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 2023-11
September 2023

Abstract: This paper develops a simple approach to overcome the shortcomings of using a standard, single treatment–effect event study to assess the ability of an empirical model to measure heterogeneous treatment effects. Equally as important, we discuss how the standard errors reported in a typical event-study analysis for the posttreatment event-time effects are, without additional information, of limited use for assessing posttreatment variations in the treatment effects. The simple reformulation of the standard event-study approach described and illustrated with artificially constructed data in this paper overcomes the limitations of conventional event-study analyses.

JEL classification: C22, C23

Key words: event studies, heterogeneous treatment effects

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

Please address questions regarding content to Laura M. Argys, Department of Economics, 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

The event study is often used to provide a visual overview of the impact of a change in policy on behavior and outcomes. According to Currie, Kleven, and Zwiers (2020), it has become increasingly more common have an events study analysis accompany a difference-in-difference analysis. They found that in 2005 less than two percent of papers in top five journals and in the NBER working paper series included event studies, but by 2018 that share had increased to approximately seven percent.

Most event-study analyses analyze the impact of a single policy event, with the primary focus on identifying pre-trends and a baseline comparison for post-treatment effects. Miller (2023) provides an overview of the features and the decision process of the event study model. However, with heterogeneous impacts of a treatment across treated units, as is likely in many economic program evaluations, diagnostic information obtained from a simple, standard event study analysis can often be inadequate or misleading. In situations with only a small number of unique (and treatment-unit assignable) effects, one could simply introduce several new groups of event-time effects, one for each unique treatment effect, in the analysis to provide accurate information, as suggested by Miller (2023). This approach, while fairly general, could possibly introduce considerable noise because of the multiplicity of event-time study effects. The individual group-specific event-time effects could be quite imprecise, and thus uninformative estimates of trends in the outcomes occurring before and after the introduction of the treatment. In the case of treatment effects depending on continuously observed covariates, an approach that allows for a different "event study" for each level of the treatment effect would seldom be feasible.

In this paper, we discuss a simple 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. Equally as important, we discuss how the standard errors reported in a typical event study analysis for the post-treatment event-time effects are, without additional information, of limited use for assessing post-treatment variations in the treatment effects. The simple reformulation of the standard event study approach described below overcomes the limitation of the conventionally reported standard error bounds presented in typical event study analyses.

The approach we suggest for evaluating event-time effects directly carries over to generalized difference-in-differences (GDiD) models where treatment effects are not modeled as simple separable effects that do not vary across time. In these GDiD models, one allows for the realistic possibility that the impacts of the treatments could depend on measured variations in individual and location characteristics over time. The impact of a job training program on employment, for example, 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 in a meaningful fashion. More substantively, should the "treatment effect" singled out as the impact of a legal change in a simple event study be the impact on school attendance for a seventeen-year-old white youth from a poorly-educated family in a high-unemployment period, or the impact on a similar Hispanic youth during a low-unemployment regime, or a somewhat arbitrary

"average" effect estimated from a statistical model that ignores economic and policy motivations for aggregating the heterogeneous treatment effects?

The fact that there is no unchanging single effect, however, does not imply that event study analyses cannot provide important information. However, rather than 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 is likely to be more informative and less arbitrary to focus instead on the primary purpose of using an event study analysis. That purpose is to uncover any persistent patterns or trends in the outcome that are not captured by the empirical model related to the timing of the treatment.

In this paper, we provide an overview of the methodology and use artificially-constructed data sets to illustrate the advantages and disadvantages of various event-study techniques. We begin by precisely defining an event study formulation with a simple, conventional single-treatment effect specification. We then present a slightly different formulation of the event-study that estimates "different" event-study time effects but is otherwise a statistically and substantively identical model specification. In this reformulated simple 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] We then expand the analysis to situations where there is no simple homogeneous treatment effect to demonstrate the usefulness of the event-study reformulation. We close this paper by applying these event-study approaches to artificial data sets where there are heterogeneous "treatment effects" that depend upon exogenous characteristics.

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

II. A Basic Event Study Formulation

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

$$y(s,t) = \beta_0^* + \beta_1^* T(s,t) + \beta_2^{*\prime} x(s,t) + \eta^*(s,t) \quad (1)$$

where the treatment unit is indexed by *s* and time periods by *t*.[2] 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 were in effect at time period t. The vector $x(s,t)$ captures characteristics of the treatment units and time periods impacting the outcome that could measure time- or unit-invariant variables, collections of fixed effects, or varying effects by unit and/or time. All unit-level fixed effects and 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.[3] To 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, or, equivalently, $[T(s,t) = 1] \Rightarrow [T(s,t+1) = 1]$.

Define D(s) as the date at which the treatment becomes in effect for unit s.[4] 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

[2] It is straightforward to introduce multiple observations per unit *s* and time period *t*. For this discussion, that only adds an unnecessary level of abstraction. Note that data do not need to be "balanced."

[3] To simply apply the approaches suggested here, it is crucial for one to understand the precise meaning of the "event-time" effects in the presence of variables in $x(s,t)$ that might be perfectly colinear with some collection of the event time variables.

[4] We assume that the treatment eventually starts in each unit *s* within the period of observation. This means that we do not need to somehow assign, perhaps probabilistically, event-times to units never observed being treated.

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 $EV_r(s,t)$ range from -B to A.[5]

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

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

Note that each coefficient $\tilde{\beta}_r^{EV}$ in equation (2) measures how the estimated "intercept" for event 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.

