Center for Economic Institutions Working Paper Series

No. 2026-2

"Event-Study Designs for Discrete Outcomes under Transition Independence"

Young Ahn and Hiroyuki Kasahara

June, 2026

Institute of Economic Research
Hitotsubashi University
2-1 Naka, Kunitachi, Tokyo, 186-8603 JAPAN
https://cei.ier.hit-u.ac.jp/English/index.html
Tel:+81-42-580-8405/Fax:+81-42-580-8333

Event-Study Designs for Discrete Outcomes under Transition Independence*

Young Ahn
Department of Economics
University of Pennsylvania
youngahn@sas.upenn.edu

Hiroyuki Kasahara
Vancouver School of Economics
University of British Columbia
hkasahar@mail.ubc.ca

June 14, 2026

Abstract

We develop a new identification strategy for average treatment effects on the treated (ATT) in panel data with discrete outcomes. Standard difference-in-differences (DiD) relies on parallel trends, which is frequently violated in categorical settings due to mean reversion, out-of-bounds counterfactuals, and ill-defined trends for multi-category outcomes. We propose an alternative identification strategy with *transition independence*: absent treatment, transition dynamics conditional on pre-treatment outcomes are identical between control and treated groups. To capture unobserved heterogeneity, we introduce a latent-type Markov structure delivering type-specific and aggregate treatment effects from short panels. Three empirical applications yield ATT estimates substantially different from conventional DiD.

Keywords: Difference-in-differences, discrete outcomes, latent heterogeneity, treatment effects, finite mixture models

JEL Classification: C14, C23, C33

1 Introduction

Many empirical questions in economics involve outcomes that are inherently discrete and categorical: whether a worker is employed, unemployed, or out of the labor force; or which occupation a worker transitions into. Difference-in-differences (DiD) designs are routinely applied in these settings—researchers binarize the categorical labels into indicators and compare mean differences between treated and control groups under the parallel trends assumption (e.g., Bustos, 2011; Anderson and Sallee, 2011; Hvide and Jones, 2018; Charoenwong et al., 2019). Yet the properties of DiD when applied to discrete outcomes have received surprisingly little scrutiny.

*This paper supersedes earlier versions of the manuscript titled "Event Studies for Discrete Outcomes with Latent Transition Heterogeneity."

When outcomes are discrete, the parallel trends assumption is logically inconsistent with the data-generating process of bounded outcomes, invalidating the identification strategy. As Roth and Sant'Anna (2023) demonstrate, treatment effect estimates under parallel trends can be sensitive to functional form, a concern that is especially pronounced for discrete outcomes, where three specific failures arise. First, discrete outcomes evolve through state transitions rather than continuous movements, so when treated and control groups differ in their baseline distributions, mean reversion alone can generate divergent trends even absent any treatment effect. Second, DiD may produce counterfactual means outside the $[0, 1]$ range—a logical impossibility for a probability—undermining the coherence, not just the precision, of the resulting estimates. Third, for multi-category outcomes such as occupation or labor-force status, the notion of a single "trend" is ill-defined. The parallel trends assumption offers no coherent basis for comparing the joint temporal evolution of categorical distributions between treated and control groups.

This paper provides a resolution. We propose to replace the parallel trends assumption with *transition independence*—the condition that transition dynamics, conditional on pre-treatment outcome paths, would be identical between treated and control units in the absence of treatment. By operating directly on the transition structure of discrete outcomes, identification with transition independence avoids the logical failures of parallel trends while accommodating the bounded, categorical nature of the data.

Our paper makes three main contributions:

1. We show that the parallel trends assumption is generically violated for discrete, bounded outcomes: it can generate out-of-bounds counterfactual probabilities and confounds mean reversion with treatment effects. We propose transition independence as an alternative credible strategy: like parallel trends, the plausibility of transition independence can be assessed using pre-treatment data.

2. We extend the framework to incorporate a discrete latent-type structure under which outcomes follow latent-type-specific Markov chains. This extension addresses unobserved heterogeneity in transition dynamics and mitigates the curse of dimensionality arising from long pre-treatment histories. We establish identification of both *latent-type-specific average treatment effects on the treated (LTATTs)* and aggregate ATTs from short panel data.

3. We demonstrate the practical importance of our approach in three empirical applications—the Dodd-Frank Act, Norway's patent reform, and the Americans with Disabilities Act of 1990 (ADA)—where our estimator produces substantively different results relative to conventional DiD.

The empirical consequences are immediate: in our reanalysis of the Dodd-Frank data, the counterfactual complaint rates implied by parallel trends fall below zero in post-treatment periods, a quantity that cannot be a probability. This is not a finite-sample artifact but a logical consequence of applying linear extrapolation to a bounded outcome. The resulting sign reversal—DiD reports an increase in complaints where our method finds a decrease—illustrates that the parallel trends assumption can produce not just imprecise but fundamentally misleading estimates for discrete outcomes.

Transition independence avoids these failures by constructing counterfactuals from transition probabilities rather than mean levels, thereby respecting the bounded, discrete nature of the outcome and accounting for differential mean reversion driven by baseline differences. Under transition independence, counterfactuals are built by applying control-group transition probabilities to the treated group's pre-treatment distribution, predicting the outcome evolution of treated units absent treatment. ATTs are then identified by comparing observed outcomes against these constructed counterfactuals.

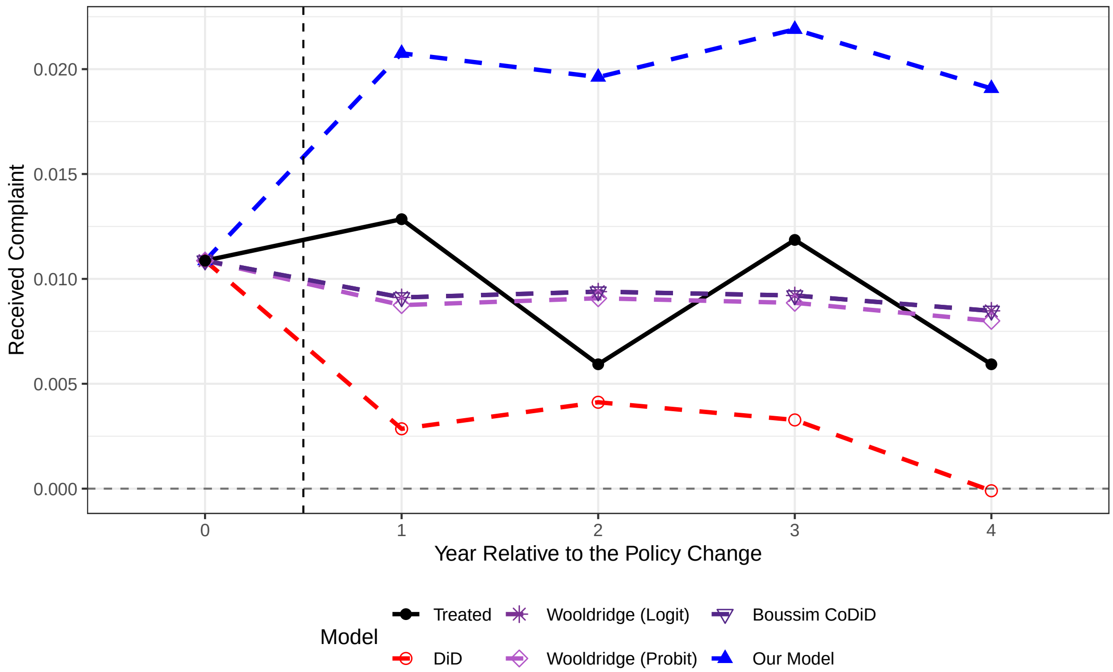
In practice, two challenges arise. First, unobserved heterogeneity may simultaneously affect treatment assignment and outcome transitions, violating transition independence at the population level. Second, the number of possible outcome histories grows exponentially with the number of pre-treatment periods, creating finite-sample limited overlap concerns. We address these challenges with two distinct devices: latent types accommodate unobserved heterogeneity by allowing transition independence to hold conditionally within types, while a low-order Markov restriction reduces the conditioning set from the full pre-treatment history to a small number of recent outcomes, resolving the dimensionality problem. The combined mixture-of-low-order-Markov approach delivers a tractable estimation framework that addresses both issues simultaneously.

