Difference-in-Differences with a Continuous Treatment

Brantly Callaway[*]   Andrew Goodman-Bacon[†]   Pedro H.C. Sant'Anna[‡]

July 13, 2021

Abstract

This paper analyzes difference-in-differences setups with a continuous treatment. We show that treatment effect on the treated-type parameters can be identified under a generalized parallel trends assumption that is similar to the binary treatment setup. However, interpreting differences in these parameters across different values of the treatment can be particularly challenging due to treatment effect heterogeneity. We discuss alternative, typically stronger, assumptions that alleviate these challenges. We also provide a variety of treatment effect decomposition results, highlighting that parameters associated with popular two-way fixed-effect specifications can be hard to interpret, *even* when there are only two time periods. We introduce alternative estimation strategies that do not suffer from these drawbacks. Our results also cover cases where (i) there is no available untreated comparison group and (ii) there are multiple periods and variation in treatment timing, which are both common in empirical work.

---

[*]University of Georgia. Email: brantly.callaway@uga.edu

[†]Federal Reserve Bank of Minneapolis and NBER. Email: andrew.j.goodman-bacon@vanderbilt.edu

[‡]Microsoft and Vanderbilt University. Email: pedro.h.santanna@vanderbilt.edu

1 Introduction

The canonical difference-in-differences (DiD) research design compares outcomes between a treatment and comparison group (difference one) before and after that treatment begins (difference two). But many DiD applications study treatments that do not simply turn "on", they have a "dose" or operate with varying intensity. Pollution dissipates across space, affecting locations near its source more severely than locations far away. Localities spend different amounts on public goods and services, or set different minimum wages. Students choose how long to stay in school.

A continuous treatment may offer practical advantages over a binary treatment. Variation in treatment intensity makes it possible to evaluate treatments that lack untreated comparison units because all units are treated to some extent. A clear "dose-response" relationship between outcomes and treatment intensity can bolster the case for a causal interpretation. In his 1965 presidential address to the Royal Society of Medicine, Sir Austin Bradford Hill, a pioneer in the study of smoking and cancer, included among his criteria for inferring causality from observational data, "a biological gradient, or dose-response curve" and argued that "we should look most carefully for such evidence" (Hill, 1965). Finally, we may care more about the effect of changes in treatment intensity (e.g., increased funding, pollution abatement, or expanded eligibility) than about the effect of the existence of a program that already exists.

Despite how conceptually useful and practically common it is to have continuous treatments, relatively little theoretical research focuses on identification and interpretation of dose-response DiD designs. This paper analyzes DiD models in which units move from no treatment to some non-zero dose, and outlines potential pitfalls of two-way fixed effects (TWFE) regression estimators. Our first contribution is to provide a clear bridge between the parameters of interest, identifying assumptions and interpretation of canonical binary DiD and dose-response DiD models, highlighting when and why additional assumptions are required. Second, we use these results to evaluate TWFE estimators in the dose-response context.

We build on two types of causal effects that arise in a non-binary DiD setting corresponding to levels and slopes of the dose-response relationship. The level effect is the *treatment effect* of "dose" d, which equals the difference between a unit's potential outcome under treatment d and its untreated potential outcome. Following Angrist and Imbens (1995), the slope effect is the *causal response* to an incremental change in the "dose" at d. The distinction between treatment effects and causal responses is crucial because they can have meaningfully different interpretations and has no analogy in binary DiD (in this case the treatment effect and causal response are the same). One theme of our paper is that different causal parameters require different assumptions. Thus, researchers attempting to answer causal questions should pay particular attention to whether the given estimation strategy is actually suitable for the given application.

We begin with a simple setup with exactly two time periods, where no units receive the treatment in the pre-period, and where some units become treated with different doses in the post-period. The average treatment effect of dose d among units that receive dose d, $ATT(d|d)$, is nonparametrically

identified under a parallel trends assumption on untreated potential outcomes only.[1] This is a simple extension of binary DiD: one simply compares outcome changes among units that experienced a particular dose to outcome changes for the untreated units.

Average causal response parameters, however, are not identified under parallel trends involving untreated outcomes alone because they must come from comparisons between higher- and lower-dose units. These comparisons mix together (i) the average causal response on the treated of dose d for units who receive dose d, $ACRT(d|d)$ and (ii) differences in the treatment effects between the two groups at the lower dose ("selection bias"). Even if lower-dose units have the same evolution of untreated potential outcomes as higher-dose units (parallel trends), they are only a good counterfactual for higher-dose units if the evolution of outcomes *at the lower dose* would have been the same, too.[2] We propose an alternative "strong" parallel trends assumption that imposes these restrictions. Under strong parallel trends, we show that comparisons of causal effect parameters across different values of the doses can themselves be interpreted as causal responses. However, strong parallel trends is likely to be a much stronger assumption than the sort of parallel trends that is frequently invoked in DiD applications with a binary treatment. Moreover, pre-trend tests commonly used to detect violations of parallel trends cannot distinguish between "standard" and "strong" parallel trends.

We use the identification results to evaluate the most common estimator for dose-response designs, a TWFE regression that includes time fixed effects (θ_t), unit fixed effects (η_i), and the interaction of a dummy for the post-treatment period ($Post_t$) with a variable that measures unit i's dose or treatment intensity, D_i:

$$Y_{it} = \theta_t + \eta_i + \beta^{twfe} \cdot D_i \cdot Post_t + v_{it} \tag{1}$$

We show that, under a parallel trends assumption on untreated potential outcomes, β^{twfe} can be decomposed into different weighted sums of different treatment effect parameters, highlighting the importance of appropriately (a priori) choosing the "building block" of the analysis. Interestingly, if one adopts the (rescaled) average treatment effect of dosage d as the building block of the decomposition, β^{twfe} is equal to a weighted sum of the (rescaled) $ATT(d|d)$, though the weights are non-convex and can be negative; see also de Chaisemartin and D'Haultfœuille (2020). If one adopts the average causal response of dosage d as the building block of the decomposition, though, β^{twfe} is equal to a weighted average of the $ACRT(d|d)$'s, where all the weights are guaranteed to

---

[1]Note that we do not refer to "the" ATT of dose d because there are many different treatment groups, each of which has an average treatment effect at dose d in theory. We use the notation $ATT(a|b)$ to denote the average effect of administering dose a to the group that actually received dose b. This is similar to Callaway and Sant'Anna (2020) who study staggered DiD designs and define $ATT(g,t)$ as the average treatment effect for timing group g at time t.

[2]To give an example, suppose that for doses $b > a$, higher-dose units have a larger average treatment effect than lower-dose units: $ATT(b|b) > ATT(a|a)$. This does not imply that the effect of experiencing dose b is greater than the effect of experiencing dose a on average for all units or even for any subpopulation. This is an important limitation of dose-response DiD designs and occurs because standard parallel trends assumptions are not strong enough to distinguish between cases where receiving more treatment generally increases outcomes *or* where differences in $ATT(b|b)$ and $ATT(a|a)$ are due to other differences between the units that experience treatment amount b relative to the units that experience treatment amount a — or due to combinations of these two possibilities.

be non-negative, plus another positively-weighted average of "selection bias" terms that come from heterogeneous treatment effect functions across dose groups. Even under strong parallel trends, which eliminates "selection bias", TWFE estimates do not equal an average of ACR parameters weighted by the population distribution of the dose (which is arguably the natural target parameter in this case). Instead, the TWFE weighting scheme typically puts more weight on doses near the average and less on the tails. If the highest or lowest doses create relatively extreme causal responses (for example if the effect is zero below a certain dose), then TWFE estimates can be a misleading summary of the overall average causal response.

We extend these baseline results to a setup with more than two time periods and where treatment varies in intensity as well as timing. In this case, the TWFE estimator is composed of (i) comparisons of the path of outcomes for units treated at the same time but with different doses, (ii) comparisons of paths of outcomes in early-treated groups relative to later-treated groups in periods before the later-treated groups become treated, (iii) comparisons of paths of outcomes of later-treated groups relative to already-treated groups, and (iv) comparisons of paths of outcomes between early-treated groups and later-treated groups in their common pre-treatment and post-treatment periods. Under a version of strong parallel trends, the first two sets of comparisons turn out to be equal to averages of causal response parameters, but the third and fourth set of comparisons contain extra terms that arise in the presence of treatment effect dynamics or heterogeneous causal responses across groups. These results on TWFE regressions with multiple periods and variation in treatment timing generalize the results in de Chaisemartin and D'Haultfœuille (2020) and Goodman-Bacon (2021) to the case with a continuous treatment.

In this context, we propose an alternative strategy that delivers overall average causal response parameters under strong parallel trends as well as event-study type parameters that do not require ruling out treatment effect dynamics or heterogeneous causal effects across groups. This strategy expands the group-time average treatment effects approach in Callaway and Sant'Anna (2020) to the case where the treatment can be continuous.

Related literature: Our paper build on different strands of the literature. First, the interpretation issues that we point out regarding comparisons of $ATT(d|d)$ at different values of d are related to existing points made on comparing "local" treatment effect parameters to each other (e.g., (Oreopoulos, 2006; Angrist and Fernandez-Val, 2013; Mogstad, Santos, and Torgovitsky, 2018) in the context of local average treatment effects or (Frölich, 2004; Fricke, 2017) in the context of difference-in-differences with multiple treatments). Second, our paper is broadly related to the literature on continuous treatments in cross-sectional setups, see, e.g., Hirano and Imbens (2004), Flores (2007), Flores, Flores-Lagunes, Gonzalez, and Neumann (2012), Galvao and Wang (2015), Kennedy, Ma, McHugh, and Small (2017), Su, Ura, and Zhang (2019), and Callaway and Huang (2020) in the context of an unconfounded treatment; and Florens, Heckman, Meghir, and Vytlacil (2008) in the context of an endogenous treatment. Our paper is also related to D'Haultfoeuille, Hoderlein, and Sasaki (2021), as they considered setups with repeated cross-sectional data. Our paper complements theirs as our setups greatly differ: we consider identifying assumptions based

on parallel trends while they rely on rank-invariance assumptions on the time trend as in Athey and Imbens (2006). Besides not being in the same DiD context as our paper, most of these aforementioned papers identify ATE-type parameters rather than ATT-type parameters, which implies that they do not run into the same interpretation issues that we consider here.

We also contribute to the more recent literature highlighting potential problems with TWFE specifications when one deviates from the canonical two-groups two-periods DiD setup, such as such as "fuzzy" designs (de Chaisemartin and D'Haultfœuille, 2018) or staggered designs (Borusyak and Jaravel, 2017; Goodman-Bacon, 2021; de Chaisemartin and D'Haultfœuille, 2020; Sun and Abraham, 2020).[3] Our results contribute to this literature by showing that, in setups with variation in treatment intensity, TWFE regressions do not recover easy-to-interpret causal parameters *even* in the two-period case. With variation in treatment timing, the problems are exacerbated as (i) TWFE is not robust against treatment effect dynamics (as already-treated units serve as the comparison group for late-treated units, just like the binary DiD case), and (ii) TWFE regressions are confounded in the presence of heterogeneous causal responses across groups, which is not an issue in the case with a binary treatment. We note that the supplemental appendix of de Chaisemartin and D'Haultfœuille (2020) briefly discuss some potential problems interpreting TWFE coefficients in the presence of a multi-valued discrete treatment. Our results complement theirs as we consider continuous treatments and provide different decompositions that rely on different building blocks than theirs. Interestingly, we show that the problem of "negative weighting" is not present when one uses average causal responses as building blocks of the decomposition in the two-period setup.

2 The Sources and Uses of Continuous/Multi-valued Treatments

The development of causal inference methods has focused, often for expository reasons, on binary (discrete) treatment variables. The history of DiD, dating to John Snow's 1855 analysis, also builds almost entirely on cases in which some units become treated and others do not. The focus on binary treatments, in addition to simplifying theoretical analyses, cements the analogy between quasi-experiments and simple randomized controlled trials.

But continuous treatment variables frequently arise and can provide much richer information— sometimes fundamentally different information—about treatment effects than a binary variable can. Price and income elasticities for example determine optimal policies like tax rates, tax bases, subsidies, and regulations in many economic models (Hendren, 2016), but they are continuous concepts that can only be estimated accurately with continuous variation.

Establishing a "dose-response" relationship between gradations of exposure and outcomes can also help support a causal interpretation and point to potential mechanisms. Meyer (1995), for example, argues that "differences in the intensity of the treatment across different groups allow one to examine if the changes in outcomes differ across treatment levels in the expected direction" (pg.

---

[3]de Chaisemartin and D'Haultfœuille (2020) also consider non-staggered designs, where treatment can "turn on" and "turn off".

158).[4] While dose-response effects can act as a kind of falsification exercise in some contexts, in others a non-linear or non-monotonic effect of treatment intensity can shed light on mechanisms. In a monopsonistic labor market, for example, a minimum wage can either raise or lower employment depending on whether it is set below or above the competitive wage. Aggregate outcomes may improve as a treatment expands to people who need it most then worsen if that treatment is forced on people who are hurt by it (Heckman and Vytlacil (2005)).

Variation in the dose or treatment intensity also makes it possible to evaluate treatments for which binary DiD is not feasible. Card (1992) exploits geographic differences in the "bite" of a 1991 federal minimum wage increase. In a statutory sense, the federal change affected all workers, so there is no untreated group to use in a binary DiD. While the minimum wage increase should have affected lower-wage workers more directly than higher-wage workers, comparing these groups in a binary DiD would require longitudinal data. Instead, Card regresses the change in each state's teen employment rate on the share of teens in that state who earned less than the new minimum wage in the pre-period and are thus "eligible" for a statutory wage increase. Converting the analysis to the state-level DiD creates a continuous treatment variable that can identify effects of a federal policy.

In practice, DiD designs with a continuous treatment are almost always estimated with a TWFE regression. Applications typically justify dose-response DiD designs by extending the logic of canonical binary DiD models, arguing that, under a parallel trends assumption on untreated potential outcomes, regression estimators that compare units treated with different doses identify the average treatment effect per unit of the dose. Wooldridge (2005) observes that a two-period DiD regression estimator "can be easily modified to allow for continuous, or at least non-binary, 'treatments' " (pg. 132). Angrist and Pischke (2008) emphasize "a second advantage of regression DD is that it facilitates the study of policies other than those that can be described by a dummy...the minimum wage is therefore a variable with differing treatment intensity across states and over time" (p. 234). The reason TWFE is so ubiquitous is that it makes it easy to "extend" regression estimators for binary DiD models, whose properties are better understood, to the more complex case with variation in treatment intensity. Yet details about what treatment effect parameter these specifications recover and under what assumptions remain unknown. We now turn to our analysis of these two central issues.

---

[4]Hill (1965) makes this point in the context of smoking and cancer:

> The fact that the death rate from cancer of the lung rises linearly with the number of cigarettes smoked daily, adds a very great deal to the simpler evidence that cigarette smokers have a higher death rate than non-smokers.

He also notes that more deaths among light rather than heavy smokers would weaken the causal claim unless one could "envisage some much more complex relationship to satisfy the cause-and-effect hypothesis." Toxicologists refer to non-monotonicity as "hormesis" and questions about whether observed non-monotonic relationship between an exposure and outcomes really represent causal effects versus offsetting confounders that dominate at low versus high doses are ongoing for factors like radioactivity (Thayer et al. (2005)).

3 Baseline Case: A New Treatment with Two Periods

We first consider a simple case that matches the canonical DiD setup except that we allow for a multi-valued or continuous treatment rather than a binary one. Later, we extend these results to more complicated data structures that are commonly encountered in empirical work.

3.1 Notation for the Baseline Case

We suppose that a researcher has access to two periods of panel data denoted by t and $t-1$. In the first period, no unit is treated. In the second period, units receive a treatment "dose" denoted by D_i. We denote the support of D by $\mathcal{D}$. We define potential outcomes for unit i in period $s \in \{t-1, t\}$ by $Y_{is}(d)$. This is the outcome that unit i would experience in period s under dose d. We also make the following assumptions

Assumption 1 (Random Sampling). *The observed data consists of $\{Y_{it}, Y_{it-1}, D_i\}_{i=1}^n$ which is independent and identically distributed.*

Assumption 2 (Support). *The support of D, $\mathcal{D} = \{0\} \cup \mathcal{D}_+$. In addition, $\mathrm{P}(D = 0) > 0$ and $dF_D(d) > 0$ for all $d \in \mathcal{D}_+$. No units are treated in period $t-1$.*

Assumption 3 (Observed Outcomes / No Anticipation). *For all units,*

$$Y_{it-1} = Y_{it-1}(0) \quad and \quad Y_{it} = Y_{it}(D_i).$$

Assumption 1 says that we observe two periods of *iid* panel data. Assumption 2 implies that there is a group of units that do not receive any dose in any period and is general enough to allow for a binary, multi-valued discrete, or continuous treatment. For some results below, we specialize this condition to explicitly make the dose continuous but most of our identification will hold under very general treatment regimes. Assumption 3 says that we observe untreated potential outcomes for all units in the first period and that, in the second period, we observe the potential outcome corresponding to the actual dose that unit i experienced. It rules out that units change their pre-treatment outcomes in response to the post-treatment dose.

For some of our results below, we provide specialized results to cases where the treatment is either continuous or multi-valued discrete. In these cases, we sometimes require more specialized versions of Assumption 2.

Assumption 2-Cont (Continuous Treatment). *(a) The support of D, $\mathcal{D} = \{0\} \cup \mathcal{D}_+$ where $\mathcal{D}_+ = [d_L, d_U]$ with $0 < d_L < d_U < \infty$. In addition, $\mathrm{P}(D = 0) > 0$ and $f_D(d) > 0$ for all $d \in \mathcal{D}_+$.*

(b) $\mathbb{E}[\Delta Y | D = d]$ is continuously differentiable on $\mathcal{D}_+$.

Assumption 2-MV (Multi-Valued Treatment). *The support of D, $\mathcal{D} = \{0\} \cup \mathcal{D}_+$ where $\mathcal{D}_+ = \{d_1, d_2, \ldots, d_J\}$ where $0 < d_1 < d_2 < \cdots < d_J$. In addition, $\mathrm{P}(D = d) > 0$ for all $d \in \mathcal{D}$.*

Assumption 2-Cont formalizes that the treatment consists of a mass of units that do not participate in the treatment and an otherwise continuous treatment though it allows for the smallest value of the treatment to be strictly larger than 0 which is common in applications. Assumption 2-MV is essentially analogous but for the case when the treatment is discrete.

3.2 Parameters of Interest with a Continuous or Multi-Valued Treatment

For a binary treatment, the effect of being treated at all is the same as the effect of an "increase" in treatment because both involve a change in the treatment variable from $D = 0$ to $D = 1$. But when the treatment can take more values, treated units experience different changes in the dose. Therefore, continuous/multi-valued treatments define more parameters of interest than binary ones. The treatment effect of dose level d in time period t for a given unit equals the potential outcomes when $D = d$ minus the untreated potential outcome: $Y_t(d) - Y_t(0)$.[5] We also define that unit's causal response at d as $Y'_t(d)$, the derivative of the potential outcome function,[6] (when d is continuous) or as the difference in potential outcomes between adjacent doses, $Y_t(d_j) - Y_t(d_{j-1})$ (when d is discrete). These two types of treatment effects—the *level* of $Y_t(d) - Y_t(0)$ or its *slope*, $Y'_t(d)$—define the parameters of interest in this case, and connects to well-known results in the instrumental variables (IV) literature on multi-valued or continuous endogenous variables (Angrist and Imbens, 1995, Angrist, Graddy, and Imbens, 2000).

Average Level Effects: What was the effect of dose d?

Treatment effects in levels extend the definitions of average treatment effects from the binary case so that they refer to the average effect of being treated with a particular dose. Like the case with a binary treatment, we can define treatment on the treated type parameters as well as overall treatment effect parameters. However, there are some more possibilities here. In particular, we define

$$ATT(a|b) = \mathbb{E}[Y_t(a) - Y_t(0)|D = b] \quad \text{and} \quad ATE(d) = \mathbb{E}[Y_t(d) - Y_t(0)].$$
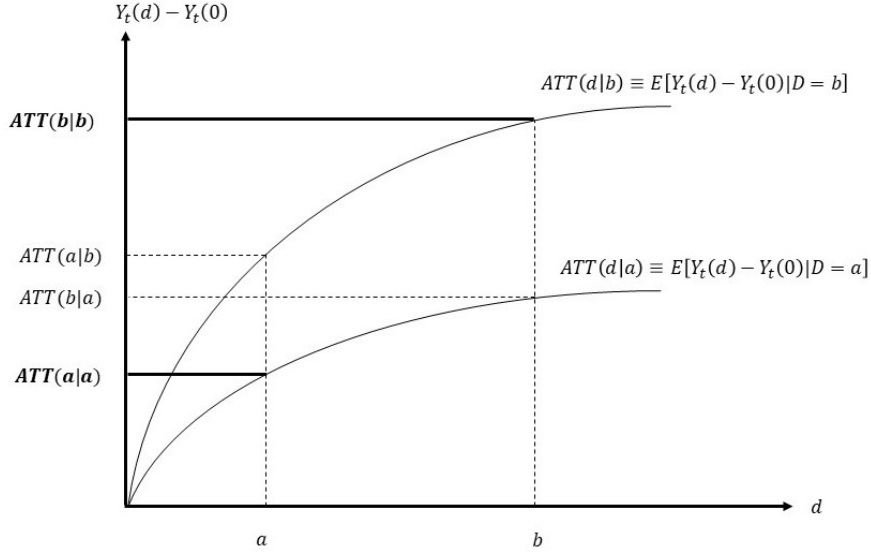
$ATT(a|b)$ is the average effect of dose a on units that actually experienced dose b. We often consider the specialized version of this parameter, $ATT(d|d)$, which is the average effect of dose d among units that actually experienced dose d. $ATE(d)$ is the mean difference between potential outcomes under dose d relative to untreated potential outcomes across all units, not just those that experienced dose d.

Figure 1 shows a graphical example of the ATT parameters with two doses, $a < b$. The definition of potential outcomes means that each unit can have its own set of treatment effects, $Y_{it}(d) - Y_{it}(0)$. The concave lines in the figure represent the average treatment effects at all doses for the groups

---

[5]We include i subscripts for units in expressions that refer to sample quantities, but not in theoretical expressions of population quantities. We also sometimes refer to $Y_t(d) - Y_t(0)$ as the treatment effect function.

[6]This is a slight abuse of notation as we do not require $Y_t(d)$ to be differentiable, but rather we mean here the effect of a marginal change in the dose on a unit's outcome: $\lim_{h \to 0^+} (Y_t(d + h) - Y_t(d))/h$.

Figure 1: Average Treatment Effects on the Treated, Two Doses

Notes: The figure plots $ATT(d|a)$ (the average effect of experiencing dose d among units that actually experienced dose a) and $ATT(d|b)$ (the average effect of experiencing dose d among units that actually experienced dose b).

actually treated with dose a or dose b. There are four potential ATT parameters in this case. Two refer to doses and groups that are actually observed — $ATT(a|a)$ and $ATT(b|b)$. As in the binary treatment case, $ATT(a|a)$ and $ATT(b|b)$ are "local" to units that experienced dose a or b. The other two refer to the effect of doses on groups that do not actually receive that particular dose — $ATT(a|b)$ and $ATT(b|a)$.

Average Slope Effects: What was the incremental effect of the dth dose unit?

With continuous treatments researchers often care about the effect of an increment in the dose. Following Angrist and Imbens (1995) we define a unit's causal response function as $Y'_t(d)$ when treatment is continuous (see Angrist, Graddy, and Imbens, 2000) and as $Y_t(d_j) - Y_t(d_{j-1})$ when treatment is discrete/multi-valued. We focus on identifying average causal response parameters. For continuous treatments they are
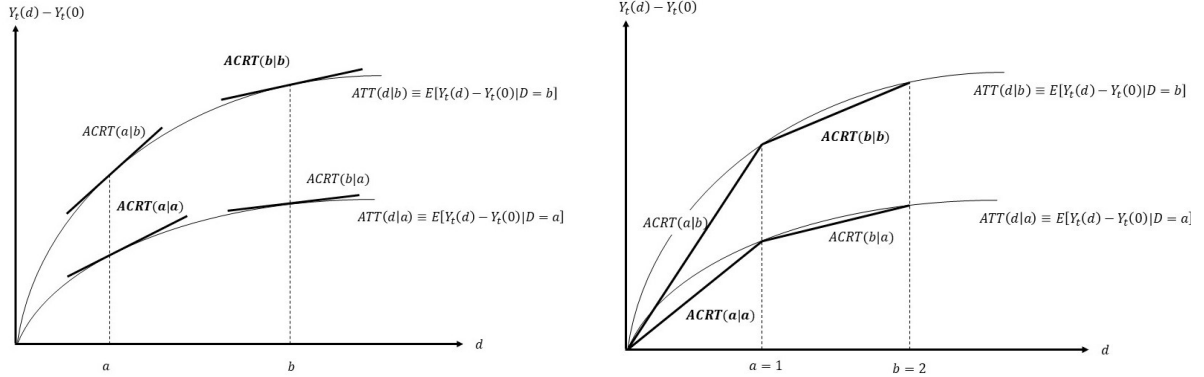
$$ACRT(d|d) = \left.\frac{\partial \mathbb{E}[Y_t(l)|D=d]}{\partial l}\right|_{l=d} \quad \text{and} \quad ACR(d) = \frac{\partial \mathbb{E}[Y_t(d)]}{\partial d}.$$

For discrete treatments the average causal response parameters are

$$ACRT(d_j|d_j) = \mathbb{E}[Y_t(d_j) - Y_t(d_{j-1})|D=d_j] \quad \text{and} \quad ACR(d_j) = \mathbb{E}[Y_t(d_j) - Y_t(d_{j-1})].$$

The average causal response on the treated, $ACRT(d|d)$, is the average difference between potential outcomes under dose d compared to potential outcomes under a marginal change in the dose for

Figure 2: Average Causal Responses, Two Doses

Notes: The figure plots $ACRT$'s as derivatives of the ATT curve (when the treatment is continuous) or as the slope of the line connecting adjacent ATT's (when the treatment is discrete).

the group of units who actually experience dose d. For multi-valued treatments $ACRT(d|d)$ equals the difference in potential outcomes between dose level d_j and the next lowest dose d_{j-1} (no matter how big the gap between d_j and d_{j-1} is).[7] $ACR(d)$ is the overall average causal response of a small change in dose; it is an average across the entire population not just units that experienced dose d.

Figure 2 depicts the causal response parameters for two doses continuing the same example from Figure 1. In panel A, we show the $ACRT$ parameters at doses a and b when d is continuous. Continuity is reflected in the fact that the slopes are tangent to the $ATT(d|d)$ functions. In panel B, we show the $ACRT$ parameters when d is discrete and when $a = 1$ and $b = 2$ so that the increment between doses equals 1. The two groups not only have different level effects, they also have different slopes at every dose: $ACRT(d|a) \leq ACRT(d|b)$. Two slopes refer to the marginal effect of doses actually received. $ACRT(a|a)$ is the average effect of the ath dose unit for group a, while $ACRT(b|b)$ is the effect of the bth dose unit for group b. The other two slopes are also average causal responses but for the dose not actually experienced by each group. $ACRT(a|b)$ is the average effect of the ath dose unit for group b who did not actually receive that much dose. $ACRT(b|a)$ is the effect of the bth dose unit for group a who actually received more than that dose. Since the average causal response is higher for group b than for group a at every dose, $ACRT(a|a) < ACR(a)$ and $ACRT(b|b) > ACR(b)$.

Another thing worth pointing out is that the shape of the average treatment effect function, $ATT(d|d)$, drives the interpretation of the slope parameters. All $ATT(d|d)$ parameters may be positive while some $ACRT(d|d)$ parameters are zero or negative. For example, some treatment may be better than none, but additional doses do not always raise outcomes. This underscores why it is important to be clear about an ATT versus an $ACRT$ interpretation in the dose-response context. Besides that, comparisons between units treated with different doses may not allow inference about

---

[7]Differences in $ATT(d|d)$ between doses that are farther apart than, say, one unit in the discrete case or differ by a finite amount in the continuous case equal averages of the $ACRT$ between the doses in question.

the ATT for any group because neither group can be used to estimate an untreated counterfactual. But these comparisons can, under conditions stated below, be informative about causal response parameters that describe how outcomes respond to small changes in the dose.

Sometimes it is worth considering average derivative versions of $ACRT$ and ACR that average the average causal responses across doses. These are given by

$$ACRT^* = \mathbb{E}[ACRT(D|D)|D > 0] \qquad \text{and} \qquad ACR^* = \mathbb{E}[ACR(D)|D > 0].$$

$ACRT^*$ and ACR^* are natural single (i.e., non-functional) target parameters for understanding the causal response of the treatment. These parameters do not show up in the case with a binary treatment because D only takes one non-zero value.

3.3 Identification with a Continuous or Multi-Valued Treatment in the Baseline Case

This section discusses identification of level effects ($ATT(d|d)$ and $ATE(d)$) and slope effects ($ACRT(d|d)$ and $ACR(d)$) under parallel trends assumptions.

Levels: Identification of Average Treatment Effects

We begin by discussing identification of the $ATT(d|d)$ because the approach and assumptions follow closely from the traditional binary treatment case. We make the following assumption.

Assumption 4 (Parallel Trends). *For all $d \in \mathcal{D}$,*

$$\mathbb{E}[Y_t(0) - Y_{t-1}(0)|D = d] = \mathbb{E}[Y_t(0) - Y_{t-1}(0)|D = 0].$$

Assumption 4 is very similar to parallel trends assumptions in binary DiD. First, it only involves the path of untreated potential outcomes. Second, it says that the path of outcomes that units with any dose d would have experienced if they had not participated in the treatment is the same as the path of outcomes that units in the untreated group actually experienced — this is very similar to the intuition for binary parallel trends except that now d can take on many values instead of just being treated or untreated.

The following result shows that under Assumption 4, $ATT(d|d)$ is identified.

Theorem 1. *Under Assumptions 1 to 4, $ATT(d|d)$ is identified for all $d \in \mathcal{D}$, and it is given by*

$$ATT(d|d) = \mathbb{E}[\Delta Y_t|D = d] - \mathbb{E}[\Delta Y_t|D = 0].$$

The proof of Theorem 1 is provided in Appendix D, but it is worth pointing out that Theorem 1 holds using essentially the same arguments as for the case with a binary treatment. $ATT(d|d)$ equals the difference between the change in outcomes for units treated with dose d and untreated units

because Assumption 4 ensures that the path of outcomes for untreated units is the same as the path of outcomes that treated units would have experienced absent the treatment.

The next result shows that, in general, $ATE(d)$ is *not* identified under Assumptions 1 to 4 alone.

Proposition 1. *Under Assumptions 1 to 4, $ATE(d)$ is not identified*

The proof of Proposition 1 is provided in Appendix D. Intuitively, the result holds by first noticing that $ATE(d) = \mathbb{E}[ATT(d|D)]$. In other words, $ATE(d)$ can be obtained by averaging $ATT(d|k)$ across all values of k — where $ATT(d|k)$ is the average effect of experiencing dose d among units that actually experienced dose k. However, for $k \neq d$, $ATT(d|k)$ is not generally identified under Assumption 4, so $ATE(d)$ is not point identified under Assumption 4 either. Additionally, the result in Proposition 1 is not surprising as, even in the binary treatment case, ATE is not identified under the analogous parallel trends assumption.

Finally in this section, we impose a parallel trends assumption that is strong enough to identify $ATE(d)$.

Assumption 5 (Strong Parallel Trends). *For all $d \in \mathcal{D}$,*

$$\mathbb{E}[Y_t(d) - Y_{t-1}(0)] = \mathbb{E}[Y_t(d) - Y_{t-1}(0)|D = d].$$

Assumption 5 says that, for all doses, the average change in outcomes over time across all units if they had been assigned that amount of dose is the same as the average change in outcomes over time for all units that experienced that dose. Assumption 5 is notably different from Assumption 4 especially because it involves potential outcomes under different doses d rather than only untreated potential outcomes. The usefulness of this assumption arises because the term on the left hand side of the equation in Assumption 5 is not identified (and, it will turn out that identifying this term is a key ingredient to identifying $ATE(d)$) while the term on the right hand side is the observed change in outcomes over time for units that experienced dose d.