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 for patterns in the post-

[5] For each $EV_r(s,t)$ for r ranging between -B and A, we assume there is at least one unit *s* that has $EV_r(s,t) = 1$ for all values of t. We do this mostly to simplify the notation. Depending on the configuration of the explanatory variables $x(s,t)$, however, additional assumptions may be needed to ensure that "event-time effects" identifiable, in the sense of measuring exactly pure time effects about the commencement of the timing of the treatment. Note, importantly, the definitions of the event-time coefficients will crucially depend upon any arbitrary normalization(s) that would be chosen in such instances.

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 (1), a mathematically and statistically equivalent specification for the event study presented in equation (2), when all coefficients are interpreted correctly, is given by:

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

(3).

Note that all of the coefficients in equation (3) are identical to the corresponding coefficients in equation (2), 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).[6] In particular, the coefficient on the treatment dummy variable in equation (3) is identical to the coefficient on the dummy variable for the event-time equaling 0 in equation (2), i.e., $\hat{\beta}_0^{EV}$ is identical to the $\tilde{\beta}_0^{EV}$ multiplying the term $EV_0(s,t)$ in equation (2). The coefficients $\hat{\beta}_r^{EV}$, for r > 0, in equation (3) equal exactly the differences $(\tilde{\beta}_r^{EV} - \tilde{\beta}_0^{EV})$ in the coefficients from equation (2),[7] so they measure how the

[6] 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.

[7] Using equation (3), a test of no change in treatment effect by event-time following treatment initiation would only require a joint test that all of the $\hat{\beta}_r^{EV}$ coefficients for r>0 equaling zero. The same test, when using equation (2), would require one to test that each of the coefficients $\tilde{\beta}_r^{EV}$ for r>0 equals the coefficient corresponding to the treatment effect in the first year of treatment, $\tilde{\beta}_0^{EV}$, or its logical equivalent.

treatment effects during each of the post treatment time periods differs from the initial treatment effect.

The coefficients on the $EV_r(s,t)$ for r>0 in equation (3), the $\hat{\beta}_r^{EV}$ for r>0, have a different interpretation than the corresponding coefficients in equation (2). This is the case because equation (3) controls for the treatment dummy variable $T(s,t)$, and because of the "fact" that the treatment, once initiated, does not stop during the period of observation. Specifically, $T(s,t) = 1$ for all time periods post-initiation of the treatment for unit s. The coefficient $\hat{\beta}_r^{EV}$ for r>0 measures how the intercept for event-time r (for r>0) differs from the intercept of the first period when the treatment is in effect, i.e., the intercept for event-time r=0. If the specification in equation (1) were correct, that is there is a constant treatment effect through time once the treatment commences, then the coefficients $\hat{\beta}_r^{EV}$ for r>0 in equation (3) should cluster around 0. Estimates from equation (2) highlight the magnitude to which post-treatment effects differ from pre-treatment outcome level at time r=-1. Equation (3), on the other hand, is not designed directly to uncover the magnitude of the treatment effect, but rather to 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 one of the examples presented in Section IV.

Another key difference between the event studies defined by equations (2) and (3) concerns the standard errors used to define the confidence intervals around the estimated event-time effects. In equation (2), 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 the appropriate ones to use for simple hypothesis tests. Equation (3) provides these exact standard errors.

III. Event Study Formulations for Heterogeneous Treatment Effects

The utility of the normalization used in equation (3) becomes most apparent when the effect of the treatment is no longer just a simple, single effect. Suppose, as discussed briefly above, that the treatment differs depending on the variables $x(s,t)$. Those are key features that should be incorporated into any evaluation of the treatment. The effect of compulsory schooling on teen labor force participation, for example, could be different for 16- and 17-year-olds even when both are subject to the mandate. Additionally, the types of jobs teens 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, 2023).

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 representation/example of these two different regression models. 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 $1[T(s,t)=0] = 1 - 1[T(s,t)=1]$

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

(4)

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, $\text{Effect}[x(s,t),\theta]$, 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 (2). That approach, however, often would 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 (3), 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 (4). Adopting some of the same notation as in equation (3) above, one could augment the regression model in equation (4) to yield

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

(5)

In equation (5), 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 (3). 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 misspecified model. Similarly, any patterns associated with the $\hat{\beta}_r^{EV}$ for r>0 would be indicative of a failure to model well the evolution of the outcomes $y(s,t)$ with the chosen functional forms. Additionally, it would be 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. In particular, consider some subgroup of the data that is defined by a particular configuration of their $x(s,t)$ values. Suppose this subgroup's outcome had been trending differently than other non-treated groups in the pre-treatment period. By focusing on only one combined set of event-time effects as in equation (5), estimation of the empirical model might not put much emphasis on this one subgroup's deficiencies for identifying effects.[8] That could result in the single, aggregate event study failing 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 categorize the data into multiple subgroups and incorporate separate sets of event-time effects for each subgroup. A single joint test that all of the subgroups' event-time effects satisfy the conditions for model adequacy might

[8] If such disparate trends were incorporated appropriately into the $g_o[x(s,t),\theta_0]$ and $g_1[x(s,t),\theta_0]$ functions, then there would be no reason for the event-time coefficients to detect a model misspecification; and this discussion would be moot.

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 subgroup(s) might have led to the overall rejection of the model, however, might be quite imprecise and inexact unless one appropriately controlled for the pre-testing and multiple hypothesis issues.