We show that the aggregate ATT can be identified as a weighted average of LTATTs, where the weights correspond to the latent-type probabilities among treated units and are identified from data. We establish identification of both LTATTs and weights from short panel data, enabling us to account for unobserved heterogeneity in transition dynamics even when the number of pre-treatment periods is limited.

A further distinctive contribution of the transition-based framework is the *flow decomposition* of treatment effects. Because our approach models outcome dynamics through state-to-state transition probabilities, we can decompose the ATT on any given outcome state into inflow and outflow components, identifying which specific transition channels drive the treatment effect (Remark 8). This decomposition is not available in standard DiD, which operates on outcome levels rather than transitions. In the ADA application, for example, the flow decomposition reveals that the negative employment effect operates primarily through increased transitions from employment directly into out-of-labor-force status, rather than through changes in job-search outcomes, a mechanism not apparent from level-based analysis.

Empirical illustration. We illustrate the full framework—transition independence, latent heterogeneity, and flow decomposition—with three applications in Section 5, each demonstrating a distinct failure mode of conventional DiD with discrete outcomes. First, DiD can produce out-of-bounds counterfactuals: in Charoenwong et al. (2019)'s analysis of the Dodd-Frank Act's 2012 reform, the parallel trends assumption implies counterfactual complaint rates below zero, violating

Figure 1: Counterfactual Complaint Rates from the Dodd-Frank Act.



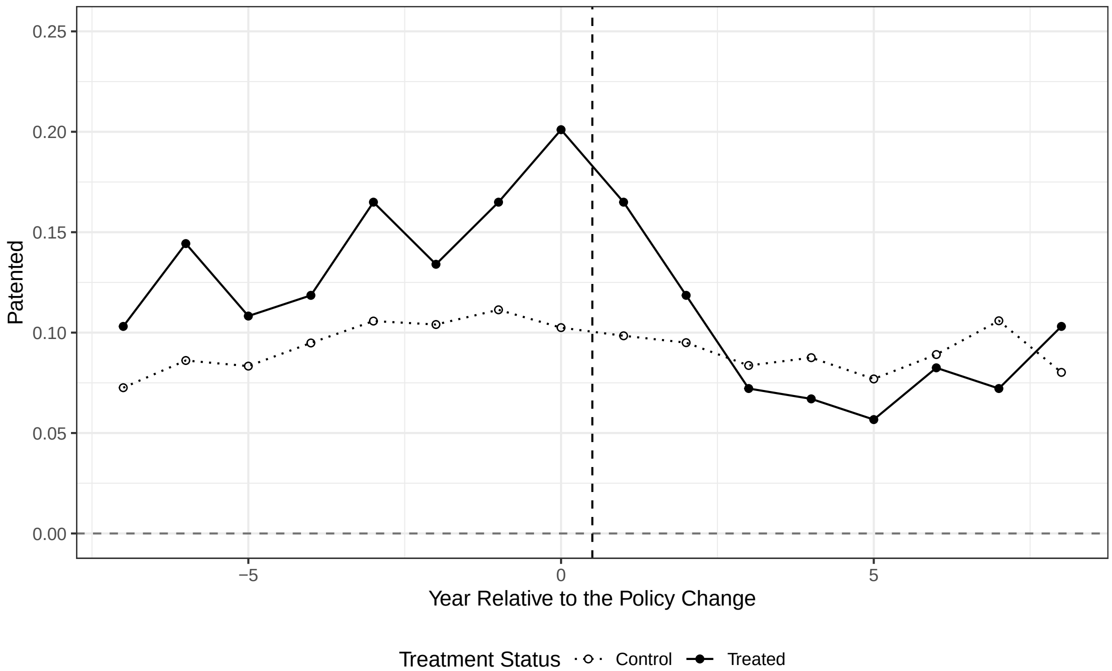
Notes: This plot reports average counterfactual complaint rates of treated units if they were not treated, using the data of Charoenwong et al. (2019). The black solid line represents the observed average complaint rates from treated units. The red dashed line represents the counterfactual complaint rates implied by the parallel trends assumption for difference-in-differences (DiD). The blue dashed line represents the counterfactual complaint rates implied from our proposed method. See Figure 7 for full comparison across different model specifications.

their probabilistic interpretation (Figure 1). Our transition-based approach avoids this out-of-bounds issue by construction, yielding estimates that indicate an increase in service quality, in contrast to the decrease reported by DiD.

Second, DiD is susceptible to mean-reversion bias when treated and control groups differ in baseline levels. In Hvide and Jones (2018)'s study of Norway's 2003 patent law reform, university inventors (treated) had nearly twice the patenting rate of non-university inventors (control) immediately before the reform (Figure 2). Because treated units had more room to decline, mean reversion causes the DiD estimator to overstate the negative impact, reporting a 4.5% decline. Our transition-based approach, which directly accounts for state-dependent dynamics, finds no significant change in patenting rates.

Third, our framework reveals treatment effect mechanisms that DiD cannot detect. Using monthly labor-force status data from the 1990 SIPP panel and following Acemoglu and Angrist (2001) and Lise et al. (2023), we compare disabled (treated) and non-disabled (control) working-age adults around the ADA. The flow decomposition shows that the negative employment effect operates primarily through increased transitions from employment directly into out-of-labor-force

Figure 2: Patenting Rates Around the Norwegian Law Reform.

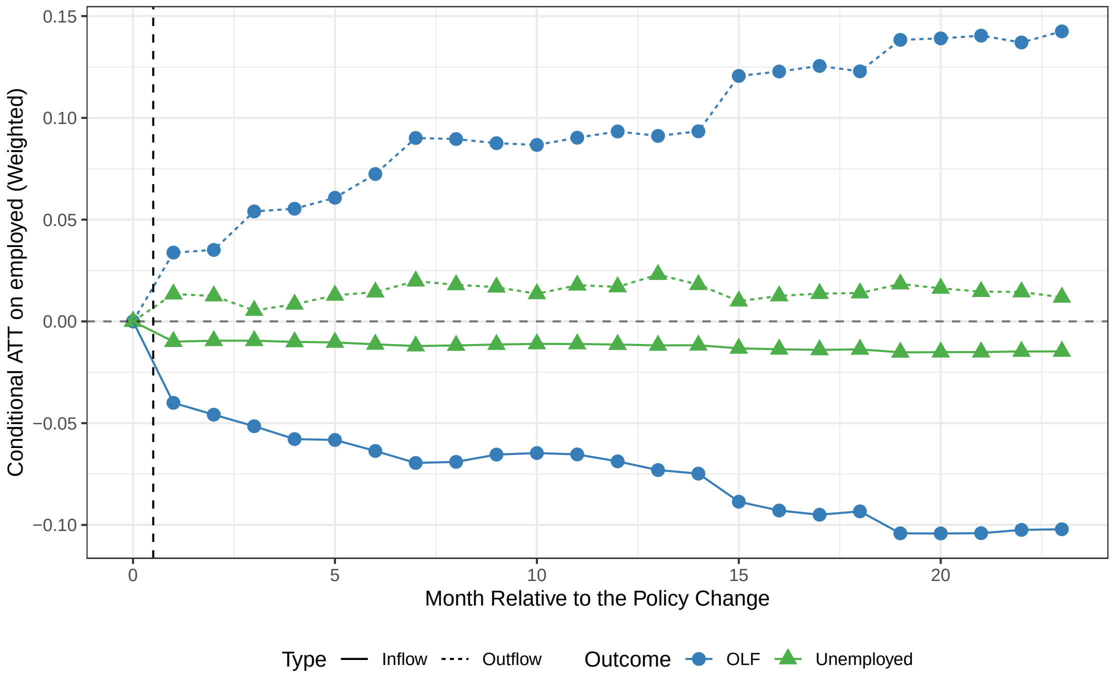


Notes: This plot presents the patenting rates of university inventors (treated group) and non-university inventors (control group) in the years surrounding the introduction of a policy reform that altered patenting rights for university researchers. The x-axis represents years relative to the policy change, with year zero indicating the implementation of the reform. The y-axis shows the annual patenting rates, measured as the proportion of inventors filing at least one patent in a given year. The dashed vertical line marks the year the policy was introduced. Note the substantial baseline difference in patenting rates between the two groups: university inventors exhibited nearly twice the patenting rate of non-university inventors in the years immediately preceding the reform.

status, rather than through changes in job-search outcomes (Figure 3), an insight into the underlying channel-specific transition mechanism unavailable from standard DiD. Furthermore, while conventional DiD fails to detect statistically significant employment effects, our transition-based methods reveal significant short-term employment reductions (Figure 4). This finding is consistent with prior evidence of unintended short-term consequences of the reform (Acemoglu and Angrist, 2001), and further suggests that the effects may be larger once mean reversion and heterogeneous transition dynamics are taken into account.