It turns out that Assumption 5 is not strictly stronger than Assumption 4; however, it is likely to be stronger (and potentially much stronger) in most applications. Assumption 5 is also related to, but weaker than assuming that *all* dose groups would have experienced the same path of outcomes had they been assigned the same dose (which would rule out any selection into a particular dose) or that $ATE(d) = ATT(d|d)$ (which is a kind of treatment effect homogeneity condition). Compared to this, Assumption 5 allows for some selection into a particular dose but requires that, on average across all doses, there is no selection into a particular dose. We discuss these points more precisely in Appendix A. The next result shows that $ATE(d)$ is identified under Assumption 5.

Theorem 2. *Under Assumptions 1 to 3 and 5 and for all $d \in \mathcal{D}$, $ATE(d)$ is identified, and it is given by*

$$ATE(d) = \mathbb{E}[\Delta Y_t | D = d] - \mathbb{E}[\Delta Y_t | D = 0].$$

The proof of Theorem 2 is provided in Appendix D. Interestingly, the estimands for $ATT(d|d)$ (under Assumption 4) and $ATE(d)$ (under Assumption 5) are identical. The stricter assumptions in this case only change the interpretation of the estimand, not its form. Another important implication of the estimands in Theorems 1 and 2 being the same is that conventional pre-tests (in this case, testing that $\mathbb{E}[\Delta Y_s|D = d] - \mathbb{E}[\Delta Y_s|D = 0] = 0$ for all $d \in \mathcal{D}_+$ where s is some pre-treatment time period) cannot distinguish between pre-treatment versions of Assumptions 4 and 5.

Another difference between the case with a binary treatment and when the treatment is continuous (or multi-valued) is that, in the latter case, researchers are often interested in comparing the effect of different doses on the outcome. Next, we provide a result for understanding comparisons of $ATT(d|d)$ and $ATE(d)$ across different values of the dose.

Proposition 2. *Under Assumptions 1 to 3 and for two different doses, $(d, d') \in \mathcal{D} \times \mathcal{D}$,*

(1) If Assumption 4 holds, then

$$ATT(d|d) - ATT(d'|d') = \underbrace{ATT(d|d) - ATT(d'|d)}_{(A)} + \underbrace{ATT(d'|d) - ATT(d'|d')}_{\text{"selection bias"}}.$$

(2) If Assumption 5 holds, then

$$ATE(d) - ATE(d') = \mathbb{E}[Y_t(d) - Y_t(d')].$$

Proposition 2 is a key result in the paper. Part (1) says that, although a standard parallel trends assumption is enough to identify $ATT(d|d)$, it is not strong enough to justify comparisons of $ATT(d|d)$ across different values of d as being causal effects of different amounts of the dose. Rather, under Assumption 4, comparisons of $ATT(d|d)$ across different values of d involve two terms. Term (A) is the causal effect of dose d relative to an alternative dose d' among units that experienced dose d. The second term is a "selection bias" term that comes from the difference between the effect of dose d' on outcomes among units that experienced dose d relative to the same effect among units that experienced dose d'. These are not restricted to be the same under Assumption 4. It is also interesting to relate the "selection bias" term to paths of outcomes over time. Notice that, under Assumption 4,

$$ATT(d'|d) - ATT(d'|d') = \mathbb{E}[Y_t(d') - Y_{t-1}(0)|D = d] - \mathbb{E}[Y_t(d') - Y_{t-1}(0)|D = d'].$$

The second term here, $\mathbb{E}[Y_t(d') - Y_{t-1}(0)|D = d']$, is the observed average path of outcomes over time among units with dose d'. The first term, $\mathbb{E}[Y_t(d') - Y_{t-1}(0)|D = d]$ is the average path of outcomes that units that actually experienced dose d would have experienced if they had instead experienced dose d'. This path is not observed, but, more importantly, Assumption 4 does not put restrictions on this path of outcomes (this holds because Assumption 4 only restricts untreated potential outcomes).

On the other hand, the second part of Proposition 2 shows that comparisons of $ATE(d)$ across different doses do yield differences in average causal effects due to differences in the the dose.

Thus, overall, Proposition 2 shows that, in DiD designs with a continuous/multi-valued treatment, interpreting differences in treatment effect parameters across different values of the dose is likely to require (substantially) stronger assumptions than researchers typically think they are imposing. The intuitive explanation for this result is that standard DiD assumptions on paths of untreated potential outcomes identify "local" effects of participating in the treatment relative to not participating in the treatment. However, comparing local treatment effect parameters is not so simple and generally requires imposing stronger assumptions in order to justify interpreting differences in local treatment effects themselves as causal effects.

Slopes: Identification of Average Causal Responses

Identifying ACR parameters is different from identifying ATT parameters because it additionally requires comparing units treated with slightly different doses. This section presents our main results on the difficulties in moving from an interpretation of level effects, $ATT(d|d)$, to slope effects, $ACRT(d|d)$, in a DiD framework with a multi-valued/continuous treatment. In the previous section, we showed that, depending on the particular parallel trends assumption invoked by the researcher, both $ATT(d|d)$ and $ATE(d)$ were equal to $\mathbb{E}[\Delta Y_t|D = d] - \mathbb{E}[\Delta Y_t|D = 0]$. This suggests

$$\frac{\partial \mathbb{E}[\Delta Y_t|D = d]}{\partial d} \qquad \text{or} \qquad \mathbb{E}[\Delta Y_t|D = d_j] - \mathbb{E}[\Delta Y_t|D = d_{j-1}].$$

as reasonable candidates for interpreting as average causal response parameters in the case where the treatment is continuous or multi-valued discrete, respectively.[8] The following proposition relates these expressions to ACRT and ACR.

Proposition 3. *Under Assumptions 1 to 3 and for $d \in \mathcal{D}_+$ (when d is continuous) or $d_j \in \mathcal{D}_+$ (when d is discrete),*

(a) If Assumption 4 holds, then

$$\frac{\partial \mathbb{E}[\Delta Y_t|D = d]}{\partial d} = ACRT(d|d) + \underbrace{\frac{\partial ATT(d|l)}{\partial l}\Big|_{l=d}}_{\text{"selection bias"}}, \qquad \text{(a-Cont)}$$

$$\mathbb{E}[\Delta Y_t|D = d_j] - \mathbb{E}[\Delta Y_t|D = d_{j-1}] = ACRT(d_j|d_j) + \underbrace{ATT(d_{j-1}|d_j) - ATT(d_{j-1}|d_{j-1})}_{\text{"selection bias"}},$$
$$\text{(a-MV)}$$

where Equation (a-Cont) holds when the dose is continuous and Equation (a-MV) holds when the dose is discrete.

---

[8]Notice that $\mathbb{E}[\Delta Y_t|D = 0]$ is common across doses and cancels when taking differences across doses.

(b) If Assumption 5 holds, then

$$\frac{\partial \mathbb{E}[\Delta Y_t | D = d]}{\partial d} = ACR(d), \qquad \text{(b-Cont)}$$

$$\mathbb{E}[\Delta Y_t | D = d_j] - \mathbb{E}[\Delta Y_t | D = d_{j-1}] = ACR(d_j), \qquad \text{(b-MV)}$$

where Equation (b-Cont) holds when the dose is continuous and Equation (b-MV) holds when the dose is discrete.

Proposition 3 is an immediate implication of Proposition 2 when applied to neighboring doses. It says that under a standard version of parallel trends (as in Assumption 4), local comparisons of paths of outcomes mix together (i) $ACRT(d|d)$ and (ii) a "selection bias" type of term that comes from differences in treatment effects across groups. Intuitively, in the case with a discrete treatment, the difference between the change in outcomes among units that experience dose d_j relative to those that experience dose d_{j-1} comes both from the fact that $d_j > d_{j-1}$, a causal response, and from differences in the groups' treatment effect of the same dose, d_{j-1}, "selection bias". To give an example, if units know their own hump-shaped treatment effects and can costlessly choose their dose, the first-order conditions for the unconstrained maximization problem is to choose d such that $Y'(d) = 0$. No unit benefits from marginal changes in their dose but $ATT(d|d)$ may vary across d because of treatment effect heterogeneity. Therefore derivatives of $\mathbb{E}[\Delta Y_t | D = d]$ may be non-zero even though $ACRT(d|d)$ is zero for all units. In practice, this suggests that it is hard to tell the difference between true causal responses to the treatment relative to "selection bias".
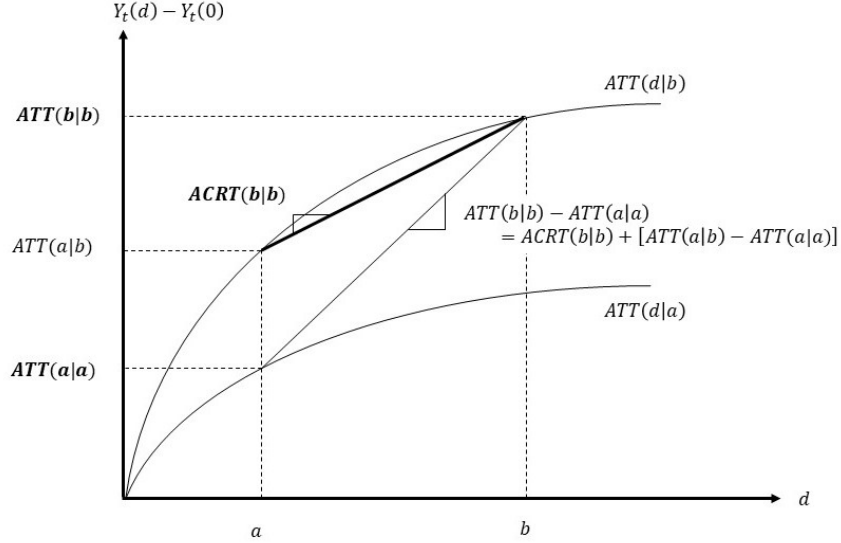
On the other hand, strong parallel trends implies that local comparisons of paths of outcomes deliver $ACR(d)$.

Figure 3 illustrates the result in Proposition 3 in the two-dose example. The slope of the line that connects the points $(a, ATT(a|a))$ and $(a + 1, ATT(a + 1|a + 1))$ is steeper than the average causal response of interest, $ACR(a+1|a+1)$ because it jumps from one ATT function to the other.[9] This is captured by the "selection bias" term, which equals the difference in treatment effects at the lower dose: $ATT(a|a + 1) - ATT(a|a)$. "Selection bias" is the same as selection-on-gains, but unlike in binary models where selection-on-gains alters the interpretation of the causal estimand, here it breaks the causal interpretation. The "selection bias" is not identified because we do not observed $Y_t(a)$ for units that experienced dose $a + 1$.

Remark 1. *An interesting intermediate assumption between Assumption 4 and Assumption 5 would be to directly assume that the "selection bias" term in Proposition 3 (i.e., $\partial ATT(d|l)/\partial l|_{l=d}$) is equal to 0. This would imply that $ACRT(d|d)$ is identified. Another interesting intermediate assumption is that for $d \in \mathcal{D}_s$ where $\mathcal{D}_s \subset \mathcal{D}_+$, $\mathbb{E}[Y_t(d) - Y_{t-1}(0)|D \in \mathcal{D}_s] = \mathbb{E}[Y_t(d) - Y_{t-1}(0)|D = d]$. This would imply that one could identify parameters such as $\mathbb{E}[Y_t(d) - Y_t(0)|D \in \mathcal{D}_s]$ for $d \in \mathcal{D}_s$ (as well as its derivative). These types of assumptions might be appealing in applications where there is substantial variation in the dose, and the researcher is willing to assume that there is no "selection*

---

[9]Because we are considering one unit increments, the "bias" can be seen on the y-axis as well. $ATT(a|a + 1) - ATT(a|a)$ is "bias" and $ATT(a + 1|a + 1) - ATT(a|a + 1)$ is the $ACRT(a + 1|a + 1)$.

Figure 3: Non-Identification of Average Causal Response with Treatment Effect Heterogeneity, Two Discrete Doses

Notes: The figure shows that comparing adjacent $ATT(d|d)$ estimates equals an $ACRT$ parameter (the slope of the higher-dose group's ATT function) and "selection bias" (the difference between the two groups' ATT functions at the lower dose).

bias" among units that selected similar doses, but the researcher is unwilling to assume that there is no "selection bias" among units that select substantially different doses.

Remark 2. *Most of our results so far have been rather negative — suggesting that identifying causal effect parameters in the case with a continuous or multi-valued treatment require stronger assumption than are typically imposed in DiD setups. One interesting and positive result of Proposition 3, however, is that, for a continuous/multi-valued treatment, identifying causal response parameters (unlike identifying the treatment effect parameters in the previous section) does not necessarily require having access to a group that does not participate in the treatment.*

3.4 What Parameter Does TWFE Estimate in the Baseline Case?

In order to estimate the effect of a continuous/multi-valued treatment on some outcome, researchers commonly estimate the two-way fixed effects regression in Equation (1). In our baseline case the TWFE estimator is equivalent to the coefficient from a univariate regression of the change in outcomes, $\Delta Y_i = Y_{it} - Y_{it-1}$, on D_i. In this section, we consider how to interpret this sort of regression in the context of DiD with a continuous/multi-valued treatment. To start with, we provide a result relating the TWFE regression to derivatives of $\mathbb{E}[\Delta Y_t | D = d]$ in the case when these derivatives may vary across d.

Proposition 4. *Consider β^{twfe} in Equation (1) and suppose that Assumption 1 holds.*

(1) If Assumption 2-Cont also holds, then

$$\beta^{twfe} = \int_{d_L}^{d_U} w_1(l)\frac{\partial \mathbb{E}[\Delta Y_t | D = l]}{\partial l}\, dl + w_0 \frac{\mathbb{E}[\Delta Y_t | D = d_L] - \mathbb{E}[\Delta Y_t | D = 0]}{d_L},$$

where

$$w_1(l) := \frac{(\mathbb{E}[D | D \geq l] - \mathbb{E}[D])\mathrm{P}(D \geq l)}{\mathrm{var}(D)} \qquad and \qquad w_0 := \frac{(\mathbb{E}[D | D > 0] - \mathbb{E}[D])\mathrm{P}(D > 0)d_L}{\mathrm{var}(D)},$$

which are weights that satisfy (i) $w_1(l) \geq 0 \;\; \forall l \in \mathcal{D}$, $w_0 > 0$, and (ii) $\int_{d_L}^{d_U} w_1(l)\, dl + w_0 = 1$.

(2) If Assumption 2-MV also holds, then

$$\beta^{twfe} = \sum_{d_j \in \mathcal{D}_+} w_1(d_j)\frac{\mathbb{E}[\Delta Y_t | D = d_j] - \mathbb{E}[\Delta Y_t | D = d_{j-1}]}{d_j - d_{j-1}},$$

where w_1 are the same weights defined in Part (1) which satisfy (i) $w_1(d_j) \geq 0 \;\; \forall \; d_j \in \mathcal{D}_+$ and (ii) $\sum_{d_j \in \mathcal{D}_+} w_1(d_j) = 1$.

Proposition 4 shows that β^{twfe} can be decomposed into a weighted average of derivatives of the conditional expectation function. All of the weights are positive and they integrate (or sum in the discrete case) to one. This result is similar to one in Yitzhaki (1996) in a different context. Proposition 4 is a mechanical decomposition of β^{twfe} in the sense that it does not invoke any of the DiD related assumptions.

This mechanical decomposition, however, just describes how β^{twfe} is computed from the data. It is not very clear on whether β^{twfe} has a causal interpretation and, if so, what causal parameter it estimates. Researchers commonly interpret β^{twfe} as the causal effect of a unit change in the dose—that is, as an average causal response. Furthermore, our identification results show that comparisons between two treated groups can include "selection bias" and need not have a causal interpretation. The rest of this section shows how to interpret the TWFE estimand under both the traditional parallel trends and strong parallel trends assumptions.

The following is a main result for interpreting TWFE regressions in this context.

Theorem 3. *Under Assumptions 1 to 3,*

(1) If, in addition, Assumption 4 holds, then

$$\beta^{twfe} = \int_{d_L}^{d_U} w_1(l)\left[ACRT(l|l) + \frac{\partial ATT(l|h)}{\partial h}\bigg|_{h=l} \right] dl + w_0 \frac{ATT(d_L|d_L)}{d_L}, \qquad \text{(1-Cont)}$$

$$\beta^{twfe} = \sum_{d_j \in \mathcal{D}_+} w_1(d_j)\left(\frac{ACRT(d_j|d_j)}{d_j - d_{j-1}} + \frac{\left(ATT(d_{j-1}|d_j) - ATT(d_{j-1}|d_{j-1})\right)}{d_j - d_{j-1}} \right), \qquad \text{(1-MV)}$$

where Equation (1-Cont) holds when the dose is continuous (i.e., under Assumption 2-Cont) and Equation (1-MV) holds when the dose is discrete (i.e., under Assumption 2-MV).

(2) If, in addition, Assumption 5 holds, then

$$\beta^{twfe} = \int_{d_L}^{d_U} w_1(l) ACR(l)\, dl + w_0 \frac{ATE(d_L)}{d_L}, \qquad \text{(2-Cont)}$$

$$\beta^{twfe} = \sum_{d_j \in \mathcal{D}_+} w_1(d_j) \frac{ACR(d_j)}{d_j - d_{j-1}}, \qquad \text{(2-MV)}$$

where Equation (2-Cont) holds when the dose is continuous (i.e., under Assumption 2-Cont) and Equation (2-MV) holds when the dose is discrete (i.e., under Assumption 2-MV).

For both Part (1) and Part (2), w_1 and w_0 are the same weights as in Proposition 4 and satisfy the same properties mentioned in that proposition.

The proof of Theorem 3 is provided in Appendix D. To gain intuition for how to interpret β^{twfe}, we begin with part (2) of Theorem 3. Under Assumption 5 there is no "selection bias" and the TWFE coefficient equals a positively weighted average of $ACR(d)$ parameters across doses. But what weighted average? First, because the TWFE weights $w_1(d)$ do not generally equal $f_{D|D>0}(d)$, the density of D conditional on $D > 0$, β^{twfe} does not equal the overall average causal response, ACR^*. If the treatment effect function is non-linear or non-monotonic so that $ACR(d)$ parameters vary widely across doses, then the TWFE estimand may differ meaningfully from ACR^*. It is therefore important to understand the TWFE weights. The TWFE weighting scheme is relatively simple to describe, however. Differentiating $w_1(d)$ shows that the weights are centered around the mean dose because $w_1'(d) = f_{D|D>0}(d)\big(\mathbb{E}[D] - d\big)$ has the sign of $\mathbb{E}[D] - d$. The weights are hump-shaped and always put the most weight on $ACR(\mathbb{E}[D])$.

That the TWFE weights are hump-shaped and centered on $\mathbb{E}[D]$ provides some intuition about what parameter TWFE estimates and how it compares to ACR^*, especially when $f_{D|D>0}(d)$ has a different shape than $w_1(d)$. If D is distributed $U(0,1)$, for example, then relative to ACR^*, the TWFE estimator puts more weight on $ACR(d)$ parameters close to the mean and less weight on $ACR(d)$s closer to 0 or 1.[10] For declining distributions like the exponential, TWFE puts less weight on the most common doses below the mean, and more weight on the rarer doses above the mean.[11] It can be the case that TWFE puts the most weight on ACR parameters for doses that are quite rare. For a bimodal distribution of D there may be very few observations with doses close to $\mathbb{E}[D]$, yet because $w_1(d)$ is always hump-shaped and centered on $\mathbb{E}[D]$ these ACR parameters will get the most weight despite potentially applying to the least common doses. If D were normally distributed,[12] we would have $w_1(d) = f_{D|D>0}(d)$ (Yitzhaki, 1996). In general, when the distribution

---

[10]For a uniformly distributed dose we have $w_1(d) = 6d(1-d)$. Therefore, the difference in weight on $ACR(d)$ across the two weighting schemes is $f_D(d) - w_1(d) = 1 - 6d(1-d)$. This function is concave up and has roots at $1/2 \pm \sqrt{3}/6$, which are about 0.21 and 0.79, so TWFE puts more weight on parameters in the middle of the distribution and less weight on the ends.

[11]For $f_D(d) = \lambda e^{-\lambda d}, \ \ d \geq 0$, we have $w_1(d) = \lambda \ell f_D(d)$. The difference between the distribution of D weights and the TWFE weights is $f_D(d) - w_1(d) = f_D(d)\lambda(\frac{1}{\lambda} - d)$. This shows that TWFE under-weights ACRs at doses below the mean $(d < \frac{1}{\lambda})$ and over-weights them at doses above the mean $(d > \frac{1}{\lambda})$.

[12]However, technically speaking, this case is ruled out in our context, as we assume that $D \geq 0$ with probability one.

of the dose is symmetric and closer to normal, TWFE weights can be close or even identical to weighting $ACR(d)$ parameters by the distribution of D. But when the distribution of the dose is skewed, TWFE weights $ACR(d)$ parameters close to the mean dose more than their population weights.

Part (1) of Theorem 3, however shows that, when we use the ACRT as the building block of the analysis, the TWFE estimator's most important potential problem is not its weighting scheme, but the fact that under parallel trends alone it includes "selection bias". The sign of this "bias" depends on how treatment effects vary across groups at a given dose; "selection on gains". The extent of "selection bias" in β^{twfe} depends on the sign, size, and weight placed on the $\left. \frac{\partial ATT(l|h)}{\partial h} \right|_{h=l}$ terms. If, for example, units with higher doses had larger positive treatment effects at every dose than units with lower doses, then TWFE estimate would be positively "biased" relative to the average of the ACRs that appear in Theorem 3. Note that "selection bias" could lead to TWFE estimates with the opposite sign as the average $ACRT(d|d)$.

In general, the sign and magnitude of "selection bias" depends on the underlying treatment effect heterogeneity that determines the strength of any selection across doses, the way that selection relates to the dose, and the TWFE weights, $w_1(d)$. The first two factors are not inherently statistical. They come from the "technology" that generates treatment effects–do ACRs vary and by how much?–and the allocation mechanism for the dose–how is the ATT function related to the observed dose? The weights determine which "selection bias" matters most and/or the extent to which any opposite signed $\left. \frac{\partial ATT(l|h)}{\partial h} \right|_{h=l}$ terms cancel. As discussed, for the TWFE estimate, "selection bias" around the average will always matter more than around extreme doses.

We conclude this section by emphasizing the importance of carefully selecting the "building block" of the analysis when decomposing the β^{twfe} estimand into causal quantities of interest. More precisely, the decomposition results in Theorem 3 presumes that researchers are potentially interested in recovering weighted averages of ACRT/ACR-type parameters. Alternatively, researchers may want to recover weighted averages of (rescaled) ATT/ATE-type parameters, which would motivate alternative decompositions for the β^{twfe} estimand, which may potentially have very different interpretations. Although we perceive the ACRT/ACR as being the most natural building blocks of the analysis in our context, we provide in Appendix B three additional decompositions for β^{twfe} in terms of other causal parameters.

First, we show that β^{twfe} can be written as a variance weighted average of all possible 2×2 comparisons among units that experience different doses. As in Theorem 3, the weights in this expression are all positive and integrate (or sum) to one. This result extends Theorem 1 in Goodman-Bacon (2021) and provides a connection between continuous and binary DiD under staggered treatment adoption.

Second, we provide an expression involving weighted averages of $ATT(d|d)/d$ (or $ATE(d)/d$ depending on which version of parallel trends is invoked). These rescaled average treatment effects are an alternative version of an average causal response in the sense that they measure the average treatment effect per dose unit of dose level d for dose group d, which is simply an average of

$ACRT(d|d)$ parameters. In the case of a discrete dose, this result is similar to the one in Theorem S3 of the Supplementary Appendix of de Chaisemartin and D'Haultfœuille (2020). In this case, it is possible for the weights at some values of d to be negative though they integrate (or sum) to one.

Finally, we show that β^{twfe} can be written as a weighted combination of $ATT(d|d)$ (or $ATE(d)$) across different values of d. In this case, the weights integrate (or sum) to 0 indicating that β^{twfe} should not be interpreted as approximating the *level* of any kind of causal effect parameter in this context.

In our view, the main take-way message from these alternative decompositions is that the same TWFE estimand β^{twfe} may have very different interpretations, depending on the underlying causal-parameter of interest. As so, researchers should decide and discuss the type of causal parameter they are interested in a given application, as the type of potential problem associated with a given estimation strategy is "question" specific. By following this route, one would be able to assess whether "negative weighting" or "non-intuitive weighting" is a main concern in DiD setups with two-time periods and variation in treatment intensity.

Remark 3. *As in Remark 2, in many applications, all units begin untreated but end up treated; there are no "never-treated" units. Proposition 4 and Theorem 3 both continue to apply in this case by noting that* $P(D = 0) = 0$.
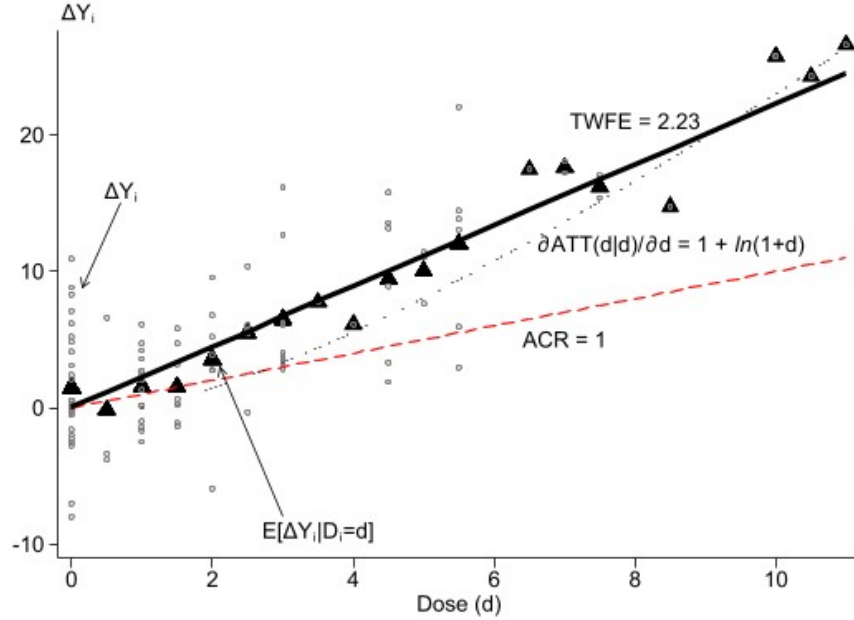
3.5 Simulated Example

Next, we illustrate the theoretical results from the previous section using simulated data. There are 100 units with doses drawn from an exponential distribution with $\lambda = 3$. We set the lowest quartile of doses to zero and, for simplicity, round remaining doses to the nearest 0.5. This matches the fact that many doses are continuous but measured in discrete intervals. We build in selection into higher doses through the following treatment effect function: $(1+d_i)ln(1+D_i)$. The multiplier $1 + d_i$ builds in positive "selection bias" since units with higher doses have higher treatment effects at all doses. The function $ln(1+D_i)$ means that all groups have a concave treatment effect function. There are two time periods and the data generating process for Y_{it} is: $Y_{it} = \alpha_{i\cdot} + Post_t/3 + (1 + d_i)ln(1 + D_i)Post_t + \epsilon_{it}$ with $\epsilon_{it} \sim N(0, 9)$.

Figure 4 plots the change in outcomes (ΔY_i in gray circles) as well as the average change in outcomes for each value of the dose ($\Delta\bar{Y}_d$ in black triangles) against the dose d_i. Figure 4 also clearly highlights the well-known fact that in this simple case, $\hat{\beta}^{twfe}$ is just the best fit line connecting ΔY_i and d_i.

Figure 5 illustrates the weighting scheme for our simulated example. It plots the empirical pdf of positive doses in the solid gray line against the TWFE weights from Theorem 3 in black dashed line. TWFE puts more weight on doses close to the average ($\mathbb{E}[D] = 2.52$) and does not have the obvious spike at low doses that $f_{D|D>0}$ has.

In our simulated example every unit has an $ACRT$ of one at its true dose because $\partial(1+d_i)ln(1+D_i)/\partial D_i = (1 + d_i)/(1 + D_i)$ which is 1 when $D_i = d_i$. But the derivative of $\mathbb{E}[\Delta Y_i|D_i = d_i]$ with respect to d_i equals $1 + ln(1 + d_i) > 1$. Therefore "bias" in the TWFE estimate can be seen clearly
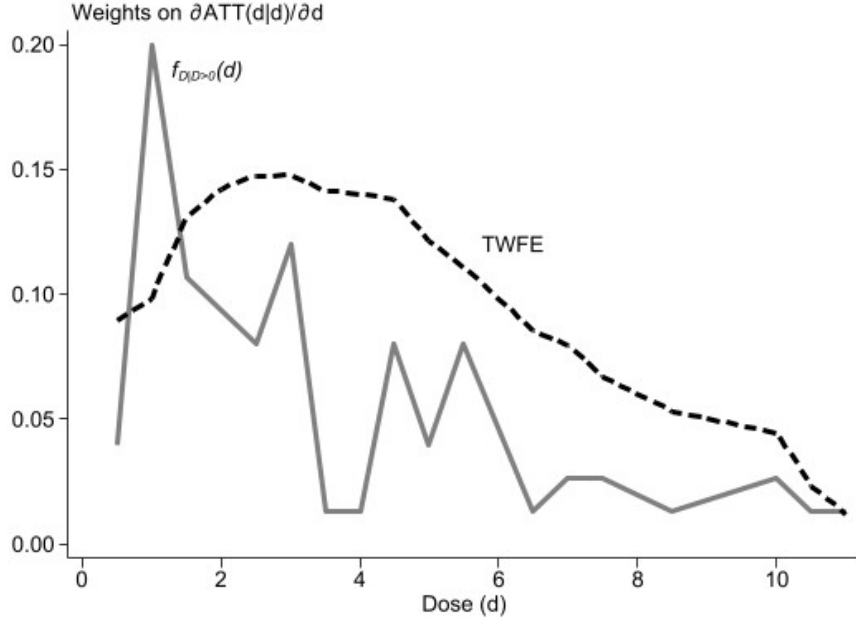
Figure 4: Baseline Case Scatter Plot: Exponential Dose and Positive "Selection Bias"

Notes: The figure plots simulated data. There are 100 units with doses drawn from an exponential distribution with $\lambda = 3$, with all draws below the 25th percentile set to zero, and remaining positive doses rounded to the nearest 0.5 (for readability). Each unit's treatment effect function is $d_i \times ln(D_i)$, where the d_i represents "selection bias" (larger treatment effects for higher dose groups) and the $ln(D_i)$ means that treatment effects are concave in the dose. This specification means that each unit's $ACRT(d_i|d_i)$ at its actual dose is 1, "selection bias" is $ln(d_i)$, so the derivative of $ATT(d_i|d_i)$ is $1 + ln(d_i)$. Outcomes are: $Y_{it} = \alpha_{i\cdot} + \alpha_{\cdot t} + d_i ln(D_i) Post_t + \epsilon_{it}$ with $\epsilon_{it} \sim N(0,9)$. The figure scatters the change in outcomes ΔY_i in gray open circles and the average change in outcomes for each observed dose $\mathbb{E}[\Delta Y_i | D_i = d]$ and against d. The red dashed line has a slope of 1 representing the true $ACRT$. The gray dotted line has a slope of $1 + ln(d)$ representing the observed relationship between $ATT(d|d)$ and d, and the solid black line is the TWFE estimate which is a line with a slope of 1.87.

in Figure 4 by the extent to which the TWFE slope deviates from one (plotted in the red dashed line). "Selection bias" in this case makes the TWFE estimate twice as large as the true $ACRT$.

Figure 5: Baseline Case Estimand Weights: Two-Way Fixed Effects Weights versus Treatment Distribution Weights

Notes: The figure plots the density of positive doses in gray and the TWFE weights from Theorem 3 in black. TWFE puts more weight on slopes for doses near $\mathbb{E}[D]$ compared to the right skewed exponential distribution of positive doses. Because "selection bias" in this example is positive and rising with the dose (is equals $ln(d)$) this weighting scheme puts more weight on doses with larger "bias" relative to a weighting scheme based on the sample distribution of observed doses.