If one did have 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 (2023). They allow for two different event study sets of effects: one for comparisons of those currently subject to Graduated Driver Licensing restrictions when compared to those never covered; the other for those who only formerly faced driving restrictions compared to those who never faced driving restrictions. But even without such prior information, the single set of event study effects described in equation (5) should allow one to uncover many empirical models' inadequacies. In the following section we simulate data to illustrate these points.

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

We create 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 units observed over 20 time periods, which, for ease of exposition, we now label states and years. We include 10 observations within each state-year, but those multiple observations are not crucial for the issues discussed here. 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. Precise details of the data-generating process for the collection of explanatory variables are contained in the Stata do-file "make_locality_data.do", available in the online appendix.

In the first of our two data-generating processes (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. 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 first column in Table 1 contains the true parameters defining the regression model for this first set of three treatment effects, and the second column displays the regression output from one simulated data set generated by the DGP using the exact regression model used in the DGP. 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 from a model that incorrectly imposes a single treatment effect that applies to all states.

The estimates in the second column of 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, 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 only know this, of course, because we made up the data and know precisely the form of the true model.

Table 1

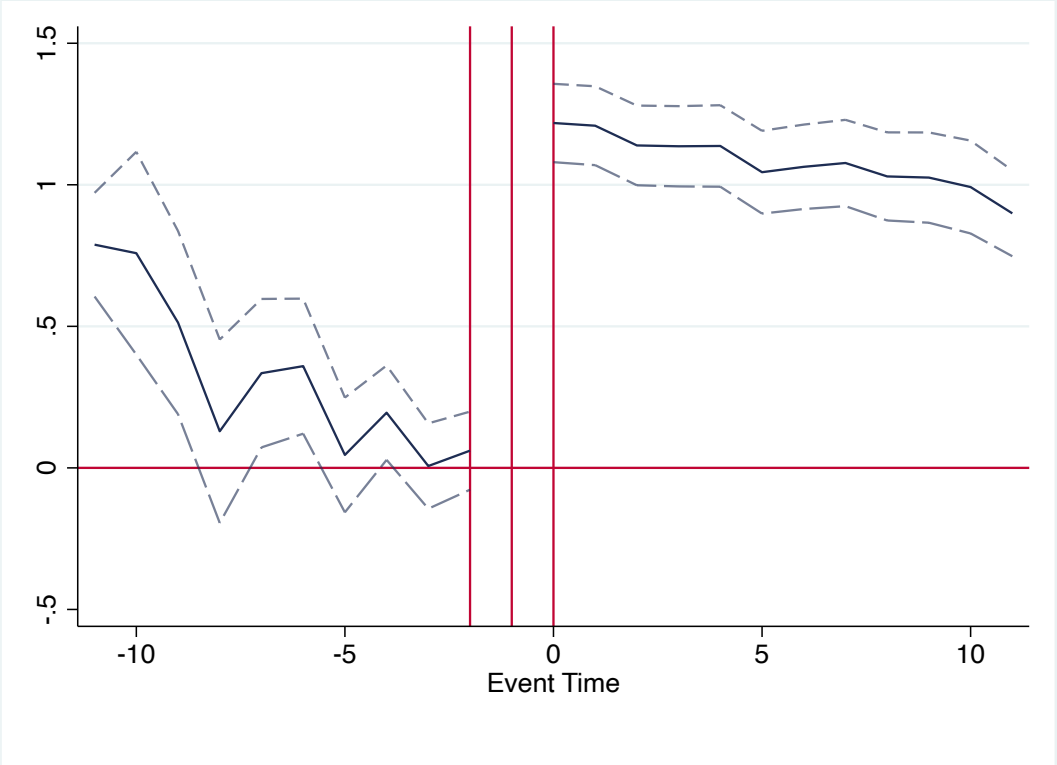
Model Parameters and a Sample Regression

Variables	Data Generating Process	Correct Model	Incorrect Model
Group 1	0.0	0.0810	-2.304***
		(0.0588)	(0.0298)
Group 2	0.0	0.0647	-1.212***
		(0.0612)	(0.0299)
Group* Treatment			
Group 1	0.0	-0.0357	
		(0.0396)	
Group 2	1.5	1.445***	
		(0.0446)	
Group 3	3.0	3.000***	
		(0.0590)	
Treatment	NA		1.014***
			(0.0335)
t	0.2	0.199***	0.207***
		(0.00220)	(0.00243)
x	1.0	0.994***	0.996***
		(0.0105)	(0.0116)
Constant	0.0	-0.0270	1.506***
		(0.0503)	(0.0335)
Observations		10,000	10,000
R-squared		0.805	0.762

Standard errors in parentheses; *** p<0.01, ** p<0.05, * p<0.1

Next, we present a standard event study graph corresponding to the "single treatment effect" estimates in Table 1 column (3). To do this we follow equation (2). We replace the treatment effect variable, treat, 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 event-time dummy variable, and we group 11 or more years post-treatment into a single dummy variable. Figure 1 displays this event time study.

Figure 1
A Standard Event Study



Standard event study that assumes a single treatment effect for the coefficients reported in Column 3 of Table 1. Period 0 indicates the treatment initiation and period -1 is the base. Small dashed shaded lines are the upper bound and the long-dashed shaded lines are the lower bound of the 95% confidence intervals.