Related literature. Our framework connects two strands of the program evaluation literature: matching on pre-treatment outcomes, and difference-in-differences with nonlinear dynamics. The first relevant approach comes from the classic literature on program evaluation using matching methods (Card and Sullivan, 1988; Heckman et al., 1997, 1998). For instance, Heckman et al. (1997) use labor force statuses from two periods in their matching procedure for the JTPA program: the month in which a unit becomes eligible and the six months prior to eligibility. We complement

Figure 3: Decomposing the ADA's Effect on Employment by Transition Channels.

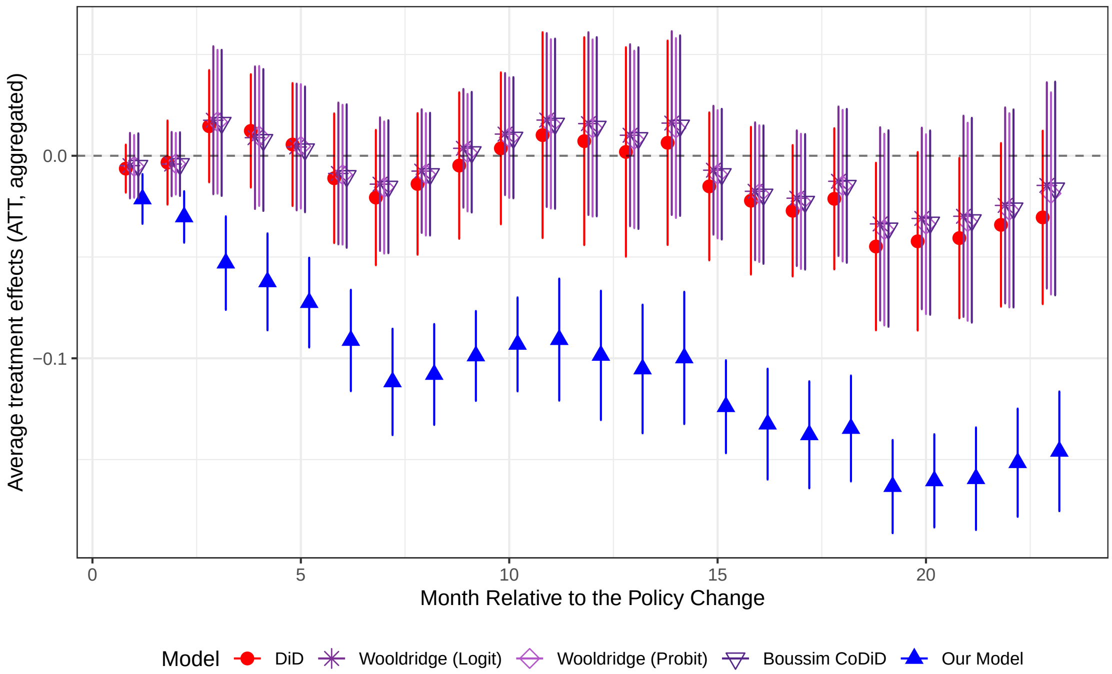


Notes: This figure decomposes the ADA's effect on employment into inflow and outflow transition channels. Solid lines show inflow effects (transitions into employment from unemployment and from out of the labor force), and dashed lines show outflow effects (transitions out of employment). The dominant channel is an increase in transitions from employment directly into out-of-labor-force status. The x-axis measures calendar months relative to the ADA, and the y-axis displays weighted conditional ATTs on employment. See Figure 14 for the full analysis in Section 5.3.

their approach by proposing a new flexible method to construct counterfactuals by identifying and matching groups with similar transition dynamics from past outcomes, accounting for latent heterogeneity that can affect transition dynamics and selection.

The second strand arises from the DiD literature on incorporating pre-treatment outcomes and allowing for nonlinear outcome dynamics (Roth et al., 2023). Earlier studies (Angrist and Pischke, 2009, Chapter 5.4; Ding and Li, 2019) show how conditioning on pre-treatment outcomes yields identifying assumptions distinct from unconditional parallel trends. We complement the literature by providing a practical solution to limited support issues when outcomes are discrete by developing a flexible framework that accommodates both latent heterogeneity and the inherent support restrictions of categorical data; see Section 2.2 for discussion. The change-in-changes model of Athey and Imbens (2006) replaces linear parallel trends with rank invariance on a continuous latent variable; our transition independence assumption instead directly models discrete state-to-state transitions without requiring a continuous latent index. More recent work further relaxes linearity: Wooldridge (2023) proposes nonlinear conditional mean functions and Di Francesco and Mellace (2025) develop probability-shift methods for qualitative outcomes. Our framework differs from these approaches

Figure 4: ATT Estimates of the ADA on Employment: DiD vs Our Method.



Notes: This figure compares the estimated average treatment effects on the treated (ATTs) of the Americans with Disabilities Act (ADA) on employment outcomes between the conventional difference-in-differences (DiD) estimator and our proposed transition-based method ($J = 1$). The x-axis measures calendar months relative to the implementation of the ADA, and the y-axis reports ATT estimates. Red circles represent DiD estimates, and blue circles represent estimates from our proposed method. Vertical bars indicate 95% bootstrap uniform confidence intervals. See Figure 13 for full comparison across model specifications.

by replacing parallel trends entirely with transition independence, extending naturally to multi-category outcomes through transition matrices, and explicitly incorporating latent heterogeneity via a finite mixture structure. In health science, Graves et al. (2022) independently propose using transition matrices for DiD with categorical outcomes, providing an applied framework without formal identification results; we extend this idea by developing nonparametric identification theory under explicit assumptions, incorporating latent heterogeneity via finite mixtures, and deriving flow decomposition of treatment effects with associated asymptotic theory. In particular, the latent-type structure is a distinguishing feature of our framework: it enables identification of type-specific treatment effects from short panels while addressing unobserved heterogeneity in transition dynamics that none of the aforementioned approaches accommodate.

Two recent nonlinear DiD frameworks share our goal of credible counterfactuals for discrete or compositional outcomes but pursue it by transforming the scale on which parallel trends is imposed, rather than by modeling transitions. Wooldridge (2023) assumes common trends in a known link transformation of the conditional *mean* (logit for binary or fractional outcomes, log for counts), while Boussim (2025) assumes "parallel growth"—common trends in the *log-quantities* of each category, equivalent under a random-utility model to parallel trends in latent utilities and to parallel trends in the log-odds of category shares. Both restrictions inherit the functional-form sensitivity of Roth and Sant'Anna (2023): parallel trends on a transformed scale is not invariant to the transformation, so the choice of link is a substantive and largely untestable modeling decision, and the assumption can rarely hold on more than one scale at once. The log-quantity formulation faces the additional difficulty that it is undefined when a category quantity is zero, a recurrent problem for sparse categorical cells (Chen and Roth, 2024). Transition independence instead restricts the untreated transition kernel directly: it requires no link or single-index choice, keeps every counterfactual within the probability simplex by construction, and accommodates unobserved latent-type heterogeneity in transition dynamics. We note, however, that Boussim (2025) additionally identifies treatment effects on category *totals* (the extensive margin), an object our share-based transition framework does not target.

For multi-category outcomes, nonlinear versions of the parallel trends assumption face the same fundamental limitation as standard approaches: the notion of a single "trend" is ill-defined. Applying DiD with a nonlinear link function such as logit or probit accommodates nonlinearity but handles multi-category outcomes by binarizing them—modeling each category separately—thereby discarding the joint transition structure. The identifying restriction is a common trend in the transformed marginal mean and the counterfactual is built from group-by-period marginal means, not from the transition law governing state-to-state movements; nor does this framework incorporate latent heterogeneity in transition dynamics (see Remark 13). Furthermore, with discrete outcomes, accounting for latent heterogeneity in transition dynamics is essential given the limited support. Our transition-based framework addresses these challenges by directly capturing the nonlinear nature of discrete outcomes and explicitly modeling latent heterogeneity. The copula invariance approach in DiD (Callaway et al., 2018; Callaway and Li, 2019; Ghanem et al., 2023) provides a related notion of dependency between outcomes, analogous to transition probabilities; however, with discrete outcomes the limited support necessitates introducing latent types over pre-treatment outcome paths to effectively capture heterogeneity in state dependence and propensity to be treated.