4   Multiple Periods and Variation in Treatment Timing and Dose

DiD applications often use more than two time periods in which case treatments, whether binary or not, can turn on at different times for different units. This section extends the results from the previous sections to allow for multiple time periods ($t = 1, ..., \mathcal{T}$) with variation in the time when units become treated ($G = g \in \mathcal{G}$). By convention, we set $G = \mathcal{T} + 1$ for units that remain untreated across all time periods, and we exclude units that are treated in the first period so that $\mathcal{G} \subseteq \{2, \ldots, \mathcal{T} + 1\}$.[13] Treated units receive dose $D = d \in \mathcal{D}$. We also focus on the case that treatment is an absorbing state (or where units do not "forget" their treatment experience).

   In this section, we somewhat modify the potential outcomes notation from the previous section to allow for variation in treatment timing. For each unit, we define potential outcomes $Y_{it}(g, d)$ indexed by both treatment timing and dose. Note that treated potential outcomes at time t depend on when a unit first becomes treated—i.e., $Y_{it}(g, d)$ may not equal $Y_{it}(g', d)$ for $g \neq g'$— which allows for general treatment effect dynamics. $Y_{it}(\mathcal{T} + 1, 0)$ is the outcome that unit i would experience if they did not participate in the treatment in any period. For simplicity, we define

---

[13]We could alternatively use $G = \infty$ for units that remain untreated across all time periods.

$Y_{it}(0) := Y_{it}(\mathcal{T} + 1, 0)$ and refer to this as a unit's untreated potential outcome.[14] We also define the variable $W_{it} := D_i \mathbf{1}\{t \geq G_i\}$ which is the amount of dose that unit i experiences in time period t; $W_{it} = 0$ for all units that are not yet treated by time period t.

Throughout this section, we make the following assumptions.

Assumption 1-MP (Random Sampling). *The observed data consists of* $\{Y_{i1}, \ldots, Y_{i\mathcal{T}}, D_i, G_i\}_{i=1}^{n}$ *which is independent and identically distributed.*

Assumption 2-MP (Support). *(a) The support of* D, $\mathcal{D} = \{0\} \cup \mathcal{D}_+$. *In addition,* $\mathrm{P}(D = 0) > 0$ *and* $dF_{D|G}(d|g) > 0$ *for all* $(g, d) \in (\mathcal{G} \setminus \{\mathcal{T} + 1\}) \times \mathcal{D}_+$.

(b) $\mathcal{D}_+ = [d_L, d_U]$ *with* $0 < d_L < d_U < \infty$.

(c) For all $g \in (\mathcal{G} \setminus \{\mathcal{T} + 1\})$ *and* $t = 2, \ldots, \mathcal{T}$, $\mathbb{E}[\Delta Y_t | G = g, D = d]$ *is continuously differentiable in* d *on* $\mathcal{D}_+$.

Assumption 3-MP (No Anticipation / Staggered Adoption). *(a) For all* $g \in \mathcal{G}$ *and* $t = 1, \ldots, \mathcal{T}$ *with* $t < g$ *(i.e., in pre-treatment periods),* $Y_{it}(g, d) = Y_{it}(0)$.

(b) $W_{i1} = 0$ *almost surely and for* $t = 2, \ldots, \mathcal{T}$, $W_{it-1} = d$ *implies that* $W_{it} = d$.

Assumption 1-MP says that we have access to $\mathcal{T}$ periods of panel data and observe each unit's dose and treatment timing. Assumption 2-MP extends our definitions of the support of D to the case with multiple periods and variation in treatment timing. As in earlier sections, many of our identification results only require part (a) (which allows for very general treatment regimes) while some of our results are specialized to the continuous case as in parts (b) and (c).[15] Assumption 2-MP also imposes a kind of common support of the dose across timing groups, though it allows for the distribution of the dose to vary across timing groups in otherwise unrestricted ways; that said, it appears to be straightforward to relax this part of the assumption at the cost of additional notation.

Assumption 3-MP(a) rules out that units anticipate experiencing the treatment in ways that affect their outcomes before they actually participate in the treatment. It would be relatively straightforward to extend our arguments in this section to allow for anticipation along the lines of Callaway and Sant'Anna (2020) (in the case of a binary treatment). Assumption 3-MP(b) implies that we consider the case with staggered adoption which means that once units become treated with dose d they remain treated with dose d in all subsequent periods. This allows us to fully categorize a unit by the timing of their treatment adoption and the amount of dose that they experience.

---

[14]The analysis in this section could be extended to allow for units to be "treated" at time g but with $d = 0$. For example, units may live in a jurisdiction that implements a program at time g for which they are not eligible. Similarly, we could allow for units to have dose d but remain untreated $g = \mathcal{T} + 1$. This would make sense if a program's dose was based on a known formula so that it was possible to observe d even for units not actually selected for treatment.

[15]For the results in this section that are specialized to the case where the treatment is continuous, it is straightforward to adjust them to allow for a multi-valued discrete treatment along the same lines as in the previous section.

For each unit, we observe their outcome in period t, Y_{it}, which is given by

$$Y_{it} = Y_{it}(0)\mathbf{1}\{t < G_i\} + Y_{it}(G_i, D_i)\mathbf{1}\{t \geq G_i\}.$$

In other words, we observe a unit's untreated potential outcomes in time periods before they participate in the treatment, and we observe treated potential outcomes in post-treatment time periods that can depend on the timing of the treatment and the amount of the dose.

4.1 Parameters of Interest with a Staggered Continuous Treatment

The causal parameters of interest are the same as in our baseline case except that they are separately defined for each timing group and in each post-treatment time period. The average treatment effect parameters of dose d, for group g, in time period t are:

$$ATT(g,t,d|g,d) := \mathbb{E}[Y_t(g,d) - Y_t(0)|G = g, D = d] \quad \text{and} \quad ATE(g,t,d) := \mathbb{E}[Y_t(g,d) - Y_t(0)|G = g].$$

$ATT(g,t,d|g,d)$ is the average effect of dose d, for timing group g, in time period t, among units in group g that experienced dose d. $ATE(g,t,d)$ is the average effect of dose d among all units in timing group g (not all units in the population though), in time period t. $ATT(g,t,d|g,d)$ and $ATE(g,t,d)$ are similar to the group-time average treatment effects discussed in Callaway and Sant'Anna ([2020](#)) except they are also specific to a dose, and allow for the effect of dose to vary arbitrarily across timing groups and time periods.

ACR parameters are similarly defined as the effect of a marginal change in the dose on the outcomes of timing group g in period t. For continuous treatments the ACR parameters are:

$$ACRT(g,t,d|g,d) := \left.\frac{\partial \mathbb{E}\left[Y_t(g,l)|G = g, D = d\right]}{\partial l}\right|_{l=d},$$

$$ACR(g,t,d) := \frac{\partial \mathbb{E}\left[Y_t(g,d)|G = g\right]}{\partial d}.$$

For discrete treatments the ACR parameters are:

$$ACRT(g,t,d_j|g,d_j) := \mathbb{E}[Y_t(g,d_j) - Y_t(g,d_{j-1})|D = d_j, G = g],$$
$$ACR(g,t,d_j) := \mathbb{E}[Y_t(g,d_j) - Y_t(g,d_{j-1})|G = g].$$

The two parameters—$ACRT(g,t,d|g,d)$ and $ACR(g,t,d)$—correspond to $ATT(g,t,d|g,d)$ and $ATE(g,t,d)$ in that they are either local to a specific timing group and dose or refer to the entire population.

In many applications, $ACR(g,t,d)$ is relatively high-dimensional and challenging to report. There are a number of possible aggregations that reduce dimensionality and result in parameters that are easier to interpret. We focus on aggregations into an overall causal response across doses, timing groups, and treated periods, as well as into an event study; see Callaway and Sant'Anna

(2020) for additional possible aggregations along these lines. Also, note that the aggregations considered below are identified if $ACR(g,t,d)$'s are identified.

To start with, we define an overall causal response of experiencing dose d, for timing group g, across all post-treatment time periods

$$ACR^{group}(g,d) = \frac{1}{\mathcal{T} - g + 1} \sum_{t=g}^{\mathcal{T}} ACR(g,t,d).$$

These can be further aggregated by averaging across timing groups,

$$ACR^{overall}(d) = \sum_{g \in \mathcal{G}} ACR^{group}(g,d)P(G = g|G \leq \mathcal{T}, D = d)$$

$ACR^{overall}(d)$ is the average causal response of dose d across all timing groups that participate in the treatment in any period. It averages $ACR(g,t,d)$ across all observed doses, groups, and treated periods (in other words, all doses at each event-time and then across all event-times). This is a natural analogue of $ACR(d)$ in the two period case.

Another further aggregation is to average across the distribution of the dose (of all timing groups that participate in the treatment)

$$ACR^{*,mp} = \mathbb{E}\left[ACR^{overall}(D)|G \leq \mathcal{T}\right].$$

$ACR^{*,mp}$ is the overall average causal response (averaged across doses and and over all timing groups that participate in the treatment in any time period). $ACR^{*,mp}$ is a single number; it is the analogue of ACR^* from the two period case and is arguably a natural target parameter for a TWFE regression.

Next, we consider an event study type of aggregation.

$$ACR^{es}(e,d) = \sum_{g \in \mathcal{G}} \mathbf{1}\{g + e \leq \mathcal{T}\}ACR(g,g+e,d)\mathrm{P}(G = g|G + e \leq \mathcal{T}, D = d).$$

$ACR^{es}(e,d)$ is the average causal response of dose d among units that have been exposed to the treatment for exactly e periods. This can be further aggregated across the distribution of the dose

$$ACR(e) = \mathbb{E}[ACR^{es}(e,D)|G \leq \mathcal{T}],$$

which is the average partial effect (averaged across all doses) among units that have been exposed to the treatment for exactly e periods. Importantly, this keeps the distribution of the dose constant across different lengths of exposure to the treatment; the distribution of the dose is set to be equal to the distribution of the dose among the group of units that ever participate in the treatment. For values of $e \geq 0$, $ACR(e)$ can be interpreted as dynamic effects; but it is also interesting to consider cases where $e < 0$ which can be interpreted as a pre-test of the parallel trends assumption.

Remark 4. *The aggregations above are related to ACR(g,t,d), but similar arguments would apply to other parameters discussed in the paper including $ATT(g,t,d|g,d)$, $ATE(g,t,d)$, and $ACRT(g,t,d|g,d)$.*

4.2 Identification with a Continuous Treatment and Staggered Timing

With multiple time periods and variation in treatment timing, there are several possible versions of parallel trends and strong parallel trends assumptions that one could make because there are many ways to compare groups with different changes in their dose over time.

We focus on a version of strong parallel trends in this section and we provide a number of alternative parallel trends assumptions (and corresponding identification results) in Appendix C.

Assumption 5-MP (Strong Parallel Trends with Multiple Periods and Variation in Treatment Timing). *For all $g \in \mathcal{G}$, $t = 2, \ldots, \mathcal{T}$, and $d \in \mathcal{D}$, $\mathbb{E}[Y_t(g,d) - Y_{t-1}(0)|G = g, D = d] = \mathbb{E}[Y_t(g,d) - Y_{t-1}(0)|G = g]$ and $\mathbb{E}[\Delta Y_t(0)|G = g, D = d] = \mathbb{E}[\Delta Y_t(0)|D = 0]$.*

Assumption 5-MP is an extension of Assumption 5 to the case with multiple time periods. In particular, it restricts paths of treated potential outcomes (not just paths of untreated potential outcomes) so that all dose groups treated at time g would have had the same path of potential outcomes at every dose.

Theorem 4. *Under Assumptions 1-MP, 2-MP(a), 3-MP, and 5-MP, and for all $g \in \mathcal{G}$, $t = 2, \ldots, \mathcal{T}$ such that $t \geq g$, and for all $d \in \mathcal{D}_+$.*

$$ATE(g,t,d) = \mathbb{E}[Y_t - Y_{g-1}|G = g, D = d] - \mathbb{E}[Y_t - Y_{g-1}|W_t = 0].$$

If, in addition, Assumption 2-MP(b) and (c) hold, then, for all $d \in \mathcal{D}_+$,

$$ACR(g,t,d) = \frac{\partial \mathbb{E}[Y_t - Y_{g-1}|G = g, D = d]}{\partial d}.$$

The proof of Theorem 4 is provided in Appendix D. The result is broadly similar to the one in the case with two periods. It says that $ATE(g,t,d)$ can be recovered by a DiD comparison between the path of outcomes from period $g - 1$ to period t for units in group g treated with dose d and the path of outcomes among units that have not participated in the treatment yet. Relative to the case with two time periods, the main difference is that the "pre-period" is $g - 1$. The reason to use the base period $g - 1$ is that this is the most recent time period when the researcher observes untreated potential outcomes for units in group g. Thus, the result is very much like the case with two time periods: take the most recent untreated potential outcomes for units in a particular group, impute the path of outcomes that they would have experienced in the absence of participating in the treatment from the group of not-yet-treated units (these steps yield mean untreated potential outcomes that units in group g would have experienced in time period t) and compare this to the outcomes that are actually observed for units in group g that experienced dose d.

Remark 5. *Theorem 4 identifies $ATE(g,t,d)$ and $ACR(g,t,d)$ under a version of strong parallel trends. In Appendix C, we discuss identifying $ATT(g,t,d|g,d)$ and $ACRT(g,t,d|g,d)$ under a version of parallel trends that only involves untreated potential outcomes; in this case, like in the two period case, $ATT(g,t,d|g,d)$ is identified, comparisons of $ATT(g,t,d|g,d)$ across different values of d do not deliver a causal effect of moving from one dose to another (as they additionally include "selection bias" terms), and derivative of paths of outcomes over time do not recover $ACRT(g,t,d|g,d)$ due to the same "selection bias" terms.*

Remark 6. *It is natural to estimate $ATE(g,t,d)$ by simply replacing the population averages in Theorem 4 by their sample counterpart. This approach is very simple and intuitive, but in some cases, it may be possible to develop more efficient estimators using GMM. See the discussion in Marcus and Sant'Anna (2021) in the context of a binary treatment. When treatment d is continuous, some smoothing is required. However, one can use off-the-shelf standard nonparametric estimations procedures based on kernels or sieves to estimate these target causal parameters.*

4.3 TWFE estimators with multiple time periods and variation in treatment timing

In applications with multiple periods and variation in treatment timing, empirical researchers almost always estimate a TWFE regression

$$Y_{it} = \theta_t + \eta_i + \beta^{twfe}W_{it} + v_{it}. \tag{2}$$

Equation (2) is exactly the same as the TWFE regression in the baseline case with two periods in Equation (1) only with the notation slightly adjusted to match this section. The results in this section generalize the results in several recent papers on TWFE estimates including Goodman-Bacon (2021) and de Chaisemartin and D'Haultfœuille (2020) to our DiD setup with variation in treatment intensity.

To start with, write population versions of TWFE adjusted variables by

$$\ddot{W}_{it} := (W_{it} - \bar{W}_i) - \left(\mathbb{E}[W_t] - \frac{1}{\mathcal{T}} \sum_{t=1}^{\mathcal{T}} \mathbb{E}[W_t] \right), \quad \text{where} \quad \bar{W}_i := \frac{1}{\mathcal{T}} \sum_{t=1}^{\mathcal{T}} W_{it}.$$

The population version of the TWFE estimator is

$$\beta^{twfe} = \frac{\dfrac{1}{\mathcal{T}} \displaystyle\sum_{t=1}^{\mathcal{T}} \mathbb{E}[Y_{it}\ddot{W}_{it}]}{\dfrac{1}{\mathcal{T}} \displaystyle\sum_{t=1}^{\mathcal{T}} \mathbb{E}[\ddot{W}_{it}^2]}. \tag{3}$$

As in the previous section, we present both a "mechanical" decomposition of the TWFE estimator and a "causal" decomposition of the estimand that relates assumptions to interpretation.

In order to define these decompositions, we introduce a bit of new notation. First, define the fraction of periods that units in group g spends treated as

$$\bar{G}_g := \frac{\mathcal{T} - (g-1)}{\mathcal{T}}.$$

For the untreated group $g = \mathcal{T} + 1$ so that $\bar{G}_{\mathcal{T}+1} = 0$.

Next, we define time periods over which averages are taken. For averaging variables across time periods, we use the following notation, for $t_1 \leq t_2$,

$$\bar{Y}_i^{(t_1,t_2)} := \frac{1}{t_2 - t_1 + 1} \sum_{t=t_1}^{t_2} Y_{it}.$$

It is also convenient to define some particular averages across time periods. For two time periods g and k, with $k > g$, (below, g and k will often index groups defined by treatment timing), we define

$$\bar{Y}_i^{PRE(g)} := \bar{Y}_i^{(1,g-1)}, \quad \bar{Y}_i^{MID(g,k)} := \bar{Y}_i^{(g,k-1)}, \quad \bar{Y}_i^{POST(k)} := \bar{Y}_i^{(k,\mathcal{T})}.$$

$\bar{Y}_i^{PRE(g)}$ is the average outcome for unit i in periods 1 to $g-1$, $\bar{Y}_i^{MID(g,k)}$ is the average outcome for unit i in periods g to $k-1$, and $\bar{Y}_i^{POST(g,k)}$ is the average outcome for unit i in periods k to $\mathcal{T}$. Below, when g and k index groups, $\bar{Y}_i^{PRE(g)}$ is the average outcome for unit i in periods before units in either group are treated, $\bar{Y}_i^{MID(g,k)}$ is the average outcome for unit i in periods after group g has become treated but before group k has been treated, and $\bar{Y}_i^{POST(k)}$ is the average outcome for unit i after both groups have become treated.
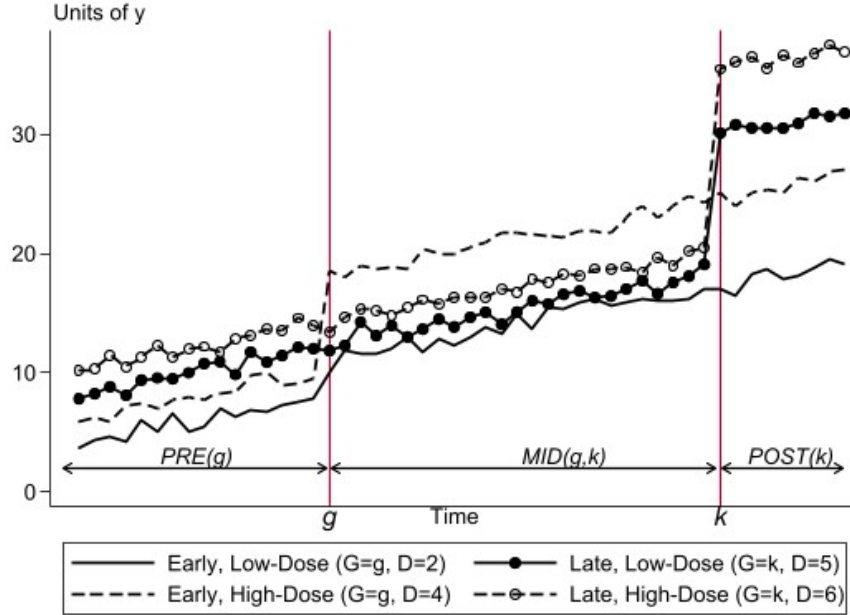
To fix ideas about how the staggered-timing/continuous treatment case works, consider a setup with two timing groups, g and k with $k > g$. Some units in the "early -treated" group have $d = 2$ and others have $d = 4$. Some units in the late treated group have $d = 5$ and others have $d = 6$. Thus, the four groups are early-treated/high-dose, early-treated/low-dose, late-treated/high-dose, and late-treated/low-dose. Figure 6 plots constructed outcomes for these groups with a treatment effect that is a one-time shift equal to $d^{1.5}$.

Following Goodman-Bacon (2021), we motivate the decomposition of the TWFE estimand by considering the four types of simple DiD estimands that can be formed using only one source of variation. The first comparison is a within-group comparison of paths of outcomes among units that experienced different amounts of the treatment.

$$\delta^{WITHIN}(g) := \frac{\mathrm{cov}(\bar{Y}^{POST(g)} - \bar{Y}^{PRE(g)}, D|G = g)}{\mathrm{var}(D|G = g)}. \tag{4}$$

This term is essentially the same as the expression for the TWFE estimand in the baseline two-period case. It equals the OLS (population) coefficient from regressing the change in average outcomes before and after g for units treated at time g on their dose, d. Figure 7 uses the four-group example to show how $\delta^{WITHIN}(g)$ and $\delta^{WITHIN}(k)$ use higher-dose units as the "treatment

Figure 6: A Simple Set-Up with Staggered Timing and Variation in the Dose

Notes: The figure plots simulated data for four groups: early-treated/high-dose, early-treated/low-dose, late-treated/high-dose, and late-treated/low-dose.

group" and lower-dose units as the "comparison group".

The second comparison is based on treatment timing. It compares paths of outcomes between a particular timing group g and a "later-treated" group k (i.e., $k > g$) in the periods after group g is treated but before group k becomes treated relative to their common pre-treatment periods.[16]

$$
\delta^{MID,PRE}(g,k) := \frac{\mathbb{E}\left[\left(\bar{Y}^{MID(g,k)} - \bar{Y}^{PRE(g)}\right)|G = g\right] - \mathbb{E}\left[\left(\bar{Y}^{MID(g,k)} - \bar{Y}^{PRE(g)}\right)|G = k\right]}{\mathbb{E}[D|G = g]}. \quad (5)
$$

Panel A of Figure 8 plots the outcomes used in this comparison with timing-group averages in black and the specific dose groups from Figure 6 in light gray. Under a parallel trends assumption, we show below that this term corresponds to a reasonable treatment effect parameter because the path of outcomes for group k (which is still in its pre-treatment period here) is what the path of outcomes would have been for group g if it had not been treated. Also note that this term encompasses comparisons of group g to the "never-treated" group.

The third comparison is between paths of outcomes for the "later-treated" group k in its post-treatment period relative to a pre-treatment period adjusted by the same path of outcomes for the

---

[16]Each of the following expressions also includes a term in the denominator. Below, this term is useful for interpreting differences across groups as partial effects of more treatment, but, for now, we largely ignore the expressions in the denominator.

Figure 7: Within-Timing-Group Comparisons Across Doses

Notes: The figure shows the within-timing group comparison between higher- and lower-dose units defined by $\delta^{WITHIN}(g)$ and $\delta^{WITHIN}(k)$.

"early -treated" group g.

$$\delta^{POST,MID}(g,k) := \frac{\mathbb{E}\left[\left(\bar{Y}^{POST(k)} - \bar{Y}^{MID(g,k)}\right)|G=k\right] - \mathbb{E}\left[\left(\bar{Y}^{POST(k)} - \bar{Y}^{MID(g,k)}\right)|G=g\right]}{\mathbb{E}[D|G=k]}. \tag{6}$$

These terms use the already-treated group g as the comparison group for group k. Panel B of Figure 8 plots the outcomes used in this term. Mechanically, the TWFE regression exploits this comparison because group g's treatment status/amount is not changing over these time periods. However, these are post-treatment periods for group g and parallel trends assumptions do not place restrictions on paths of post-treatment outcomes, which are subtracted in Equation (6). Below we discuss assumptions about treatment effect heterogeneity over time that are necessary to deal with this issue.[17]
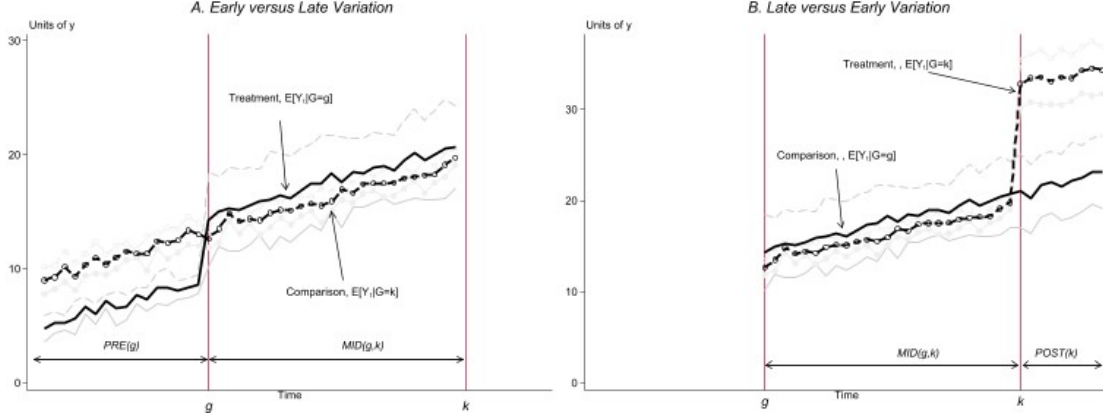
The final comparison that shows up in the TWFE estimator is between paths of outcomes between "early" and "late" treated groups in their common post-treatment periods relative to their common pre-treatment periods. In other words, this comparison only uses periods in which treatment status differs and focusing only on the "endpoints" where the two timing groups are either both untreated or both treated with potentially different average doses.

$$\delta^{POST,PRE}(g,k) := \frac{\mathbb{E}\left[\left(\bar{Y}^{POST(k)} - \bar{Y}^{PRE(g)}\right)|G=g\right] - \mathbb{E}\left[\left(\bar{Y}^{POST(k)} - \bar{Y}^{PRE(g)}\right)|G=k\right]}{\mathbb{E}[D|G=g] - \mathbb{E}[D|G=k]}. \tag{7}$$

Figure 9 shows the outcomes that determine the comparisons that show up in this term. The reason that this term shows up in β^{twfe} is that differences in the paths of outcomes between groups that

---

[17]This sort of comparison also shows up in the case with a binary, staggered treatment. See, e.g., Borusyak and Jaravel (2017), de Chaisemartin and D'Haultfœuille (2020) and Goodman-Bacon (2021).

Figure 8: Between-Timing-Group Comparisons

Notes: The figure shows the between-timing-group comparisons that average the outcomes in groups g and k across dose levels and compare the early group to the later group (panel C) or the later group to the early group (panel D).

have different distributions of the treatment are informative about β^{twfe}. For example, if more dose tends to increase outcomes and group g's dose is higher on average than group k's, then outcomes may increase more among group g than group k resulting in $\delta^{POST,PRE}(g,k)$ not being equal to 0.[18]

Next, we show how β^{twfe} weights these simple DiD terms together and discuss its theoretical interpretation under a parallel trends assumptions. To characterize the weights, first, define

$$p_{g|\{g,k\}} := P(G = g|G \in \{g,k\}),$$

which is the probability of being in group g conditional on being in either group g or k. We also define the following weights, which measure the variance of the treatment variable used to estimate each of the simple DiD terms in equations Equations (4) to (7).

$$w^{g,within}(g) := \text{var}(D|G=g)(1-\bar{G}_g)\bar{G}_g p_g \bigg/ \frac{1}{\mathcal{T}}\sum_{t=1}^{\mathcal{T}}\mathbb{E}[\ddot{W}_{it}^2],$$
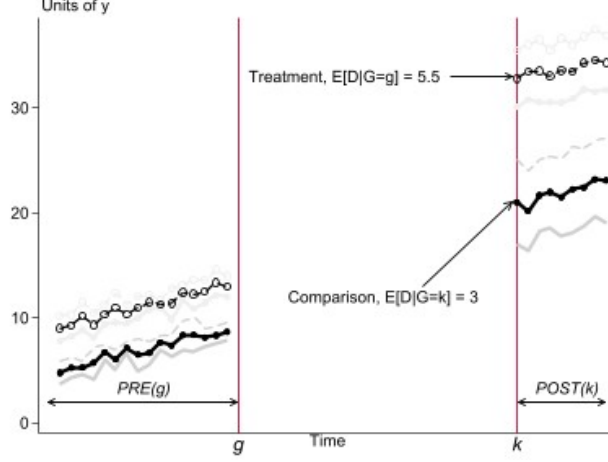
$$w^{g,post}(g,k) := \mathbb{E}[D|G=g]^2(1-\bar{G}_g)(\bar{G}_g - \bar{G}_k)(p_g + p_k)^2 p_{g|\{g,k\}}(1-p_{g|\{g,k\}}) \bigg/ \frac{1}{\mathcal{T}}\sum_{t=1}^{\mathcal{T}}\mathbb{E}[\ddot{W}_{it}^2],$$

$$w^{k,post}(g,k) := \mathbb{E}[D|G=k]^2\bar{G}_k(\bar{G}_g - \bar{G}_k)(p_g + p_k)^2 p_{g|\{g,k\}}(1-p_{g|\{g,k\}}) \bigg/ \frac{1}{\mathcal{T}}\sum_{t=1}^{\mathcal{T}}\mathbb{E}[\ddot{W}_{it}^2],$$

$$w^{long}(g,k) := (\mathbb{E}[D|G=g] - \mathbb{E}[D|G=k])^2\bar{G}_k(1-\bar{G}_g)(p_g + p_k)^2 p_{g|\{g,k\}}(1-p_{g|\{g,k\}}) \bigg/ \frac{1}{\mathcal{T}}\sum_{t=1}^{\mathcal{T}}\mathbb{E}[\ddot{W}_{it}^2].$$

---

[18]To be more precise, this term involves comparisons between groups g and k for the group with a higher dose on average to the group with a smaller dose on average. When $\mathbb{E}[D|G=g] > \mathbb{E}[D|G=k]$, this corresponds to the expression in Equation (7). When $\mathbb{E}[D|G=g] < \mathbb{E}[D|G=k]$, one can multiply both the numerator and denominator by -1 so that we effectively make a positive-weight comparison for the group that experienced more dose relative to the group that experienced less dose.

Figure 9: Long Comparisons Between Timing Groups

Notes: The figure shows the comparisons between timing groups in the $POST(k)$ window when both are treated with potentially different average doses and the $PRE(g)$ window when neither group is treated.

These weights are similar to the ones in Goodman-Bacon (2021) in the sense that they combine the size of the sample and the variance of treatment used to calculate each simple DiD term. In $w^{g,within}(g)$, for example, $\text{var}(D|G = g)$ measures how much the dose varies across units with $G = g$, $(1 - \bar{G}_g)\bar{G}_g$ measures the variance that comes from timing which falls when g is closer to 0 or $\mathcal{T}$, and p_g measures the share of units with $G = g$ (i.e.,. subsample size). Since they only compare outcomes between timing-groups, $w^{g,post}(g,k)$ and $w^{k,post}(g,k)$ do not contain a within-timing-group variance of D, but they do include $\mathbb{E}[D|G = k]^2$ which reflects the fact that timing groups with higher average doses get more weight. The rest of the timing weights have the same interpretation as in Goodman-Bacon (2021). Finally, $w^{long}(g,k)$ includes the square of the difference in mean doses between groups g and k—$(\mathbb{E}[D|G = g] - \mathbb{E}[D|G = k])^2$—which shows that the "endpoint" comparisons only influence β^{twfe} to the extent that timing groups have different average doses. Two timing groups with the same average dose do not contribute a $\delta^{POST,PRE}(g,k)$ term because there is no differential change in their doses between the $PRE(g)$ window (when both groups are untreated) and the $POST(k)$ window (when both groups have $E[D|G = g] = E[D|G = k]$).

Our next result combines the simple DiD terms and their variance weights to provide a mechanical decomposition of β^{twfe} in DiD setups with variation in treatment timing and variation in treatment intensity.

Proposition 5. *Under Assumptions 1-MP, 2-MP(a), and 3-MP, β^{twfe} in Equation (2) can be written as*

$$\beta^{twfe} = \sum_{g \in \mathcal{G}} w^{g,within}(g)\delta^{WITHIN}(g)$$

$$+ \sum_{g \in \mathcal{G}} \sum_{k \in \mathcal{G}, k > g} \left\{ w^{g,post}(g,k)\delta^{MID,PRE}(g,k) + w^{k,post}(g,k)\delta^{POST,MID}(g,k) + w^{long}(g,k)\delta^{POST,PRE}(g,k) \right\}.$$

In addition, (i) $w^{g,within}(g) \geq 0$, $w^{g,post}(g,k) \geq 0$, $w^{k,post}(g,k)$, and $w^{long}(g,k) \geq 0$ for all $g \in \mathcal{G}$ and $k \in \mathcal{G}$ with $k > g$, and (ii) $\sum_{g \in \mathcal{G}} w^{g,within}(g) + \sum_{g \in \mathcal{G}} \sum_{k \in \mathcal{G}, k > g} \{w^{g,post(g,k)}(g,k) + w^{k,post}(g,k) + w^{long}(g,k)\} = 1$.