A cursory examination of Figure 1 suggests that while the outcome does appear to be trending downwards in the early pre-treatment years, there is an immediate uptick in outcome at the time of the treatment initiation that only 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 model, we resoundingly reject the null hypothesis that all of the pre-treatment event-time effects are zero (10 restrictions; $p<0.0001$).[9] We also reject the null hypothesis that all of the post-treatment effects are the same (11 restrictions; $p=0.0048$). Not surprisingly a combined test for the two composite hypotheses rejects the combined null hypothesis (21 restrictions, $p<0.001$). Of course, these three p-values, given that we examine sequentially three related tests, are not accurate representations of the true probabilities under each stated null hypothesis.

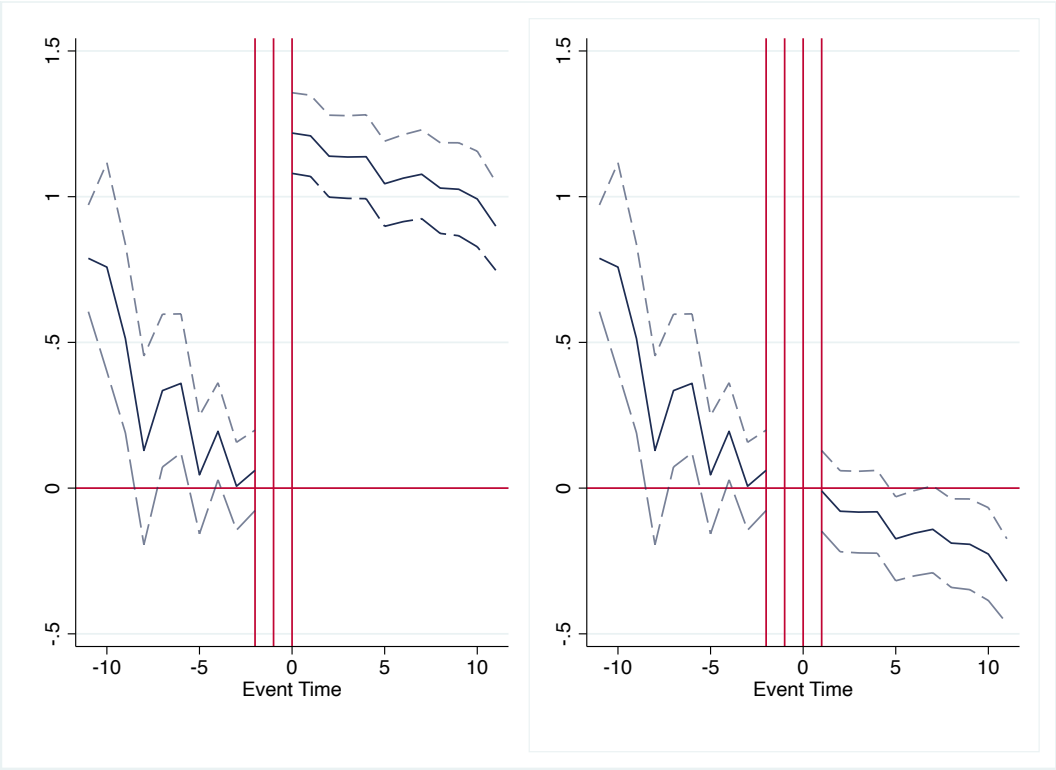
Instead of using the standard event study framework described in equation (2), this next event study utilizes the approach described in equation (3) that controls for the post-treatment effect when estimating the event-time coefficients. In this statistically-equivalent specification, the post-initiation event-time effects are measured relative to the measured impact of the treatment in the initial treatment year (instead of relative to the last year prior to the treatment). The right-hand panel of Figure 2 presents this modified event study approach while the left-hand panel merely repeats Figure 1.

[9] 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.

Figure 2
Comparison of the Two Event Study Approaches with a Single Treatment Effect

Standard Event Study
Assuming Single Treatment Effect
Column 3, Table 1

Modified Event Study
Event Time 0 Effect (se):
1.22 (0.07)



Period 0 indicates the treatment initiation and period -1 is the base. Small dashed shaded lines are the upper bound and the long-dashed shaded lines are the lower bound of the 95% confidence intervals. The event time 0 effect and standard error come from the event study estimation; see the do-file "simpler_model.do" in the appendix.

There are two major differences between these two event study graphs. First, the post-treatment initiation event study effects in the right-hand panel are measured relative to the treatment impact (intercept) at event-time 0 (the first treatment period). In the left-hand panel they are instead measured relative to the "effect" (intercept) at event-time -1. Second, and more importantly, the standard errors used to construct the pointwise confidence interval bands in the right-hand graph 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 standard errors used to construct the confidence intervals in the left-hand graph, instead, correspond to the standard errors appropriate for testing differences from the "event-time effect" estimated for the last pre-treatment time period. The confidence bands in the right-hand panel likely provide more relevant measures for assessing the adequacy of the estimated model, which would typically be the reason to apply an event study framework.[10]

Since we know there are three different groups of states with possibly different effect sizes across state-groups, we can construct different event studies for each of the three groups. To do this we construct three separate sets of pre- and post-treatment dummy variables, one for each of the three groups of states.[11] The "Group 2 event-time -5 dummy variable," for example, equals 1 only for an observation in a state belonging to Group 2 exactly five years before the beginning of the treatment in that particular state; otherwise, it is zero. Figure 3 presents these three sets of event-studies together in a single graph.

