation of this model would yield unbiased and consistent estimators of the two treatment effects β_1 and β_2, this would be the correct model for policy evaluation.

Suppose instead, as is typical in the literature, one estimates a single treatment effect. In this case, β^*, the single parameter defined by the OLS estimation procedure, does not recognize explicitly the heterogeneity in the treatment effects. The estimable regression model in this

[1] Other normalizations could be used; for example, one might instead examine deviations from some "average effect" across all post-treatment time periods.

instance becomes:

$$y_{g,i} = \gamma_0 + \beta^* T_{g,i} + \alpha^* \cdot 1(g = G_1) + \zeta^*_{g,i} \tag{2}$$

where the error term in this simplified model, when equation (1) is the true model, is, by definition, given by:

$$\zeta^*_{g,i} = \varepsilon_{g,i} + (\beta_1 - \beta^*)T_{g,i} \cdot 1(g = 1) + \ (\beta_2 - \beta^*)T_{g,i} \cdot 1(g = 2). \tag{3}$$

This implied error term is derived under the assumption that equation (1) is the true model. However, if β_1 and β_2 differ, then it must be the case that

$$E\left[\zeta^*_{g,i} | T_{g,i}, 1(g = 1), 1(g = 2)\right] \neq 0 \tag{4}$$

The standard OLS assumption for an unbiased and consistent estimator is necessarily violated whenever one estimates a "single treatment effect" when treatment effects are heterogeneous and related to any of the explanatory variables in the empirical model.[2]

The key implication of this fact is that the "single treatment effect" OLS regression model cannot be considered to represent a policy-relevant, conditional expected value. In this simple example, an application of the Frisch-Waugh-Lovell theorem reveals that the OLS estimator of β^* yields:

$$\hat{\beta}^* = \frac{\hat{\pi}_1 \hat{f}_1 \left(1 - \hat{f}_1\right)}{\hat{\pi}_1 \hat{f}_1 \left(1 - \hat{f}_1\right) + \hat{\pi}_2 \hat{f}_2 \left(1 - \hat{f}_2\right)} \hat{\beta}_1 + \frac{\hat{\pi}_2 \hat{f}_2 \left(1 - \hat{f}_2\right)}{\hat{\pi}_1 \hat{f}_1 \left(1 - \hat{f}_1\right) + \hat{\pi}_2 \hat{f}_2 \left(1 - \hat{f}_2\right)} \hat{\beta}_2 \tag{5}$$

where the $\hat{\beta}_g$ are the OLS estimators of the treatment effect obtained from separate regressions

[2] The failure of the "single effect" empirical model to satisfy the key conditional expected value assumption in the presence of heterogeneous effects has given rise to the growing literature on the inadequacies of the two-way fixed effect models. See, for example, De Chaisemartin and D'Haultfoeuille (2020), Goodman-Bacon (2021), and Sun and Abraham (2021).

for each of the two groups, the $\hat{\pi}_g$are the shares of the sample in each of the two groups, and the $\hat{f}_g$ are the fraction in each of the two groups within the sample receiving the treatment.

As Mroz (2024) points out, the way the OLS estimator implicitly "averages" the heterogeneous effects could provide misleading implications for policymakers. Consider, for example, a situation where the two groups are equally represented in the sample ($\hat{\pi}_1 = \hat{\pi}_2 = 0.5$), with 50% of group 1 receiving the treatment and 90% of group 2 receiving the treatment. In this case, the implicit group 1 weight used by OLS applied to equation (2) would be 2.78 times larger than the implicit weight assigned by OLS to group 2 ($0.7353 = \frac{.5(1-.5)}{.5(1-.5)+.9(1-.9)}$ $vs.$ $0.2647 = \frac{.9(1-.9)}{.5(1-.5)+.9(1-.9)}$). This occurs even though observations in group 2 are 80% more likely to be treated than observations in group 1. With more than two groups, and especially for empirical models including other control variables, the interpretation of the "single treatment effect" is even less straightforward.

Such a weighted average would clearly not approximate either an average treatment effect (all observations weighted equally) or an average effect of the treatment on the treated (all treated observations weighted equally). In most cases, an estimate of $\hat{\beta}^*$, as defined by equation (5), would not provide precise and useful information for guiding policy analyses. This lack of policy relevance from the OLS estimated "single treatment effect model" suggests that researchers should focus on estimating heterogeneous treatment effects. After estimating the heterogenous effects, one could construct policy-relevant weighted averages of these estimated effects to provide meaningful and useful average effects.[3] Section VI provides a more detailed

[3] It is known that standard event-study analyses can fail to provide useful information in the presence of heterogeneous treatment effects e.g. Sun and Abraham (2021).

discussion of the benefits of estimating heterogeneous treatment effects.

In this paper, we directly address this shortcoming of the standard event-study approach and provide an alternative event-study approach that one can use to assess the adequacy of an empirical model. This approach explicitly estimates heterogeneous treatment effects. Once these heterogeneous effects have been estimated and appear to be adequate (using the tests suggested below), one can then aggregate explicitly the heterogeneous effects to a single "average estimate" by using weights that would be appropriate for a particular policy analysis.

III. A Basic Event-Study Formulation

To begin, suppose the model under consideration is a simple model with separable treatment effects of the form:

$$y(s,t) = \beta_0^* + \beta_1^* T(s,t) + \beta_2^{*\prime} x(s,t) + \eta^*(s,t) \tag{6}$$

where the treatment unit is indexed by *s* and time periods by *t*.[4] Let $y(s,t)$ represent the outcome of interest that could be impacted by the treatment, measured by the dummy variable $T(s,t)$ which equals 1 if the treatment is in effect at time period t for unit s. The vector $x(s,t)$ captures characteristics of the treatment units and time periods impacting the outcome that could measure time- or unit-invariant characteristics, characteristics that could vary by unit and/or time, or collections of fixed effects. All unit-level fixed effects and calendar-time fixed effects, if any, are subsumed within $x(s,t)$. However, $x(s,t)$ does not contain any explicit "time-of-commencement-of-treatment" related variables. These will be introduced explicitly below.[5] To

[4] It is straightforward to introduce multiple observations per unit *s* and time period *t* and/or to have unbalanced data.

[5] Note, event time coefficients cannot be identified in the presence of variables in $x(s,t)$ that are

minimize the notation, we only consider the case where once the treatment takes place it remains in place throughout the end of the period of observation, i.e., $[T(s,t)=1] \Rightarrow [T(s,t+1)=1]$.