Several recent methodological advances address latent heterogeneity in panel data. Bonhomme and Manresa (2015) and Bonhomme et al. (2022) develop group fixed-effects models that identify latent types and their type-specific treatment effects using clustering methods. Similarly, Arkhangelsky et al. (2021) propose the synthetic difference-in-differences (SDiD) estimator, which accounts for heterogeneous effects by reweighting control units. Both approaches, however, require long panels for consistent estimation, whereas our proposed method allows identification of latent-type-specific treatment effects even in short panels while naturally capturing mean reversion in discrete outcomes.

Another concept related to transition independence is the *sequential exchangeability* assumption, widely used in dynamic treatment and causal panel models (e.g. Robins, 1986; Hernán and Robins, 2025; Marx et al., 2025). This condition requires that, once the complete outcome–treatment history is controlled for, treatment assignment is as good as random with respect to current potential outcomes. In contrast, our transition independence assumption restricts not the assignment mechanism but the evolution of *untreated* potential outcomes, requiring that their transition dynamics conditional on the pre-treatment path be identical across treated and control units. Although both impose conditional independence between potential outcomes and treatment, they differ in their conditioning sets: sequential exchangeability conditions on the *entire observed outcome–treatment history*, whereas transition independence conditions only on the *pre-treatment outcome history*. Consequently, neither condition implies the other (see Remark 3).

While transition independence is a natural assumption for discrete outcomes, it may be less plausible when agents are forward-looking and adjust transition behavior in anticipation of treatment, or when aggregate shocks differentially affect treated and control units' transition dynamics. In practice, we recommend assessing the plausibility of transition independence by testing for pre-treatment differences in transition probabilities between groups, as demonstrated in our empirical applications. See Remarks 4-5 and 21 for further discussion.

The remainder of this paper is organized as follows. Section 2 introduces a potential outcome model with discrete outcomes and illustrates how ATT can be identified from the transition independence assumption. We also offer several remarks, including extensions to staggered treatment allocation. In Section 3, we add latent type structures to the model to allow latent heterogeneity in transition dynamics and establish identification. In Section 4, we develop an estimator and provide a two-stage procedure to estimate latent-type-specific and aggregate ATTs. Section 5 presents three empirical applications. A standalone R package with replication codes is available at `https://github.com/bayesiahn/ak`.

2 Discrete Outcome Models

This section establishes identification of ATTs under transition independence for discrete outcomes. We show that ATTs are point-identified by comparing observed treated outcomes with counterfactuals constructed from control group transition probabilities (Proposition 1). When parallel trends fail unconditionally, we characterize the resulting bias in standard DiD estimates (Proposition 9) and establish equivalence between transition independence and a conditional version of parallel trends (Proposition 14).

2.1 Potential Outcome Model with Transition Independence

Consider a panel data model over $T = T_0 + T_1$ periods, indexed by $t \in \mathcal{T} := \{1, 2, \ldots, T_0, T_0 + 1, ..., T\}$, where T_0 and T_1 denote the numbers of pre- and post-treatment periods, respectively. Observational units are indexed by $i \in \mathcal{N} := \{1, \ldots, n\}$. Each unit may receive a binary treatment at some period or remain untreated throughout. We assume that all treated units begin treatment simultaneously at period $T_0 + 1$ and that, once a unit is treated, then it remains treated for the rest of periods.[1]

For each unit i at time t, an econometrician observes a discrete outcome with K possible categories, denoted by $Y_{it} \in \mathcal{Y} := \{\bar{y}^{(1)}, \bar{y}^{(2)}, ..., \bar{y}^{(K)}\}$, and a binary treatment indicator $D_{it} \in \{0, 1\}$. The potential outcome may depend on the entire treatment path over T periods, denoted by $Y_{it}(D_{i1}, \ldots, D_{iT})$. Let $Y_{it}(\mathbf{0}_{T_0}, \mathbf{1}_{T_1})$ denote the potential outcome for unit i at period t when first treated at T_0+1, where $\mathbf{0}_{T_0}$ and $\mathbf{1}_{T_1}$ are vectors of zeros and ones of lengths T_0 and T_1, respectively. Since treatment is absorbing, the treatment path is determined entirely by the first treatment period. We therefore simplify notation by defining $Y_{it}(1) := Y_{it}(\mathbf{0}_{T_0}, \mathbf{1}_{T_1})$ and $Y_{it}(0) := Y_{it}(\mathbf{0}_T)$. For brevity, let $D_i := D_{i,T_0+1}$ denote the treatment status at period $T_0 + 1$; then $D_{it} = 0$ for $t \leq T_0$ and $D_{it} = D_i$ for $t \geq T_0 + 1$.

For notational brevity, we drop the unit subscript i when no confusion arises; e.g., we write $Y_{it}(0)$ and D_i as $Y_t(0)$ and D, respectively.

Given this definition of $Y_t(d)$, $d \in \{0, 1\}$, define the corresponding vector of binary potential outcomes:

$$\mathbf{X}_t(d) := \begin{pmatrix} X_t^{(1)}(d) \\ \vdots \\ X_t^{(K)}(d) \end{pmatrix} := \begin{pmatrix} \mathbf{1}(Y_t(d) = \bar{y}^{(1)}) \\ \vdots \\ \mathbf{1}(Y_t(d) = \bar{y}^{(K)}) \end{pmatrix} \in \mathcal{X}, \tag{1}$$

where $\mathbf{1}(\cdot)$ denotes the indicator function and $\mathcal{X} := \{(x^{(1)}, ..., x^{(K)}) \in \{0, 1\}^K : \sum_{k=1}^{K} x^{(k)} = 1\}$. Define $\boldsymbol{X}_t$ analogously to $\boldsymbol{X}_t(d)$, replacing $Y_t(d)$ with the observed outcome Y_t. Then, the vector of observed binarized outcome $\boldsymbol{X}_t$ relates to the potential outcomes as

$$\boldsymbol{X}_t = \begin{cases} \boldsymbol{X}_t(0) & \text{if } 1 \leq t \leq T_0 \\ D\boldsymbol{X}_t(1) + (1 - D)\boldsymbol{X}_t(0) & \text{if } t \geq T_0 + 1. \end{cases} \tag{2}$$

Our primary object of interest is the vector of average treatment effects on the treated (ATTs), defined as

$$\boldsymbol{\mu}_t^{\text{ATT}} := \mathbb{E}\left[\boldsymbol{X}_t(1) - \boldsymbol{X}_t(0) \mid D = 1\right], \quad t \geq T_0 + 1, \tag{3}$$

where the kth element of $\boldsymbol{\mu}_t^{\text{ATT}}$ represents the change in the probability of belonging to category k induced by treatment.

We adopt two standard assumptions from the event-study literature: the no anticipatory effects and common support (overlap) assumptions. Let $\boldsymbol{X}_1^{T_0} := \{\boldsymbol{X}_s\}_{s=1}^{T_0}$ denote collection of pre-treatment outcomes.

Assumption NA (No anticipatory effects)**.** For all $t \in \{1, ..., T_0\}$ and $i \in \mathcal{N}$, $\boldsymbol{X}_{it}(1) = \boldsymbol{X}_{it}(0)$.

Assumption CS (Common Support)**.** For any $\boldsymbol{x}_1^{T_0} \in \mathcal{X}^{T_0}$ with $\Pr(\boldsymbol{X}_1^{T_0} = \boldsymbol{x}_1^{T_0}) > 0$, there exists a positive constant $\epsilon > 0$ such that $\epsilon \leq \Pr(D = 1 | \boldsymbol{X}_1^{T_0} = \boldsymbol{x}_1^{T_0}) < 1 - \epsilon$.

[1] Extension to staggered treatment timings can be straightforwardly managed by analyzing subgroups with identical treatment onset periods. We provide details in Remark 7 and Appendix B.