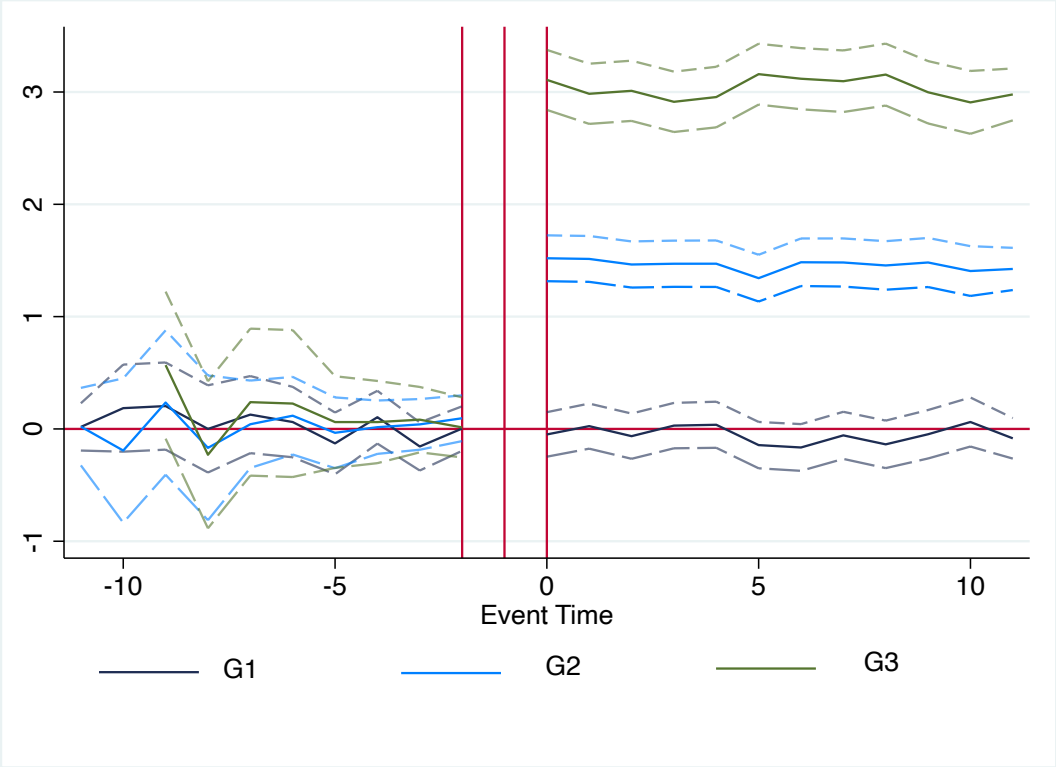
By correctly modeling the heterogeneous treatment effects, the anomalies uncovered in the single treatment effect event studies presented in Figure 2 disappear when examining Figure 3. There is no apparent evidence of any pre- or post-treatment trends for any of the three groups from a visual inspection of Figure 3, just like in the true DGP. Additionally, all hypothesis tests (separate or combined by group, and pre- or post-treatment effects separate or combined) fail to reject their corresponding null hypothesis (no pre-treatment trends and no variations in post-treatment trends)

[10] Though, as noted above, one could easily construct the more-relevant event-time effects and their standard errors for the right-hand graph from the information contained in the regression output for the left-hand panel.

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

Figure 3

A Small Number (3) of Different Treatment Effects



Period 0 indicates the treatment initiation and period -1 is the base. Small dashed shaded lines are the upper bound and the long-dashed shaded lines are the lower bound of the 95% confidence intervals. G refers to group. The event time effects and standard errors come from the event study estimation; see the do-file simpler_model.do in the appendix.

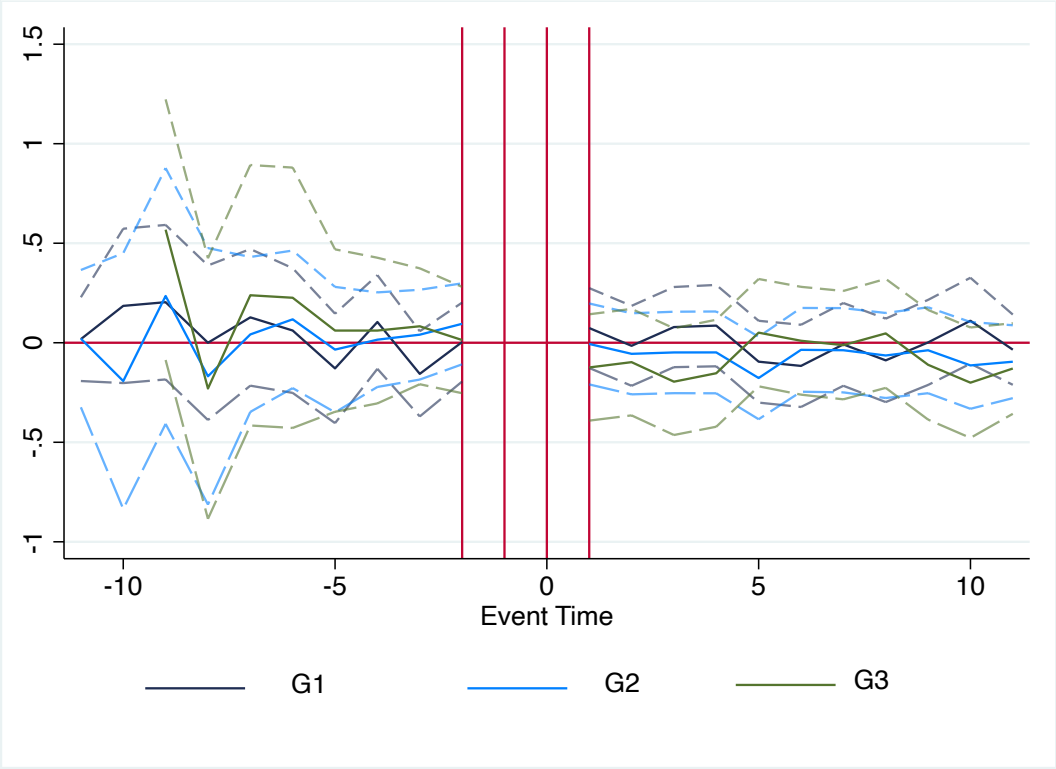
The treatment effects displayed on the right-hand side of Figure 3 appear to convey a significant amount of information visually, but that is mostly because in our made-up DGP we specified the treatment effects to be quite disparate. Additionally, in more realistic models, it might be difficult to assess from the post-treatment portions of Figure 3 whether or not 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 slightly 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 effect. That type of assessment is one of the primary purposes for examining an event study graph in the post-treatment period.