Define D(s) as the date at which the treatment begins for unit s.[6] We define the set of event-time dummy variables as $EV_r(s,t) = 1[t - D(s) = r]$, where $1[.]$ is the indicator function that equals 1 (instead of 0) if the event within the brackets is true. For example, $EV_0(s,t)$ is a dummy variable taking the value 1 in the time period when the treatment commences for unit s; $EV_{-1}(s,t)$ is a dummy variable taking the value 1 only in the last time period before the treatment commences for unit s; and $EV_1(s,t)$ is a dummy variable taking the value 1 in the second time period the treatment is in effect for unit s. Across all observations, the values of r range from -B to A, so there are $(\mathrm{A}+\mathrm{B}+1)$ dummy variables for $EV_r(s,t)$.[7]

Using these dummy variables, the standard event-study empirical model becomes:

$$y(s,t) = \tilde{\beta}_0 + \sum_{r=-B}^{-2} \tilde{\beta}_r^{EV} EV_r(s,t) + \sum_{r=0}^{A} \tilde{\beta}_r^{EV} EV_r(s,t) + \tilde{\beta}_2' x(s,t) + \tilde{\eta}(s,t) \tag{7}$$

Note that each coefficient $\tilde{\beta}_r^{EV}$ in equation (7) measures how the estimated "intercept" for event

perfectly colinear with some collection of the event-time variables.

[6] In this derivation we assume that the treatment eventually starts in each unit *s* within the period of observation. If this is not the case, then one would need to somehow assign, perhaps probabilistically, event times to units never observed being treated.

[7] Throughout this analysis, we assume all event-time effects are identified, though one could readily adapt the analysis to allow for "missing"/unidentified event-time effects by making additional or different exclusion restrictions. Interpretations of event-time coefficients always depend upon arbitrary normalizations whenever there is an intercept in the model.

time period r differs from the intercept for the excluded event-time -1, the last period prior to the introduction of the treatment. $\tilde{\beta}_0^{EV}$ is typically interpreted as the effect of the treatment in the first time period under the treatment, but formally all it measures is how the estimated intercept at event time period 0, the time period when the treatment is introduced, differs from the intercept for event time period -1, given the other explanatory variables in the econometric model.

A standard event-study graph plots the coefficients $\tilde{\beta}_r^{EV}$ against the event study time variable r. Typically, if the model is well-specified, the coefficients $\tilde{\beta}_r^{EV}$ for r < -1 should cluster around 0. The visual pattern followed by the estimates $\tilde{\beta}_{-B}^{EV}$ through $\tilde{\beta}_{-2}^{EV}$ is often used as an informal "test" for the parallel trend assumption used to validate a difference-in-difference or other related types of model specifications. Researchers frequently look at patterns in the post-treatment event-time coefficients, the $\tilde{\beta}_0^{EV}$ through $\tilde{\beta}_A^{EV}$, and interpret those patterns as a description of how the impact of the treatment evolves as a function of the duration of time since the initiation of the treatment.

There is a convenient, alternative normalization that makes it easier to track the evolution of the post-event-time coefficients; this normalization will be especially useful in instances when there is not a simple invariant "treatment effect." For the model described in equation (6), a mathematically and statistically equivalent specification for the event study presented in equation (7), when all coefficients are interpreted correctly, is given by:

$$\begin{aligned} y(s,t) = \hat{\beta}_0 + \sum_{r=-B}^{-2} \hat{\beta}_r^{EV} EV_r(s,t) + \hat{\beta}_0^{EV} T(s,t) + \sum_{r=1}^{A} \hat{\beta}_r^{EV} EV_r(s,t) + \\ \hat{\beta}_2' x(s,t) + \hat{\eta}(s,t) \end{aligned} \tag{8}$$

Note that all of the coefficients in equation (8) are identical to the corresponding coefficients in equation (7), except for the coefficients on the event-time dummy variables after the beginning of the treatment (i.e., $\hat{\beta}_r^{EV}$and the $\tilde{\beta}_r^{EV}$ for r>0).[8] In particular, the coefficient on the treatment dummy variable in equation (8) is identical to the coefficient on the dummy variable for the event time equaling 0 in equation (7), i.e., $\hat{\beta}_0^{EV}$ is identical to the $\tilde{\beta}_0^{EV}$multiplying the term $EV_0(s, t)$ in equation (7).

The coefficients on $EV_r(s, t)$ for each r>0 in equation (8) have a different interpretation than the corresponding coefficients in equation (7). Specifically, the coefficients $\hat{\beta}_r^{EV}$, for r > 0, in equation (8) equal exactly the differences $(\tilde{\beta}_r^{EV} - \tilde{\beta}_0^{EV})$ in the coefficients from equation (7); they measure how the treatment effects during each of the post treatment time periods differ from the initial treatment effect. This is the case because equation (8) controls for the treatment dummy variable $T(s, t)$. Each estimate of an event-time effect from equation (7) instead records the magnitude to which the post-treatment effect at event time r differs from pre-treatment outcome level at time r=-1.

If the specification in equation (6) were correct, i.e., there is a constant treatment effect through time once the treatment commences, then the coefficients $\hat{\beta}_r^{EV}$ for r>0 in equation (8) should cluster around 0. It is crucial to recognize that the event-time coefficients in equation (8) are not designed to uncover directly the magnitude of the treatment effect. Rather, they highlight any post-trends or patterns that are not captured in the model (relative to the treatment effect as estimated for the first treatment period, r=0, $\hat{\beta}_0^{EV}$). This feature is illustrated in some examples presented in Section V.

Another key difference between the event studies defined by equations (7) and (8)

[8] The error terms in the two equations are also identical, as the two models describe exactly the same relationship. They only differ by using different arbitrary normalizations.

concerns the standard errors used to define the confidence intervals around the estimated event-time effects. In equation (7), all the standard errors for the event-time effects, both prior to and subsequent to the commencement of the treatment, are appropriate when describing differences between the intercept at each event time and the intercept at event time -1. In many instances, however, one would like to evaluate the performance of the model in the post-treatment period and/or evaluate the evolution of the treatment effect over time. If that is the case, the standard errors corresponding to how the intercepts in post-treatment initiation periods differ from the intercept in the first treatment time period would be more appropriate standard errors to use for simple hypothesis tests related to how treatment effects vary relative to the treatment effect in the first period it is in place. Equation (8) provides these exact standard errors.[9]

IV. Event Study Formulations for Heterogeneous Treatment Effects

The utility of the normalization used in equation (8) becomes most apparent when the effect of the treatment is no longer just a simple, single effect. Suppose, as discussed briefly above, that the effect of the treatment differs depending on the variables $x(s,t)$. Those are key features that should be incorporated into any evaluation of the treatment. For example, the effect of compulsory schooling on teen labor force participation could be different for 16- and 17-year-olds, even when both are subject to the same mandate. Additionally, the types of jobs teens

[9] Using equation (8), a test of no change in treatment effect by event time following treatment initiation would only require a joint test that all the $\hat{\beta}_r^{EV}$ coefficients for r>0 are equal to zero. The same test, when using equation (7), would require one to test that all of the coefficients $\tilde{\beta}_r^{EV}$, for r>0, equal the coefficient corresponding to the treatment effect in the first year of treatment, $\tilde{\beta}_0^{EV}$.

might consider appropriate could depend on local employment conditions or their ability to drive to a job during the evening or at night (Argys, Mroz, and Pitts, 2025).