We impose the transition independence assumption, which equates the post-T_0 transition behavior of untreated potential outcomes across treated and control units, conditional on the entire pre-treatment path. Formally, we require that the transition probabilities of untreated potential outcomes be independent of treatment status:

$$\begin{aligned}&\Pr\left(\boldsymbol{X}_t(0)=\boldsymbol{x}_t \mid \boldsymbol{X}_{T_0}(0)=\boldsymbol{x}_{T_0},\ldots,\boldsymbol{X}_1(0)=\boldsymbol{x}_1, D=1\right)\\ &=\Pr\left(\boldsymbol{X}_t(0)=\boldsymbol{x}_t \mid \boldsymbol{X}_{T_0}(0)=\boldsymbol{x}_{T_0},\ldots,\boldsymbol{X}_1(0)=\boldsymbol{x}_1, D=0\right) \quad \text{for } t\geq T_0+1\end{aligned} \tag{TI}$$

for all possible pre-treatment outcome paths $\{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_{T_0}\}$.

Assumption TI (Transition independence)**.** For $t=T_0+1,...,T$, Equation (TI) holds for all $\boldsymbol{x}_t\in\mathcal{X}$ and $\boldsymbol{x}_1^{T_0}\in\mathcal{X}^{T_0}$.

Transition independence implies that the outcome dynamics of control units serve as valid counterfactuals for treated units sharing the same pre-treatment history. By matching treated and control units on their observed outcome paths, the post-treatment transitions of the control group identify what the treated group would have experienced absent treatment.

Our first proposition shows that ATTs in (3) are identified under Assumptions TI, NA, and CS.

Proposition 1. *Suppose that Assumptions TI, NA, and CS hold. Then, for $t=T_0+1,...,T$, ATTs are identified by*

$$\boldsymbol{\mu}_t^{ATT}=\mathbb{E}\left[\mathbf{X}_t-\mathbb{E}\left[\mathbf{X}_t \mid \boldsymbol{X}_1^{T_0}, D=0\right]\middle| D=1\right]. \tag{4}$$

Proposition 1 demonstrates that the ATTs are identified by the difference between the mean of observed post-treatment outcomes from the treated units and that of their *counterfactual* expected untreated potential outcomes conditional on the entire sequence of pre-treatment outcomes. This result is intuitive: the treatment effect is attributable to the difference between the observed outcomes and the counterfactual outcomes that would have been observed if the treated units had not received the treatment. The counterfactual expected untreated potential outcomes are constructed by extrapolating the transition dynamics from the control group to the treated group.

The next corollary shows that the (unconditional) ATT at time t can be written as a weighted average of *conditional* ATTs given the pre-treatment outcome history, with weights given by the treated group's distribution of pre-treatment outcomes.

Corollary 2. *Suppose that Assumptions TI, NA, and CS hold. Then, for $t\geq T_0+1$, ATTs are identified by*

$$\boldsymbol{\mu}_t^{ATT}=\sum_{\boldsymbol{x}_1^{T_0}\in\mathcal{X}^{T_0}}\underbrace{\left\{\mathbb{E}\left[\mathbf{X}_t \mid \boldsymbol{x}_1^{T_0}, D=1\right]-\mathbb{E}\left[\mathbf{X}_t \mid \boldsymbol{x}_1^{T_0}, D=0\right]\right\}\times\Pr(\boldsymbol{X}_1^{T_0}=\boldsymbol{x}_1^{T_0}|D=1)}_{\boldsymbol{X}_1^{T_0}=\boldsymbol{x}_1^{T_0}\ \textit{history-specific contribution to the ATT}}. \tag{5}$$

Corollary 2 provides a useful diagnostic tool for uncovering the mechanisms that drive treatment effects over time. (5) expresses the overall ATT as a mixture of conditional effects across strata defined by the pre-treatment outcome histories, capturing how each history-specific transition contributes to the aggregate effect. The decomposition (5) also allows one to trace how the relative importance of different histories evolves over time, since it is valid for every post-treatment period $t \geq T_0 + 1$.

Remark 3 (Relation to sequential exchangeability)**.** A concept related to transition independence is the *sequential exchangeability* assumption (e.g. Robins, 1986; Hernán and Robins, 2025; Marx et al., 2025), defined as

$$\boldsymbol{X}_t(d) \;\perp\!\!\!\perp\; D_t \,\Big|\, \big(\boldsymbol{X}_{t-1}, D_{t-1}, \ldots, \boldsymbol{X}_1, D_1\big) \quad \text{for } d \in \{0,1\} \text{ and } t = 1, ..., T. \tag{SE}$$

Sequential exchangeability (SE) and transition independence (TI) are distinct conditions that differ in their conditioning sets. Sequential exchangeability requires that the current treatment D_t be as good as random given the *entire observed history of outcomes and treatments*, $\boldsymbol{H}_{t-1} := \{\boldsymbol{X}_s, D_s\}_{s=1}^{t-1}$, whereas transition independence restricts the evolution of *potential outcomes* $\boldsymbol{X}_t(d)$ given only the *pre-treatment control outcome history* $\boldsymbol{X}_1^{T_0}(0) := \{\boldsymbol{X}_s(0)\}_{s=1}^{T_0}$ and the overall treatment status D.

SE conditions on the observed history $\boldsymbol{H}_{t-1}$, which includes post-treatment outcomes $\boldsymbol{X}_{T_0+1}^{t-1} := \{\boldsymbol{X}_s\}_{s=T_0+1}^{t-1}$ that depend on D. Thus, even if SE holds, i.e. $\boldsymbol{X}_t(d) \perp\!\!\!\perp D_t \mid \boldsymbol{H}_{t-1}$, the marginal law $\Pr(\boldsymbol{X}_t(d) \mid \boldsymbol{X}_1^{T_0}(0), D)$ will generally depend on D:

$$\Pr(\boldsymbol{X}_t(d) \mid \boldsymbol{X}_1^{T_0}(0), D) = \sum_{\boldsymbol{h}} \Pr(\boldsymbol{X}_t(d) \mid \boldsymbol{H}_{t-1} = \boldsymbol{h}, D) \, \Pr(\boldsymbol{H}_{t-1} = \boldsymbol{h} \mid \boldsymbol{X}_1^{T_0}, D).$$

Although the first term $\Pr(\boldsymbol{X}_t(d) \mid \boldsymbol{H}_{t-1} = \boldsymbol{h}, D)$ in the summand is invariant in D by SE, the second term $\Pr(\boldsymbol{H}_{t-1} = \boldsymbol{h} \mid \boldsymbol{X}_1^{T_0}, D)$ generally is not, because $\boldsymbol{H}_{t-1}$ contains post-treatment outcomes affected by treatment. Therefore, SE does not imply TI. Intuitively, in event-study settings with permanent treatment, for $t > T_0 + 1$, the SE condition is "trivially true" because D is contained in $\boldsymbol{H}_{t-1}$, and hence fails to impose the cross-group equality in counterfactual outcome dynamics that is required under TI.

Conversely, TI does not in general imply SE. Transition independence governs only the counterfactual no-treatment process, leaving the assignment and treated potential outcomes unrestricted. Even if the transition dynamics of $\boldsymbol{X}_t(0)$ are identical across treated and control units, $\boldsymbol{X}_t(1)$ may remain correlated with D, violating the exogeneity condition required by SE.[2]

Finally, Proposition 1 shows that the transition independence assumption (TI) provides an alternative identification strategy to the classical G-formula (Robins, 1986) for estimating causal

[2] To see this with a concrete example, consider a two-period model ($T_0 = 1$, $T = 2$) with binary outcomes $\mathcal{X} = \{0,1\}$. Let $\Pr(\boldsymbol{X}_2(0) = 1 \mid \boldsymbol{X}_1(0) = x_1, D = d) = 0.5$ for all (x_1, d), so TI holds trivially. However, set $\Pr(\boldsymbol{X}_2(1) = 1 \mid \boldsymbol{X}_1 = 1, D = 1) = 0.9$ while $\Pr(D = 1 \mid \boldsymbol{X}_1 = 1) = 0.8$ and $\Pr(D = 1 \mid \boldsymbol{X}_1 = 0) = 0.2$, so that $\boldsymbol{X}_2(1) \not\perp\!\!\!\perp D \mid \boldsymbol{X}_1$, violating SE.

effects in dynamic settings. While our approach shares the G-formula's reliance on the law of iterated expectations, its specialization to event-study settings allows the ATT to be identified using only pre-treatment information, thereby circumventing the need to condition on potential post-treatment confounding that is required by the G-formula and SE.