Proposition 5 generalizes the decomposition theorem for binary staggered timing designs in Goodman-Bacon (2021) to our setup with variation in treatment intensity. [19] Notice that it does not require Assumption 2-MP(b) or (c), and is therefore compatible with a binary, multi-valued, continuous, or mixed treatment. It says that β^{twfe} can be written as a weighted average of the four comparisons in Equations (4) to (7). These weights are all positive and sum to one.

Proposition 5 is a new, explicit description of what kinds of comparisons TWFE uses to compute β^{twfe}, but it does not on its own provide guidance on how to interpret TWFE estimates. Our baseline results, for example, show that simple estimators like $\delta^{WITHIN}(g)$ equal averages of $ACRT$ parameters plus "selection bias" that arises from heterogeneous treatment effect functions. Similarly, the terms that compare outcomes across timing groups necessarily average over the dose-specific treatment effects of units within that timing group. We analyze the theoretical interpretation of each of these simple DiD estimand under different assumptions and then discuss what this implies about the (arguably implicit) identifying assumptions and estimand for TWFE.

To begin we define additional weights that apply to the underlying causal parameters in the DiD terms in Equations (4) through (7):

$$w_1^{within}(g,l) := \frac{\left(\mathbb{E}[D|G=g, D \geq l] - \mathbb{E}[D|G=g]\right)}{\text{var}(D|G=g)} \mathrm{P}(D \geq l|G=g),$$

$$w_1(g,l) := \frac{\mathrm{P}(D \geq l|G=g)}{\mathbb{E}[D|G=g]}, \qquad w_0(g) := \frac{d_L}{\mathbb{E}[D|G=g]},$$

$$w_1^{across}(g,k,l) := \frac{(\mathrm{P}(D \geq l|G=g) - \mathrm{P}(D \geq l|G=k))}{(\mathbb{E}[D|G=g] - \mathbb{E}[D|G=k])}, \qquad \tilde{w}_0^{across}(g,k) := \frac{d_L}{(\mathbb{E}[D|G=g] - \mathbb{E}[D|G=k])},$$

$$\tilde{w}_1^{across}(g,k,l) := \frac{\mathrm{P}(D \geq l|G=k)}{(\mathbb{E}[D|G=g] - \mathbb{E}[D|G=k])}.$$

In addition, define the following differences in paths of outcomes over time

$$\pi^{POST(\tilde{k}),PRE(\tilde{g})}(g) := \mathbb{E}\left[(\bar{Y}^{POST(\tilde{k})} - \bar{Y}^{PRE(\tilde{g})})|G=g\right] - \mathbb{E}\left[(\bar{Y}^{POST(\tilde{k})} - \bar{Y}^{PRE(\tilde{g})})|D=0\right],$$

$$\pi^{MID(\tilde{g},\tilde{k}),PRE(\tilde{g})}(g) := \mathbb{E}\left[(\bar{Y}^{MID(\tilde{g},\tilde{k})} - \bar{Y}^{PRE(\tilde{g})})|G=g\right] - \mathbb{E}\left[(\bar{Y}^{MID(\tilde{g},\tilde{k})} - \bar{Y}^{PRE(\tilde{g})})|D=0\right],$$

$$\pi^{POST(\tilde{k}),MID(\tilde{g},\tilde{k})}(g) := \mathbb{E}\left[(\bar{Y}^{POST(\tilde{k})} - \bar{Y}^{MID(\tilde{g},\tilde{k})})|G=g\right] - \mathbb{E}\left[(\bar{Y}^{POST(\tilde{k})} - \bar{Y}^{MID(\tilde{g},\tilde{k})})|D=0\right],$$

---

[19] In particular, in the special case of a staggered, binary treatment, $w^{g,within}(g)\delta^{WITHIN}(g) = 0$ (since there is no within group variation in the dose in this case), and $w^{long}(g,k)\delta^{POST,PRE}(g,k) = 0$ (because the distribution of the dose is the same across all groups). Then, Proposition 5 collapses to Theorem 1 in Goodman-Bacon (2021) because the terms $w^{g,post(g,k)}\delta^{MID,PRE}(g,k)$ and $w^{k,post}(g,k)\delta^{POST,MID}(g,k)$ correspond exactly to between-timing-group comparisons.

and, similarly,

$$\pi_D^{POST(\tilde{k}),PRE(\tilde{g})}(g,d) := \mathbb{E}\left[(\bar{Y}^{POST(\tilde{k})} - \bar{Y}^{PRE(\tilde{g})})|G = g, D = d\right] - \mathbb{E}\left[(\bar{Y}^{POST(\tilde{k})} - \bar{Y}^{PRE(\tilde{g})})|D = 0\right],$$

$$\pi_D^{MID(\tilde{g},\tilde{k}),PRE(\tilde{g})}(g,d) := \mathbb{E}\left[(\bar{Y}^{MID(\tilde{g},\tilde{k})} - \bar{Y}^{PRE(\tilde{g})})|G = g, D = d\right] - \mathbb{E}\left[(\bar{Y}^{MID(\tilde{g},\tilde{k})} - \bar{Y}^{PRE(\tilde{g})})|D = 0\right],$$

$$\pi_D^{POST(\tilde{k}),MID(\tilde{g},\tilde{k})}(g,d) := \mathbb{E}\left[(\bar{Y}^{POST(\tilde{k})} - \bar{Y}^{MID(\tilde{g},\tilde{k})})|G = g, D = d\right] - \mathbb{E}\left[(\bar{Y}^{POST(\tilde{k})} - \bar{Y}^{MID(\tilde{g},\tilde{k})})|D = 0\right],$$

which are the same paths of outcomes but conditional on having dose d.

The following result is our main result on interpreting TWFE estimates with continuous treatment.

Theorem 5. *Under Assumptions 1-MP, 2-MP, and 3-MP,*

(1) The four comparisons in Equations (4) to (7) can be written as

$$\delta^{WITHIN}(g) = \int_{\mathcal{D}_+} w_1^{within}(g,l) \frac{\partial \pi_D^{POST(g),PRE(g)}(g,l)}{\partial l}\, dl,$$

$$\delta^{MID,PRE}(g,k) = \int_{\mathcal{D}_+} w_1(g,l) \frac{\partial \pi_D^{MID(g,k),PRE(g)}(g,l)}{\partial l}\, dl + w_0(g) \frac{\pi_D^{MID(g,k),PRE(g)}(g,d_L)}{d_L}$$
$$- w_0(g) \frac{\pi^{MID(g,k),PRE(g)}(k)}{d_L},$$

$$\delta^{POST,MID}(g,k) = \int_{\mathcal{D}_+} w_1(k,l) \frac{\partial \pi_D^{POST(k),MID(g,k)}(k,l)}{\partial l}\, dl + w_0(k) \frac{\pi^{POST(k),MID(g,k)}(k,d_L)}{d_L}$$
$$- w_0(k) \left(\frac{\pi^{POST(k),PRE(g)}(g) - \pi^{MID(g,k),PRE(g)}(g)}{d_L} \right),$$

$$\delta^{POST,PRE}(g,k) = \int_{\mathcal{D}_+} w_1^{across}(g,k,l) \frac{\partial \pi_D^{POST(k),PRE(g)}(g,l)}{\partial l}\, dl$$
$$- \left\{ \int_{\mathcal{D}_+} \tilde{w}_1^{across}(g,k,l) \left(\frac{\partial \pi_D^{POST(k),PRE(g)}(k,l)}{\partial l} - \frac{\partial \pi_D^{POST(k),PRE(g)}(g,l)}{\partial l} \right) dl \right.$$
$$\left. + \tilde{w}_0^{across}(g,k) \left(\frac{\pi_D^{POST(k),PRE(g)}(k,d_L) - \pi_D^{POST(k),PRE(g)}(g,d_L)}{d_L} \right) \right\}.$$

(2) If, in addition, Assumption 5-MP holds, then

$$\delta^{WITHIN}(g) = \int_{\mathcal{D}_+} w_1^{within}(g,l) \overline{ACR}^{POST(g)}(g,l) dl,$$

$$\delta^{MID,PRE}(g,k) = \int_{\mathcal{D}_+} w_1(g,l) \overline{ACR}^{MID(g,k)}(g,l)\, dl + w_0(g) \frac{\overline{ATE}^{MID(g,k)}(g,d_L)}{d_L},$$

33

$$\delta^{POST,MID}(g,k) = \int_{\mathcal{D}_+} w_1(k,l)\overline{ACR}^{POST(k)}(k,l)\,dl + w_0(k)\frac{\overline{ATE}^{POST(k)}(k,d_L)}{d_L}$$
$$- w_0(k)\left(\frac{\pi^{POST(k),PRE(g)}(g) - \pi^{MID(g,k),PRE(g)}(g)}{d_L}\right),$$

$$\delta^{POST,PRE}(g,k) = \int_{\mathcal{D}_+} w_1^{across}(g,k,l)\overline{ACR}^{POST(k)}(g,l)\,dl$$
$$- \left\{ \int_{\mathcal{D}_+} \tilde{w}_1^{across}(g,k,l)\left(\frac{\partial\pi_D^{POST(k),PRE(g)}(k,l)}{\partial l} - \frac{\partial\pi_D^{POST(k),PRE(g)}(g,l)}{\partial l}\right)\,dl \right.$$
$$\left. + \tilde{w}_0^{across}(g,k)\left(\frac{\pi_D^{POST(k),PRE(g)}(k,d_L) - \pi_D^{POST(k),PRE(g)}(g,d_L)}{d_L}\right)\right\}.$$

In addition, (i) $w_1^{within}(g,d) \geq 0$, $w_1(g,d) \geq 0$, and $w_0(g) \geq 0$, for all $g \in \mathcal{G}$ and $d \in \mathcal{D}_+$ and (ii) $\int_{\mathcal{D}_+} w_1^{within}(g,l)\,dl = 1$, $\int_{\mathcal{D}_+} w_1(g,l)\,dl + w_0(g) = 1$, and $\int_{\mathcal{D}_+} w_1^{across}(g,k,l)\,dl = 1$.

Part (1) of Theorem 5 links the four sets of comparisons in the TWFE estimator in Proposition 5 to derivatives of conditional expectations (this is analogous to Proposition 4 in the baseline case above) along with some additional (nuisance) paths of outcomes.

Part (2) of Theorem 5 imposes Assumption 5-MP. Under Assumption 5-MP, $\delta^{WITHIN}(g)$ and $\delta^{MID,PRE}(g,k)$ both deliver weighted averages of ACR-type parameters. However, $\delta^{POST,MID}(g,k)$ and $\delta^{POST,PRE}(g,k)$ still involve non-negligible nuisance terms. Under Assumption 5-MP, the additional term in $\delta^{POST,MID}(g,k)$ involves the difference between treatment effects for group g in group k's post-treatment periods relative to treatment effects for group g in the periods after group g is treated but before group k is treated — that is, treatment effect dynamics. Parallel trends assumptions do not imply that this term is equal to 0. And, in the special case where the treatment is binary, this term corresponds to the "problematic" term related to treatment effect dynamics in Goodman-Bacon (2021).

The additional nuisance term in $\delta^{POST,PRE}(g,k)$ involves differences in partial effects of more treatment across groups in their common post-treatment periods. Parallel trends does not restrict these partial effects to be equal to each other. This term does not show up in the case with a binary treatment because, by construction, the distribution of the dose is the same across groups. It is helpful to further consider where this expression comes from. For simplicity, temporarily suppose that the partial effect of more dose is positive and constant across groups, time, and dose. In this case, if group g has more dose on average than group k, then its outcomes should increase more from group g and k's common pre-treatment period to their common post-treatment period. This is the comparison that shows up in $\delta^{POST,PRE}(g,k)$. However, when partial effects are not the same across groups and times (which is not implied by any parallel trends assumption), then, for example, it could be the case that the partial effect of dose is positive for all groups and time periods but greater for group k relative to group g. If these differences are large enough, it could lead to the cross-group, long-difference comparisons in $\delta^{POST,PRE}(g,k)$ having the opposite sign.

Next, we engage on how one could potentially "rescue" TWFE procedures such that β^{twfe}

would always recover a weighted average of reasonable treatment effect parameters. To do so, one must further restrict different types of treatment effect heterogeneity.

Assumption 6. *(a) [No Treatment Effect Dynamics] For all $g \in \mathcal{G} \setminus (\mathcal{T} + 1)$ and $t \geq g$ (i.e, post-treatment periods for group g), $ACR(g,t,d)$ and $ATE(g,t,d_L)$ do not vary with t.*

(b) [Homogeneous Causal Responses across Groups] For all $g \in \mathcal{G} \setminus (\mathcal{T} + 1)$ with $t \geq g$ and $k \in \mathcal{G} \setminus (\mathcal{T} + 1)$ with $t \geq k$, $ACR(g,t,d) = ACR(k,t,d)$ and $ATE(g,t,d_L) = ATE(k,t,d_L)$.

(c) [Homogeneous Causal Responses across Dose] For all $g \in \mathcal{G} \setminus (\mathcal{T}+1)$ with $t \geq g$, $ACR(g,t,d)$ does not vary across d, and, in addition, $ATE(g,t,d_L)/d_L = ACR(g,t,d)$.

Assumption 6 introduces three additional conditions limiting treatment effect heterogeneity. Assumption 6(a) imposes that, within a timing-group, the causal response to the treatment does not vary across time which rules out treatment effect dynamics. Assumption 6(b) imposes that, for a fixed time period, causal responses to the treatment are constant across timing-groups. Assumption 6(c) imposes that, within timing-group and time period, the causal response to more dose is constant across different values of the dose.

Proposition 6. *Under Assumptions 1-MP, 2-MP, 3-MP, and 5-MP,*

(a) If, in addition, Assumption 6(a) holds, then

$$\delta^{POST,MID}(g,k) = \int_{\mathcal{D}_+} w_1(k,l)\overline{ACR}^{POST(k)}(k,l)\,dl + w_0(k)\frac{\overline{ATE}^{POST(k)}(k,d_L)}{d_L}.$$

(b) If, in addition, Assumption 6(b) holds, then

$$\delta^{POST,PRE}(g,k) = \int_{\mathcal{D}_+} w_1^{across}(g,k,l)\overline{ACR}^{POST(k)}(g,l)\,dl.$$

(c) If, in addition, Assumption 6(a), (b) and (c) hold, then

$$\beta^{twfe} = ACR^{*,mp}.$$

Proposition 6 provides additional conditions under which the nuisance terms in $\delta^{POST,MID}(g,k)$ and $\delta^{POST,PRE}(g,k)$ are equal to 0. For $\delta^{POST,MID}(g,k)$, these nuisance terms can be eliminated by ruling out treatment effect dynamics; that is, by assuming that, within a particular group, the causal response to more dose does not vary across time. Ruling out these sort of treatment effect dynamics is analogous to the kinds of conditions that are required to interpret TWFE estimates with a binary treatment. In order for the nuisance terms in $\delta^{POST,PRE}(g,k)$ to be equal to 0, we impose homogeneous causal responses across groups — that the causal response to more dose is the same across groups conditional on having the same amount of dose and being in the same time period. Neither of these assumptions are implied by any of the parallel trends assumptions that we have considered, and they are both potentially very strong. Therefore, under both Assumption 6(a) and

(b), β^{twfe} is equal to a weighted average of average causal response parameters, but these weights continue to be driven by the TWFE estimation strategy and, like in the baseline two period case, can continue to deliver poor estimates of the overall average causal response to the treatment. Imposing Assumption 6(a), (b), and (c) implies that $ACR(g,t,d)$ does not vary by timing group, time period, or the amount of dose, and part (c) of Proposition 6 says that β^{twfe} is equal to the overall average causal response under these additional, strong conditions.

Remark 7. *The results in Part (2) of Theorem 5 and in Proposition 6 relied on the multi-period version of strong parallel trends in Assumption 5-MP. In Theorem 5-Extended in Appendix D, we additionally show that, under a multi-period version of standard parallel trends (this is analogous to Assumption 4 in the two period case and details are provided in Assumption 4-MP(a) in Appendix C), similar results hold except that $\overline{ACR}^{.}(\cdot,d)$ should be replaced by $\overline{ACRT}^{.}(\cdot,d|\cdot,d)+\frac{\partial\overline{ATT}^{.}(\cdot,d|\cdot,l)}{\partial l}\Big|_{l=d}$ where the second term is a "selection bias" term, and $\overline{ATE}^{.}(\cdot,d_L)$ should be replaced by $\overline{ATT}^{.}(\cdot,d_L|\cdot,d_L)$. This implies that, under a standard version of parallel trends, all four comparisons in Equations (4) to (7) include "selection bias" terms.*

4.4 Discussion

The results in this section suggest three important weaknesses of TWFE estimands in a difference-in-differences framework with multiple time periods, and variation in treatment intensity and timing of adoptions. First, like the TWFE estimands considered above in the case with two time periods, TWFE estimands have weights that are driven by the estimation method. As above, these weights may have undesirable properties in setups where treatment effect heterogeneity is the rule rather than the exception.

Second, in addition to reasonable treatment effect parameters, TWFE estimands also include undesirable components due to treatment effect dynamics and heterogeneous causal responses across groups and time periods. That these show up in the TWFE estimand is potentially problematic and can possibly lead to very poor performance of the TWFE estimator. Ruling out these problems requires substantially stronger conditions in addition to any kind of parallel trends assumption.

Finally, even when these extra conditions hold (i.e., the best case scenario for TWFE), if a researcher invokes a standard parallel trends assumption, the TWFE estimand delivers weighted averages of derivatives of *ATT*-type parameters which are themselves hard to interpret because, like in the two period case, they include both actual causal responses and "selection bias" terms.

Of these three weaknesses, the first two can be completely avoided by using the estimands presented in Theorem 4. These estimands rely only on parallel trends assumptions; in particular, they are available without imposing any conditions on treatment effect dynamics or how causal responses vary across groups. The third weakness, though, is a more fundamental challenge of difference-in-differences approaches with variation in treatment intensity as comparing treatment effect parameters across different values of the dose appears to fundamentally require imposing stronger assumptions that rule out some forms of selection into different amounts of the treatment. Although undesirable, we are not aware of any other practical solution to this empirically

relevant DiD problem. Thus, we urge practitioners to transparently discuss their assumptions, potentially exploiting context-specific knowledge to justify the plausibility of a stronger parallel trends assumption in the given application.

5   Conclusion

In this paper, we have studied difference-in-differences approaches to identifying and estimating causal effects of a multi-valued or continuous treatment. The paper has a number of results that are potentially useful to empirical researchers, and, to conclude the paper, we briefly summarize these results.

First, while *ATT*-type parameters can be identified under a standard parallel trends assumption, a fundamental complication in the case with a multi-valued/continuous treatment is that comparisons across different amounts of the treatment are confounded by "selection bias" type terms that make these sorts of comparisons very difficult to interpret. This kind of bias carries over to identifying average causal responses of more dose. These sorts of difficulties can be avoided by invoking alternative parallel trends assumptions, but these assumptions are likely to be substantially stronger than the ones most researchers have in mind when they are using a difference-in-differences identification strategy. In addition, pre-tests commonly used in DiD applications are not able to distinguish between these two types of parallel trends assumptions.

Furthermore, two way fixed effects regressions that are commonly used by empirical researchers have a number of drawbacks. In a baseline case with two periods, TWFE regressions deliver a weighted average of causal responses to the treatment. The weights are all positive, but they are driven by the estimation procedure which can result in misleading results in a number of realistic cases. Moreover, in cases where there are multiple time periods, variation in treatment timing and in treatment intensity (which are common in applications), TWFE regressions are additionally sensitive to (i) treatment effect dynamics and (ii) heterogeneous causal responses across timing groups. We propose an identification and estimation strategy that is straightforward to implement and does not suffer from these drawbacks.

References

[1] Joshua D Angrist. "Grouped-data estimation and testing in simple labor-supply models". *Journal of Econometrics* 47.2-3 (1991), pp. 243–266.

[2] Joshua D Angrist and Ivan Fernandez-Val. "ExtrapoLATE-ing: External validity and overidentification in the LATE framework". *Advances in Economics and Econometrics: Volume 3, Econometrics: Tenth World Congress*. Vol. 51. Cambridge University Press. 2013, p. 401.

[3] Joshua D Angrist, Kathryn Graddy, and Guido W Imbens. "The interpretation of instrumental variables estimators in simultaneous equations models with an application to the demand for fish". *The Review of Economic Studies* 67.3 (2000), pp. 499–527.

[4] Joshua D Angrist and Guido W Imbens. "Two-stage least squares estimation of average causal effects in models with variable treatment intensity". *Journal of the American statistical Association* 90.430 (1995), pp. 431–442.

[5] Joshua D Angrist and Jorn-Steffen Pischke. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press, 2008.

[6] Susan Athey and Guido Imbens. "Identification and inference in nonlinear difference-in-differences models". *Econometrica* 74.2 (2006), pp. 431–497.

[7] Kirill Borusyak and Xavier Jaravel. "Revisiting event study designs". Working Paper. 2017.

[8] Brantly Callaway and Weige Huang. "Distributional effects of a continuous treatment with an application on intergenerational mobility". *Oxford Bulletin of Economics and Statistics* 82.4 (2020), pp. 808–842.

[9] Brantly Callaway and Pedro H.C. Sant'Anna. "Difference-in-differences with multiple time periods". *Journal of Econometrics* Forthcoming (2020).

[10] David Card. "Using regional variation in wages to measure the effects of the federal minimum wage". *Industrial and Labor Relations Review* 46.1 (1992), pp. 22–37.

[11] Xavier D'Haultfoeuille, Stefan Hoderlein, and Yuya Sasaki. "Nonlinear difference-indifferences in repeated cross sections with continuous treatments". Working Paper. 2021.

[12] Clement de Chaisemartin and Xavier D'Haultfœuille. "Fuzzy differences-in-differences". *The Review of Economic Studies* 85.2 (2018), pp. 999–1028.

[13] Clement de Chaisemartin and Xavier D'Haultfœuille. "Two-way fixed effects estimators with heterogeneous treatment effects". *American Economic Review* 110.9 (2020), pp. 2964–2996.

[14] Jean-Pierre Florens, James J Heckman, Costas Meghir, and Edward Vytlacil. "Identification of treatment effects using control functions in models with continuous, endogenous treatment and heterogeneous effects". *Econometrica* 76.5 (2008), pp. 1191–1206.

[15] Carlos A Flores. "Estimation of dose-response functions and optimal doses with a continuous treatment". Working Paper. 2007.

[16] Carlos A Flores, Alfonso Flores-Lagunes, Arturo Gonzalez, and Todd C Neumann. "Estimating the effects of length of exposure to instruction in a training program: the case of job corps". *Review of Economics and Statistics* 94.1 (2012), pp. 153–171.

[17] Hans Fricke. "Identification based on difference-in-differences approaches with multiple treatments". *Oxford Bulletin of Economics and Statistics* 79.3 (2017), pp. 426–433.

[18] Markus Frölich. "Programme evaluation with multiple treatments". *Journal of Economic Surveys* 18.2 (2004), pp. 181–224.

[19] Antonio F Galvao and Liang Wang. "Uniformly semiparametric efficient estimation of treatment effects with a continuous treatment". *Journal of the American Statistical Association* 110.512 (2015), pp. 1528–1542.

[20] Andrew Goodman-Bacon. "Difference-in-differences with variation in treatment timing". *Journal of Econometrics* forthcoming (2021).

[21] James Heckman and Edward Vytlacil. "Structural equations, treatment effects, and econometric policy evaluation1". *Econometrica* 73.3 (2005), pp. 669–738.

[22] Nathaniel Hendren. "The policy elasticity". *Tax Policy and the Economy* 30.1 (2016), pp. 51–89.

[23] Sir Austin Bradford Hill. "The environment and disease: association or causation?" *Journal of the Royal Society of Medicine* 58.5 (1965), pp. 295–300.

[24] Keisuke Hirano and Guido W Imbens. "The propensity score with continuous treatments". *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives* 226164 (2004), pp. 73–84.

[25] Edward H Kennedy, Zongming Ma, Matthew D McHugh, and Dylan S Small. "Non-parametric methods for doubly robust estimation of continuous treatment effects". *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 79.4 (2017), pp. 1229–1245.

[26] Michelle Marcus and Pedro HC Sant'Anna. "The role of parallel trends in event study settings: An application to environmental economics". *Journal of the Association of Environmental and Resource Economists* 8.2 (2021), pp. 235–275.

[27] Bruce D. Meyer. "Natural and Quasi-Experiments in Economics". *Journal of Business & Economic Statistics* 13.2 (1995), pp. 151–161.

[28] Magne Mogstad, Andres Santos, and Alexander Torgovitsky. "Using instrumental variables for inference about policy relevant treatment parameters". *Econometrica* 86.5 (2018), pp. 1589–1619.

[29] Philip Oreopoulos. "Estimating average and local average treatment effects of education when compulsory schooling laws really matter". *American Economic Review* 96.1 (2006), pp. 152–175.

[30] Liangjun Su, Takuya Ura, and Yichong Zhang. "Non-separable models with high-dimensional data". *Journal of Econometrics* 212.2 (2019), pp. 646–677.

[31] Liyang Sun and Sarah Abraham. "Estimating dynamic treatment effects in event studies with heterogeneous treatment effects". *Journal of Econometrics* forthcoming (2020).

[32] Kristina A. Thayer, Ronald Melnick, Kathy Burns, Devra Davis, and James Huff. "Fundamental Flaws of Hormesis for Public Health Decisions". *Environmental Health Perspectives* 113.10 (2005), pp. 1271–1276.

[33] Jeffrey M Wooldridge. "Fixed-effects and related estimators for correlated random-coefficient and treatment-effect panel data models". *Review of Economics and Statistics* 87.2 (2005), pp. 385–390.

[34] Shlomo Yitzhaki. "On using linear regressions in welfare economics". *Journal of Business & Economic Statistics* 14.4 (1996), pp. 478–486.

A Comparing Alternative Parallel Trends Assumptions

It is worth thinking more carefully about the differences between Assumption 4 and Assumption 5. In this section, we show that Assumption 5 is not strictly stronger than Assumption 4 though it is likely to be *stronger in practice* in most applications.

To see that Assumption 5 is not strictly stronger, consider the case where there are two doses d_1 and d_2. In this case, Assumption 4 is equivalent to the following conditions

$$\mathbb{E}[\Delta Y_t(0)|D = d_1] = \mathbb{E}[\Delta Y_t(0)|D = d_2] = \mathbb{E}[\Delta Y_t(0)|D = 0] \tag{8}$$

while Assumption 5 is equivalent to

$$\mathbb{E}[\Delta Y_t(0)] = \mathbb{E}[\Delta Y_t(0)|D = 0] \tag{Comp-0}$$

$$\mathbb{E}[Y_t(d_1) - Y_{t-1}(0)] = \mathbb{E}[Y_t(d_1) - Y_{t-1}(0)|D = d_1] \tag{Comp-1}$$

$$\mathbb{E}[Y_t(d_2) - Y_{t-1}(0)] = \mathbb{E}[Y_t(d_2) - Y_{t-1}(0)|D = d_2] \tag{Comp-2}$$

Assumption 4 does not place any restrictions on any potential outcomes besides untreated potential outcomes, and therefore the "extra" conditions in Equations (Comp-1) and (Comp-2) imply that Assumption 5 is not weaker than Assumption 4.

On the other hand, Equation (Comp-0) does not imply Equation (8); rather, it implies that

$$\mathbb{E}[\Delta Y_t(0)|D = 0] = \mathbb{E}[\Delta Y_t(0)|D = d_1]\frac{\mathrm{P}(D = d_1)}{\mathrm{P}(D = d_1) + \mathrm{P}(D = d_2)} + \mathbb{E}[\Delta Y_t(0)|D = d_2]\frac{\mathrm{P}(D = d_2)}{\mathrm{P}(D = d_1) + \mathrm{P}(D = d_2)}$$

In other words, the trend in untreated potential outcomes does not have to be exactly the same for all doses, but, instead, they have to be the same on average.

In practice, this potentially allows for some units to select their amount of dose on the basis of the path of their untreated potential outcomes (which is not allowed under the standard parallel trends assumption in Assumption 4), but that the amount of selection has to average out across doses to be equal to zero. It seems hard to think of realistic cases where Assumption 5 would be weaker than Assumption 4 in practice.

A related alternative assumption is

Assumption 5-Alt (Alternative Strong Parallel Trends Assumption). *For all $d \in \mathcal{D}$ and $l \in \mathcal{D}$,*

$$\mathbb{E}[Y_t(d) - Y_{t-1}(0)|D = l] = \mathbb{E}[Y_t(d) - Y_{t-1}(0)|D = d]$$

Assumption 5-Alt is a stronger, but closely related version of the strong parallel trends assumption in Assumption 5. Assumption 5-Alt says that, across all potential doses d, the path of potential outcomes $Y_t(d) - Y_{t-1}(0)$ (which is the path of outcomes that a unit would experience if they experienced dose d in time period t and were untreated in period $t - 1$) is, on average, (i) the same across all actual doses experienced, l, and (ii) is equal to the average path of outcomes for units that actually experienced dose d. Further, note that $\mathbb{E}[Y_t(d) - Y_{t-1}(0)|D = l]$ is not identified from the sampling process except in the case where $l = d$ (i.e., the left hand side of the equation in Assumption 5-Alt is not identified from the sampling process, but the right hand side is). It is immediately clear that this assumption implies both Assumptions 4 and 5. While it does not place restrictions on the levels of untreated potential outcomes in period $t - 1$, it does place (substantial) restrictions on treated potential outcomes and on treatment effect heterogeneity which is demonstrated in the next proposition.

Proposition 7. *Assumption 5-Alt implies that*

$$ATE(d) = ATT(d|d)$$

Proof. Starting with the definition of $ATE(d)$,

$$
\begin{aligned}
ATE(d) &= \mathbb{E}[Y_t(d) - Y_t(0)] \\
&= \mathbb{E}[Y_t(d) - Y_{t-1}(0)] - \mathbb{E}[Y_t(0) - Y_{t-1}(0)] \\
&= \mathbb{E}[Y_t(d) - Y_{t-1}(0)|D = d] - \mathbb{E}[Y_t(0) - Y_{t-1}(0)|D = d] \\
&= \mathbb{E}[Y_t(d) - Y_t(0)|D = d] \\
&= ATT(d|d)
\end{aligned}
$$

where the second equality holds by adding and subtracting $\mathbb{E}[Y_{t-1}(0)]$, the third equality holds by Assumption 5-Alt (also, notice that this equality does not hold under Assumption 4, nor is $ATE(d)$ generally equal to $ATT(d|d)$ under Assumption 4 alone), the fourth equality holds by canceling the two $\mathbb{E}[Y_{t-1}(0)|D = d]$ terms, and the remaining term in that equality is $ATT(d|d)$ $\square$

Proposition 7 shows that Assumption 5-Alt implies that the overall average effect of dose d is equal to the average effect of dose d for units who actually experienced dose d. The implication of this result is that Assumption 5-Alt rules out many forms of selection into a particular dose d on the basis of the effect of that amount of dose.

B   Alternative Decompositions for TWFE Regression

In this section, we provide three alternative decompositions of the TWFE regression estimator in Equation (1) in the baseline case with two periods, where no unit is treated yet in the first period, and where some units remain untreated in the second period.

The first decomposition is one where β^{twfe} equals a weighted average of 2×2 DiD comparisons between pairs of dose groups scaled by the difference in their doses:

Proposition 8. *Consider β^{twfe} in Equation (1) and suppose that Assumption 1 holds.*

(1) If Assumption 2-Cont also holds, then

$$
\begin{aligned}
\beta^{twfe} = &\int_{\mathcal{D}_+} \int_{\mathcal{D}, h>l} w_1^{2\times2,cont}(l,h) \frac{(m_\Delta(h) - m_\Delta(l))}{(h-l)} \, dh \, dl \\
&+ \int_{\mathcal{D}, h>0} w_0^{2\times2,cont}(h) \frac{m_\Delta(h) - m_\Delta(0)}{h} \, dh
\end{aligned}
$$

where

$$
\begin{aligned}
w_1^{2\times2,cont}(l,h) &:= (h-l)^2 (f_D(h) + f_D(l))^2 f_{D|\{h,l\}}(h) f_{D|\{h,l\}}(l)/\mathrm{var}(D) \\
w_0^{2\times2,cont}(h) &:= h^2 (f_D(h) + p_0^D)^2 f_{D|\{h,0\}}(h) p_{0|\{h,0\}}^D/\mathrm{var}(D)
\end{aligned}
$$

and

$$f_{D|\{h,l\}}(h) := f_D(h)/(f_D(h) + f_D(l))$$
$$f_{D|\{h,l\}}(l) := f_D(l)/(f_D(h) + f_D(l))$$
$$f_{D|\{h,0\}}(h) := f_D(h)/(f_D(h) + p_0^D)$$
$$p_{0|\{h,0\}}^D := p_0^D/(f_D(h) + p_0^D)$$

In addition, $w_1^{2\times2,cont}(l,h) \geq 0$ *for all* $(l,h) \in \mathcal{D}_+ \times \mathcal{D}_{h>l}$, $w_0^{2\times2,cont}(h) \geq 0$ *for all* $h \in \mathcal{D}_+$, *and* $\int_{\mathcal{D}_+}\int_{\mathcal{D},h>l} w_1^{2\times2,cont}(l,h)\,dh\,dl + \int_{\mathcal{D}_+} w_0^{2\times2,cont}(h)\,dh = 1$.