One direct way to capture such differential effects is to allow there to be different functions describing the outcome during the pre-treatment regime and under the treatment regime. Let $g_o[x(s,t),\theta_0]$ be the regression function describing the outcome in the absence of the treatment and $g_1[x(s,t),\theta_1]$ be the regression function describing the outcome in the presence of the treatment. The model in equation (1) is an extremely simple example of these two different regression models, where $g_o[x(s,t),\theta_0] = \beta_0^* + \beta_2^{*\prime}x(s,t)$ and $g_1[x(s,t),\theta_1] = \beta_0^* + \beta_1^* + \beta_2^{*\prime}x(s,t)$

Using this new notation, the regression model describing the impacts of the treatment is given by

$$y(s,t) = g_o[x(s,t),\theta_0]\cdot 1[T(s,t)=0] \; + g_1[x(s,t),\theta_1]\cdot 1[T(s,t)=1] \; + \eta^*(s,t)$$

Or, since $1[T(s,t)=0] = 1 - 1[T(s,t)=1]$:

$$y(s,t) = g_o[x(s,t),\theta_0] + \text{Effect}[x(s,t),\theta]\cdot 1[T(s,t)=1] \; + \eta^*(s,t), \qquad (9)$$

where

$$\text{Effect}[x(s,t),\theta] = \{\, g_1[x(s,t),\theta_1] - \; g_0[x(s,t),\theta_0]\,\}.$$

In this formulation, there is no single treatment effect. Rather, the effect of the treatment, Effect$[x(s,t),\theta]$, is a function of the vector of characteristics $x(s,t)$. The treatment effects could vary through time as well as by the value of observable unit-specific characteristics. One could, in principle, construct a different event time set of dummy variables for every relevant combination of the elements in the vector $x(s,t)$ and use those to specify a high-dimensional event study in the spirit of equation (7). That approach, however, often might be infeasible or yield mostly noise, especially when the number of units s and/or time periods t is small relative to the number of unique, relevant values of the vectors $x(s,t)$.

The event study formulation in equation (8), however, could easily be adapted to assess whether there are variations in the outcome $y(s,t)$, such as non-parallel trends prior to the initiation of the treatment, that are not captured well by the model in equation (9). Adopting some of the same notation as in equation (8) above, one could augment the regression model in equation (9) to yield the following event-study formulation:

$$\begin{aligned} y(s,t) = \hat{g}_0\left[x(s,t),\hat{\theta}_0\right] + \widehat{\text{Effect}}\left[x(s,t),\hat{\theta}\right] \cdot 1[T(s,t)=1] \\ + \sum_{r=-B}^{-2} \hat{\beta}_r^{EV} EV_r(s,t) + \sum_{r=1}^{A} \hat{\beta}_r^{EV} EV_r(s,t) + \hat{\eta}(s,t). \end{aligned} \qquad (10)$$

In equation (10), the interpretations of the parameters $\hat{\beta}_r^{EV}$ for all values of r (not equal to -1 or 0, given the imposed normalization) would be identical to those discussed for equation (8). The $\hat{\beta}_r^{EV}$ for r<-1 could be used to identify pre-treatment trends not captured by the functional form; the presence of such trends would suggest a mis-specified model. Similarly, any patterns associated with the $\hat{\beta}_r^{EV}$ for r>0 would be indicative of a failure to model the evolution of the outcomes $y(s,t)$ post-treatment with the chosen functional forms. Additionally, it is simple to test the null hypothesis that the regression model is correctly specified by testing the joint hypothesis that all the $\hat{\beta}_r^{EV} = 0$, for r=-B,...,-2,1,...A.

There is a cost of moving from a high-dimensional collection of event studies and their corresponding event-time effects (say separate sets of event-time coefficients, one for each age and/or education level) to a single set of homogeneous event-time effects. Specifically, consider some subgroup of the data that is defined by a particular configuration of their $x(s,t)$ values. Then, suppose this subgroup's outcome had been trending differently than that of other non-treated groups in the pre-treatment period. By focusing on only one combined set of event-time effects as in equation (10), estimation of the empirical model might put little emphasis on this one subgroup's deficiencies for identifying effects. Thus, this single, aggregate event study could fail to uncover the model's deficiencies.

There are alternatives to reducing the number of possible event studies to just one set of event-time effects. One could use economic reasoning to categorize the data into multiple subgroups and incorporate separate sets of event-time effects for each subgroup. A single joint test that all subgroups' event-time effects satisfy the conditions for model adequacy could provide a more powerful test of the null hypothesis that the model is appropriate for describing the effects of the treatment than the test from a single aggregated event study. Attempts to later test which specific subgroup(s) might have led to the overall rejection of the model, however, could be imprecise and inexact because of pre-testing and multiple hypothesis issues.

Alternatively, if one has some a priori information that some particular subgroup(s) might be differentially problematic, that information should be incorporated explicitly into the specification of the event studies. That is the approach used in Argys, Mroz, and Pitts (2025) to examine the impact of Graduated Driver Licensing (GDL) restrictions on teen labor supply. They allow for two different event-study sets of effects: one for comparisons of those currently subject to GDL restrictions when compared to those never covered; the other for those who formerly faced driving restrictions compared to those who never faced driving restrictions. In the following section we simulate data to illustrate these points.

V. Simulated Examples of Event Studies in the Presence of Heterogenous Effects

We generate artificial data to illustrate comparisons across the different approaches for modeling the event-study time effects. The artificial data sets constructed for this exercise contain a collection of 50 potentially treated "states" observed over 20 time periods (years). We include 10 observations within each state-year. Each state is observed for at least two years prior to the introduction of a non-reversible treatment, and the propensity to start the treatment in state s is stochastically related to the magnitude of the potential treatment effect for the state.[10]

In the first of our DGPs built upon this framework, we allow the outcome variable (y) to be impacted by a common time trend (t) and a time-varying state-specific explanatory variable (x). We specify and identify three groups of states differentiated by their time trends in the exogenous, but stochastic propensity to initiate the treatment. There is no variation in the treatment effect within each of the three state groupings. Thus, there are exactly three different treatment effects. [11]

The first column in Appendix Table 1 contains the true parameters used in the DGP to define the regression model for this first set of three treatment effects. The second column displays regression output from a simulated data set generated by the DGP using

[10] Precise details of the data-generating process (DGP) for the collection of explanatory variables are in the Stata do-file "make_locality_data.do", available in the online appendix.