Remark 4 (Testing for transition independence in pre-treatment periods)**.** The assumption of transition independence (TI) is inherently non-testable. However, analogous to the "pre-trends" test in the DiD design (Roth, 2022), we can test if transition independence holds in pre-treatment periods by testing the following null hypothesis:

$$\begin{aligned} H_0 : &\Pr\left(\boldsymbol{X}_{T_0} = \boldsymbol{x}_{T_0} | \boldsymbol{X}_1^{T_0-1} = \boldsymbol{x}_1^{T_0-1}, D = 0\right) \\ &= \Pr\left(\boldsymbol{X}_{T_0} = \boldsymbol{x}_{T_0} | \boldsymbol{X}_1^{T_0-1} = \boldsymbol{x}_1^{T_0-1}, D = 1\right) \quad \text{for all } \boldsymbol{x}_1^{T_0} \in \mathcal{X}^{T_0}. \end{aligned} \tag{6}$$

Testing the null hypothesis in equation (6) may face practical difficulties due to limited support in the conditioned outcomes $\boldsymbol{x}_1^{T_0-1}$, especially when the pre-treatment period is long. One alternative approach is to assume that transition independence holds by conditioning on outcomes from only a limited number of pre-treatment periods. We discuss how to implement this approach while incorporating latent heterogeneity in transition probabilities in Section 3 (Remark 21).

Remark 5 (Placebo test)**.** Alternatively, the null hypothesis (6) can be tested using a procedure analogous to the "placebo" ATT estimates proposed by Callaway and Sant'Anna (2021) in the DiD setting. Specifically, construct a sample analogue of (4) with the outcome period set to $t = T_0$ and the conditioning set shifted back by one period. This yields a test of whether $\boldsymbol{\mu}_{T_0}^{\text{ATT}} = \boldsymbol{0}$, since under the no anticipatory effects assumption (Assumption NA) and transition independence at T_0, we have $\boldsymbol{\mu}_{T_0}^{\text{ATT}} = \boldsymbol{0}$.

Remark 6 (Transition independence with covariates)**.** If there are additional covariates available, we may relax Assumption TI by considering the following extension with discrete control variables $V_t \in \mathcal{V}$.

Assumption TI' (Transition independence conditional on covariates)**.** For all $t = T_0 + 1, .., T$,

$$\Pr\left(\boldsymbol{X}_t(0) = \boldsymbol{x}_t \mid \boldsymbol{X}_1^{T_0} = \boldsymbol{x}_1^{T_0}, V_t = v_t, D = 1\right) = \Pr\left(\boldsymbol{X}_t(0) = \boldsymbol{x}_t \mid \boldsymbol{X}_1^{T_0} = \boldsymbol{x}_1^{T_0}, V_t = v_t, D = 0\right) \tag{7}$$

holds for all $\boldsymbol{x}_1^{T_0} \in \mathcal{X}^{T_0}$ and $v_t \in \mathcal{V}$.

An extension of Proposition 1 to the case with covariates is straightforward. Once we condition on covariates for the transition probabilities for counterfactual untreated potential outcomes, we may identify the ATT by replacing the transition probabilities in Proposition 1 with the conditional transition probabilities.

Remark 7 (Staggered treatment adoption)**.** Our identification strategy can be extended to staggered treatment adoption by replacing a treatment group D_i with a treatment *cohort*: the period in which treatment is given, $G_i \in \mathcal{G}$ with $\mathcal{G} \subset \mathcal{T}$, defined by $G_i = \min\{t \mid D_{it} = 1\}$ for treated units. For control units, we may use never-treated or (and) not-yet-treated groups to construct counterfactual untreated potential outcomes as in Callaway and Sant'Anna (2021). We can then extend the transition independence assumption for each treatment cohort, allowing us to identify the ATTs for each treatment cohort using a similar argument as in Proposition 1. The aggregate ATT is then identified by the weighted average of the ATTs for each treatment cohort. A formal treatment of the extension is provided in Appendix B.

Remark 8 (Decomposition by flows)**.** When transition independence holds with one lag of conditioning, the ATT admits a flow decomposition by inflows to and outflows from each outcome state, which is useful for investigating the underlying mechanism through which the treatment dynamically affects outcomes.

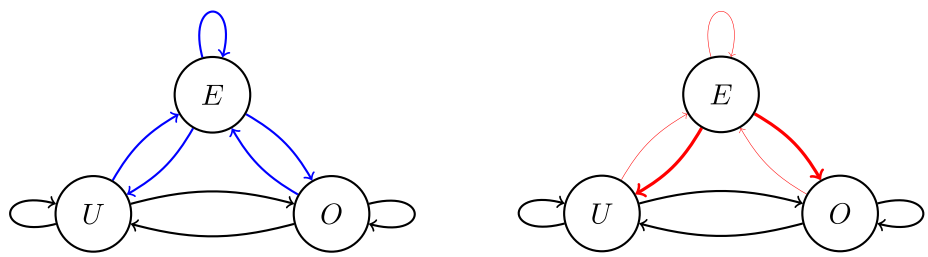
Specifically, from Corollary 2, we can express the ATT on the kth outcome as

$$
\begin{aligned}
&\mathbb{E}\big[X_t^{(k)}(1) - X_t^{(k)}(0) \mid D = 1\big] \\
&= \sum_{y \neq \bar{y}^{(k)}} \underbrace{\Big\{ \Pr(Y_t = \bar{y}^{(k)} \mid Y_{T_0} = y, D = 1) - \Pr(Y_t = \bar{y}^{(k)} \mid Y_{T_0} = y, D = 0) \Big\} \Pr(Y_{T_0} = y \mid D = 1)}_{\text{inflow effect from } y} \\
&\quad - \sum_{y \neq \bar{y}^{(k)}} \underbrace{\Big\{ \Pr(Y_t = y \mid Y_{T_0} = \bar{y}^{(k)}, D = 1) - \Pr(Y_t = y \mid Y_{T_0} = \bar{y}^{(k)}, D = 0) \Big\} \Pr(Y_{T_0} = \bar{y}^{(k)} \mid D = 1)}_{\text{outflow effect to } y} .
\end{aligned}
\tag{8}
$$

The flow decomposition (8) implies that changes in the probability of remaining in the focal state $\bar{y}^{(k)}$ (e.g., employed-to-employed) can be understood as the net contribution of transition flows to and from all alternative states. For illustration, consider the ADA example with three outcome states: Employment (E), Unemployment (U), and Out of Labor Force (O). The introduction of the ADA may induce changes in the transition dynamics across these three outcome states. Under transition independence with the Markov assumption, such impacts on transition probabilities are illustrated in Figure 5, where the ADA induces a larger net outflow from E to U or O by increasing the outflow from E to U or O and decreasing the inflow from U or O to E.

The flow-based decomposition (8) highlights underlying mobility patterns that conventional ATTs on a given outcome could mask. For example, an increase in employment may be driven primarily by reduced separations (lower outflow from E), by enhanced hiring from U or O (higher inflow into E), or by a combination of both. Furthermore, the decomposition provides insight into the underlying mechanism. Many policies, including the ADA, operate through specific channels such as job retention, accommodation costs, or hiring practices. Estimating inflow and outflow components separately can thus help assess which hypothesized channels are empirically salient. When latent heterogeneity is present ($J > 1$), the decomposition extends to each latent type; see

Figure 5: ADA's Effect on Employment Transitions.



$Y_{it}(0)$: Untreated Potential Outcome $\quad$ $Y_{it}(1)$: Treated Potential Outcome

Notes: This figure illustrates how the ADA affects employment rates through changes in inflow and outflow effects in (8) from unemployment (U) and out-of-labor-force (O) states to employment (E) state. Colored (blue and red) arrows indicate the transitions that drive changes in employment rates under treatment. Arrow thickness reflects the relative strength of each flow (from the magnitudes of inflow and outflow effects from each channel estimated in Section 5).

Corollary 25 in the Appendix.