(2) *If Assumption 2-MV also holds, then*

$$\beta^{twfe} = \sum_{l\in\mathcal{D}}\sum_{h\in\mathcal{D},h>l} w^{2\times2,disc}(l,h)\frac{(m_\Delta(h) - m_\Delta(l))}{(h-l)}$$

where

$$w^{2\times2,disc}(l,h) := (h-l)^2(p_l^D + p_h^D)^2 p_{l|\{g,h\}}^D(1 - p_{l|\{g,h\}}^D)/\mathrm{var}(D)$$
$$p_{l|\{l,h\}}^D := \mathrm{P}(D = l|D \in \{l,h\})$$

and $p_h^D := \mathrm{P}(D = h)$, $p_l^D := \mathrm{P}(D = l)$. *In addition,* $w^{2\times2,disc}(l,h) \geq 0$ *for all* $(l,h) \in \mathcal{D}^2$ *and* $\sum_{l\in\mathcal{D}}\sum_{h\in\mathcal{D},h>l} w^{2\times2,disc}(l,h) = 1$.

Proof. From the proof of Proposition 4, we have that

$$\begin{aligned}
\beta &= \mathbb{E}\left[\frac{(D - \mathbb{E}[D])}{\mathrm{var}(D)}m_\Delta(D)\right]\\
&= \frac{1}{\mathrm{var}(D)}\int_{\mathcal{D}}(h - \mathbb{E}[D])m_\Delta(h)\,dF_D(h)\\
&= \frac{1}{\mathrm{var}(D)}\int_{\mathcal{D}}\left(h - \int_{\mathcal{D}} l\,dF_D(l)\right)m_\Delta(h)\,dF_D(h)\\
&= \frac{1}{\mathrm{var}(D)}\int_{\mathcal{D}}\int_{\mathcal{D}}(h - l)m_\Delta(h)\,dF_D(h)\,dF_D(l)\\
&= \frac{1}{\mathrm{var}(D)}\int_{\mathcal{D}}\int_{\mathcal{D},h>l}(h - l)(m_\Delta(h) - m_\Delta(l))\,dF_D(h)\,dF_D(l)\\
&= \frac{1}{\mathrm{var}(D)}\int_{\mathcal{D}}\int_{\mathcal{D},h>l}(h - l)^2\frac{(m_\Delta(h) - m_\Delta(l))}{(h-l)}\,dF_D(h)\,dF_D(l) \quad (9)
\end{aligned}$$

where the second equality holds by writing the expectation as an integral, the third equality write $\mathbb{E}[D]$ as an integral, the fourth equality rearranges terms, the fifth equality holds because the integrations are symmetric, and the last equality holds by multiplying and dividing by $(h - l)$.

The above arguments hold if treatment is continuous or discrete. Under Assumption 2-Cont,

$$\begin{aligned}
\text{Equation (9)} &= \frac{1}{\mathrm{var}(D)}\int_{\mathcal{D}_+}\int_{\mathcal{D},h>l}(h-l)^2\frac{(m_\Delta(h) - m_\Delta(l))}{(h-l)}f_D(h)f_D(l)\,dh\,dl\\
&+ \frac{1}{\mathrm{var}(D)}\int_{\mathcal{D},h>0}h^2\frac{m_\Delta(h) - m_\Delta(0)}{h}f_D(h)p_0^D\,dh
\end{aligned}$$

which holds by splitting up the first integral in Equation (9) by whether $l \in \mathcal{D}_+$ or $l = 0$. Then, the result for part (1) holds by multiplying and dividing the first line by $(f_D(h) + f_D(l))^2$ and by the definition $f_{D|\{h,l\}}$ and multiplying and dividing the second line by $(f_D(h) + p_0^D)^2$ and by the definitions of $f_{D|\{h,0\}}$ and $p_{0|\{h,0\}}^D$.

Under Assumption 2-MV,

$$\text{Equation (9)} = \frac{1}{\text{var}(D)} \sum_{l \in \mathcal{D}} \sum_{h \in \mathcal{D}, h > l} (h - l)^2 \frac{(m_\Delta(h) - m_\Delta(l))}{(h - l)} p_h^D p_l^D$$

$$= \frac{1}{\text{var}(D)} \sum_{l \in \mathcal{D}} \sum_{h \in \mathcal{D}, h > l} (h - l)^2 \frac{(m_\Delta(h) - m_\Delta(l))}{(h - l)} (p_l^D + p_h^D)^2 p_{l|\{g,h\}}^D (1 - p_{l|\{g,h\}}^D)$$

where the first equality holds immediately and the second equality holds by multiplying and dividing by $(p_l^D + p_h^D)^2$ and by the definition of $p_{l|\{g,h\}}^D$.

That the weights are all positive holds immediately by their definitions. That the weights integrate to one holds because

$$\int_{\mathcal{D}_+} \int_{\mathcal{D}, h > l} w_1^{2 \times 2, cont}(l, h) \, dh \, dl + \int_{\mathcal{D}_+} w_0^{2 \times 2, cont}(h) \, dh = \frac{1}{\text{var}(D)} \int_{\mathcal{D}} \int_{\mathcal{D}} \mathbf{1}\{h > l\}(h - l)^2 \, dF_D(h) \, dF_D(l)$$
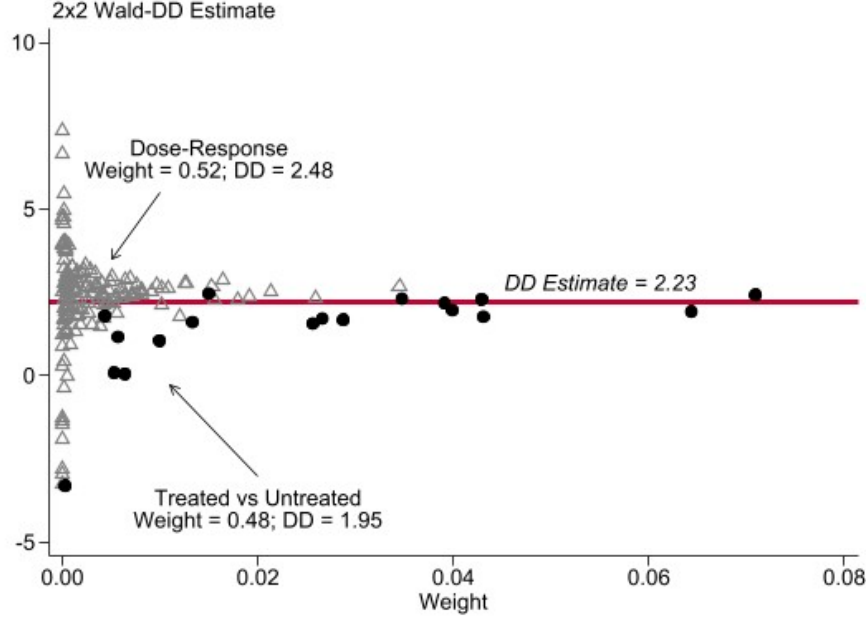
$$= 1$$

The same sort of argument holds for the discrete case as well. $\square$

Proposition 8 is analogous to the decomposition theorem for binary staggered timing designs in Goodman-Bacon (2021) in that it expresses the TWFE coefficient as a variance-weighted average of 2×2 DiD comparisons (it is also mechanically very similar to the Wald-IV theorem in Angrist (1991)). Each 2×2 term is the change in average outcomes for a group with a higher dose ($m_\Delta(h)$) minus the same difference for a group with a lower dose ($m_\Delta(l)$), divided by the difference in their doses ($h - l$). de Chaisemartin and D'Haultfœuille (2018) refer to this as a "Wald-DD" estimator. The weights combine the size of each subsample, ($p_l^D + p_h^D$), with the variance of the dose in that subsample. The variance depends on the relative size of the two groups, measured by $p_{l|\{g,h\}}^D (1 - p_{l|\{g,h\}}^D)$, and the distance between their doses, ($h - l$). This formula reflects the intuitive way researchers read a scatter plot between $m_\Delta(d)$ and d: each $\left(\frac{m_\Delta(h) - m_\Delta(l)}{k - l}\right)$ is the slope of a line connecting two points and large groups and groups with very different doses (i.e., far apart on the x-axis) have the most influence on the slope.

Using the same simulated data as in Section 4, Figure 10 represents the decomposition result for β^{twfe} in a different way by plotting each 2×2 Wald-DiD against its weight as in Figure 6 of Goodman-Bacon (2021). Comparisons between each treated group and the untreated group are in black circles and comparisons between two treated groups are in gray triangles. With K non-zero doses and some untreated units there are $(K + 1)K/2$ Wald-DiD comparisons in Proposition 8. With 18 non-zero doses our example has 171 Wald-DiD terms. Because the untreated group is so large (a quarter of the sample), comparisons to the untreated group get about half the weight in this example even though there are just 18 of them, one for each observed dose.

Next, we consider a decomposition that is based on $(m_\Delta(d) - m_\Delta(0))/d$. Under, for example, Assumption 5, this expression is equal to $ATE(d)/d$ which is an alternative way to define an average causal response.

Figure 10: **Baseline Case Decomposition: Two-Way Fixed Effects Estimator as a Weighted Average of Wald-DiDs**

Notes: The figure plots each 2×2 Wald DiD estimate against its weight from Proposition 8. Closed black circles are comparisons between one dose group and untreated observations: $(\Delta \bar{Y}_h - \Delta \bar{Y}_0)/h$. Open gray triangles are comparisons between two dose groups: $(\Delta \bar{Y}_h - \Delta \bar{Y}_\ell)/(h - \ell)$. The weights are proportional to the share of observations in each subsample $(n_h + n_\ell)^2$ and the variance of the dose in each subsample. The variance of the dose equals the relative size of the two groups $(n_{h\ell}(1 - n_{h\ell}))$, and the square of the distance between their doses $(h - \ell)^2$.

Proposition 9. *Consider β^{twfe} in Equation* (1) *and suppose that Assumption* 1 *holds.*

(1) If Assumption 2-Cont also holds, then

$$\beta^{twfe} = \int_{\mathcal{D}_+} w_1^{alt\text{-}acr,cont}(l) \frac{(m_\Delta(l) - m_\Delta(0))}{l} \, dl$$

where

$$w_1^{alt\text{-}acr,cont}(l) := \frac{(l - \mathbb{E}[D])l}{\text{var}(D)} f_D(l)$$

In addition, $\int_{\mathcal{D}} w^{alt\text{-}acr,cont}(l) \, dl = 1$, but $w^{alt\text{-}acr,cont}(l)$ can be negative for some values of $l \in \mathcal{D}$.

(2) If Assumption 2-MV also holds, then

$$\beta^{twfe} = \sum_{l \in \mathcal{D}_+} w^{alt\text{-}acr,disc} \frac{(m_\Delta(l) - m_\Delta(0))}{l}$$

where

$$w^{alt\text{-}acr,disc}(l) := \frac{(l - \mathbb{E}[D])l}{\mathrm{var}(D)} p_l^D$$

In addition, $\sum_{l \in \mathcal{D}_+} w^{alt\text{-}acr,disc}(l) = 1$, but $w^{alt\text{-}acr,disc}$ can be negative for some values of $l \in \mathcal{D}$.

Proof. From the proof of Proposition 4, we have that

$$
\begin{aligned}
\beta^{twfe} &= \frac{\mathrm{P}(D > 0)}{\mathrm{var}(D)} \mathbb{E}[(D - \mathbb{E}[D])(m_\Delta(D) - m_\Delta(0))|D > 0] \\
&= \frac{\mathrm{P}(D > 0)}{\mathrm{var}(D)} \int_{\mathcal{D}_+} (l - \mathbb{E}[D])(m_\Delta(l) - m_\Delta(0)) \, dF_{D|D>0}(l) \\
&= \frac{\mathrm{P}(D > 0)}{\mathrm{var}(D)} \int_{\mathcal{D}_+} (l - \mathbb{E}[D]) l \frac{(m_\Delta(l) - m_\Delta(0))}{l} \, dF_{D|D>0}(l) \\
&= \frac{1}{\mathrm{var}(D)} \int_{\mathcal{D}_+} (l - \mathbb{E}[D]) l \frac{(m_\Delta(l) - m_\Delta(0))}{l} f_D(l) \, dl \\
&= \int_{\mathcal{D}_+} w^{alt\text{-}acr,\mathrm{cont}}(l) \frac{(m_\Delta(l) - m_\Delta(0))}{l} \, dl
\end{aligned}
$$

where the second equality holds by writing the expectation as an integral, the third equality holds by multiplying and dividing by l, the fourth equality holds under Assumption 2-Cont, and the last equality holds by the definition of $w^{alt\text{-}acr,\mathrm{cont}}$.

For part (2), the first three equalities above continue to hold. The fourth equality replaces the integral with a summation and $f_D(l)$ with p_l^D; then the result holds by the definition of $w^{alt\text{-}acr,\mathrm{disc}}$.

In both cases, the weights can be negative because it is possible that $l < \mathbb{E}[D]$ for some values of $l \in \mathcal{D}_+$. That the weights integrate to 1 holds because

$$
\begin{aligned}
\int_{\mathcal{D}_+} w^{alt\text{-}acr,\mathrm{cont}}(l) \, dl &= \left(\int_{\mathcal{D}_+} (l - \mathbb{E}[D]) l \, dF_D(l) + (0 - \mathbb{E}[D]) 0 P(D = 0) \right) \Big/ \mathrm{var}(D) \\
&= \left(\int_{\mathcal{D}} (l - \mathbb{E}[D]) l \, dF_D(l) \right) \Big/ \mathrm{var}(D) \\
&= 1
\end{aligned}
$$

An analogous argument applies for $w^{alt\text{-}acr,\mathrm{disc}}$. $\qquad \square$

Finally, we provide a decomposition in terms of levels of paths of outcomes: $m_\Delta(d) - m_\Delta(0)$.

Proposition 10. *Consider β^{twfe} in Equation (1) and suppose that Assumption 1 holds.*

(1) If Assumption 2-Cont also holds, then

$$\beta^{twfe} = \int_{\mathcal{D}} w^{levels,cont}(l)(m_\Delta(l) - m_\Delta(0)) \, dl$$

where

$$w^{levels,cont}(l) := \frac{(l - \mathbb{E}[D])}{\mathrm{var}(D)} f_D(l)$$

In addition, $\int_{\mathcal{D}} w^{levels,cont}(l)\, dl = 0$, and $w^{levels,cont}(l)$ can be negative for some values of $l \in \mathcal{D}$.

(2) If Assumption 2-MV also holds, then

$$\beta^{twfe} = \sum_{l \in \mathcal{D}_+} w^{levels,disc}(m_\Delta(l) - m_\Delta(0))$$

where

$$w^{levels,disc}(l) := \frac{(l - \mathbb{E}[D])}{\text{var}(D)} p_l^D$$

In addition, $\sum_{l \in \mathcal{D}_+} w^{levels,disc}(l) = 0$, but $w^{levels,disc}$ can be negative for some values of $l \in \mathcal{D}$.

Proof. From the proof of Proposition 4, we have that

$$\begin{aligned}
\beta^{twfe} &= \frac{\text{P}(D > 0)}{\text{var}(D)} \mathbb{E}[(D - \mathbb{E}[D])(m_\Delta(D) - m_\Delta(0))|D > 0] \\
&= \frac{\text{P}(D > 0)}{\text{var}(D)} \int_{\mathcal{D}_+} (l - \mathbb{E}[D])(m_\Delta(l) - m_\Delta(0))\, dF_{D|D>0}(l) \\
&= \frac{1}{\text{var}(D)} \int_{\mathcal{D}_+} (l - \mathbb{E}[D])(m_\Delta(l) - m_\Delta(0)) f_D(l)\, dl \\
&= \int_{\mathcal{D}_+} w^{levels,cont}(l)(m_\Delta(l) - m_\Delta(0))\, dl
\end{aligned}$$

where the second equality holds by writing the expectation as an integral, the third equality holds under Assumption 2-Cont, and the last equality holds by the definition of $w^{levels,cont}$.

For part (2), the first two equalities above continue to hold. For the third equality, replace the integral with a summation and $f_D(l)$ with p_l^D; then the result holds by the definition of $w^{levels,disc}$.

In both cases, the weights can be negative since l can be less than $\mathbb{E}[D]$. That the weights integrate to 0 holds because

$$\begin{aligned}
\int_{\mathcal{D}_+} w^{levels,cont}(l)\, dl &= \left(\int_{\mathcal{D}_+} (l - \mathbb{E}[D])\, dF_D(l) + (0 - \mathbb{E}[D])0\text{P}(D = 0) \right) \Big/ \text{var}(D) \\
&= \left(\int_{\mathcal{D}} (l - \mathbb{E}[D])\, dF_D(l) \right) \Big/ \text{var}(D) \\
&= (\mathbb{E}[D] - \mathbb{E}[D])/\text{var}(D) \\
&= 0
\end{aligned}$$

An analogous argument applies for $w^{levels,disc}$. $\qquad\square$

Proposition 10 suggests that it would be inappropriate to interpret β^{twfe} as approximating the level effect of the dose.

C Additional Details for Multiple Periods and Variation in Treatment Timing

In this section, we consider alternative identifying assumptions for treatment effect parameters of interest in the case with multiple periods, variation in treatment timing, and where the dose can

vary across units.

As in the baseline case with two time periods, identifying ATT parameters involves untreated groups that serve as a valid counterfactual for treated groups. We first define a parallel trend assumption similar to Assumption 4 whose parts correspond to different comparison groups and time periods where one may believe that parallel trends in untreated potential outcomes holds.

Assumption 4-MP (Parallel Trends with Multiple Periods and Variation in Treatment Timing).

(a) For all $g \in \mathcal{G}$, $t = 2, \ldots, \mathcal{T}$, $d \in \mathcal{D}$, $\mathbb{E}[\Delta Y_t(0)|G = g, D = d] = \mathbb{E}[\Delta Y_t(0)|D = 0]$

(b) For all $g \in \mathcal{G}$, $t = g, \ldots, \mathcal{T}$, $d \in \mathcal{D}$, $\mathbb{E}[\Delta Y_t(0)|G = g, D = d] = \mathbb{E}[\Delta Y_t(0)|D = 0]$

(c) For all $g \in \mathcal{G}$, $t = g, \ldots, \mathcal{T}$, $d \in \mathcal{D}$, $\mathbb{E}[\Delta Y_t(0)|G = g, D = d] = \mathbb{E}[\Delta Y_t(0)|G = k]$ for all groups $k \in \mathcal{G}$ such that $t < k$ (i.e., pre-treatment periods for group k).

Assumption 4-MP(a) is the strongest assumption about paths of untreated potential outcomes. It says that paths of untreated potential outcomes are the same for all groups and for all doses across all time periods. Assumption 4-MP(b) says that the path of outcomes for group g in post treatment time periods is the same as the path of untreated potential outcomes among never-treated units. Parallel pre-trends need not hold under part (b). Assumption 4-MP(c) says that the path of outcomes for group g in post treatment time periods is the same as the path of outcomes among all groups that are not treated yet in that period — this includes both the untreated group as well as groups that will eventually be treated but that are not treated yet. Based on the results in earlier sections, note that each parallel trends assumption in Assumption 4-MP is directed towards identifying $ATT(g, t, d|g, d)$ rather than $ATE(g, t, d)$.

Next, we provide an analogous set of assumptions that target identifying $ATE(g, t, d)$.

Assumption 5-MP-Extended (Strong Parallel Trends with Multiple Periods and Variation in Treatment Timing).

(a) For all $g \in \mathcal{G}$, $t = 2, \ldots, \mathcal{T}$, and $d \in \mathcal{D}$, $\mathbb{E}[Y_t(g, d) - Y_{t-1}(0)|G = g, D = d] = \mathbb{E}[Y_t(g, d) - Y_{t-1}(0)|G = g]$ and $\mathbb{E}[\Delta Y_t(0)|G = g, D = d] = \mathbb{E}[\Delta Y_t(0)|D = 0]$

(b) For all $g \in \mathcal{G}$, $t = g, \ldots, \mathcal{T}$, $d \in \mathcal{D}$, $\mathbb{E}[Y_t(g, d) - Y_{t-1}(0)|G = g, D = d] = \mathbb{E}[Y_t(g, d) - Y_{t-1}(0)|G = g]$ and $\mathbb{E}[\Delta Y_t(0)|G = g, D = d] = \mathbb{E}[\Delta Y_t(0)|D = 0]$

(c) For all $g \in \mathcal{G}$, $t = g, \ldots, \mathcal{T}$, $d \in \mathcal{D}$, $\mathbb{E}[Y_t(g, d) - Y_{t-1}(0)|G = g, D = d] = \mathbb{E}[Y_t(g, d) - Y_{t-1}(0)|G = g]$ and $\mathbb{E}[\Delta Y_t(0)|G = g, D = d] = \mathbb{E}[\Delta Y_t(0)|G = k]$ for all groups $k \in \mathcal{G}$ such that $t < k$ (i.e., pre-treatment periods for group k).

Parts (a), (b), and (c) of the assumption correspond to the same parts in Assumption 4-MP and differ based on which group is used as the comparison group in terms of untreated potential outcomes. Part (a) additionally corresponds to Assumption 5-MP in the main text. Finally, the reason that there are two parts to these assumptions rather than just one as in Assumption 4-MP is that, in the setup of this section, conditional on being in group g with $t \geq g$, there are no untreated units in the group; thus, the second part of the assumption handles untreated potential outcome slightly differently than treated potential outcomes.

Theorem 6. *Under Assumptions 1-MP, 2-MP(a), and 3-MP, and for all $g \in \mathcal{G}$, $t = 2, \ldots, \mathcal{T}$ such that $t \geq g$, and for all $d \in \mathcal{D}$,*

(1a) If, in addition, either Assumption 4-MP(a) or (c) holds, then

$$ATT(g,t,d|g,d) = \mathbb{E}[Y_t - Y_{g-1}|G = g, D = d] - \mathbb{E}[Y_t - Y_{g-1}|W_t = 0]$$

(1b) If, in addition, Assumption 4-MP(b) holds, then

$$ATT(g,t,d|g,d) = \mathbb{E}[Y_t - Y_{g-1}|G = g, D = d] - \mathbb{E}[Y_t - Y_{g-1}|D = 0]$$

(2a) If, in addition, either Assumption 5-MP-Extended(a) or (c) holds, then

$$ATE(g,t,d) = \mathbb{E}[Y_t - Y_{g-1}|G = g, D = d] - \mathbb{E}[Y_t - Y_{g-1}|W_t = 0]$$

(2b) If, in addition, Assumption 5-MP-Extended(b) holds, then

$$ATE(g,t,d) = \mathbb{E}[Y_t - Y_{g-1}|G = g, D = d] - \mathbb{E}[Y_t - Y_{g-1}|D = 0]$$

The proof of Theorem 6 is provided in Appendix E. Part (1a) of Theorem 6 says that $ATT(g,t,d|g,d)$ — the average effect of participating in the treatment in time period t among units who became treated in period g and experienced dose d — is identified under a parallel trends assumption and that it is equal to the average path of outcomes experienced by units in group g under dose d adjusted by the average path of outcomes experienced among units that are not-yet-treated by period t. The results in the other parts are similar as well. For part (1b), the weaker parallel trends assumption in Assumption 4-MP(b) implies that the never-treated group should be used as the comparison group (this is a smaller comparison group relative to the not-yet-treated group). Parts (2a) and (2b) show that under Assumption 5-MP-Extended the same estimands identify $ATE(g,t,d)$.

Remark 8. *The parallel trends assumptions in Assumption 4-MP are not the only possible ones. Interestingly, with a multi-valued/continuous treatment, there are some possible (and reasonable) comparison groups that are available that are not available with a binary treatment. For example, one could assume that*

For all $g \in \mathcal{G}$, $t = g, \ldots, \mathcal{T}$, $d \in \mathcal{D}$, $\mathbb{E}[\Delta Y_t(0)|G = g, D = d] = \mathbb{E}[\Delta Y_t(0)|G = k, D = d]$ for all groups $k \in \mathcal{G}$ such that $t < k$ (i.e., pre-treatment periods for group k).

This sort of assumption amounts to using as a comparison group the set of units that are not yet treated but will eventually experience the same dose. It is straightforward to adapt the approach described in Theorem 6 to this sort of case and propose related estimators that can deliver consistent estimates of $ATT(g,t,d|g,d)$.

Remark 9. *If a researcher is interested in targeting a particular $ATT(g,t,d|g,d)$ or $ATE(g,t,d)$, it is generally possible to weaken Assumption 4-MP or 5-MP-Extended. For example, one could make parallel trends directly about long differences, $(Y_t - Y_{g-1})$, rather than all short differences (this sort of assumption is generally weaker), or, in part (c) of each assumption, use more aggregated comparison groups instead of imposing parallel trends for all possible comparison groups (which is also weaker), or alternatively only make parallel trends assumptions for the particular dose being considered.*

D   Proofs

D.1   Proofs of Results in Section 3.3

This section contains the proofs of the results in Section 3.3 on identifying $ATT(d|d)$ and $ATE(d)$ under parallel trends assumptions and with a multi-valued/continuous treatment.

Proof of Theorem 1

Proof. To show the result, notice that

$$
\begin{aligned}
ATT(d|d) &= \mathbb{E}[Y_t(d) - Y_t(0)|D = d] \\
&= \mathbb{E}[Y_t(d) - Y_{t-1}(0)|D = d] - \mathbb{E}[Y_t(0) - Y_{t-1}(0)|D = d] \\
&= \mathbb{E}[Y_t(d) - Y_{t-1}(0)|D = d] - \mathbb{E}[Y_t(0) - Y_{t-1}(0)|D = 0] \\
&= \mathbb{E}[\Delta Y_t|D = d] - \mathbb{E}[\Delta Y_t|D = 0]
\end{aligned}
$$

where the second equality holds by adding and subtracting $\mathbb{E}[Y_{t-1}(0)|D = d]$, the third equality holds by Assumption 4, and the last equality holds because $Y_t(d)$ and $Y_{t-1}(0)$ are observed potential outcomes when $D = d$ and $Y_t(0)$ and $Y_{t-1}(0)$ are observed potential outcomes when $D = 0$.   □

Proof of Proposition 1

Proof. To show the result, notice that

$$
\begin{aligned}
ATE(d) &= \mathbb{E}[Y_t(d) - Y_t(0)] \\
&= \mathbb{E}\Big[\mathbb{E}[Y_t(d) - Y_t(0)|D]\Big] \\
&= \int_{\mathcal{D}} ATT(d|l)\, dF_D(l)
\end{aligned}
$$

where the second equality holds by the law of iterated expectations, and the third equality holds by the definition of $ATT(d|l)$. Then, the result holds because $ATT(d|l)$ is only identified under Assumption 4 when $d = l$.   □

Proof of Theorem 2

Proof. Notice that

$$
\begin{aligned}
ATE(d) &= \mathbb{E}[Y_t(d) - Y_t(0)] \\
&= \mathbb{E}[Y_t(d) - Y_{t-1}(0)] - \mathbb{E}[Y_t(0) - Y_{t-1}(0)] \\
&= \mathbb{E}[Y_t(d) - Y_{t-1}(0)|D = d] - \mathbb{E}[Y_t(0) - Y_{t-1}(0)|D = 0] \\
&= \mathbb{E}[\Delta Y_t|D = d] - \mathbb{E}[\Delta Y_t|D = 0]
\end{aligned}
$$

where the second equality holds by adding and subtracting $\mathbb{E}[Y_{t-1}(0)]$, the third equality holds by Assumption 5, and the fourth equality holds because $Y_t(d)$ and $Y_{t-1}(0)$ are observed outcomes when $D = d$.   □

Proof of Proposition 2

Proof. For part (1), the result holds by adding and subtracting $ATT(d'|d)$. For part (2), the result holds immediately by the definition of $ATE(d)$.

$\square$

Proof of Proposition 3

Proof. For Equation (a-Cont), notice that, for $d \in \mathcal{D}_+$ and $(d+h) \in \mathcal{D}_+$,

$$\frac{\mathbb{E}[\Delta Y_t | D = d] - \mathbb{E}[\Delta Y_t | D = d + h]}{h} = \frac{ATT(d|d) - ATT(d+h|d+h)}{h}$$

$$= \frac{ATT(d|d) - ATT(d+h|d)}{h} + \frac{ATT(d+h|d) - ATT(d+h|d+h)}{h}$$

where the first equality holds by Theorem 1 and the second equality holds by Proposition 2. The result holds by taking the limit as $h \to 0$ and the definition of $ACRT(d|d)$.

For Equation (a-MV) and for $d_j \in \mathcal{D}_+$,

$$\mathbb{E}[\Delta Y_t | D = d_j] - \mathbb{E}[\Delta Y_t | D = d_{j-1}] = \Big(ATT(d_j|d_j) - ATT(d_{j-1}|d_j)\Big) + \Big(ATT(d_{j-1}|d_j) - ATT(d_{j-1}|d_{j-1})\Big)$$

$$= ACRT(d_j|d_j) + \Big(ATT(d_{j-1}|d_j) - ATT(d_{j-1}|d_{j-1})\Big)$$

where the first equality holds by Theorem 1 and Proposition 2 and the second equality holds by the definition of $ACRT(d_j|d_j)$.

Similarly, the result in Equation (b-Cont) holds by noting that for $d \in \mathcal{D}_+$ and $(d+h) \in \mathcal{D}_+$,

$$\frac{\mathbb{E}[\Delta Y_t | D = d] - \mathbb{E}[\Delta Y_t | D = d + h]}{h} = \frac{ATE(d) - ATE(d+h)}{h}$$

which follows from Proposition 2 and then by following the same arguments as for Equation (a-Cont).

For Equation (b-MV) and for $d_j \in \mathcal{D}_+$,

$$\mathbb{E}[\Delta Y_t | D = d_j] - \mathbb{E}[\Delta Y_t | D = d_{j-1}] = \Big(ATE(d_j) - ATT(d_{j-1})\Big)$$

which holds by Proposition 2. $\square$

D.2   Proofs of Results from Section 3.4

This section contains the proofs of the results in Section 3.4 on interpreting TWFE regressions with a multi-valued/continuous treatment.

Proof of Proposition 4

To conserve on notation, we define

$$m_\Delta(d) := \mathbb{E}[\Delta Y | D = d]$$

Proof. First, notice that Equation (1) is equivalent to

$$\Delta Y_i = (\theta_t - \theta_{t-1}) + \beta^{twfe} D_i + \Delta v_{it} \tag{10}$$

which holds by taking first differences and because all units are untreated in the first period. Therefore, it immediately follows that

$$
\begin{aligned}
\beta^{twfe} &= \frac{\mathbb{E}[\Delta Y(D - \mathbb{E}[D])]}{\mathrm{var}(D)} \\
&= \mathbb{E}\left[\frac{(D - \mathbb{E}[D])}{\mathrm{var}(D)}(m_\Delta(D) - m_\Delta(0))\right] \\
&= \mathbb{E}\left[\frac{(D - \mathbb{E}[D])}{\mathrm{var}(D)}(m_\Delta(D) - m_\Delta(0))\Big|D > 0\right]\mathrm{P}(D > 0) \\
&= \mathbb{E}\left[\frac{(D - \mathbb{E}[D])}{\mathrm{var}(D)}(m_\Delta(D) - m_\Delta(d_L))\Big|D > 0\right]\mathrm{P}(D > 0) + \mathbb{E}\left[\frac{(D - \mathbb{E}[D])}{\mathrm{var}(D)}(m_\Delta(d_L) - m_\Delta(0))\Big|D > 0\right]\mathrm{P}(D > \\
&:= A_1 + A_2
\end{aligned}
$$

where the first equality holds because Equation (10) is a simple linear regression of ΔY on an intercept and D, the second equality holds because $\mathbb{E}[(D - \mathbb{E}[D])m_\Delta(0)] = 0$, the third equality holds because $\mathbb{E}[m_\Delta(D) - m_\Delta(0)|D = 0] = 0$, and the fourth equality holds by adding and subtracting $m_\Delta(d_L)$.