[11] Details on the exact model specification and Stata code for this first data set and the first set of graphs that follow can be found in the Stata do-file "simpler_model.do." in the online appendix.

the same regression function that generated the data. Prior to the introduction of the treatment in each state, there are no systematic differences in the outcome across groups that are not explained by the exogenous variable x and the time trend t. The third column contains estimates using the same data set with a regression model that incorrectly imposes a single treatment effect that applies to all states instead if three group-specific treatment effects.

The estimates in the second column of Appendix Table 1 closely correspond to the true regression coefficients specified in Column 1 for the actual DGP. When we impose the restriction that the three treatment effects are identical (Column 3), the estimated coefficients on the explanatory variable x and the time trend change little. The state-group membership coefficients, however, do differ substantially from their true zero values. The single estimated treatment effect falls within the range of the three treatment effects, but it does not represent an easily interpretable average effect (e.g., Callaway and Sant'Anna, 2021; Goodman-Bacon, 2021; Sun and Abraham, 2021). This reflects the model misspecification due to the assumption of a single treatment effect rather than any bias due to the staggered treatments. We know this to be true because we generated the data and know precisely the form of the true model.

Next, we present a standard event-study graph corresponding to the "single treatment effect" (incorrect model) estimates in Appendix Table 1 column (3). To do this, we follow equation (7), which is the standard event-study approach. We replace the treatment effect variable, Treatment, with a sequence of dummy variables indicating the time since the introduction of the treatment (event-time dummy variables), using the excluded event-time dummy for the last pre-treatment period (r=-1) as the base event-time. We also combine 11 or more years pre-treatment into a single dummy variable, and we group 11 or more years post-treatment into a single event-time dummy variable. Figure 1a displays this event study.

A cursory examination of Figure 1a suggests that while the outcome does appear to be trending downwards in the early pre-treatment years, there is an immediate uptick in the outcome at the time of the treatment initiation that diminishes slightly the longer the treatment has been in effect. Since we made up these data, however, we know there are no such features in the true DGP corresponding to any of the pre- or post-treatment trends. In fact, in this incorrectly specified regression model applied to this one data set, we resoundingly reject the null hypothesis that all the pre-treatment event-time effects are zero (10 restrictions; $p<0.0001$).[12] We also reject that all post-treatment effects are the same (11 restrictions; $p=0.0048$).

Instead of using the standard event-study framework described in equation (7), the event study presented in Figure 1b utilizes the approach described by equation (8). In this statistically equivalent specification, the post-initiation event-time effects are measured relative to the measured impact of the treatment in the initial treatment year (instead of relative to the year just prior to the treatment). All test statistics are identical for consistently-defined event study effects in Figures 1a and 1b. However, Figure 1b directly displays how the post-period event treatment effects differ from the initial treatment effect.

[12] The power to reject in these examples, however, is extremely arbitrary as we set the accuracy of the model in our specification of the DGP. We do not address here any concerns about Type I and Type II errors.

It is important to emphasize the two major differences between the two event-study graphs in Figure 1. First, in Figure 1a all the event study effects are measured relative to the "effect" at event time -1. In Figure 1b the post-treatment initiation event-study effects are instead measured relative to the estimated treatment impact of 1.22 in the first treatment period, i.e. at event time 0. Second, the standard errors used to construct the confidence intervals in Figure 1a correspond to the standard errors appropriate for testing differences from the "event time effect" estimated for the last pre-treatment time period. The standard errors used in Figure 1b for the post-treatment periods correspond to the standard errors for testing hypotheses about how the effects for event times +1 and later differ from the initial treatment effect at event time 0. The confidence bands in the right-hand panel likely provide more relevant measures for assessing directly the adequacy of the estimated model to capture post-treatment variation in effects, which would typically be a primary reason to apply an event-study framework in the post-treatment period.[13]

In Figure 2 we correctly model the three different treatment effects. We also incorporate three separate sets of pre- and post-treatment dummy variables, one for each of the three groups of states, to assess the adequacy of the statistical model.[14] The "Group 2 event-time -5" dummy variable, for example, equals 1 only for an observation in a state belonging to Group 2 exactly

[13] The more-relevant event-time effects and their standard errors for Figure 1b could also be constructed from the information contained in the regression output for Figure 1a.

[14] For the DGP used here, the ranges of event-times observed separately by the three different groups differ. That is obvious in Figures 2a and 2b where there are fewer pre-treatment event-time effects for Group 3 than for Groups 1 and 2.

five years before the beginning of the treatment in that particular state; otherwise, it is zero. Using a slight modification of the standard event study in equation (7) to allow for multiple effects and event times, Figure 2a presents these three sets of event studies together in a single graph.

By correctly modeling the heterogeneous treatment effects, the misleading trends uncovered in the single treatment effect event studies presented in Figure 1 disappear in Figure 2a. There is no apparent evidence of any pre- or post-treatment trends within any of the three groups from a visual inspection of Figure 2a, just as in the true DGP. Additionally, in this single data set generated by the DGP, all hypothesis tests (separate or combined by group, and pre- or post-treatment effects separate or combined) fail to reject their corresponding null hypothesis of no pre-treatment trends and no variations in post-treatment trends.

The treatment effects displayed on the right-hand side of Figure 2a convey a significant amount of information visually, but that is mostly because in our DGP we specified the treatment effects to be quite disparate. In more realistic models, it might be difficult to assess from the post-treatment portions of Figure 2a whether there are significant deviations from the constant effects in the post-treatment period without a solid reference point.

We begin to address these shortcomings by first respecifying the event-study regression model in a way that allows one to better assess visually whether there are significant deviations from the modeled treatment effects, similar to what was done in Figure 1b. We present this alternative specification in Figure 2b. This is accomplished by replacing the three event-time 0 dummy variables that were used in the regression model underlying Figure 2a with three different group-treatment dummy variables; this removes the event time 0 effects from the post-treatment event study effects as in equation (8) but with multiple sets of event study effects.

By removing the event time 0 effects, the pointwise confidence bands in Figure 2b overlap considerably. Unfortunately, this makes it difficult to visually inspect the three separate sets of event-time effects. With only three "treatment effects" in this model, one could easily plot out three separate graphs, one for each set of event-time coefficients.[15] However, in models with numerous heterogenous treatment effects, such a strategy would likely be noisy or infeasible. In such circumstances, reducing the clutter could make the event study more useful. To accomplish this, we apply a single set of event-time dummy variables along with the three different estimated treatment effects, as described in equation (10), to this same set of data. Figure 2c applies this approach to the same data set used in Figures 1, 2a, and 2b. The same result of no misspecification that was indicated by the three separates lines in Figure 2b is shown in Figure 2c. Specifically, note the absence of trends like those portrayed in Figure 1 which imposed a single treatment effect.