In Figure 4 we apply the approach used in equation (3) adapted to difference out the three separate, event-time r=0 treatment effects from the post-treatment event time-effects displayed in Figure 3. We do this by replacing the three event-time 0 dummy variables that were used in the regression model underlying Figure 3, with the three different treatment dummy variables.

Figure 4

Event Study for Three Treatment Effects Using Modified Equation (3)



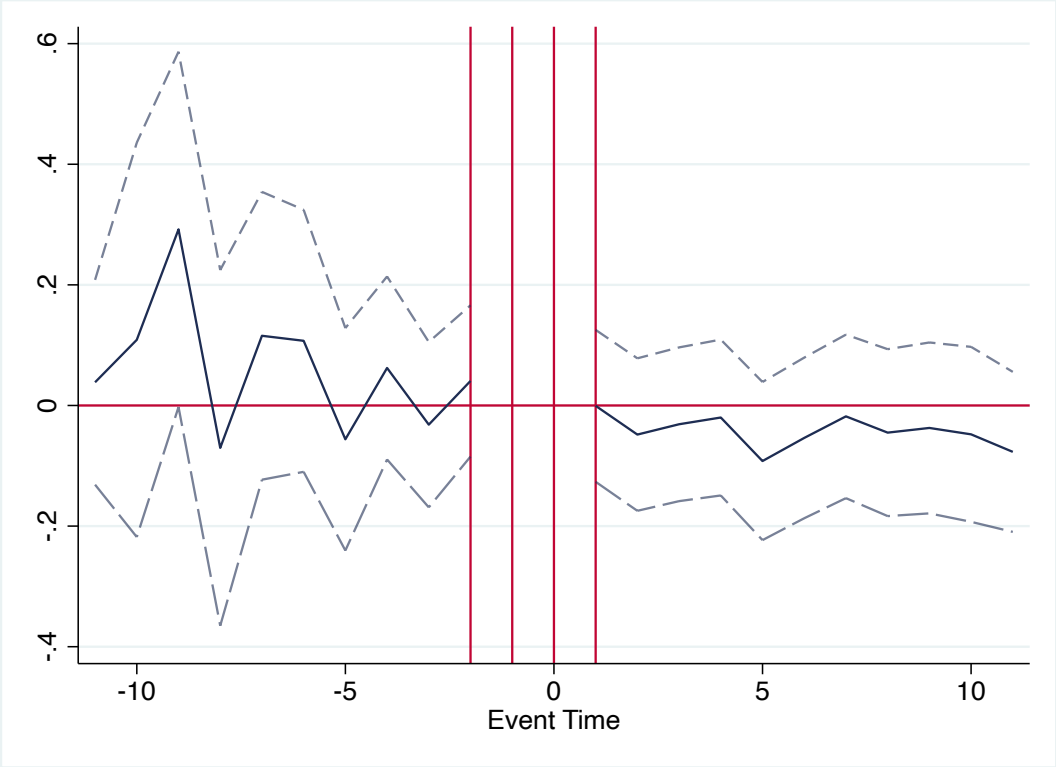
Period 0 indicates the treatment initiation and period -1 is the base period for the pre-treatment event study effects and the time 0 event effects; period 0 is the "base" for the post-treatment event-time effects. Small dashed shaded lines are the upper bound and the long-dashed shaded lines are the lower bound of the 95% confidence intervals. Different treatment and event time effects are estimated for each group (G). Event time 0 effects (se): G1: -0.049 (0.101); G2: 1.520 (0.104); G3: 3.108 (0.136). These come from the event study estimation in the do-file simpler_model.do in the appendix.

By removing the event-time r=0 effects, the pointwise confidence bands overlap considerably, making 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. And, of course, a simple F-test of all the event-time effects in Figure 4 equaling zero would provide a test of the adequacy of the regression model that incorporates the group-specific treatment effects.

To reduce the clutter in Figure 4, we apply the single set of event-times approach described in equation (5) to this same set of data. In a real data set, where there could be model inadequacies related to the treatment groups or other factors, this approach could help one to uncover issues with the specified econometric model. However, if as mentioned above, one has an a priori notion that some subgroup might be differentially subject to model misspecifications than other subgroups, then the information obtained by using the simplification found in equation (5), instead of using equation (3) with multiple sets of event-time effects, might be a less useful approach. Of course, in this made-up data set, where we know there are no peculiarities in the DGP related to event-times, there is no real information loss from examining only a single set of event-time effects. As expected, Figure 5 displays this feature of the true DGP from an estimated model with only a single set of event-time "effects."