2.2 Comparison with the DiD Estimator

2.2.1 The DiD Estimator and the Parallel Trends Assumption

Many empirical studies apply the DiD estimator when the outcome of interest is binary (Bustos, 2011; Anderson and Sallee, 2011; Hvide and Jones, 2018; Charoenwong et al., 2019). Even when the outcome is discrete with multiple categories, we may apply the DiD estimator to each of binary outcomes constructed from a discrete variable in (1), where the DiD estimator identifies the population quantity given by

$$\boldsymbol{\mu}_t^{\text{DiD}} = \mathbb{E}[\boldsymbol{X}_t - \boldsymbol{X}_{T_0} | D = 1] - \mathbb{E}[\boldsymbol{X}_t - \boldsymbol{X}_{T_0} | D = 0] \quad \text{for } t = T_0 + 1, ..., T.$$

The key assumption for the DiD estimator is the following mean change independence assumption, also known as parallel trends. Recall $\boldsymbol{X}_t(0) := (X_t^{(1)}(0), ..., X_t^{(K)}(0))^\top$ with $X_t^{(k)}(0) := \mathbf{1}(Y_t(0) = \bar{y}^{(k)})$.

Assumption PT (Parallel trends)**.** For $t = T_0 + 1, ..., T$,

$$\mathbb{E}\left[\boldsymbol{X}_t(0) - \boldsymbol{X}_{T_0}(0) \mid D = 0\right] = \mathbb{E}\left[\boldsymbol{X}_t(0) - \boldsymbol{X}_{T_0}(0) \mid D = 1\right]. \tag{PT}$$

The DiD estimator identifies the ATTs under the parallel trends assumption, as documented in the previous literature (Roth et al., 2023).

Proposition 9. *Suppose that Assumptions NA, CS, and PT hold. Then,* $\boldsymbol{\mu}_t^{ATT} = \boldsymbol{\mu}_t^{DiD}$.

However, for bounded discrete outcomes the parallel trends assumption (PT) is difficult to sustain whenever the pre-treatment outcome distributions differ between treated and control units.

Because $\boldsymbol{X}_t(0)$ is confined to the probability simplex, imposing a common mean change across the two groups can force the implied untreated counterfactual for the treated outside the unit interval—an out-of-bounds prediction that cannot describe any distribution of a discrete outcome. The mechanism behind this failure is mean reversion: groups with higher initial rates have less room to rise, so a change that is feasible for one group need not be feasible for the other.

To illustrate, consider the following example with binary employment outcomes. Let $X_t \in \{0, 1\}$ represent binary employment status across two periods, $t = 1, 2$. Initially, 50% of treated units and 25% of control units are employed ($\Pr(X_1(0) = 1 \mid D = 1) = 0.5$ and $\Pr(X_1(0) = 1 \mid D = 0) = 0.25$). By period 2, 87.5% of treated units are employed. We assume that employed individuals remain employed.

As illustrated in Figure 6, suppose that two-thirds of unemployed control units become employed ($\Pr(X_2(0) = 1 \mid X_1(0) = 0, D = 0) = 2/3$). This corresponds to a 50 percentage-point increase in their employment rate, which—under the parallel trends assumption—requires the treated group's employment rate to also rise by 50 percentage points in the absence of treatment. Such an implication is implausible: it would require all unemployed treated units to become employed, yielding 100% employment ($\Pr(X_2(0) = 1 \mid X_1(0) = 0, D = 1) = 1$). This extreme prediction arises because parallel trends ignores mean reversion: groups with higher initial rates have limited room to improve further.

On the other hand, the transition independence assumption constructs counterfactuals directly from transition probabilities, thereby avoiding the implausible implications induced by parallel trends in the presence of mean reversion. In contrast to the parallel trends assumption, our Assumption TI requires that the transition probability function of the potential outcome $\boldsymbol{X}_t(0)$, rather than its mean change over time, be independent of treatment status D as $\Pr(X_2(0) = 1 \mid X_1(0) = 0, D = 0) = \Pr(X_2(0) = 1 \mid X_1(0) = 0, D = 1) = 2/3$. Notably, these assumptions yield opposite ATT estimates in this example: negative under parallel trends but positive under transition independence. Given the implausibility of the parallel trends assumption for bounded outcomes, the transition independence assumption provides a useful alternative for making causal inferences with discrete outcome models.

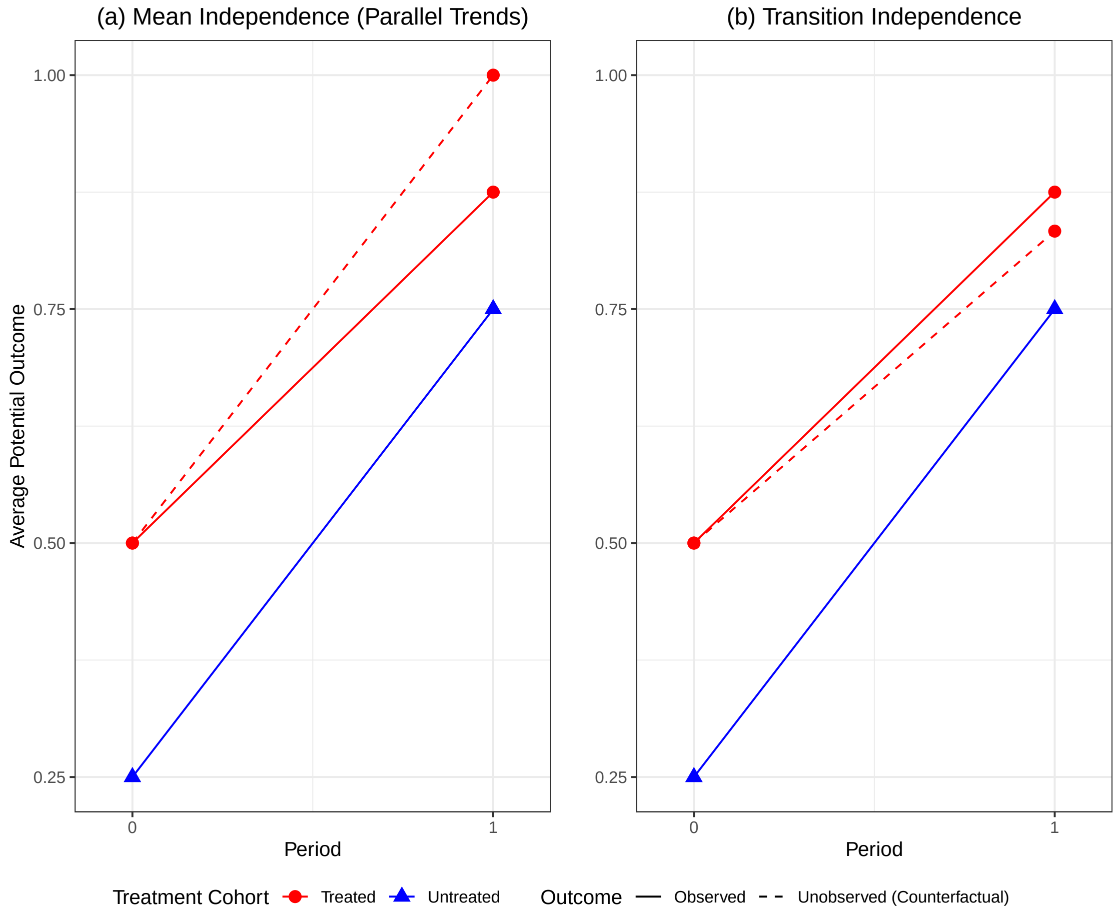
We may characterize the bias of the DiD estimator when the transition independence assumption holds as follows.

Proposition 10. *Suppose that Assumptions TI, NA, and CS hold. Then, the bias of the DiD estimator is given by*

$$\boldsymbol{\mu}_t^{DiD} - \boldsymbol{\mu}_t^{ATT} = \sum_{\boldsymbol{x}_1^{T_0} \in \mathcal{X}^{T_0}} \mathbb{E}[\boldsymbol{X}_t - \boldsymbol{x}_{T_0} | \boldsymbol{X}_1^{T_0} = \boldsymbol{x}_1^{T_0}, D = 0] \times \left\{ \Pr\left(\boldsymbol{X}_1^{T_0} = \boldsymbol{x}_1^{T_0} \mid D = 1 \right) - \Pr\left(\boldsymbol{X}_1^{T_0} = \boldsymbol{x}_1^{T_0} \mid D = 0 \right) \right\} \tag{9}$$

for $t = T_0 + 1, ..., T$.