To build on the example shown in Figure 2, we turn to a more complicated set of treatment effects, where the impact of the treatment varies over time and across groups and as a function of observed continuous exogenous variables. In this second DGP we include three features to the model: (1) differential group-specific impacts on the level (intercept) of the treatment effects; (2) explanatory variables with trends impacting the outcome differentially by group that change because the treatment commences (the WVAR in Appendix Table 2); and (3) time trends whose effects shift differentially by group because of the start of the treatment (the

[15] A simple F-test of all the event-time effects in Figure 3 equaling zero would provide a test of the adequacy of the regression model that incorporates the group-specific treatment effects.

TVAR in Appendix Table 2).[16] The true regression parameters for this second DGP are displayed in the first column of Appendix Table 2, and the second column displays regression output from a simulated data set generated by this second DGP using a correctly-specified model.

Figure 3, similar to Figure 2c, contains a single set of event-time dummy variables for the more complicated regression model presented in column 2 of Appendix Table 2 and as described by equation (10). Not surprisingly, since the regression model corresponds exactly to the true DGP, there is no evidence of model misspecification in either the pre-or post-treatment periods. All tests of pre- and post-treatment event-study coefficients (i.e., those estimated with this one dataset using the true DGP) fail to reject the null hypothesis of zero event-time effects, whether tested pre-treatment only, post-treatment only, or all event-time effects tested jointly. Figure 3 provides a concise summary of the regression model's performance, even though there is a different treatment effect associated with each unit within each group that varies across the post-treatment periods as a function of discrete and continuous exogenous variables and time trends.

To illustrate the performance of this approach when the model is misspecified, we create a third DGP that alters the second DGP slightly to incorporate post-treatment, state-specific shifts in the outcome that have a 10% hazard of taking place after the treatment has been in effect for three time periods. The idea behind incorporating these post-treatment shifts is that there could be unmodeled policy changes impacting the outcome taking place after the start of the initial treatment. The regression results displayed in the third column of Appendix Table 2 contain the estimates when the regression model used in the second column is applied to the third DGP that has these unmodeled, randomly starting outcome shifts that commence post-treatment

[16] See Stata do-file less_simple_model.do.

and vary differentially for units within the same groups.

A comparison of coefficients in columns 2 and 3 of Appendix Table 2 demonstrates the importance of the model misspecification due to the additional post-treatment shifts in the outcome variable. Almost all of the estimated coefficients in the third column are within one standard error of the estimates in the second column, and in only one instance out of the 18 estimated coefficients is the difference as large as two standard errors. A cursory examination of the table might suggest that there is no evidence of model misspecification in Column 3.

However, the event-study analysis presented in Figure 4 tells a different story. There is some evidence of a post-treatment commencement uptick in the "treatment effect." That visual observation is confirmed by an F-test that rejects the hypothesis that all the event-time effects jointly equal zero. Even though the model accounts for detailed and varying treatment effects across groups and units and variations within units, a single set of event study effects is able to pick up the model misspecification.

VI. What is "The Treatment Effect?"

The presence of numerous treatment effects is a departure from the current literature and thus could benefit from a more detailed discussion. Most research has a single treatment effect, and researchers use a simple event-study analysis to assess the reliability of the estimate. The explicit introduction of heterogeneous treatment effects in the econometric regression model requires adopting a different approach for carrying out event-study analyses.

The event-study coefficients corresponding to the pre-treatment time periods are straightforward; they should all be close to zero, provided one has a well-specified statistical model for the pre-intervention periods regardless of whether or not there are heterogeneous treatment effects. The interpretation of the simple event-study effects in the post-treatment regime, however, is where this work departs from the prior literature. In a typical event study, one often interprets the coefficients on the post-treatment dummy variables as measuring over time variations in the (single) treatment effect that represent delays in rolling out the treatment, learning, or some other set of factors that cause the treatment's impact to vary over time. Researchers tend to be agnostic about the source of these post-treatment effect variations, though not infrequently they do provide heuristic explanations.

The econometric specifications we consider here, however, do not allow for such unmodeled treatment effects. Rather, the techniques developed and discussed here assume that researchers have a well-developed behavioral model of how treatment effects might vary across time (and other dimensions) and implements an econometric model that can capture policy-relevant treatment effect heterogeneity. If researchers believe they have an accurate model of treatment effects, then the univariate event-time effects suggested here provide an important tool for one to assess the adequacy of a specific model. All event-time effects should be null in both the pre- and post-treatment time periods, given an adequate econometric specification. In addition, if researchers suspect there might be a model misspecification along some other dimension than simple time effects, it is straightforward to extend the "event-study model" suggested here to examine the model's ability to adequately capture the real-world data.[17]

[17] If one uses measures of time-since-initiation of treatment to capture how the treatment effects vary over time, then some of the coefficients on the post-treatment event-time dummy variables would likely be unidentified. In such instances, one would need to choose arbitrary normalizations for the event-time coefficients and use tests based on the "identified" event-time coefficients for the specification test.

Researchers typically desire a single value for “the treatment effect,” which can potentially mask important and observable treatment effect heterogeneity. Almost ubiquitously, they estimate econometric models that report such a single “treatment effect”. However, the explicit heterogeneous effect models we introduce in this research (e.g., described in the discussion of equation (9) and in the example set of estimates displayed in Appendix Table 2) allow for treatment effects to vary by individual and environmental characteristics. If one requires the calculation of a single “treatment effect,” then one must take an appropriate weighted average of the heterogenous effects. The question researchers would like to answer should determine the specific set of weights.

In standard two-way fixed effect models, this “single treatment effect” approach has given rise to a large literature demonstrating that a single-valued effect estimated by OLS is a weighted average of some set of the heterogeneous treatment effects where many of the weights can be negative (see, for example, Goodman-Bacon, 2021). Some proposed solutions to the “negative weight” issue have focused on OLS estimators of single-valued treatment effects estimated from particular subsets of the available data where the OLS estimates of the single-valued effect do not implicitly rely upon any negative weights for the implicit heterogeneous treatment effects (e.g., Callaway and Sant’Anna, 2021; Sun and Abraham, 2021). This comes at the cost of ignoring a wide swath of the population’s heterogeneous treatment effects. Such data-deletion “fixes” do eliminate complex, philosophical interpretation issues arising from negative weights for constructed “average effects”. However, it is by no means clear that the implicit positive weights imposed by such OLS estimations of a single treatment effect parameter, given the elimination of many of the heterogeneous treatment effects from the analysis, would be appropriate for conducting meaningful policy evaluations.