Figure 5

Event Study for Three Treatment Effects Using Equation (5)



Period 0 indicates the treatment initiation and 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 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. Different treatment and event time effects are estimated for each group (G). Event time 0 effects (se): G1: 0.020 (0.071); G2: 1.491 (0.072); G3: 3.037 (0.081). These come from the event study estimation in the do-file simpler_model.do in the appendix.

We next turn to a more complicated set of treatment effects where the impact of the treatment varies over time and across groups as a function of observed exogenous variables. In this DGP (see Stata do-file less_simple_model.do) there are differential group-specific impacts on the level (intercept) of the treatment effects, explanatory variables with trends impacting the outcome differentially by group that change after the treatment commences (WVAR), and time trends whose effects shift differentially by group at the start of the treatment (TVAR). The true regression parameters for this DGP are displayed in the first column of Table 2, and the second

column contains estimates from one simulated data set using the (correct) regression model as specified in the DGP. We also create a third DGP that alters the second DGP slightly to allow for post-treatment, state-specific shifts in the outcome that have a 10% hazard of taking place after the treatment has been in effect for 3 time periods. The idea behind incorporating these post-treatment shifts is that there could be related unmodeled policy changes taking place after the start of the initial treatment. The regression results displayed in the third column of Table 2 contain the point estimates when the regression model used in the second column is applied to the DGP that has these unmodeled, randomly starting outcome shifts that can commence post-treatment and vary differentially for units with the same groups.

Figure 6 contains the event study results for the regression model presented in column 2 of Table 2 as described by equation (5). Not surprisingly, since the regression model corresponds exactly to the true DGP, there is no evidence of a model misspecification in either the pre-treatment period or in the post-treatment period. All tests of pre- and post-treatment event study coefficients fail to reject the null hypothesis of zero event-time effects, whether tested pre-treatment as a group, post-treatment as a group, or all event-time effects tested jointly. Figure 6 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 exogenous variables and time trends.

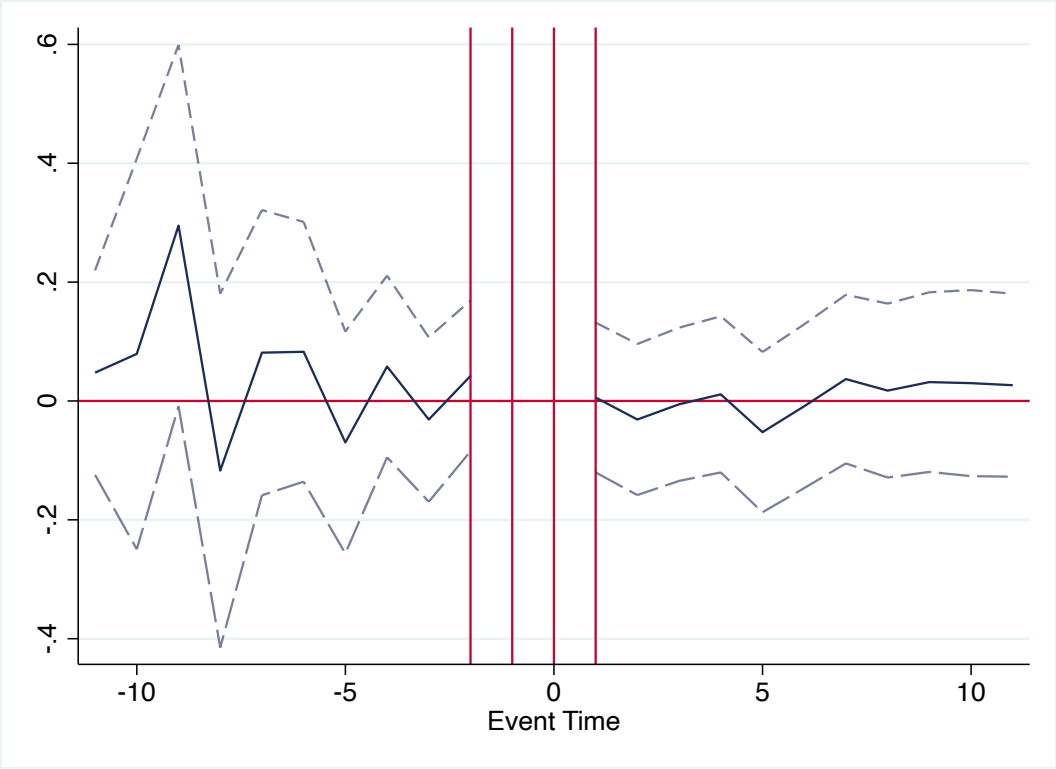
Table 2
Data Generating Process with Treatment Effects depending on Groups with Group-Specific Responses to Time Trends and Trended Exogenous Variables