We consider A_1 and A_2 separately next. First, for A_1,

$$
\begin{aligned}
A_1 &= \mathbb{E}\left[\frac{(D - \mathbb{E}[D])}{\mathrm{var}(D)}(m_\Delta(D) - m_\Delta(d_L))\Big|D > 0\right]\mathrm{P}(D > 0) \\
&= \frac{\mathrm{P}(D > 0)}{\mathrm{var}(D)}\int_{d_L}^{d_U}(k - \mathbb{E}[D])(m_\Delta(k) - m_\Delta(d_L))\,dF_{D|D>0}(k) \\
&= \frac{\mathrm{P}(D > 0)}{\mathrm{var}(D)}\int_{d_L}^{d_U}(k - \mathbb{E}[D])\int_{d_L}^{k} m'_\Delta(l)\,dl\,dF_{D|D>0}(k) \\
&= \frac{\mathrm{P}(D > 0)}{\mathrm{var}(D)}\int_{d_L}^{d_U}(k - \mathbb{E}[D])\int_{d_L}^{d_U} \mathbf{1}\{l \le k\}m'_\Delta(l)\,dl\,dF_{D|D>0}(k) \\
&= \frac{\mathrm{P}(D > 0)}{\mathrm{var}(D)}\int_{d_L}^{d_U} m'_\Delta(l)\int_{d_L}^{d_U}(k - \mathbb{E}[D])\mathbf{1}\{l \le k\}\,dF_{D|D>0}(k)\,dl \\
&= \frac{\mathrm{P}(D > 0)}{\mathrm{var}(D)}\int_{d_L}^{d_U} m'_\Delta(l)\mathbb{E}[(D - \mathbb{E}[D])\mathbf{1}\{l \le D\}|D > 0]\,dl \\
&= \frac{\mathrm{P}(D > 0)}{\mathrm{var}(D)}\int_{d_L}^{d_U} m'_\Delta(l)\mathbb{E}[(D - \mathbb{E}[D])|D \ge l]\mathrm{P}(D \ge l|D > 0)\,dl \\
&= \int_{d_L}^{d_U} m'_\Delta(l)\frac{(\mathbb{E}[D|D \ge l] - \mathbb{E}[D])\mathrm{P}(D \ge l)}{\mathrm{var}(D)}\,dl \tag{11}
\end{aligned}
$$

where the first equality is the definition of A_1, the second equality holds by rearranging terms and writing the expectation as an integral, the third equality holds by the fundamental theorem of calculus, the fourth equality rewrites the inner integral so that it is over d_U to d_L, the fifth equality holds by changing the order of integration and rearranging terms, the sixth equality holds by rewriting the inner integral as an expectation, the seventh equality holds by the law of iterated

52

expectations (and since $D \geq l \implies D > 0$), and the last equality holds by combining terms.

Next, for A_2, it immediately holds that

$$A_2 = \mathbb{E}\left[\frac{(D - \mathbb{E}[D])}{\text{var}(D)}(m_\Delta(d_L) - m_\Delta(0))|D > 0\right]\text{P}(D > 0)$$

$$= \frac{(\mathbb{E}[D|D > 0] - \mathbb{E}[D])\text{P}(D > 0)d_L}{\text{var}(D)}\frac{(m_\Delta(d_L) - m_\Delta(0))}{d_L} \tag{12}$$

where the first equality is the definition of A_2, and the second equality holds by multiplying and dividing by d_L.

Then, the first result in Proposition 4 holds by combining Equations (11) and (12). That the weights are all positive holds immediately since $(\mathbb{E}[D|D \geq l] - \mathbb{E}[D]) > 0$ for all $l \geq d_L$, $\text{P}(D \geq l) > 0$ for all $l \geq d_L$, $(\mathbb{E}[D|D > 0] - \mathbb{E}[D]) > 0$, $\text{P}(D > 0) > 0$, and $\text{var}(D) > 0$.

Next, we next show that $\int_{d_L}^{d_U} w_1(l) \, dl + w_0 = 1$. First, notice that

$$\int_{d_L}^{d_U} w_1(l) \, dl + w_0 = \frac{1}{\text{var}(D)}\left\{ \int_{d_L}^{d_U} \mathbb{E}[D|D \geq l]\text{P}(D \geq l) \, dl \right.$$

$$- \mathbb{E}[D]\int_{d_L}^{d_U} \text{P}(D \geq l) \, dl$$

$$+ \mathbb{E}[D|D > 0]\text{P}(D > 0)d_L$$

$$\left. - \mathbb{E}[D]\text{P}(D > 0)d_L \right\}$$

$$:= \frac{1}{\text{var}(D)}\left\{ B_1 - B_2 + B_3 - B_4 \right\}$$

and we consider $B_1, B_2, B_3,$ and B_4 in turn.

For B_1, first notice that for all $l \in \mathcal{D}_+$,

$$\mathbb{E}[D|D \geq l]\text{P}(D \geq l) = \mathbb{E}[D\mathbf{1}\{D \geq l\}|D \geq l]\text{P}(D \geq l)$$

$$= \mathbb{E}[D\mathbf{1}\{D \geq l\}] \tag{13}$$

which holds by the law of iterated expectations and implies that

$$B_1 = \int_{d_L}^{d_U} \mathbb{E}[D|D \geq l]\text{P}(D \geq l) \, dl$$

$$= \int_{d_L}^{d_U} \int_{\mathcal{D}} d\mathbf{1}\{d \geq l\} \, dF_D(d) \, dl$$

$$= \int_{\mathcal{D}} d\left(\int_{d_L}^{d_U} \mathbf{1}\{l \leq d\} \, dl\right) dF_D(d)$$

$$= \int_{\mathcal{D}} d(d - d_L) \, dF_D(d)$$

$$= \mathbb{E}[D^2] - \mathbb{E}[D]d_L \tag{14}$$

where the first line is the definition of B_1, the second equality holds by Equation (13), the third equality holds by changing the order of integration, the fourth equality holds by carrying out the

inner integration, and the last equality holds by rewriting the integral as an expectation.

Next, for term B_2,

$$B_2 = \mathbb{E}[D] \int_{d_L}^{d_U} \mathrm{P}(D \geq l) \, dl$$

$$= \mathbb{E}[D]\mathrm{P}(D > 0) \int_{d_L}^{d_U} \mathrm{P}(D \geq l | D > 0) \, dl$$

$$= \mathbb{E}[D]\mathrm{P}(D > 0) \int_{d_L}^{d_U} \int_{d_L}^{d_U} \mathbf{1}\{d \geq l\} \, dF_{D|D>0}(d) \, dl$$

$$= \mathbb{E}[D]\mathrm{P}(D > 0) \int_{d_L}^{d_U} \left(\int_{d_L}^{d_U} \mathbf{1}\{l \leq d\} \, dl \right) dF_{D|D>0}(d)$$

$$= \mathbb{E}[D]\mathrm{P}(D > 0) \int_{d_L}^{d_U} (d - d_L) \, dF_{D|D>0}(d)$$

$$= \mathbb{E}[D]\mathrm{P}(D > 0)\Big(\mathbb{E}[D|D > 0] - d_L\Big)$$

$$= \mathbb{E}[D]^2 - \mathbb{E}[D]\mathrm{P}(D > 0)d_L \tag{15}$$

where the first equality is the definition of B_2, the second equality holds by the law of iterated expectations, the third equality holds by writing $\mathrm{P}(D \geq l | D > 0)$ as an integral, the fourth equality changes the order of integration, the fifth equality carries out the inside integration, the sixth equality rewrites the integral as an expectation, the last equality holds by combining terms and by the law of iterated expectations.

Next,

$$B_3 = \mathbb{E}[D|D > 0]\mathrm{P}(D > 0)d_L$$

$$= \mathbb{E}[D]d_L \tag{16}$$

which holds by the law of iterated expectations. And finally, recall that

$$B_4 = \mathbb{E}[D]\mathrm{P}(D > 0)d_L \tag{17}$$

Thus, from Equations (14) to (17), it follows that

$$B_1 - B_2 + B_3 + B_4 = \mathbb{E}[D^2] - \mathbb{E}[D]^2 = \mathrm{var}(D)$$

which implies the result.

For part (2), the proof is similar as for part (1), but we provide the details here for completeness. Notice that,

$$\beta^{twfe} = \mathbb{E}\left[\frac{(D - \mathbb{E}[D])}{\mathrm{var}(D)} (m_\Delta(D) - m_\Delta(0)) \right]$$

$$= \frac{1}{\mathrm{var}(D)} \sum_{d \in \mathcal{D}} (d - \mathbb{E}[D])(m_\Delta(d) - m_\Delta(0))p_d^D$$

$$= \frac{1}{\mathrm{var}(D)} \sum_{d \in \mathcal{D}} (d - \mathbb{E}[D])p_d^D \sum_{d_j \in \mathcal{D}_+} \mathbf{1}\{d_j \leq d\}(m_\Delta(d_j) - m_\Delta(d_{j-1}))$$

$$= \frac{1}{\text{var}(D)} \sum_{d_j \in \mathcal{D}_+} (m_\Delta(d_j) - m_\Delta(d_{j-1})) \sum_{d \in \mathcal{D}} (d - \mathbb{E}[D]) \mathbf{1}\{d \ge d_j\} p_d^D$$

$$= \sum_{d_j \in \mathcal{D}_+} (m_\Delta(d_j) - m_\Delta(d_{j-1})) \frac{(\mathbb{E}[D|D \ge d_j] - \mathbb{E}[D])\mathrm{P}(D \ge d_j)}{\text{var}(D)}$$

$$= \sum_{d_j \in \mathcal{D}_+} w_1(d_j)(d_j - d_{j-1}) \frac{(m_\Delta(d_j) - m_\Delta(d_{j-1}))}{(d_j - d_{j-1})}$$

where the second equality holds by writing the expectation as a summation, the third equality holds by adding and subtracting $m_\Delta(d_j)$ for all d_j's between 0 and d, the fourth equality holds by changing the order of the summations, the fifth equality writes the second summation as an expectation, and the last equality holds by the definition of the weights and by multiplying and dividing by $(d_j - d_{j-1})$. That $w_1(d_j)(d_j - d_{j-1}) > 0$ holds immediately since $w_1(d_j) \ge 0$ for all $d_j \in \mathcal{D}_+$ and $d_j > d_{j-1}$. Further,

$$\sum_{d_j \in \mathcal{D}_+} w_1(d_j)(d_j - d_{j-1}) = \left(\sum_{d_j \in \mathcal{D}_+} \mathbb{E}[\mathbf{1}\{D \ge d_j\}D](d_j - d_{j-1}) - \mathbb{E}[D] \sum_{d_j \in \mathcal{D}_+} \mathrm{P}(D \ge d_j)(d_j - d_{j-1}) \right) / \text{var}(D)$$

$$:= (A - B)/\text{var}(D)$$

We consider each of these terms in turn:

$$A = \sum_{d_j \in \mathcal{D}_+} \sum_{d_k \in \mathcal{D}} \mathbf{1}\{d_k \ge d_j\} d_k p_{d_k}^D (d_j - d_{j-1})$$

$$= \sum_{d_k \in \mathcal{D}} p_{d_k}^D d_k \sum_{d_j \in \mathcal{D}_+, d_j \le d_k} (d_j - d_{j-1})$$

$$= \sum_{d_k \in \mathcal{D}} p_{d_k}^D d_k (d_k - 0)$$

$$= \mathbb{E}[D^2]$$

where the first equality holds by writing the expectation for Term A as a summation, the second equality holds by re-ordering the summations, the third equality holds by canceling all the duplicate d_j terms across summations (and because $d_0 = 0$), and the last equality holds by the definition of $\mathbb{E}[D^2]$.

Next,

$$B = \mathbb{E}[D] \sum_{d_j \in \mathcal{D}_+} \sum_{d_k \in \mathcal{D}} \mathbf{1}\{d_k \ge d_j\} p_{d_k}^D (d_j - d_{j-1})$$

$$= \mathbb{E}[D] \sum_{d_k \in \mathcal{D}} p_{d_k}^D \sum_{d_j \in \mathcal{D}_+, d_j \le d_k} (d_j - d_{j-1})$$

$$= \mathbb{E}[D] \sum_{d_k \in \mathcal{D}} d_k p_{d_k}^D$$

$$= \mathbb{E}[D]^2$$

where the first equality holds by writing the expectation for Term B as a summation, the second equality holds by re-ordering the summations, the third equality holds by canceling all the duplicate

55

d_j terms across summations (and because $d_0 = 0$), and the last equality holds by the definition of $\mathbb{E}[D]$.

This implies that $A - B = \text{var}(D)$ which implies that the weights sum to 1. $\qquad\square$

Proof of Theorem 3

Proof. The result holds immediately by plugging in the result in Proposition 3 into the result in Proposition 4 as well as noting that $\mathbb{E}[\Delta Y_t | D = d_L] - \mathbb{E}[\Delta Y_t | D = 0] = ATT(d_L | d_L)$ (under Assumption 4) and that $\mathbb{E}[\Delta Y_t | D = d_L] - \mathbb{E}[\Delta Y_t | D = 0] = ATE(d_L)$ (under Assumption 5). $\qquad\square$

E   Proofs of Results from Section 4

This section contains the proofs of results from Section 4 on DiD with a multi-valued/continuous treatment and with multiple periods and variation in treatment timing.

Proof of Theorems 4 and 6

This section proves Theorem 6; note that Theorem 4, in the main text, corresponds to Part (2a) of Theorem 6 (under Assumption 5-MP-Extended(a)).

For part (1a), we show the result under Assumption 4-MP(c) which is strictly weaker than Assumption 4-MP(a). First, notice that,

$$
\begin{aligned}
ATT(g,t,d|g,d) &= \mathbb{E}[Y_t(d) - Y_t(0)|G = g, D = d] \\
&= \mathbb{E}[Y_t(d) - Y_{g-1}(0)|G = g, D = d] - \mathbb{E}[Y_t(0) - Y_{g-1}(0)|G = g, D = d] \\
&= \mathbb{E}[Y_t(d) - Y_{g-1}(0)|G = g, D = d] - \sum_{s=g}^{t} \mathbb{E}[Y_s(0) - Y_{s-1}(0)|G = g, D = d] \quad (18) \\
&= \mathbb{E}[Y_t(d) - Y_{g-1}(0)|G = g, D = d] - \sum_{s=g}^{t} \mathbb{E}[Y_s(0) - Y_{s-1}(0)|W_t = 0] \\
&= \mathbb{E}[Y_t(d) - Y_{g-1}(0)|G = g, D = d] - \mathbb{E}[Y_t(0) - Y_{g-1}(0)|W_t = 0] \\
&= \mathbb{E}[Y_t - Y_{g-1}|G = g, D = d] - \mathbb{E}[Y_t - Y_{g-1}|W_t = 0]
\end{aligned}
$$

where the first equality is the definition of $ATT(g,t,d|g,d)$, the second equality holds by adding and subtracting $\mathbb{E}[Y_{g-1}(0)|G = g, D = d]$, the third equality holds by adding and subtracting $\mathbb{E}[Y_s(0)|G = g, D = d]$ for $s = g, \ldots, (t-1)$, the fourth equality holds under Assumption 4-MP(c), the fifth equality holds by canceling all the terms involving $\mathbb{E}[Y_s(0)|W_t = 0]$ for $s = g, \ldots, (t-1)$ (i.e., from the reverse of the argument for the third equality), and the last equality holds from writing the potential outcomes in terms of their observed counterparts.

For part (1b), in Equation (18), $\sum_{s=g}^{t} \mathbb{E}[Y_s(0) - Y_{s-1}(0)|G = g, D = d] = \sum_{s=g}^{t} \mathbb{E}[Y_s(0) - Y_{s-1}(0)|D = 0]$ under Assumption 4-MP(b). Then, the result holds by otherwise following the same arguments as in part (1a).

For part (2a), we show the result under Assumption 5-MP-Extended(c) which is strictly weaker

than Assumption 5-MP-Extended(a). First, notice that

$$
\begin{aligned}
ATE(g,t,d) &= \mathbb{E}[Y_t(g,d) - Y_t(0)|G=g] \\
&= \mathbb{E}[Y_t(g,d) - Y_{t-1}(0)|G=g] - \mathbb{E}[Y_t(0) - Y_{t-1}(0)|G=g] \\
&= \mathbb{E}[Y_t(g,d) - Y_{t-1}(0)|G=g, D=d] - \mathbb{E}[Y_t(0) - Y_{t-1}(0)|G=g] \\
&= \mathbb{E}[Y_t(g,d) - Y_{g-1}(0)|G=g, D=d] - \mathbb{E}[Y_{t-1}(0) - Y_{g-1}(0)|G=g, D=d] \\
&\quad - \Big(\mathbb{E}[Y_t(0) - Y_{g-1}(0)|G=g] - \mathbb{E}[Y_{t-1}(0) - Y_{g-1}(0)|G=g] \Big) \\
&= \mathbb{E}[Y_t(g,d) - Y_{g-1}(0)|G=g, D=d] - \sum_{s=g}^{t} \mathbb{E}[Y_s(0) - Y_{s-1}(0)|G=g] \quad\quad (19) \\
&\quad - \sum_{s=g}^{t-1} \Big(\mathbb{E}[Y_s(0) - Y_{s-1}(0)|G=g, D=d] - \mathbb{E}[Y_s(0) - Y_{s-1}(0)|G=g] \Big) \\
&= \mathbb{E}[Y_t(g,d) - Y_{g-1}(0)|G=g, D=d] - \sum_{s=g}^{t} \mathbb{E}[Y_s(0) - Y_{s-1}(0)|W_t=0] \\
&= \mathbb{E}[Y_t(g,d) - Y_{g-1}(0)|G=g, D=d] - \mathbb{E}[Y_t(0) - Y_{g-1}(0)|W_t=0] \\
&= \mathbb{E}[Y_t - Y_{g-1}|G=g, D=d] - \mathbb{E}[Y_t - Y_{g-1}|W_t=0]
\end{aligned}
$$

where the first equality holds by the definition of $ATE(g,t,d)$, the second equality adds and subtracts $\mathbb{E}[Y_{t-1}(0)|G=g]$, the third equality holds by Assumption 5-MP-Extended(c), the fourth equality adds and subtracts both $\mathbb{E}[Y_{g-1}(0)|G=g, D=d]$ and $\mathbb{E}[Y_{g-1}(0)|G=g]$, the fifth equality holds by writing "long differences" as summations over "short differences" and by rearranging terms, the sixth equality holds by Assumption 5-MP-Extended(c) and by canceling terms, the seventh equality holds by rewriting the sum of short differences as a long difference, and the last equality holds by writing potential outcomes in terms of their corresponding observed outcomes and is the result.

The expression for $ACR(g,t,d)$ comes from taking the partial derivative of $ATE(g,t,d) = \mathbb{E}[Y_t - Y_{g-1}|G=g, D=d] - \mathbb{E}[Y_t - Y_{g-1}|W_t=0]$ with respect to d and by noting that $\mathbb{E}[Y_t - Y_{g-1}|W_t=0]$ does not depend on d.

Finally, for part (2b), in Equation (19), $\sum_{s=g}^{t} \mathbb{E}[Y_s(0) - Y_{s-1}(0)|G=g] = \sum_{s=g}^{t} \mathbb{E}[Y_s(0) - Y_{s-1}(0)|D=0]$ under Assumption 5-MP-Extended(b). The result then follows using the same subsequent arguments as in part (2a).

E.1 Proofs of Proposition 5, Theorem 5, and Proposition 6

This section contains the proofs for interpreting TWFE regressions in the case with a continuous treatment, multiple periods, and variation in treatment timing as in Section 4.

Before proving the main results in this section, we introduce some additional notation.

$$
v(g,t) := \mathbf{1}\{t \geq g\} - \bar{G}_g \quad\quad (20)
$$

where the term $\mathbf{1}\{t \geq g\}$ is equal to one in post-treatment time periods for units in group g and recalling that we defined $\bar{G}_g = \frac{\mathcal{T} - g + 1}{\mathcal{T}}$ which is the fraction of periods that units in group g are exposed to the treatment (and notice that this latter term does not depend on the particular time period t). Further, notice that $v(g,t)$ is positive in post-treatment time periods and negative in pre-treatment time periods for units in a particular group. Finally, also note that, for the "never-

treated" group, $g = \mathcal{T} + 1$ (which we set by convention and is helpful to unify the notation in this section) so that both terms in the expression for v are equal to 0 for the "never-treated" group.

Furthermore, recall that, for $1 \leq t_1 \leq t_2 \leq \mathcal{T}$, we defined

$$\bar{Y}_i^{(t_1,t_2)} := \frac{1}{t_2 - t_1 + 1} \sum_{t=t_1}^{t_2} Y_{it}$$

where below (and following the notation used throughout the paper), we sometimes leave the subscript i implicit.

We next state and prove some additional results that are helpful for proving the main results. The first lemma re-writes (overall) expected dose experienced in period t adjusted by the overall expected dose (across periods and units) in a form that is useful in proving later results.

Lemma 1. *Under Assumptions 1-MP, 2-MP(a), and 3-MP,*

$$\mathbb{E}[W_t] - \frac{1}{\mathcal{T}} \sum_{s=1}^{\mathcal{T}} \mathbb{E}[W_s] = \sum_{g \in \mathcal{G}} \int_{\mathcal{D}} dv(g,t) \, dF_{D|G}(d|g) p_g$$

Proof. First, notice that

$$\begin{aligned}
\mathbb{E}[W_t] &= \sum_{g \in \mathcal{G}} \int_{\mathcal{D}} \mathbb{E}[W_t | G = g, D = d] \, dF_{D|G}(d|g) p_g \\
&= \sum_{g \in \mathcal{G}} \int_{\mathcal{D}} d\mathbf{1}\{t \geq g\} \, dF_{D|G}(d|g) p_g
\end{aligned} \tag{21}$$

where the first equality holds by the law of iterated expectations and the second equality holds because, after conditioning on group and dose, W_t is fully determined.

Thus,

$$\begin{aligned}
\mathbb{E}[W_t] - \frac{1}{\mathcal{T}} \sum_{s=1}^{\mathcal{T}} \mathbb{E}[W_s] &= \sum_{g \in \mathcal{G}} \int_{\mathcal{D}} d\mathbf{1}\{t \geq g\} \, dF_{D|G}(d|g) p_g - \frac{1}{\mathcal{T}} \sum_{s=1}^{\mathcal{T}} \sum_{g \in \mathcal{G}} \int_{\mathcal{D}} d\mathbf{1}\{s \geq g\} \, dF_{D|G}(d|g) p_g \\
&= \frac{1}{\mathcal{T}} \sum_{s=1}^{\mathcal{T}} \sum_{g \in \mathcal{G}} \int_{\mathcal{D}} d \left(\mathbf{1}\{t \geq g\} - \mathbf{1}\{s \geq g\} \right) dF_{D|G}(d|g) p_g \\
&= \sum_{g \in \mathcal{G}} \int_{\mathcal{D}} d \left\{ \frac{1}{\mathcal{T}} \sum_{s=1}^{\mathcal{T}} \mathbf{1}\{t \geq g\} - \mathbf{1}\{s \geq g\} \right\} dF_{D|G}(d|g) p_g \\
&= \sum_{g \in \mathcal{G}} \int_{\mathcal{D}} d \left\{ \mathbf{1}\{t \geq g\} - \frac{\mathcal{T} - g + 1}{\mathcal{T}} \right\} dF_{D|G}(d|g) p_g \\
&= \sum_{g \in \mathcal{G}} \int_{\mathcal{D}} dv(g,t) \, dF_{D|G}(d|g) p_g
\end{aligned}$$

where the first equality applies Equation (21) to both terms, the second equality combines terms by averaging the first term across time periods, the third equality re-orders the summations/integrals, the fourth equality holds because $\mathbf{1}\{t \geq g\}$ does not depend on s and by counting the fraction of periods where $s \geq g$, and the last equality holds by the definition of $v(g,t)$. $\qquad \square$

The next lemma provides an intermediate result for the expression for the numerator of β^{twfe} in Equation (2).

Lemma 2. *Under Assumptions 1-MP, 2-MP(a), and 3-MP,*

$$\frac{1}{\mathcal{T}}\sum_{t=1}^{\mathcal{T}}\mathbb{E}[Y_{it}\ddot{W}_{it}] = \frac{1}{\mathcal{T}}\sum_{t=1}^{\mathcal{T}}\left\{\sum_{g\in\mathcal{G}}\int_{\mathcal{D}}d\left(\mathbb{E}[Y_t|G=g,D=d]-\mathbb{E}[Y_t]\right)v(g,t)\,dF_{D|G}(d|g)p_g\right\}$$

Proof. Starting with the numerator for β^{twfe} in Equation (2)

$$\frac{1}{\mathcal{T}}\sum_{t=1}^{\mathcal{T}}\mathbb{E}[Y_{it}\ddot{W}_{it}]$$

$$=\frac{1}{\mathcal{T}}\sum_{t=1}^{\mathcal{T}}\left\{\mathbb{E}[Y_{it}W_{it}]-\mathbb{E}[Y_{it}\bar{W}_i]-\mathbb{E}[Y_t]\left(\mathbb{E}[W_t]-\frac{1}{\mathcal{T}}\sum_{s=1}^{\mathcal{T}}\mathbb{E}[W_s]\right)\right\}$$

$$=\frac{1}{\mathcal{T}}\sum_{t=1}^{\mathcal{T}}\left\{\mathbb{E}[Y_tD\mathbf{1}\{t\geq G\}]-\mathbb{E}\left[Y_tD\frac{\mathcal{T}-G+1}{\mathcal{T}}\right]-\mathbb{E}[Y_t]\left(\mathbb{E}[W_t]-\frac{1}{\mathcal{T}}\sum_{s=1}^{\mathcal{T}}\mathbb{E}[W_s]\right)\right\}$$

$$=\frac{1}{\mathcal{T}}\sum_{t=1}^{\mathcal{T}}\left\{\sum_{g\in\mathcal{G}}\int_{\mathcal{D}}\left(\mathbb{E}[Y_td\mathbf{1}\{t\geq g\}|G=g,D=d]-\mathbb{E}\left[Y_t\frac{\mathcal{T}-g+1}{\mathcal{T}}d\Big|G=g,D=d\right]\right)dF_{D|G}(d|g)p_g\right.$$

$$\left. -\mathbb{E}[Y_t]\left(\mathbb{E}[W_t]-\frac{1}{\mathcal{T}}\sum_{s=1}^{\mathcal{T}}\mathbb{E}[W_s]\right)\right\}$$

$$=\frac{1}{\mathcal{T}}\sum_{t=1}^{\mathcal{T}}\left\{\sum_{g\in\mathcal{G}}\int_{\mathcal{D}}d\Big(\mathbb{E}[Y_t|G=g,D=d]v(g,t)\Big)dF_{D|G}(d|g)p_g-\mathbb{E}[Y_t]\left(\mathbb{E}[W_t]-\frac{1}{\mathcal{T}}\sum_{s=1}^{\mathcal{T}}\mathbb{E}[W_s]\right)\right\}$$

$$=\frac{1}{\mathcal{T}}\sum_{t=1}^{\mathcal{T}}\left\{\sum_{g\in\mathcal{G}}\int_{\mathcal{D}}d\Big(\mathbb{E}[Y_t|G=g,D=d]v(g,t)\Big)dF_{D|G}(d|g)p_g-\mathbb{E}[Y_t]\left(\sum_{g\in\mathcal{G}}\int_{\mathcal{D}}dv(g,t)\,dF_{D|G}(d|g)p_g\right)\right\}$$

$$=\frac{1}{\mathcal{T}}\sum_{t=1}^{\mathcal{T}}\left\{\sum_{g\in\mathcal{G}}\int_{\mathcal{D}}d\left(\mathbb{E}[Y_t|G=g,D=d]-\mathbb{E}[Y_t]\right)v(g,t)\,dF_{D|G}(d|g)p_g\right\}$$

where the first equality holds by the definition of $\ddot{W}_{it}$, the second equality holds by plugging in for W_{it} and $\bar{W}_i$, the third equality holds by the law of iterated expectations, the fourth equality holds by the definition of $v(g,t)$, the fifth equality holds by Lemma 1, and the sixth equality just combines terms.

□

Next, based on the result in Lemma 2, we can write the numerator in the expression for β^{twfe} as

$$\frac{1}{\mathcal{T}}\sum_{t=1}^{\mathcal{T}}\mathbb{E}[Y_{it}\ddot{W}_{it}]$$

$$=\frac{1}{\mathcal{T}}\sum_{t=1}^{\mathcal{T}}\left\{\sum_{g\in\mathcal{G}}\int_{\mathcal{D}}d\left(\mathbb{E}[Y_t|G=g,D=d]-\mathbb{E}[Y_t]\right)v(g,t)\,dF_{D|G}(d|g)p_g\right\}$$

$$= \frac{1}{\mathcal{T}} \sum_{t=1}^{\mathcal{T}} \sum_{g \in \mathcal{G}} \int_{\mathcal{D}} d\Big(\mathbb{E}[Y_t | G = g, D = d] - \mathbb{E}[Y_t | G = g] \Big) v(g,t) \, dF_{D|G}(d|g) p_g \qquad (22)$$

$$+ \frac{1}{\mathcal{T}} \sum_{t=1}^{\mathcal{T}} \sum_{g \in \mathcal{G}} \int_{\mathcal{D}} d\Big(\mathbb{E}[Y_t | G = g] - \mathbb{E}[Y_t] \Big) v(g,t) \, dF_{D|G}(d|g) p_g \qquad (23)$$

where the first equality holds from Lemma 2 and the second equality holds by adding and subtracting $\mathbb{E}[Y_t | G = g]$.

The expression in Equation (22) involves comparisons between units in the same group but that have different doses. The expression in Equation (23) involves comparisons across different groups. We consider each of these terms in more detail below.