When there is heterogeneity in treatment effects that depends on the explanatory variables in a single treatment effect OLS estimation, as discussed briefly in Section II, it cannot be the case that $E(error|X) = 0$. This happens because the implicit OLS error term for a model with a single-valued treatment effect, in the presence of true heterogeneous treatment effects related to the explanatory variables, explicitly violates the fundamental assumption that the included regressors are exogenous variables within the specified regression function. This violation of the statistical exogeneity assumption can result in subsets of the treatment effects that contain higher proportions of treatment compliers receiving a smaller weight in the implicit OLS weighted average than subsets of treatment effects with a smaller fraction of compliers, as discussed above. Such weighting schemes are incompatible with important measures such as the average treatment effect and the average effect of the treatment on the treated (Mroz, 2024). For most welfare policy-motivated metrics, weights for the heterogeneous treatment effects should generally be non-negative and a non-decreasing function of the fraction treated within homogeneous sub-groups (holding sub-group sizes constant); simple OLS estimators do not automatically provide this type of weighted average. A regression model that explicitly allows for observable heterogeneity in the set of treatment effects, like the model in equation (9), overcomes these serious shortcomings. Assuming one has a set of theoretical and policy-motivated weights, it is straightforward to average across the estimated heterogeneous effects.[18]

Suppose, for example, there are two groups, teenage children and young adults, and the appropriate set of weights should put twice as much importance on the teenage children relative to the young adults. In this case, for any fixed set of other explanatory variables potentially

[18] See Argys, Mroz and Pitts (2025), for an example.

impacting the effect of the treatment, one only needs to define relative weights for teenagers that are twice the magnitude of the weights for young adults and apply these weights to construct the average of the estimated heterogeneous effects. Such a policy-relevant weighted average would typically have a larger standard error than one obtained from a single treatment effect OLS model. As Mroz (2024) demonstrates, OLS only achieves its extra statistical "efficiency" because its implicit choice of weights ignores the policy relevance and instead minimizes the variance of its single-valued estimator under the assumption of homoscedastic errors as if there were no treatment effect heterogeneity.

A simple way to calculate a policy-relevant average effect is for one to consider a set of explanatory variables $x(s_1, t_1), x(s_2, t_2), \dots, x(s_K, t_K)$. Associated with each of the K sets of explanatory variables is a policy-relevant weight w_k that reflects the researcher's or a policymaker's belief of the importance of individuals/groups with characteristics $x(s_k, t_k)$ for the calculation of the policy-relevant average effect. Using the definition of the heterogeneous treatment effects in equation (9), it is straightforward to calculate this particular policy-relevant average effect as:

$$\text{Effect}[w_1, \dots, w_K; x(s_1, t_1), \dots, x(s_K, t_K)] = \tag{11}$$

$$= \left(\frac{1}{\sum_{k=1}^{K} w_k}\right) \sum_{k=1}^{K} w_k \{\, g_1[x(s_k, t_k), \theta_1] - \; g_0[x(s_k, t_k), \theta_0] \,\}$$

The researcher or policymaker can ensure that all weights w_k are non-negative and that they reflect a reasonable and defensible policy objective. If the g-functions are linear, it is straightforward to calculate the standard error of this particular average effect (e.g., Stata's lincom command). If the g-functions are nonlinear, one can use the delta method to calculate the standard error using, for example, Stata's nlcom command.

VII. Conclusion

This new event-study approach, described in equation (10), has the ability to capture model misspecifications in the presence of heterogeneous treatment effects. Unlike the imposed, single homogeneous effect analysis displayed in Figure 1, it should be less prone to false rejections of the null hypothesis of no event-time variations in the outcome after modeling the appropriate treatment effect heterogeneity. Additionally, even in the presence of a single treatment effect, the standard errors obtained by using versions of equations (8) and (10) provide the correct information for assessing whether unmodeled variations in treatment effects remain. The standard errors from a more conventional event-study analysis, which focus on variations in event-time effects compared to the "level" in the last pre-treatment period, do not provide that correct information directly. These features will contribute to a more precise understanding of the impacts of policy changes on outcomes and behaviors.

References

Argys, Laura M., Thomas A. Mroz, and M. Melinda Pitts (2025) Driven from Work: Graduated Driver licensing Restrictions and the Decline in Teen Labor Force Participation," Federal Reserve Bank of Atlanta Working Paper 2019-6.

Callaway, Brantly and Pedro H.C. Sant'Anna. 2021. "Difference-in-Differences with multiple time periods," Journal of Econometrics, Vol. 225, Issue 2, December, pp. 200-230.

Currie, Janet, Henrik Kleven, and Esmée Zwiers. 2020. "Technology and Big Data Are Changing Economics: Mining Text to Track Methods." *AEA Papers and Proceedings*, 110: 42-48.

de Chaisemartin, Clément and Xavier D'Haultfoeuille. 2020. "Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects," American Economic Review, Vol. 110, No. 9, September, pp. 2964–96.

Goodman-Bacon, Andrew. 2021. "Difference-in-differences with variation in treatment timing," Journal of Econometrics, Vol. 225, Issue 2, December, pp. 254-277.

Miller, Douglas L. 2023. "An Introductory Guide to Event Study Models." *Journal of Economic Perspectives*, 37 (2): 203-30.

Mroz, Thomas A. (2024). "Estimating Empirical Models with Interesting Economic Effects in the Mid-21st Century," Unpublished Manuscript. Georgia State University.

Sun, Liyang and Sarah Abraham. 2021. "Estimating dynamic treatment effects in event studies with heterogeneous treatment effects," Journal of Econometrics, Vol. 225, Issue 2, December, pp. 175-199.