Variables	Data Generating Process	Correct Model	Incorrect Model
Group 1	1.00	1.038***	1.040***
		(0.156)	(0.159)
Group 2	2.00	2.134***	2.137***
		(0.163)	(0.167)
Group*Treatment			
Group 1	0.00	0.131	0.0469
		(0.0882)	(0.0900)
Group 2	1.50	1.485***	1.331***
		(0.0994)	(0.101)
Group 3	3.00	3.076***	3.005***
		(0.159)	(0.162)
Group*WVAR			
Group 1	0.60	0.563***	0.560***
		(0.111)	(0.114)
Group 2	0.80	0.851***	0.846***
		(0.111)	(0.113)
Group 3	1.00	1.147***	1.146***
		(0.166)	(0.169)
Group *Treatment* WVAR			
Group 1	0.80	0.790***	0.767***
		(0.126)	(0.129)
Group 2	1.30	1.252***	1.357***
		(0.123)	(0.125)
Group 3	1.80	1.643***	1.639***
		(0.175)	(0.179)
Group*TVAR			
Group 1	0.20	0.224***	0.224***
		(0.00715)	(0.00730)
Group 2	0.20	0.189***	0.190***
		(0.0130)	(0.0133)
Group 3	0.20	0.194***	0.194***
		(0.0262)	(0.0268)
Group*Treatment*TVAR			
Group 1	-0.03	-0.0569***	-0.0360***
		(0.00815)	(0.00832)
Group 2	0.05	0.0565***	0.0724***
		(0.0136)	(0.0139)
Group 3	0.11	0.116***	0.127***
		(0.0267)	(0.0272)
X	1.00	0.995***	0.999***

Table 2 (cont.)			
		(0.0110)	(0.0112)
Constant	0.00	-0.0930	-0.0936
		(0.142)	(0.145)
DGP has an unmodeled post treatment revision		No	Yes
Observations		10,000	10,000
R-squared		0.917	0.916

Standard errors in parentheses*** p<0.01, ** p<0.05, * p<0.1

Figure 6
Heterogeneous Effects Varying over Time and By Group



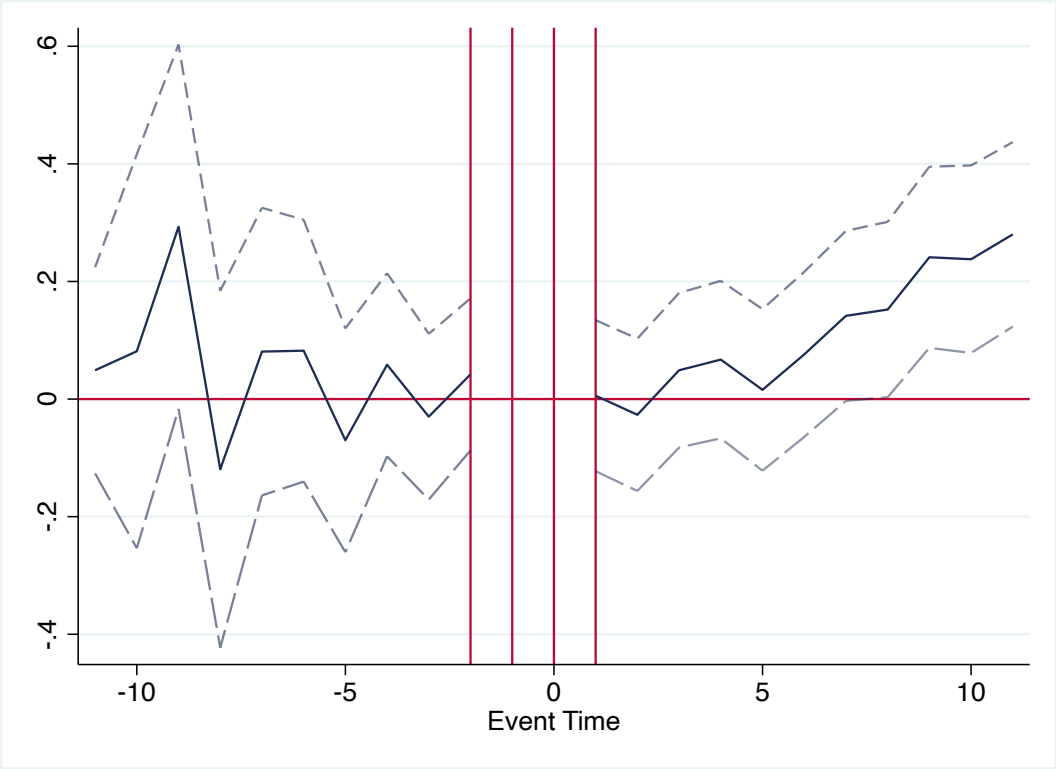
Non-traditional event study, which removes all treatment effects, trends, and interactions that are modeled as responding to the treatment. Period 0 indicates the treatment initiation and period -1 is the base period for the pre-treatment event study effects and the time 0 event effects; period 0 is the "base" for the 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.

A comparison of coefficients in columns 2 and 3 of Table 2 demonstrates the importance of the model misspecification due to the additional post-treatment shifts in the outcome variable. Almost all 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 this table might suggest that there is no evidence of model misspecification.

However, the event study analysis presented in Figure 7 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 all of the event-time effects jointly equal zero. Even though there are variable treatment effects across groups and units and across unit variations in the model not fitting well, the use of a simple event study is able to pick up the misspecifications after one removes the treatment effects as described in equation (5).

Figure 7
Heterogeneous Effects Varying over Time and By Group
Incorrect Regression Model



Non-traditional event study, which removes all treatment effects, trends, and interactions that are modeled as responding to the treatment. Period 0 indicates the treatment initiation and 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 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.

V. Conclusion

The event study approach described in equation (5) has the ability to capture model misspecifications in the presence of heterogeneous treatment effects. Unlike the imposed, single homogeneous effect analysis displayed in Figures 1 and 2, 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 or a small number of treatment effects, the standard errors obtained by using versions of equation (3) provide the correct information for assessing how treatment effects might vary post-treatment. 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. All of these features will contribute to a more precise understanding of the impacts of policy changes on outcomes and behaviors.

References

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.

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.

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.