Lemma 3. *Under Assumptions 1-MP, 2-MP(a), and 3-MP,*

$$\frac{1}{\mathcal{T}} \sum_{t=1}^{\mathcal{T}} \sum_{g \in \mathcal{G}} \int_{\mathcal{D}} d\Big(\mathbb{E}[Y_t | G = g, D = d] - \mathbb{E}[Y_t | G = g] \Big) v(g,t) \, dF_{D|G}(d|g) p_g$$

$$= \sum_{g \in \mathcal{G}} \left\{ (1 - \bar{G}_g) \bar{G}_g \mathrm{cov} \left(\bar{Y}^{POST(g)} - \bar{Y}^{PRE(g)}, D | G = g \right) \right\} p_g$$

Proof.

$$\frac{1}{\mathcal{T}} \sum_{t=1}^{\mathcal{T}} \sum_{g \in \mathcal{G}} \int_{\mathcal{D}} d\Big(\mathbb{E}[Y_t | G = g, D = d] - \mathbb{E}[Y_t | G = g] \Big) v(g,t) \, dF_{D|G}(d|g) p_g$$

$$= \frac{1}{\mathcal{T}} \sum_{t=1}^{\mathcal{T}} \left\{ \sum_{g \in \mathcal{G}} \mathbb{E}[Y_t (D - \mathbb{E}[D | G = g]) | G = g] v(g,t) p_g \right\}$$

$$= \sum_{g \in \mathcal{G}} \left\{ \frac{1}{\mathcal{T}} \sum_{t=1}^{\mathcal{T}} \mathbb{E}[Y_t (D - \mathbb{E}[D | G = g]) | G = g] v(g,t) \right\} p_g$$

$$= \sum_{g \in \mathcal{G}} \left\{ -\frac{1}{\mathcal{T}} \frac{(T - g + 1)}{T} \sum_{t=1}^{g-1} \mathbb{E}[Y_t (D - \mathbb{E}[D | G = g]) | G = g] \right.$$

$$\left. + \frac{1}{\mathcal{T}} \frac{(g-1)}{T} \sum_{t=g}^{\mathcal{T}} \mathbb{E}[Y_t (D - \mathbb{E}[D | G = g]) | G = g] \right\} p_g$$

$$= \sum_{g \in \mathcal{G}} \left\{ \frac{g-1}{\mathcal{T}} \frac{(T - g + 1)}{T} \left(\frac{1}{\mathcal{T} - g + 1} \sum_{t=g}^{\mathcal{T}} \mathbb{E}[Y_t (D - \mathbb{E}[D | G = g]) | G = g] \right. \right.$$

$$\left. \left. - \frac{1}{g-1} \sum_{t=1}^{g-1} \mathbb{E}[Y_t (D - \mathbb{E}[D | G = g]) | G = g] \right) \right\} p_g$$

$$= \sum_{g \in \mathcal{G}} \left\{ \frac{g-1}{\mathcal{T}} \frac{(T - g + 1)}{T} \left(\mathbb{E}\big[(\bar{Y}^{POST(g)} - \bar{Y}^{PRE(g)})(D - \mathbb{E}[D | G = g]) | G = g \big] \right) \right\} p_g$$

$$= \sum_{g \in \mathcal{G}} \left\{ (1 - \bar{G}_g) \bar{G}_g \left(\mathbb{E}\big[(\bar{Y}^{POST(g)} - \bar{Y}^{PRE(g)})(D - \mathbb{E}[D | G = g]) | G = g \big] \right) \right\} p_g$$

60

$$= \sum_{g \in \mathcal{G}} \left\{ (1 - \bar{G}_g) \bar{G}_g \text{cov} \left(\bar{Y}^{POST(g)} - \bar{Y}^{PRE(g)}, D | G = g \right) \right\} p_g$$

where the first equality holds by the law of iterated expectations (and combining terms involving d and Y_t), the second equality changes the order of the summations, the third equality holds by splitting the summation involving t in time period g and plugs in for $v(g,t)$ (which is constant within group g and across time periods from $1, \ldots, g-1$ and from $g, \ldots, \mathcal{T}$), the fourth equality multiplies and divides by terms so that the inside expressions can be written as averages, the fifth equality holds by changing the order of the expectation and averaging over time periods, the sixth equality holds by the definition of $\bar{G}_g$, and the last equality holds by the definition of covariance. $\square$

Lemma 3 shows that part of the TWFE estimator comes from a weighted average of post- vs. pre-treatment outcomes within group but who experienced different doses. In particular, notice that, for units in group g, $\bar{Y}_i^{POST(g)}$ is their average post-treatment outcome while $\bar{Y}_i^{PRE(g)}$ is their average pre-treatment outcome.

Next, we consider the expression from Equation (23) above which arises from differences in outcomes across groups. We handle this term over several following results.

Lemma 4. *Under Assumptions 1-MP, 2-MP(a), and 3-MP,*

$$\frac{1}{\mathcal{T}} \sum_{t=1}^{\mathcal{T}} \left\{ \sum_{g \in \mathcal{G}} \int_{\mathcal{D}} d \Big(\mathbb{E}[Y_t | G = g] - \mathbb{E}[Y_t] \Big) v(g,t) \, dF_{D|G}(d|g) p_g \right\}$$

$$= \frac{1}{\mathcal{T}} \sum_{t=1}^{\mathcal{T}} \left\{ \sum_{g \in \mathcal{G}} \sum_{k \in \mathcal{G}, k > g} \Big(\mathbb{E}[D | G = g] v(g,t) - \mathbb{E}[D | G = k] v(k,t) \Big) \Big(\mathbb{E}[Y_t | G = g] - \mathbb{E}[Y_t | G = k] \Big) p_k p_g \right\}$$

Proof. Notice that

$$\frac{1}{\mathcal{T}} \sum_{t=1}^{\mathcal{T}} \left\{ \sum_{g \in \mathcal{G}} \int_{\mathcal{D}} d \Big(\mathbb{E}[Y_t | G = g] - \mathbb{E}[Y_t] \Big) v(g,t) \, dF_{D|G}(d|g) p_g \right\}$$

$$= \frac{1}{\mathcal{T}} \sum_{t=1}^{\mathcal{T}} \left\{ \sum_{g \in \mathcal{G}} \mathbb{E}[D | G = g] \Big(\mathbb{E}[Y_t | G = g] - \mathbb{E}[Y_t] \Big) v(g,t) p_g \right\}$$

$$= \frac{1}{\mathcal{T}} \sum_{t=1}^{\mathcal{T}} \left\{ \sum_{g \in \mathcal{G}} \mathbb{E}[D | G = g] \Big(\mathbb{E}[Y_t | G = g] - \sum_{k \in \mathcal{G}} \mathbb{E}[Y_t | G = k] p_k \Big) v(g,t) p_g \right\}$$

$$= \frac{1}{\mathcal{T}} \sum_{t=1}^{\mathcal{T}} \left\{ \sum_{g \in \mathcal{G}} \sum_{k \in \mathcal{G}} \mathbb{E}[D | G = g] v(g,t) \Big(\mathbb{E}[Y_t | G = g] - \mathbb{E}[Y_t | G = k] \Big) p_k p_g \right\}$$

$$= \frac{1}{\mathcal{T}} \sum_{t=1}^{\mathcal{T}} \left\{ \sum_{g \in \mathcal{G}} \sum_{k \in \mathcal{G}, k > g} \Big(\mathbb{E}[D | G = g] v(g,t) - \mathbb{E}[D | G = k] v(k,t) \Big) \Big(\mathbb{E}[Y_t | G = g] - \mathbb{E}[Y_t | G = k] \Big) p_k p_g \right\}$$

where the first equality holds by integrating over $\mathcal{D}$, the second equality holds by the law of iterated expectations, the third equality holds by combining terms, and the last equality holds because all combinations of g and k occur twice. $\square$

Lemma 4 is helpful because it shows that the cross-group part of the TWFE estimator can be written as comparisons for each group relative to later-treated groups.

Next, we provide an important intermediate result. Before stating this result, we define the following weights

$$\tilde{w}^{g,within}(g) := \text{var}(D|G=g)(1-\bar{G}_g)\bar{G}_g p_g \qquad \tilde{w}^{g,post}(g,k) := \mathbb{E}[D|G=g]^2(1-\bar{G}_g)(\bar{G}_g-\bar{G}_k)p_k p_g$$

$$\tilde{w}^{k,post}(g,k) := \mathbb{E}[D|G=k]^2 \bar{G}_k(\bar{G}_g-\bar{G}_k)p_k p_g$$

$$\tilde{w}^{long}(g,k) := (\mathbb{E}[D|G=g]-\mathbb{E}[D|G=k])^2 \bar{G}_k(1-\bar{G}_g)p_k p_g$$

which correspond to $w^{g,post}$, $w^{k,post}$, and $w^{long}(g,k)$ in the main text except they do not divide by $\mathcal{T}^{-1}\sum_{t=1}^{\mathcal{T}}\mathbb{E}[\ddot{W}_{it}^2]$. In addition, notice that

$$\mathbb{E}[D|G=g]v(g,t) - \mathbb{E}[D|G=k]v(k,t)$$

$$= \begin{cases} -\mathbb{E}[D|G=g]\bar{G}_g + \mathbb{E}[D|G=k]\bar{G}_k & \text{for } t < g < k \\ \mathbb{E}[D|G=g](1-\bar{G}_g) + \mathbb{E}[D|G=k]\bar{G}_k & \text{for } g \leq t < k \\ \mathbb{E}[D|G=g](1-\bar{G}_g) - \mathbb{E}[D|G=k](1-\bar{G}_k) & \text{for } g < k \leq t \end{cases} \tag{24}$$

which holds by the definition of v and is useful for the proof of the following lemma.

Lemma 5. *Under Assumptions 1-MP, 2-MP(a), and 3-MP,*

$$\frac{1}{\mathcal{T}}\sum_{t=1}^{\mathcal{T}}\left\{\sum_{g\in\mathcal{G}}\int_{\mathcal{D}} d\Big(\mathbb{E}[Y_t|G=g]-\mathbb{E}[Y_t]\Big)v(g,t)\,dF_{D|G}(d|g)p_g\right\}$$

$$= \sum_{g\in\mathcal{G}}\sum_{k\in\mathcal{G},k>g}\left\{\tilde{w}^{g,post}(g,k)\left(\mathbb{E}\left[(\bar{Y}^{MID(g,k)}-\bar{Y}^{PRE(g)})|G=g\right]-\mathbb{E}\left[(\bar{Y}^{MID(g,k)}-\bar{Y}^{PRE(g)})|G=k\right]\right)\right.$$

$$+ \tilde{w}^{k,post}(g,k)\left(\mathbb{E}\left[(\bar{Y}^{POST(k)}-\bar{Y}^{MID(g,k)})|G=k\right]-\mathbb{E}\left[(\bar{Y}^{POST(k)}-\bar{Y}^{MID(g,k)})|G=g\right]\right)$$

$$\left.+ \tilde{w}^{long}(g,k)\left(\mathbb{E}\left[(\bar{Y}^{POST(k)}-\bar{Y}^{PRE(g)})|G=g\right]-\mathbb{E}\left[(\bar{Y}^{POST(k)}-\bar{Y}^{PRE(g)})|G=k\right]\right)\right\}$$

Proof. The result holds as follows

$$\frac{1}{\mathcal{T}}\sum_{t=1}^{\mathcal{T}}\left\{\sum_{g\in\mathcal{G}}\int_{\mathcal{D}} d\Big(\mathbb{E}[Y_t|G=g]-\mathbb{E}[Y_t]\Big)v(g,t)\,dF_{D|G}(d|g)p_g\right\}$$

$$= \sum_{g\in\mathcal{G}}\sum_{k\in\mathcal{G},k>g}\left\{\frac{1}{\mathcal{T}}\sum_{t=1}^{\mathcal{T}}\Big(\mathbb{E}[D|G=g]v(g,t)-\mathbb{E}[D|G=k]v(k,t)\Big)\Big(\mathbb{E}[Y_t|G=g]-\mathbb{E}[Y_t|G=k]\Big)\right\}p_k p_g$$

$$= \sum_{g\in\mathcal{G}}\sum_{k\in\mathcal{G},k>g}\left\{\frac{1}{\mathcal{T}}\Big(-\mathbb{E}[D|G=g]\bar{G}_g+\mathbb{E}[D|G=k]\bar{G}_k\Big)\sum_{t=1}^{g-1}\Big(\mathbb{E}[Y_t|G=g]-\mathbb{E}[Y_t|G=k]\Big)\right.$$

$$+ \frac{1}{\mathcal{T}}\Big(\mathbb{E}[D|G=g](1-\bar{G}_g)+\mathbb{E}[D|G=k]\bar{G}_k\Big)\sum_{t=g}^{k-1}\Big(\mathbb{E}[Y_t|G=g]-\mathbb{E}[Y_t|G=k]\Big)$$

$$\left.+ \frac{1}{\mathcal{T}}\Big(\mathbb{E}[D|G=g](1-\bar{G}_g)-\mathbb{E}[D|G=k](1-\bar{G}_k)\Big)\sum_{t=k}^{\mathcal{T}}\Big(\mathbb{E}[Y_t|G=g]-\mathbb{E}[Y_t|G=k]\Big)\right\}p_k p_g$$

$$= \sum_{g\in\mathcal{G}}\sum_{k\in\mathcal{G},k>g}\left\{(1-\bar{G}_g)\left(-\mathbb{E}[D|G=g]\bar{G}_g+\mathbb{E}[D|G=k]\bar{G}_k\right)\left(\mathbb{E}[\bar{Y}^{PRE(g)}|G=g]-\mathbb{E}[\bar{Y}^{PRE(g)}|G=k]\right)\right.$$

$$\left.+ (\bar{G}_g-\bar{G}_k)\left(\mathbb{E}[D|G=g](1-\bar{G}_g)+\mathbb{E}[D|G=k]\bar{G}_k\right)\left(\mathbb{E}[\bar{Y}^{MID(g,k)}|G=g]-\mathbb{E}[\bar{Y}^{MID(g,k)}|G=k]\right)\right.$$

62

$$+ \bar{G}_k \left(\mathbb{E}[D|G=g](1-\bar{G}_g) - \mathbb{E}[D|G=k](1-\bar{G}_k) \right) \left(\mathbb{E}[\bar{Y}^{POST(k)}|G=g] - \mathbb{E}[\bar{Y}^{POST(k)}|G=k] \right) \Bigg\} p_k p_g$$

$$= \sum_{g \in \mathcal{G}} \sum_{k \in \mathcal{G}, k>g} \Bigg\{ (1-\bar{G}_g) \left(-\mathbb{E}[D|G=g](\bar{G}_g - \bar{G}_k) + (\mathbb{E}[D|G=k] - \mathbb{E}[D|G=g])\bar{G}_k \right) \left(\mathbb{E}[\bar{Y}^{PRE(g)}|G=g] - \mathbb{E}[\bar{Y}^{PRE(g)}|G=k] \right)$$

$$+ (\bar{G}_g - \bar{G}_k) \left(\mathbb{E}[D|G=g](1-\bar{G}_g) + \mathbb{E}[D|G=k]\bar{G}_k \right) \left(\mathbb{E}[\bar{Y}^{MID(g,k)}|G=g] - \mathbb{E}[\bar{Y}^{MID(g,k)}|G=k] \right)$$

$$+ \bar{G}_k \left((\mathbb{E}[D|G=g] - \mathbb{E}[D|G=k])(1-\bar{G}_g) - \mathbb{E}[D|G=k](\bar{G}_g - \bar{G}_k) \right) \left(\mathbb{E}[\bar{Y}^{POST(k)}|G=g] - \mathbb{E}[\bar{Y}^{POST(k)}|G=k] \right) \Bigg\} p_k p_g$$

$$= \sum_{g \in \mathcal{G}} \sum_{k \in \mathcal{G}, k>g} \Bigg\{ \mathbb{E}[D|G=g](1-\bar{G}_g)(\bar{G}_g - \bar{G}_k) \left(\mathbb{E}\left[(\bar{Y}^{MID(g,k)} - \bar{Y}^{PRE(g)})|G=g \right] - \mathbb{E}\left[(\bar{Y}^{MID(g,k)} - \bar{Y}^{PRE(g)})|G=k \right] \right)$$

$$+ \mathbb{E}[D|G=k]\bar{G}_k(\bar{G}_g - \bar{G}_k) \left(\mathbb{E}\left[(\bar{Y}^{POST(k)} - \bar{Y}^{MID(g,k)})|G=k \right] - \mathbb{E}\left[(\bar{Y}^{POST(k)} - \bar{Y}^{MID(g,k)})|G=g \right] \right)$$

$$+ (\mathbb{E}[D|G=g] - \mathbb{E}[D|G=k])\bar{G}_k(1-\bar{G}_g) \left(\mathbb{E}\left[(\bar{Y}^{POST(k)} - \bar{Y}^{PRE(g)})|G=g \right] - \mathbb{E}\left[(\bar{Y}^{POST(k)} - \bar{Y}^{PRE(g)})|G=k \right] \right) \Bigg\} p_k p_g$$

$$= \sum_{g \in \mathcal{G}} \sum_{k \in \mathcal{G}, k>g} \Bigg\{ \tilde{w}^{g,post}(g,k) \left(\mathbb{E}\left[(\bar{Y}^{MID(g,k)} - \bar{Y}^{PRE(g)})|G=g \right] - \mathbb{E}\left[(\bar{Y}^{MID(g,k)} - \bar{Y}^{PRE(g)})|G=k \right] \right)$$

$$+ \tilde{w}^{k,post}(g,k) \left(\mathbb{E}\left[(\bar{Y}^{POST(k)} - \bar{Y}^{MID(g,k)})|G=k \right] - \mathbb{E}\left[(\bar{Y}^{POST(k)} - \bar{Y}^{MID(g,k)})|G=g \right] \right)$$

$$+ \tilde{w}^{long}(g,k) \left(\mathbb{E}\left[(\bar{Y}^{POST(k)} - \bar{Y}^{PRE(g)})|G=g \right] - \mathbb{E}\left[(\bar{Y}^{POST(k)} - \bar{Y}^{PRE(g)})|G=k \right] \right) \Bigg\}$$

where the first equality uses the result in Lemma 4, the second equality changes the order of the summations (splitting them at g and k where the value of $v(g,t)$ and $v(k,t)$ change) and uses Equation (24), the third equality holds by averaging over time periods (which involves multiplying and dividing by $g-1$ in the first line, multiplying and dividing by $k-g$ in the second line, and multiplying and dividing by $\mathcal{T}-k+1$ in the last line), the fourth equality rearranges the expressions for the weights, the fifth equality holds by rearranging terms with common weights, and the last equality holds by the definitions of $\tilde{w}^{g,post}$, $\tilde{w}^{k,post}$, and $\tilde{w}^{long}$ and by noticing that

$$p_k p_g = (p_g + p_k)^2 p_{g|\{g,k\}}(1 - p_{g|\{g,k\}})$$

which holds by multiplying and dividing both p_k and p_g by $(p_g + p_k)$ and by the definition of $p_{g|\{g,k\}}$. $\qquad\square$

The result in Lemma 5 is very closely related to the result on interpreting TWFE regressions with a binary treatment and multiple time periods and variation in treatment timing in Goodman-Bacon (2021).[20] In particular, it says that, even with a continuous/multi-valued treatment, the TWFE regression estimator involves comparisons between (i) the path of outcomes for units that become treated relative to the path of outcomes for units that are not treated yet, (ii) the path of outcomes for units that become treated relative to the path of outcomes for units that have already been treated, and (iii) comparisons of the paths of outcomes across groups from their common pre-treatment periods to their common post-treatment periods. Intuitively, the first set of comparisons are very much in the spirit of DiD, but the second and third sets of comparisons are not (except under additional specialized conditions). We formalize this intuition in the proof of Theorem 5

---

[20]One difference worth noting is that the weights are slightly different due to the terms involving $E[D|G=g]$ and $\mathbb{E}[D|G=k]$. With a binary treatment, these expectations are equal to each other by construction, but with a continuous treatment these terms are no longer generally equal to each other. This also implies that the third term does not show up in the case with a binary treatment.

below.

Lemma 6. *Under Assumptions 1-MP, 2-MP(a), and 3-MP,*

$$\frac{1}{\mathcal{T}}\sum_{t=1}^{\mathcal{T}}\mathbb{E}[\ddot{W}_{it}^2] = \sum_{g\in\mathcal{G}}\tilde{w}^{g,within}(g) + \sum_{g\in\mathcal{G}}\sum_{k\in\mathcal{G},k>g}\left\{\tilde{w}^{g,post}(g,k) + \tilde{w}^{k,post}(g,k) + \tilde{w}^{long}(g,k)\right\}$$

Proof. To start with, notice that $\mathbb{E}[\ddot{W}_{it}^2] = \mathbb{E}[W_{it}\ddot{W}_{it}]$. Then, we can apply the arguments of Lemmas 2 to 5 but with W_{it} replacing Y_{it}. This implies that

$$\frac{1}{\mathcal{T}}\sum_{t=1}^{\mathcal{T}}\mathbb{E}[\ddot{W}_{it}^2]$$

$$= \sum_{g\in\mathcal{G}}\tilde{w}^{g,within}(g)\frac{\mathrm{cov}(\bar{W}^{POST(g)} - \bar{W}^{PRE(g)}, D|G=g)}{\mathrm{var}(D|G=g)}$$

$$+ \sum_{g\in\mathcal{G}}\sum_{k\in\mathcal{G},k>g}\left\{\tilde{w}^{g,post}(g,k)\frac{\mathbb{E}\left[(\bar{W}^{MID(g,k)} - \bar{W}^{PRE(g)})|G=g\right] - \mathbb{E}\left[(\bar{W}^{MID(g,k)} - \bar{W}^{PRE(g)})|G=k\right]}{\mathbb{E}[D|G=g]}\right.$$

$$+ \tilde{w}^{k,post}(g,k)\frac{\mathbb{E}\left[(\bar{W}^{POST(k)} - \bar{W}^{MID(g,k)})|G=k\right] - \mathbb{E}\left[(\bar{W}^{POST(k)} - \bar{W}^{MID(g,k)})|G=g\right]}{\mathbb{E}[D|G=k]}$$

$$\left. + \tilde{w}^{long}(g,k)\frac{\mathbb{E}\left[(\bar{W}^{POST(k)} - \bar{W}^{PRE(g)})|G=g\right] - \mathbb{E}\left[(\bar{W}^{POST(k)} - \bar{W}^{PRE(g)})|G=k\right]}{\mathbb{E}[D|G=g] - \mathbb{E}[D|G=k]}\right\}$$

$$= \sum_{g\in\mathcal{G}}\tilde{w}^{g,within}(g) + \sum_{g\in\mathcal{G}}\sum_{k\in\mathcal{G},k>g}\left\{\tilde{w}^{g,post}(g,k) + \tilde{w}^{k,post}(g,k) + \tilde{w}^{long}(g,k)\right\}$$

where the last equality holds by noting that $\bar{W} = D$ in post-treatment periods and $\bar{W} = 0$ in pre-treatment periods, and then by canceling terms. □

Proof of Proposition 5

Proof. Proposition 5 immediately holds by combining the results in Lemma 2, from Equations (22) and (23), and by Lemmas 3 to 5 (which all concern the numerator in the expression for β^{twfe} in Equation (3)), and then dividing by $(1/\mathcal{T})\sum_{t=1}^{\mathcal{T}}\mathbb{E}[\ddot{W}_{it}^2]$ (which corresponds to the denominator in the expression for β^{twfe} in Equation (3)). That the weights are all positive holds immediately by their definitions. That they sum to one holds by the definitions of the weights and by Lemma 6. □

Next, we move to proving Theorem 5. To do this we provide expressions for each of the comparisons that show up in Proposition 5 in terms of derivatives of paths of outcomes. These results invoke Assumption 2-MP(b) and (c) and, therefore, use that the treatment is actually continuous, but they do not invoke any parallel trends assumptions. That said, it would be straightforward to adapt these results to the case with a discrete multi-valued treatment along the lines of the baseline two period case considered above.

It is also useful to note that

$$\frac{\partial \pi_D^{POST(\tilde{k}),PRE(\tilde{g})}(g,d)}{\partial d} = \frac{\partial \mathbb{E}\left[(\bar{Y}^{POST(\tilde{k})} - \bar{Y}^{PRE(\tilde{g})})|G=g,D=d\right]}{\partial d}$$

$$\frac{\partial \pi_D^{MID(\tilde{g},\tilde{k}),PRE(\tilde{g})}(g,d)}{\partial d} = \frac{\mathbb{E}\left[(\bar{Y}^{MID(\tilde{g},\tilde{k})} - \bar{Y}^{PRE(\tilde{g})})|G=g,D=d\right]}{\partial d}$$

$$\frac{\partial \pi_D^{POST(\tilde{k}),MID(\tilde{g},\tilde{k})}(g,d)}{\partial d} = \frac{\partial \mathbb{E}\left[(\bar{Y}^{POST(\tilde{k})} - \bar{Y}^{MID(\tilde{g},\tilde{k})})|G=g,D=d\right]}{\partial d}$$

which holds because the second parts of each π_D term do not vary with the dose.

Next, we consider a result for the main term in $\delta^{WITHIN}(g)$ in Equation (4).

Lemma 7. *Under Assumptions 1-MP, 2-MP, and 3-MP,*

$$\text{cov}\left(\bar{Y}^{POST(g)} - \bar{Y}^{PRE(g)}, D|G=g\right)$$

$$= \int_{\mathcal{D}_+} \left(\mathbb{E}[D|G=g,D\geq l] - \mathbb{E}[D|G=g]\right)\text{P}(D\geq l|G=g)\frac{\partial \mathbb{E}[\bar{Y}^{POST(g)} - \bar{Y}^{PRE(g)}|G=g,D=l]}{\partial l}\,dl$$

Proof. First, notice that

$$\text{cov}\left(\bar{Y}^{POST(g)} - \bar{Y}^{PRE(g)}, D|G=g\right) = \mathbb{E}\left[(\bar{Y}^{POST(g)} - \bar{Y}^{PRE(g)})(D - \mathbb{E}[D|G=g])|G=g\right]$$

Then, the proof follows essentially the same arguments as in Theorem 3 with $\bar{Y}^{POST(g)} - \bar{Y}^{PRE(g)}$ replacing ΔY and the other arguments relating to the distribution of the dose holding conditional on being in group g. The second term, involving d_L, in Theorem 3 does not show up here as, by construction, there are no untreated units in group g. $\qquad \square$

Lemma 7 says that part of $\delta^{WITHIN}(g)$ in the TWFE regression estimator comes from a weighted average of $\frac{\partial \mathbb{E}[\bar{Y}^{POST(g)} - \bar{Y}^{PRE(g)}|G=g,D=d]}{\partial d}$.

Next, we consider the main term in the expression for $\delta^{MID,PRE}(g,k)$ in Equation (5). This term is quite similar to the baseline two-period case considered in Theorem 3 because units in group k have not been treated yet.

Lemma 8. *Under Assumptions 1-MP, 2-MP, and 3-MP, and for $k > g$,*

$$\mathbb{E}\left[(\bar{Y}^{MID(g,k)} - \bar{Y}^{PRE(g)})|G=g\right] - \mathbb{E}\left[(\bar{Y}^{MID(g,k)} - \bar{Y}^{PRE(g)})|G=k\right]$$

$$= \int_{\mathcal{D}_+} \text{P}(D\geq l|G=g)\frac{\partial \mathbb{E}[\bar{Y}^{MID(g,k)} - \bar{Y}^{PRE(g)}|G=g,D=l]}{\partial l}\,dl$$

$$+ d_L \frac{\mathbb{E}[\bar{Y}^{MID(g,k)} - \bar{Y}^{PRE(g)}|G=g,D=d_L] - \mathbb{E}[\bar{Y}^{MID(g,k)} - \bar{Y}^{PRE(g)}|D=0]}{d_L}$$

$$- d_L \frac{\mathbb{E}[\bar{Y}^{MID(g,k)} - \bar{Y}^{PRE(g)}|G=k] - \mathbb{E}[\bar{Y}^{MID(g,k)} - \bar{Y}^{PRE(g)}|D=0]}{d_L}$$

Proof. To start with, notice that

$$\mathbb{E}\left[\left(\bar{Y}^{MID(g,k)} - \bar{Y}^{PRE(g)}\right)|G = g\right] - \mathbb{E}\left[\left(\bar{Y}^{MID(g,k)} - \bar{Y}^{PRE(g)}\right)|G = k\right]$$

$$= \mathbb{E}\left[\left(\bar{Y}^{MID(g,k)} - \bar{Y}^{PRE(g)}\right)|G = g\right] - \mathbb{E}\left[\left(\bar{Y}^{MID(g,k)} - \bar{Y}^{PRE(g)}\right)|D = 0\right]$$

$$- \left(\mathbb{E}\left[\left(\bar{Y}^{MID(g,k)} - \bar{Y}^{PRE(g)}\right)|G = k\right] - \mathbb{E}\left[\left(\bar{Y}^{MID(g,k)} - \bar{Y}^{PRE(g)}\right)|D = 0\right]\right)$$

$$= \int_{\mathcal{D}_+} \mathrm{P}(D \geq l|G = g)\frac{\partial\mathbb{E}[\bar{Y}^{MID(g,k)} - \bar{Y}^{PRE(g)}|G = g, D = l]}{\partial l}\, dl$$

$$+ d_L\frac{\mathbb{E}[\bar{Y}^{MID(g,k)} - \bar{Y}^{PRE(g)}|G = g, D = d_L] - \mathbb{E}[\bar{Y}^{MID(g,k)} - \bar{Y}^{PRE(g)}|D = 0]}{d_L}$$

$$- d_L\frac{\mathbb{E}[\bar{Y}^{MID(g,k)} - \bar{Y}^{PRE(g)}|G = k] - \mathbb{E}[\bar{Y}^{MID(g,k)} - \bar{Y}^{PRE(g)}|D = 0]}{d_L}$$

where the first equality holds by adding and subtracting $\mathbb{E}\left[\left(\bar{Y}^{MID(g,k)} - \bar{Y}^{PRE(g)}\right)|D = 0\right]$. For the second equality, notice that

$$\mathbb{E}\left[\left(\bar{Y}^{MID(g,k)} - \bar{Y}^{PRE(g)}\right)|G = g\right] - \mathbb{E}\left[\left(\bar{Y}^{MID(g,k)} - \bar{Y}^{PRE(g)}\right)|D = 0\right]$$

$$= \mathbb{E}\left[\left(\bar{Y}^{MID(g,k)} - \bar{Y}^{PRE(g)}\right)|G = g\right] - \mathbb{E}\left[\left(\bar{Y}^{MID(g,k)} - \bar{Y}^{PRE(g)}\right)|G = g, D = d_L\right]$$

$$+ \mathbb{E}\left[\left(\bar{Y}^{MID(g,k)} - \bar{Y}^{PRE(g)}\right)|G = g, D = d_L\right] - \mathbb{E}\left[\left(\bar{Y}^{MID(g,k)} - \bar{Y}^{PRE(g)}\right)|D = 0\right]$$

Moreover,

$$\mathbb{E}\left[\left(\bar{Y}^{MID(g,k)} - \bar{Y}^{PRE(g)}\right)|G = g\right] - \mathbb{E}\left[\left(\bar{Y}^{MID(g,k)} - \bar{Y}^{PRE(g)}\right)|G = g, D = d_L\right]$$

$$= \int_{\mathcal{D}_+} \mathbb{E}\left[\left(\bar{Y}^{MID(g,k)} - \bar{Y}^{PRE(g)}\right)|G = g, D = d\right] - \mathbb{E}\left[\left(\bar{Y}^{MID(g,k)} - \bar{Y}^{PRE(g)}\right)|G = g, D = d_L\right]dF_{D|G}(d|g)$$

$$= \int_{\mathcal{D}_+}\int_{\mathcal{D}_+} \mathbf{1}\{l \leq d\}\frac{\partial\mathbb{E}\left[\left(\bar{Y}^{MID(g,k)} - \bar{Y}^{PRE(g)}\right)|G = g, D = l\right]}{\partial l}\, dl\, dF_{D|G}(d|g)$$

$$= \int_{\mathcal{D}_+} \mathrm{P}(D \geq l|G = g)\frac{\partial\mathbb{E}[\bar{Y}^{MID(g,k)} - \bar{Y}^{PRE(g)}|G = g, D = l]}{\partial l}\, dl$$

where the first equality holds by the law of iterated expectations, the second equality holds by the fundamental theorem of calculus, and the last equality holds by changing the order of integration and simplifying.

Combining the above expressions implies the result. □

Next, we consider the main term for $\delta^{POST,MID}(g,k)$ in Equation (6) which comes from comparing paths of outcomes for newly treated groups relative to already-treated groups.