Figure 1: Comparison of the Two Event Study Approaches Assuming a Single Treatment Effect

1a: Standard Event Study 1b: Modified Event Study

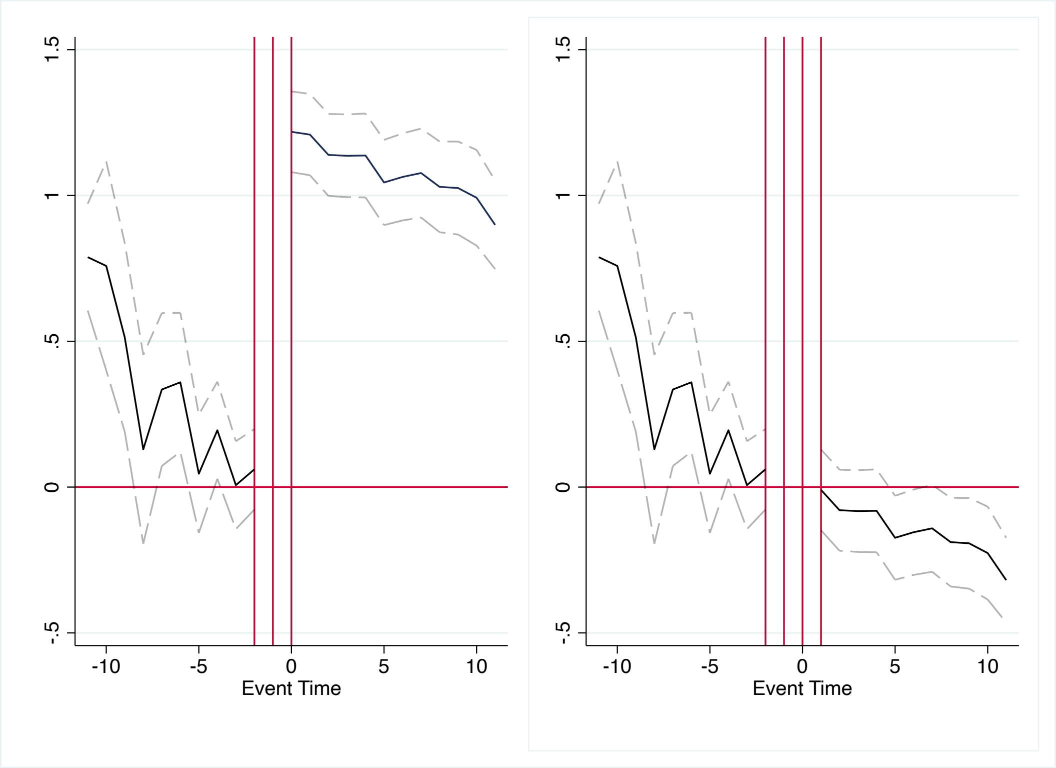
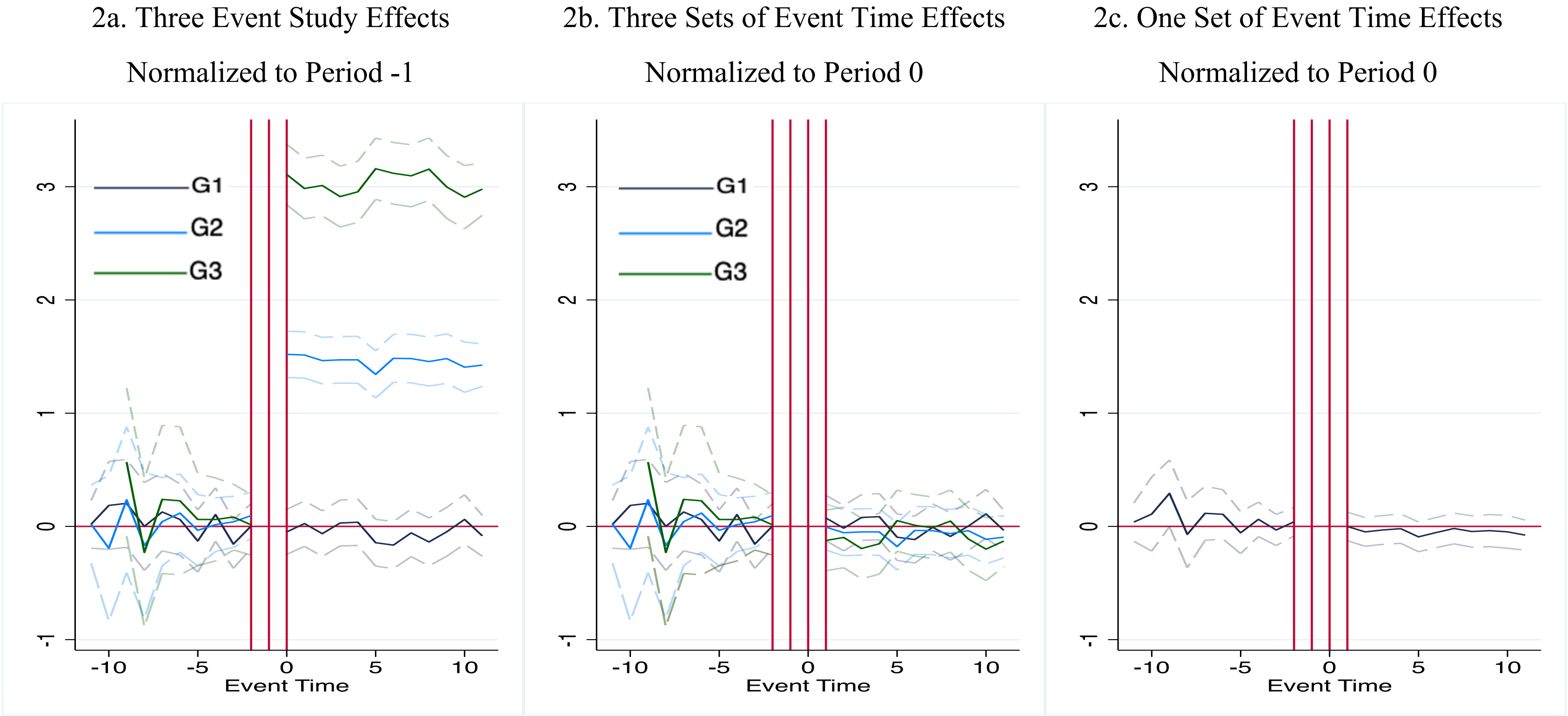


Figure 1 is from the model of effects presented in Column 3 of Appendix Table 1. Event time 0 indicates the treatment initiation. Pre- and post-treatment effects are relative to event time -1 in 1a while the post-treatment effects are relative to time 0 in 1b. The shaded lines bound the 95% confidence intervals. Event time 0 effect (1.22) and standard error (0.07) come from the event-study estimation; see the do-file "simpler_model.do" in the appendix.

Figure 2. Event Studies for Three Estimated Treatment Effects



2a. Three Event Study Effects

Normalized to Period -1

2b. Three Sets of Event Time Effects

Normalized to Period 0

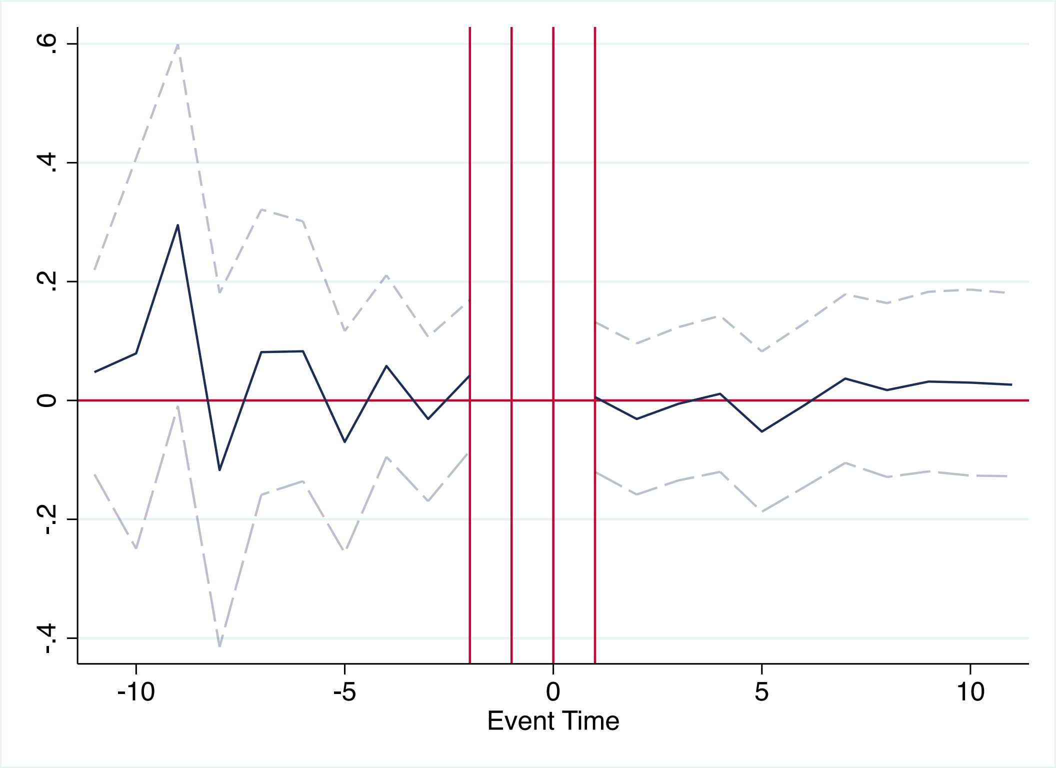
2c. One Set of Event Time Effects

Normalized to Period 0

Figure 2 is from the (correct) model of effects presented in Column 2 of Appendix Table 1. G refers to group. Period 0 indicates the treatment initiation. In 2a period -1 is the base for the pre- and post-treatment periods, while period 0 is the base for the post-treatment period in both 2b and 2c. Figure 2c utilizes Equation 10 when recognizing three different treatment effects with a single set of event study effects. The shaded lines bound the 95% confidence intervals. The event-time effects and standard errors come from the event-study estimation; see the do-file simpler_model.do in the appendix.

Figure 3

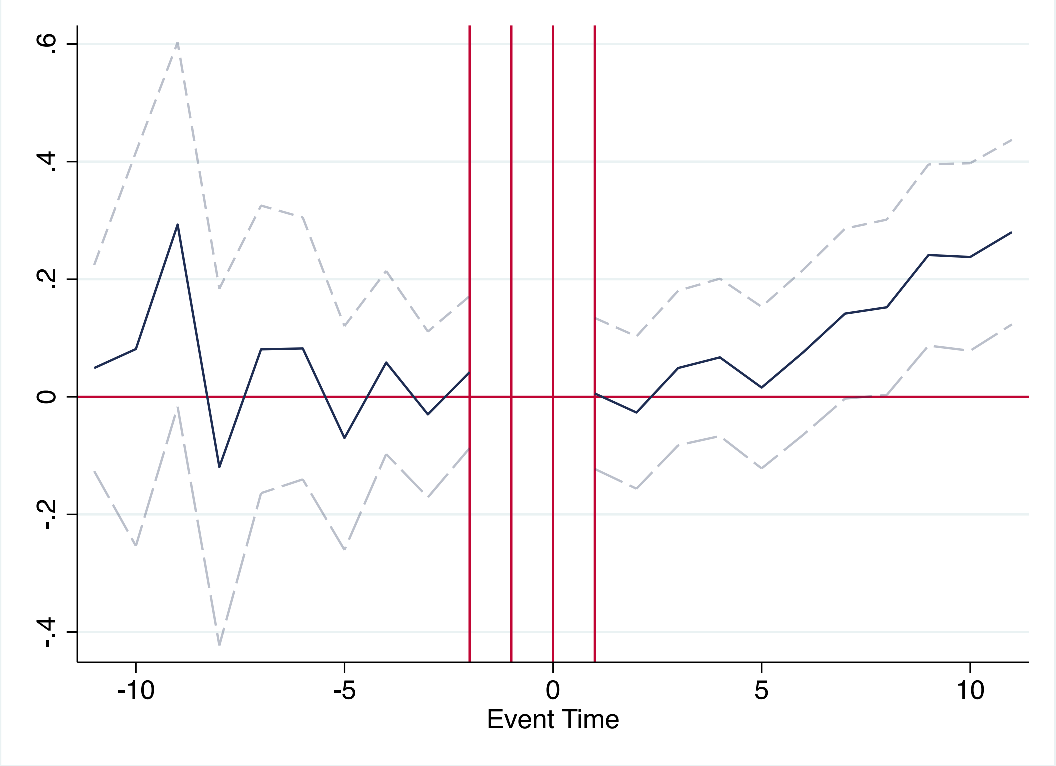
Event Study for Estimated Treatment Effects that Vary with Time, Group, and Continuous Exogenous Variables in a Correctly Specified Econometric Model



This is a non-traditional event study which removes all treatment effects, trends, and interactions that are modeled as impacting the treatment. This event study indicates that there are no non-modeled effects related to the treatment. Period 0 indicates the treatment initiation; period -1 is the base period for the pre-treatment event-study effects and the time 0 event effects; period 0 is the “base” for the remaining post-treatment event-time effects. Small dashed shaded lines are the upper bound and the long-dashed shaded lines are the lower bound of the 95% confidence intervals.

Figure 4

Event Study for Estimated Treatment Effects that Vary with Time, Group, and Continuous Exogenous Variables in an Incorrectly Specified Econometric Model



Non-traditional event study, which removes all treatment effects, trends, and interactions that are modeled as impacting the treatment. Period 0 indicates the treatment initiation; period -1 is the base period for the pre-treatment event-study effects and the time 0 event effects; period 0 is the "base" for the remaining post-treatment event-time effects. Small dashed shaded lines are the upper bound and the long-dashed shaded lines are the lower bound of the 95% confidence intervals.