Lemma 9. *Under Assumptions 1-MP, 2-MP, and 3-MP, and for $k > g$,*

$$\mathbb{E}\left[\left(\bar{Y}^{POST(k)} - \bar{Y}^{MID(g,k)}\right)|G = k\right] - \mathbb{E}\left[\left(\bar{Y}^{POST(k)} - \bar{Y}^{MID(g,k)}\right)|G = g\right]$$

$$= \int_{\mathcal{D}_+} \mathrm{P}(D \geq l|G = k)\frac{\partial\mathbb{E}[\bar{Y}^{POST(k)} - \bar{Y}^{MID(g,k)}|G = k, D = l]}{\partial l}\, dl$$

$$+ d_L \frac{\mathbb{E}[\bar{Y}^{POST(k)} - \bar{Y}^{MID(g,k)}|G=k, D=d_L] - \mathbb{E}[\bar{Y}^{POST(k)} - \bar{Y}^{MID(g,k)}|D=0]}{d_L}$$

$$- \left\{ \mathbb{E}[\bar{Y}^{POST(k)} - \bar{Y}^{PRE(g)}|G=g] - \mathbb{E}[\bar{Y}^{POST(k)} - \bar{Y}^{PRE(g)}|D=0] \right.$$

$$\left. - \left(\mathbb{E}[\bar{Y}^{MID(g,k)} - \bar{Y}^{PRE(g)}|G=g] - \mathbb{E}[\bar{Y}^{MID(g,k)} - \bar{Y}^{PRE(g)}|D=0] \right) \right\}$$

Proof. Notice that

$$\mathbb{E}\left[(\bar{Y}^{POST(k)} - \bar{Y}^{MID(g,k)})|G=k \right] - \mathbb{E}\left[(\bar{Y}^{POST(k)} - \bar{Y}^{MID(g,k)})|G=g \right]$$

$$= \left(\mathbb{E}\left[(\bar{Y}^{POST(k)} - \bar{Y}^{MID(g,k)})|G=k \right] - \mathbb{E}\left[(\bar{Y}^{POST(k)} - \bar{Y}^{MID(g,k)})|D=0 \right] \right)$$

$$- \left(\mathbb{E}\left[(\bar{Y}^{POST(k)} - \bar{Y}^{MID(g,k)})|G=g \right] - \mathbb{E}\left[(\bar{Y}^{POST(k)} - \bar{Y}^{MID(g,k)})|D=0 \right] \right)$$

$$= \left(\mathbb{E}\left[(\bar{Y}^{POST(k)} - \bar{Y}^{MID(g,k)})|G=k \right] - \mathbb{E}\left[(\bar{Y}^{POST(k)} - \bar{Y}^{MID(g,k)})|D=0 \right] \right) \qquad (25)$$

$$- \left\{ \left(\mathbb{E}\left[(\bar{Y}^{POST(k)} - \bar{Y}^{PRE(g)})|G=g \right] - \mathbb{E}\left[(\bar{Y}^{POST(k)} - \bar{Y}^{PRE(g)})|D=0 \right] \right) \right.$$

$$\left. - \left(\mathbb{E}\left[(\bar{Y}^{MID(g,k)} - \bar{Y}^{PRE(g)})|G=g \right] - \mathbb{E}\left[(\bar{Y}^{MID(g,k)} - \bar{Y}^{PRE(g)})|D=0 \right] \right) \right\}$$

$$= \int_{\mathcal{D}_+} \mathrm{P}(D \geq l|G=k) \frac{\partial \mathbb{E}[\bar{Y}^{POST(k)} - \bar{Y}^{MID(g,k)}|G=k, D=l]}{\partial l} \, dl \qquad (26)$$

$$+ d_L \frac{\mathbb{E}[\bar{Y}^{POST(k)} - \bar{Y}^{MID}(g,k)|G=g, D=d_L] - \mathbb{E}\left[(\bar{Y}^{POST(k)} - \bar{Y}^{MID(g,k)})|D=0 \right]}{d_L}$$

$$- \left\{ \left(\mathbb{E}\left[(\bar{Y}^{POST(k)} - \bar{Y}^{PRE(g)})|G=g \right] - \mathbb{E}\left[(\bar{Y}^{POST(k)} - \bar{Y}^{PRE(g)})|D=0 \right] \right) \right.$$

$$\left. - \left(\mathbb{E}\left[(\bar{Y}^{MID(g,k)} - \bar{Y}^{PRE(g)})|G=g \right] - \mathbb{E}\left[(\bar{Y}^{MID(g,k)} - \bar{Y}^{PRE(g)})|D=0 \right] \right) \right\}$$

where the first equality holds by adding and subtracting $\mathbb{E}\left[(\bar{Y}^{POST(k)} - \bar{Y}^{MID(g,k)})|D=0 \right]$, the second equality holds by adding and subtracting both $\mathbb{E}\left[\bar{Y}^{PRE(g)}|G=g \right]$ and $\mathbb{E}\left[\bar{Y}^{PRE(g)}|D=0 \right]$, and the last equality holds by applying the same sort of arguments as in the proof of Lemma 8. $\quad\square$

The expression in Lemma 9 appears complicated and is worth explaining in some more detail. Consider Equation (25) in the proof of Lemma 9. There are three parts of this expression. The first part compares the path of outcomes in post-treatment periods relative to some pre-treatment periods for units in group k to the path of outcomes for units that never participate in the treatment. This sort of comparison is very much in the spirit of DiD and will correspond to a reasonable treatment effect parameter under appropriate parallel trends assumptions. Similarly, under suitable parallel trends assumptions, the terms in the second and third lines will correspond to treatment effects for group g between periods k and $\mathcal{T}$ (the second line) and treatment effects for group g between periods g and $k-1$ (the third line). Therefore, the difference between these terms can be

thought of as some form of treatment effect dynamics. That means, in general, for this overall term to correspond to a treatment effect parameter for group k, there needs to be no treatment effect dynamics for group g. Ruling out treatment effect dynamics is not implied by any sort of parallel trends assumption and therefore involves an additional (and potentially very strong) assumption.

Finally, we consider the main term for $\delta^{POST,PRE}(g,k)$ in Equation (7).

Lemma 10. *Under Assumptions 1-MP, 2-MP, and 3-MP, and for $k > g$,*

$$
\mathbb{E}\left[(\bar{Y}^{POST(k)} - \bar{Y}^{PRE(g)})|G = g\right] - \mathbb{E}\left[(\bar{Y}^{POST(k)} - \bar{Y}^{PRE(g)})|G = k\right]
$$

$$
= \int_{\mathcal{D}_+} (\mathrm{P}(D \geq l|G = g) - \mathrm{P}(D \geq l|G = k)) \frac{\partial \mathbb{E}[\bar{Y}^{POST(k)} - \bar{Y}^{PRE(g)}|G = g, D = l]}{\partial l} \, dl
$$

$$
- \left\{ \int_{\mathcal{D}_+} \mathrm{P}(D \geq l|G = k) \left(\frac{\partial \mathbb{E}[\bar{Y}^{POST(k)} - \bar{Y}^{PRE(g)}|G = k, D = l]}{\partial l} - \frac{\partial \mathbb{E}[\bar{Y}^{POST(k)} - \bar{Y}^{PRE(g)}|G = g, D = l]}{\partial l} \right) \, dl \right.
$$

$$
+ d_L \frac{\mathbb{E}[\bar{Y}^{POST(k)} - \bar{Y}^{PRE(g)}|G = k, D = d_L] - \mathbb{E}[\bar{Y}^{POST(k)} - \bar{Y}^{PRE(g)}|D = 0]}{d_L}
$$

$$
\left. - d_L \frac{\mathbb{E}[\bar{Y}^{POST(k)} - \bar{Y}^{PRE(g)}|G = g, D = d_L] - \mathbb{E}[\bar{Y}^{POST(k)} - \bar{Y}^{PRE(g)}|D = 0]}{d_L} \right\}
$$

Proof. First, by adding and subtracting terms

$$
\mathbb{E}\left[(\bar{Y}^{POST(k)} - \bar{Y}^{PRE(g)})|G = g\right] - \mathbb{E}\left[(\bar{Y}^{POST(k)} - \bar{Y}^{PRE(g)})|G = k\right]
$$

$$
= \mathbb{E}\left[(\bar{Y}^{POST(k)} - \bar{Y}^{PRE(g)})|G = g\right] - \mathbb{E}\left[(\bar{Y}^{POST(k)} - \bar{Y}^{PRE(g)})|D = 0\right]
$$

$$
- \left(\mathbb{E}\left[(\bar{Y}^{POST(k)} - \bar{Y}^{PRE(g)})|G = k\right] - \mathbb{E}\left[(\bar{Y}^{POST(k)} - \bar{Y}^{PRE(g)})|D = 0\right] \right)
$$

Then, using similar arguments as in Lemma 8 above, one can show that

$$
\mathbb{E}\left[(\bar{Y}^{POST(k)} - \bar{Y}^{PRE(g)})|G = g\right] - \mathbb{E}\left[(\bar{Y}^{POST(k)} - \bar{Y}^{PRE(g)})|D = 0\right]
$$

$$
= \int_{\mathcal{D}_+} \mathrm{P}(D \geq l|G = g) \frac{\partial \mathbb{E}[\bar{Y}^{POST(k)} - \bar{Y}^{PRE(g)}|G = g, D = l]}{\partial l} \, dl
$$

$$
+ d_L \frac{\mathbb{E}[\bar{Y}^{POST(k)} - \bar{Y}^{PRE(g)}|G = g, D = d_L] - \mathbb{E}[\bar{Y}^{POST(k)} - \bar{Y}^{PRE(g)}|D = 0]}{d_L}
$$

and that

$$
\mathbb{E}\left[(\bar{Y}^{POST(k)} - \bar{Y}^{PRE(g)})|G = k\right] - \mathbb{E}\left[(\bar{Y}^{POST(k)} - \bar{Y}^{PRE(g)})|D = 0\right]
$$

$$
= \int_{\mathcal{D}_+} \mathrm{P}(D \geq l|G = k) \frac{\partial \mathbb{E}[\bar{Y}^{POST(k)} - \bar{Y}^{PRE(g)}|G = k, D = l]}{\partial l} \, dl
$$

$$
+ d_L \frac{\mathbb{E}[\bar{Y}^{POST(k)} - \bar{Y}^{PRE(g)}|G = k, D = d_L] - \mathbb{E}[\bar{Y}^{POST(k)} - \bar{Y}^{PRE(g)}|D = 0]}{d_L}
$$

Then, the result holds by adding and subtracting $\int_{\mathcal{D}_+} \mathrm{P}(D \geq l|G = k) \frac{\partial \mathbb{E}[\bar{Y}^{POST(k)} - \bar{Y}^{PRE(g)}|G=g, D=l]}{\partial l} \, dl$ and combining terms. $\square$

Proof of Part (1) of Theorem 5

Proof. Starting from the result in Proposition 5, the expression for $\delta^{WITHIN}(g)$ comes from its definition, the result in Lemma 7, and the definition of the weights $w_1^{within}(g,l)$. The expression for $\delta^{MID,PRE}(g,k)$ comes from its definition, the result in Lemma 8, and the definitions of $w_1(g,l)$ and $w_0(g)$. The expression for $\delta^{POST,MID}(g,k)$ comes from combining its definition with the result in Lemma 9, and the definitions of $w_1(k,l)$ and $w_0(k)$. Finally, the expression for $\delta^{POST,PRE}(g,k)$ comes from its definition, the result in Lemma 10, and the definitions of $w_1^{across}(g,k,l)$, $\tilde{w}_1^{across}(g,k,l)$, and $\tilde{w}_0^{across}(g,k)$.

That $w_1^{within}(g,d) \geq 0$, $w_1(g,0) \geq 0$, $w_0(g) \geq 0$ for all $g \in \mathcal{G}$ and $d \in \mathcal{D}_+$ all hold immediately from the definitions of the weights. That $\int_{\mathcal{D}_+} w_1^{within}(g,l)\,dl = 1$, $\int_{\mathcal{D}_+} w_1(g,l)\,dl + w_0(g) = 1$, and $\int_{\mathcal{D}_+} w_1^{across}(g,k,l)\,dl = 1$ hold from the same sorts of arguments used to show that the weights integrate to 1 in the proof of Proposition 4. $\square$

Notice that none of the previous results have invoked any sort of parallel trends assumption. Next, we push forward the previous results once a researcher invokes parallel trends assumptions; in the main text, we considered the case where the researcher invoked Assumption 5-MP, but here we consider both that assumption and Assumption 4-MP(a). To further understand this, for $1 \leq t_1 < t_2 \leq \mathcal{T}$ define

$$\bar{Y}_i^{(t_1,t_2)}(g,d) = \frac{1}{t_2 - t_1 + 1} \sum_{t=t_1}^{t_2} Y_{it}(g,t,d)$$

which averages potential outcomes from time periods t_1 to t_2 for unit i if they were in group g and experienced dose d. Note that $\bar{Y}_i^{(t_1,t_2)} = \bar{Y}_i^{(t_1,t_2)}(G_i, D_i)$. Next, for $t_1 \leq t_2$, define

$$\overline{ATT}^{(t_1,t_2)}(g,d|g,d) := \frac{1}{t_2 - t_1 + 1} \sum_{t=t_1}^{t_2} ATT(g,t,d|g,d)$$

which is the average treatment effect experienced by units in group g who experienced dose d averaged across periods from t_1 to t_2. Likewise, define

$$\overline{ATE}^{(t_1,t_2)}(g,d) := \frac{1}{t_2 - t_1 + 1} \sum_{t=t_1}^{t_2} ATE(g,t,d)$$

which is the average treatment effect of dose d among all units in group g averaged across periods from t_1 to t_2. An alternative expression for $\overline{ATT}^{(t_1,t_2)}(g,d|g,d)$ is given by

$$\overline{ATT}^{(t_1,t_2)}(g,d|g,d) = \mathbb{E}\left[\bar{Y}^{(t_1,t_2)}(g,d) - \bar{Y}^{(t_1,t_2)}(0)|G=g, D=d\right]$$

which holds by the definition of $ATT(g,t,d|g,d)$ and changing the order of the expectation and the average over time periods; here, $\mathbb{E}[\bar{Y}^{(t_1,t_2)}(0)|G=g, D=d]$ is the average outcome that units in group g that experienced dose d would have experienced if they had not participated in the treatment between time periods t_1 and t_2. Similarly, for $\overline{ATE}^{(t_1,t_2)}(g,d)$,

$$\overline{ATE}^{(t_1,t_2)}(g,d) = \mathbb{E}\left[\bar{Y}^{(t_1,t_2)}(g,d) - \bar{Y}^{(t_1,t_2)}(0)|G=g\right]$$

69

In addition, define

$$\overline{ACRT}^{(t_1,t_2)}(g,d|g,d) := \left.\frac{\partial \overline{ATT}(g,l|g,d)}{\partial l}\right|_{l=d} \quad \text{and} \quad \overline{ACR}^{(t_1,t_2)}(g,d) := \frac{\partial \overline{ATE}(g,d)}{\partial d}$$

which are the average causal response to a marginal increase in the dose among units in group g conditional on having dose experienced dose d (for $\overline{ACRT}(g,d|g,d)$) and the average causal response to a marginal increase in the dose among all units in group g.

The next result connects derivatives of conditional expectations to $ACRT$ and ACR parameters under parallel trends assumptions.

Lemma 11. *Under Assumptions 1-MP, 2-MP, and 3-MP, and for $1 \le t_1 \le t_2 < g \le t_3 \le t_4 \le \mathcal{T}$ (i.e., t_1 and t_2 are pre-treatment periods for group g, and t_3 and t_4 are post-treatment periods for group g), and for $d \in \mathcal{D}_+$,*

(1) If, in addition, Assumption 4-MP(a) holds, then

$$\frac{\partial \mathbb{E}\left[\bar{Y}^{(t_3,t_4)} - \bar{Y}^{(t_1,t_2)}|G=g,D=d\right]}{\partial d} = \overline{ACRT}^{(t_3,t_4)}(g,d|g,d) + \left.\frac{\partial \overline{ATT}^{(t_3,t_4)}(g,d|g,l)}{\partial l}\right|_{l=d}$$

(2) If, in addition, Assumption 5-MP holds, then

$$\frac{\partial \mathbb{E}\left[\bar{Y}^{(t_3,t_4)} - \bar{Y}^{(t_1,t_2)}|G=g,D=d\right]}{\partial d} = \overline{ACR}^{(t_3,t_4)}(g,d)$$

Proof. For part (1), notice that, for $1 \le t_1 \le t_2 < g \le t_3 \le t_4 \le \mathcal{T}$ (i.e., for group g, t_1 and t_2 are pre-treatment time periods while t_3 and t_4 are post treatment time periods), we can write

$$\begin{aligned}
\mathbb{E}\left[\bar{Y}^{(t_3,t_4)} - \bar{Y}^{(t_1,t_2)}|G=g,D=d\right] &= \mathbb{E}\left[\bar{Y}^{(t_3,t_4)}(g,d) - \bar{Y}^{(t_1,t_2)}(0)|G=g,D=d\right] \\
&= \mathbb{E}\left[\bar{Y}^{(t_3,t_4)}(g,d) - \bar{Y}^{(t_3,t_4)}(0)|G=g,D=d\right] \\
&\quad - \mathbb{E}\left[\bar{Y}^{(t_3,t_4)}(0) - \bar{Y}^{(t_1,t_2)}(0)|G=g,D=d\right] \\
&= \overline{ATT}^{(t_3,t_4)}(g,d|g,d) \\
&\quad - \mathbb{E}\left[\bar{Y}^{(t_3,t_4)}(0) - \bar{Y}^{(t_1,t_2)}(0)|G=g,D=d\right]
\end{aligned}$$

where the first equality holds by writing observed outcomes in terms of their corresponding potential outcomes, the second equality holds by adding and subtracting $\mathbb{E}\left[\bar{Y}^{(t_3,t_4)}(0)|G=g,D=d\right]$, and the last equality holds by the definition of $\overline{ATT}^{(t_3,t_4)}(g,d|g,d)$.

This equation looks very similar to DiD-type equations in simpler cases such as when there are two periods and two groups. The left hand side is immediately identified. The right hand side involves a causal effect parameter of interest and an unobserved path of untreated potential outcomes that would typically be handled using a parallel trends assumption.

In particular, under Assumption 4-MP(a) (though notice that Assumption 4-MP(b) and (c) are not generally strong enough here),

$$\mathbb{E}\left[\bar{Y}^{(t_3,t_4)}(0) - \bar{Y}^{(t_1,t_2)}(0)|G=g,D=d\right] = \mathbb{E}\left[\bar{Y}^{(t_3,t_4)}(0) - \bar{Y}^{(t_1,t_2)}(0)|D=0\right]$$

which, importantly, does not vary across d or g.

This suggests that, under Assumption 4-MP(a),

$$\mathbb{E}\left[\bar{Y}^{(t_3,t_4)} - \bar{Y}^{(t_1,t_2)}|G=g, D=d\right] = \overline{ATT}^{(t_3,t_4)}(g,d|g,d) - \mathbb{E}\left[\bar{Y}^{(t_3,t_4)}(0) - \bar{Y}^{(t_1,t_2)}(0)|D=0\right]$$

Taking derivatives of both sides of the previous equation with respect to d implies the result.

For part (2), notice that,

$$\begin{aligned}
\mathbb{E}\left[\bar{Y}^{(t_3,t_4)} - \bar{Y}^{(t_1,t_2)}|G=g, D=d\right] &= \mathbb{E}\left[\bar{Y}^{(t_3,t_4)}(g,d) - \bar{Y}^{(t_1,t_2)}(0)|G=g, D=d\right] \\
&= \mathbb{E}\left[\bar{Y}^{(t_3,t_4)}(g,d) - \bar{Y}^{(t_1,t_2)}(0)|G=g\right] \\
&= \mathbb{E}\left[\bar{Y}^{(t_3,t_4)}(g,d) - \bar{Y}^{(t_3,t_4)}(0)|G=g\right] \\
&\quad + \mathbb{E}\left[\bar{Y}^{(t_3,t_4)}(0) - \bar{Y}^{(t_1,t_2)}(0)|G=g\right] \\
&= \overline{ATE}^{(t_3,t_4)}(g,d) + \mathbb{E}\left[\bar{Y}^{(t_3,t_4)}(0) - \bar{Y}^{(t_1,t_2)}(0)|D=0\right]
\end{aligned}$$

where the first equality holds by writing observed outcomes in terms of their corresponding potential outcomes, the second equality holds by Assumption 5-MP, the third equality holds by adding and subtracting $\mathbb{E}[\bar{Y}^{(t_3,t_4)}(0)|G=g]$, and the last equality holds by the definition of $\overline{ATE}^{(t_3,t_4)}(g,d)$ and by Assumption 5-MP. Taking derivatives of both sides implies the result for part (2). $\qquad\square$

The result in Lemma 11 says that, under Assumption 4-MP(a), the derivative of the path of outcomes (averaged over some post-treatment periods) relative to some pre-treatment periods corresponds to $ACRT(g,t,d|g,d)$ plus the derivative of a selection bias-type term with respect to d across some post-treatment time periods for units in group g. Similarly, under Assumption 5-MP, the derivative of the averaged path of outcomes over time in some post-treatment periods relative to the same average path of outcomes in some pre-treatment periods corresponds to an average of $ACR(g,d)$ with respect to d across the same post-treatment time periods.

The intuition for this sort of result is very similar to that of Proposition 3 in the baseline case with two time periods.

Lemma 12. *Under Assumptions 1-MP, 2-MP, and 3-MP, and for $1 \le t_1 \le t_2 < g \le t_3 \le t_4 < k$ (i.e., t_1 and t_2 are pre-treatment periods for both groups g and k, group g is treated before group k, and t_3 and t_4 are post-treatment periods for group g but pre-treatment periods for group k),*

(1) If, in addition, Assumption 4-MP(a) holds, then

$$d_L \frac{\mathbb{E}\left[\bar{Y}^{(t_3,t_4)} - \bar{Y}^{(t_1,t_2)}|G=g, D=d_L\right] - \mathbb{E}\left[\bar{Y}^{(t_3,t_4)} - \bar{Y}^{(t_1,t_2)}|G=k\right]}{d_L} = d_L \frac{\overline{ATT}^{(t_3,t_4)}(g,d_L|g,d_L)}{d_L}$$

(2) If, in addition, Assumption 5-MP holds, then

$$d_L \frac{\mathbb{E}\left[\bar{Y}^{(t_3,t_4)} - \bar{Y}^{(t_1,t_2)}|G=g, D=d_L\right] - \mathbb{E}\left[\bar{Y}^{(t_3,t_4)} - \bar{Y}^{(t_1,t_2)}|G=k\right]}{d_L} = d_L \frac{\overline{ATE}^{(t_3,t_4)}(g,d_L)}{d_L}$$

Proof. For part (1), notice that

$$\mathbb{E}\left[\bar{Y}^{(t_3,t_4)} - \bar{Y}^{(t_1,t_2)}|G = g, D = d_L\right] - \mathbb{E}\left[\bar{Y}^{(t_3,t_4)} - \bar{Y}^{(t_1,t_2)}|G = k\right]$$

$$= \mathbb{E}\left[\bar{Y}^{(t_3,t_4)}(g,d_L) - \bar{Y}^{(t_1,t_2)}(0)|G = g, D = d_L\right] - \mathbb{E}\left[\bar{Y}^{(t_3,t_4)}(0) - \bar{Y}^{(t_1,t_2)}(0)|G = k\right]$$

$$= \mathbb{E}\left[\bar{Y}^{(t_3,t_4)}(g,d_L) - \bar{Y}^{(t_3,t_4)}(0)|G = g, D = d_L\right]$$

$$+ \left\{\mathbb{E}\left[\bar{Y}^{(t_3,t_4)}(0) - \bar{Y}^{(t_1,t_2)}(0)|G = g, D = d_L\right] - \mathbb{E}\left[\bar{Y}^{(t_3,t_4)}(0) - \bar{Y}^{(t_1,t_2)}(0)|G = k\right]\right\}$$

$$= \overline{ATT}^{(t_3,t_4)}(g,d_L)$$

where the first equality holds by writing observed outcomes in terms of their corresponding potential outcomes, the second equality holds by adding and subtracting $\mathbb{E}\left[\bar{Y}^{(t_3,t_4)}(0)|G = g, D = d_L\right]$, and the last equality holds by the definition of $\overline{ATT}^{(t_3,t_4)}(g,d_L)$ and because the difference between the two terms involving paths of untreated potential outcomes on the second line of the previous equality is equal to 0 under Assumption 4-MP(a). Then, the result holds by multiplying and dividing by d_L.

For part (2),

$$\mathbb{E}\left[\bar{Y}^{(t_3,t_4)} - \bar{Y}^{(t_1,t_2)}|G = g, D = d_L\right] - \mathbb{E}\left[\bar{Y}^{(t_3,t_4)} - \bar{Y}^{(t_1,t_2)}|G = k\right]$$

$$= \mathbb{E}\left[\bar{Y}^{(t_3,t_4)}(g,d_L) - \bar{Y}^{(t_1,t_2)}(0)|G = g, D = d_L\right] - \mathbb{E}\left[\bar{Y}^{(t_3,t_4)}(0) - \bar{Y}^{(t_1,t_2)}(0)|G = k\right]$$

$$= \mathbb{E}\left[\bar{Y}^{(t_3,t_4)}(g,d_L) - \bar{Y}^{(t_1,t_2)}(0)|G = g\right] - \mathbb{E}\left[\bar{Y}^{(t_3,t_4)}(0) - \bar{Y}^{(t_1,t_2)}(0)|G = k\right]$$

$$= \mathbb{E}\left[\bar{Y}^{(t_3,t_4)}(g,d_L) - \bar{Y}^{(t_3,t_4)}(0)|G = g\right]$$

$$+ \left\{\mathbb{E}\left[\bar{Y}^{(t_3,t_4)}(0) - \bar{Y}^{(t_1,t_2)}(0)|G = g\right] - \mathbb{E}\left[\bar{Y}^{(t_3,t_4)}(0) - \bar{Y}^{(t_1,t_2)}(0)|G = k\right]\right\}$$

$$= \overline{ATE}^{(t_3,t_4)}(g,d_L)$$

where the first equality holds by writing observed outcomes in terms of their corresponding potential outcomes, the second equality holds by Assumption 5-MP, the third equality holds by adding and subtracting $\mathbb{E}[\bar{Y}^{(t_3,t_4)}(0)|G = g]$, and the last equality holds by Assumption 5-MP. The result holds by multiplying and dividing by d_L. $\square$

Proof of Part (2) of Theorem 5

Proof. The result holds immediately by using the results of Lemmas 11 and 12 in each of the expressions for $\delta^{WITHIN}(g)$, $\delta^{MID,PRE}(g,k)$, $\delta^{POST,MID}(g,k)$, and $\delta^{POST,PRE}(g,k)$ in part (1) of Theorem 5. $\square$

Proof of Proposition 6

Proof. For part (a), using similar arguments as in Lemma 8 and then under Assumption 5-MP, it follows that

$$\mathbb{E}\left[\bar{Y}^{POST(k)} - \bar{Y}^{PRE(g)}|G = g\right] - \mathbb{E}\left[\bar{Y}^{POST(k)} - \bar{Y}^{PRE(g)}|D = 0\right]$$

$$= \int_{\mathcal{D}_+} \mathrm{P}(D \geq l | G = g) \overline{ACR}^{POST(k)}(g, l) \, dl + d_L \frac{\overline{ATE}^{POST(k)}(g, d_L)}{d_L}$$

and that

$$\mathbb{E}\left[\bar{Y}^{MID(g,k)} - \bar{Y}^{PRE(g)} | G = g\right] - \mathbb{E}\left[\bar{Y}^{MID(g,k)} - \bar{Y}^{PRE(g)} | D = 0\right]$$

$$= \int_{\mathcal{D}_+} \mathrm{P}(D \geq l | G = g) \overline{ACR}^{MID(g,k)}(g, l) \, dl + d_L \frac{\overline{ATE}^{MID(g,k)}(g, d_L)}{d_L}$$

Under Assumption 6(a), $ACR(g, t, d)$ and $ATE(g, t, d_L)$ do not vary over time which implies that, for all $g \in \mathcal{G}$ and $k \in \mathcal{G}$ with $k > g$, $\overline{ACR}^{POST(k)}(g, l) = \overline{ACR}^{POST(k)}(g, l)$ for all $l \in \mathcal{D}_+$ and $\overline{ATE}^{POST(k)}(g, d_L) = \overline{ATE}^{MID(g,k)}(g, d_L)$. This implies that $\mathbb{E}\left[\bar{Y}^{POST(k)} - \bar{Y}^{PRE(g))} | G = g\right] = \mathbb{E}\left[\bar{Y}^{MID(g,k)} - \bar{Y}^{PRE(g))} | G = g\right]$ which implies the result for part (a).

For part (b), notice that, under Assumption 5-MP,

$$\frac{\partial \pi_D^{POST(k),PRE(g)}(k, l)}{\partial l} - \frac{\partial \pi_D^{POST(k),PRE(g)}(g, l)}{\partial l} = \overline{ACR}^{POST(k)}(k, l) - \overline{ACR}^{POST(k)}(g, l)$$

$$= 0$$

for $l \in \mathcal{D}_+$ and where the second equality holds by Assumption 6(b) (which implies that, for a particular time period, $ACR(g, t, d)$ does not vary across groups).

The same sort of arguments imply that

$$\frac{\pi_D^{POST(k),PRE(g)}(k, d_L) - \pi_D^{POST(k),PRE(g)}(g, d_L)}{d_L} = \frac{\overline{ATE}^{POST(k)}(k, d_L) - \overline{ATE}^{POST(k)}(g, d_L)}{d_L}$$

$$= 0$$

Finally, for part (c), under Assumption 6(a), (b), and (c), $ACR(g, t, d)$ does not vary across groups, time periods, or dose; since this does not vary, we denote it by ACR for the remainder of the proof. Moreover, from Theorem 5, we have that $\int_{\mathcal{D}_+} w_1^{within}(g, l) \, dl = 1$, $\int_{\mathcal{D}_+} w_1(g, l) \, dl + w_0(g) = 1$, and that $\int_{\mathcal{D}_+} w_1^{across}(g, k, l) = 1$. From the first two parts of the current result, we also have that the nuisance paths of outcomes in $\delta^{POST,MID}(g, k)$ and $\delta^{POST,PRE}(g, k)$ are both equal to 0 under Assumption 6(a) and (b). This implies that, under the conditions for part (c), $\delta^{WITHIN}(g) = \delta^{MID,PRE}(g, k) = \delta^{POST,MID}(g, k) = \delta^{POST,PRE}(g, k) = ACR$. Finally, from Proposition 5, we have that β^{twfe} is a weighted average of $\delta^{MID,PRE}(g, k)$, $\delta^{POST,MID}(g, k)$, $\delta^{POST,MID}(g, k)$, and $\delta^{POST,PRE}(g, k)$. That these are all equal to each other implies that $\beta^{twfe} = ACR = ACR^{*,mp}$. $\square$

Next, we provide a version of Theorem 5 extended to the case where Assumption 4-MP(a) (which is the multi-period version of standard parallel trends that only involves untreated potential outcomes) holds.

Theorem 5-Extended. *Under Assumptions 1-MP, 2-MP, 3-MP, and 4-MP(a),*

$$\delta^{WITHIN}(g) = \int_{\mathcal{D}_+} w_1^{within}(g,l) \left(\overline{ACRT}^{POST(g)}(g,l|g,l) + \frac{\partial \overline{ATT}^{POST(g)}(g,l|g,h)}{\partial h} \bigg|_{h=l} \right) dl$$

$$\delta^{MID,PRE}(g,k) = \int_{\mathcal{D}_+} w_1(g,l) \left(\overline{ACRT}^{MID(g,k)}(g,l|g,l) + \frac{\partial \overline{ATT}^{MID(g,k)}(g,l|g,h)}{\partial h} \bigg|_{h=l} \right) dl$$
$$+ w_0(g) \frac{\overline{ATT}^{MID(g,k)}(g,d_L|g,d_L)}{d_L}$$

$$\delta^{POST,MID}(g,k) = \int_{\mathcal{D}_+} w_1(k,l) \left(\overline{ACRT}^{POST(k)}(k,l|k,l) + \frac{\partial \overline{ATT}^{POST(k)}(k,l|k,h)}{\partial h} \bigg|_{h=l} \right) dl$$
$$+ w_0(k) \frac{\overline{ATT}^{POST(k)}(k,d_L|k,d_L)}{d_L} - w_0(k) \left(\frac{\pi^{POST(k),PRE(g)}(g) - \pi^{MID(g,k),PRE(g)}(g)}{d_L} \right)$$

$$\delta^{POST,PRE}(g,k) = \int_{\mathcal{D}_+} w_1^{across}(g,k,l) \left(\overline{ACRT}^{POST(k)}(g,l|g,l) + \frac{\partial \overline{ATT}(g,l|g,h)}{\partial h} \bigg|_{h=l} \right) dl$$
$$- \left\{ \int_{\mathcal{D}_+} \tilde{w}_1^{across}(g,k,l) \left(\frac{\partial \pi_D^{POST(k),PRE(g)}(k,l)}{\partial l} - \frac{\partial \pi_D^{POST(k),PRE(g)}(g,l)}{\partial l} \right) dl \right.$$
$$\left. + \tilde{w}_0^{across}(g,k) \left(\frac{\pi_D^{POST(k),PRE(g)}(k,d_L) - \pi_D^{POST(k),PRE(g)}(g,d_L)}{d_L} \right) \right\}$$

where the weights are the same as in Theorem 5 and satisfy the same properties.

Proof. The result holds immediately by plugging in the results of part (1) of Lemmas 11 and 12 for $\delta^{WITHIN}(g)$, $\delta^{MID,PRE}(g,k)$, $\delta^{POST,MID}(g,k)$, and $\delta^{POST,PRE}(g,k)$ in part (1) of Theorem 5. □