Figure 6: A Numerical Example: Parallel Trends and Transition Independence.



Notes: The left plot describes the case where the parallel trends assumption holds, yet does not satisfy the transition independence assumption for binary outcome variables. The right plot describes the case where the transition independence assumption holds, yet the parallel trends assumption does not hold. The corresponding data generating process is described in Section 2.

Proposition 10 highlights the two sources of bias for the DiD estimator under the transition independence assumption. The bias arises from differences in pre-treatment outcome distributions between treated and control units, $\Pr(\boldsymbol{X}_1^{T_0} = \boldsymbol{x}_1^{T_0} \mid D = 1) - \Pr(\boldsymbol{X}_1^{T_0} = \boldsymbol{x}_1^{T_0} \mid D = 0)$, multiplied by trends in untreated potential outcomes, $\mathbb{E}[\boldsymbol{X}_t - \boldsymbol{x}_{T_0} | \boldsymbol{X}_1^{T_0} = \boldsymbol{x}_1^{T_0}, D = 0]$. See Ding and Li (2019) for a similar decomposition with continuous outcomes and Angrist and Pischke (2009) for linear models in the context of conditional parallel trends. The DiD estimator may thus be substantially biased when pre-treatment outcome distributions differ between groups, especially when control units experience large outcome changes over time. Since each term is identifiable from the data, the potential bias and its source can be investigated directly.

The same decomposition is the natural vehicle for comparing the standard DiD estimator with the nonlinear DiD of Wooldridge (2023) (see also Blundell et al., 2004; Blundell and Costa Dias, 2009; Puhani, 2012; Athey and Imbens, 2006), which replaces linear parallel trends with parallel trends in a known, strictly increasing link G (logit or probit for binary outcomes). It is convenient to express all three estimators through the untreated counterfactual mean they impute for the treated: writing $\boldsymbol{\mu}_t^{\bullet} = \mathbb{E}[\boldsymbol{X}_t \mid D = 1] - m_t^{\bullet}$, transition independence imputes the counterfactual that identifies the ATT,

$$m_t^{\mathrm{TI}} = \sum_{\boldsymbol{x}_1^{T_0} \in \mathcal{X}^{T_0}} \mathbb{E}\Big[\boldsymbol{X}_t \mid \boldsymbol{X}_1^{T_0} = \boldsymbol{x}_1^{T_0}, D = 0\Big] \Pr\Big(\boldsymbol{X}_1^{T_0} = \boldsymbol{x}_1^{T_0} \mid D = 1\Big),$$

the linear DiD imputes the identity-link counterfactual $m_t^{\mathrm{DiD}} = \mathbb{E}[\boldsymbol{X}_{T_0} \mid D = 1] + \mathbb{E}[\boldsymbol{X}_t - \boldsymbol{X}_{T_0} \mid D = 0]$, and the nonlinear DiD imputes m_t^W, defined componentwise by $G^{-1}(m_t^W) = G^{-1}\big(\mathbb{E}[\boldsymbol{X}_{T_0} \mid D = 1]\big) + G^{-1}\big(\mathbb{E}[\boldsymbol{X}_t \mid D = 0]\big) - G^{-1}\big(\mathbb{E}[\boldsymbol{X}_{T_0} \mid D = 0]\big)$ for binary or fractional margins, and by the simplex-preserving reference-odds form $m_{t,k}^W \propto \mathbb{E}[X_{T_0}^{(k)} \mid D = 1]\ \mathbb{E}[X_t^{(k)} \mid D = 0] / \mathbb{E}[X_{T_0}^{(k)} \mid D = 0]$, normalized across k to sum to one, for $K > 2$. The identity link $G(z) = z$ returns m_t^{DiD}, so linear DiD is the special case of nonlinear DiD with no transformation. Throughout we take the relevant marginal means to be interior, so that G^{-1} and the reference odds are well defined; unlike m_t^{TI} and m_t^W, which are proper distributions, the linear counterfactual m_t^{DiD} need not lie in the simplex—the out-of-bounds defect noted above. The bias of the nonlinear DiD estimator then admits the following decomposition.

Proposition 11 (Bias of nonlinear DiD under transition independence)**.** *Suppose that Assumptions TI, NA, and CS hold and the relevant marginal means are interior. Then, for $t = T_0{+}1, \ldots, T$,*

$$\boldsymbol{\mu}_t^W - \boldsymbol{\mu}_t^{\mathrm{ATT}} = \underbrace{\big(\boldsymbol{\mu}_t^{\mathrm{DiD}} - \boldsymbol{\mu}_t^{\mathrm{ATT}}\big)}_{\textit{composition bias}} + \underbrace{\big(m_t^{\mathrm{DiD}} - m_t^W\big)}_{\textit{index-link curvature}} = m_t^{\mathrm{TI}} - m_t^W, \tag{10}$$

where the composition-bias term is exactly the linear DiD bias of Proposition 10, given in (9). Both terms are identified from the data.

Proposition 11 unifies the linear and nonlinear DiD estimators within a single bias characterization, and shows that the nonlinear link does not remove the source of bias identified in Proposition 10. The nonlinear DiD estimator inherits the same composition bias—the control trend weighted by the difference in pre-treatment history distributions across groups—and merely adds an index-link curvature term that re-expresses the trend on the G^{-1} scale. Three implications follow. First, the identity link sets the curvature term to zero and recovers Proposition 10. Second, equal pre-treatment marginal means across groups also annihilate the curvature term, but not the composition bias, which depends on the full pre-treatment distribution; both terms vanish when the entire pre-treatment distributions coincide ($f_1 = f_0$), the leading case in which linear DiD is itself unbiased. Third, with $K > 2$ categories the gap widens, because m_t^W depends on the control transition structure only through its post-treatment marginal and is therefore blind to the flows between categories, whereas m_t^{TI} propagates the treated baseline distribution through the full transition probabilities (see Remark 15).

The binary case makes the wedge transparent and shows that a nonlinear link can be powerless.

Corollary 12 (Binary illustration)**.** *Let the outcome be binary, $X_t \in \{0,1\}$, with one pre-treatment period T_0 and one post-treatment period T_0+1, and suppose the untreated process is first-order Markov. Write the baseline means $a = \Pr(X_{T_0}(0)=1 \mid D=0)$ and $b = \Pr(X_{T_0}(0)=1 \mid D=1)$, the control transitions $q_0 = \Pr(X_{T_0+1}(0)=1 \mid X_{T_0}(0)=0, D=0)$ and $q_1 = \Pr(X_{T_0+1}(0)=1 \mid X_{T_0}(0)=1, D=0)$, and the control post-treatment mean $c = (1-a)q_0 + aq_1$, with $a,b,c \in (0,1)$. Then the transition-independence and linear/nonlinear DiD counterfactuals are*

$$m^{\mathrm{TI}} = c + (b-a)(q_1-q_0), \qquad m^{\mathrm{DiD}} = b + c - a, \qquad \Lambda^{-1}(m^W) = \Lambda^{-1}(b) + \Lambda^{-1}(c) - \Lambda^{-1}(a),$$

with $\Lambda(z) = e^z/(1+e^z)$ the logistic link. The transition-independence wedge $m^{\mathrm{TI}} - c = (b-a)(q_1-q_0)$ is the product of the composition imbalance $b-a$ and the degree of state dependence $q_1 - q_0$; it is invisible to both linear and nonlinear DiD, whose counterfactuals depend on the data only through the marginal means (a,b,c). The curvature term $m^{\mathrm{DiD}} - m^W$ vanishes if and only if $(a-b)(a-c)(b+c-1) = 0$, so whenever $a = c$ or $b+c = 1$ the logit link returns exactly the linear DiD counterfactual, even though both are biased.

The numerical example of Figure 6 is of this form, with control baseline $a = 1/4$, treated baseline $b = 1/2$, control transitions $q_0 = 2/3$ and $q_1 = 1$ (employed units stay employed), so that $c = 3/4$ and the observed treated post-treatment mean is 0.875. Transition independence imputes $m^{\mathrm{TI}} = 5/6 \approx 0.833$; linear DiD imputes $m^{\mathrm{DiD}} = b + c - a = 1$, the implausible 100%-employment counterfactual flagged above, lying on the boundary of the simplex; and logit DiD imputes $m^W = 0.9$. The implied ATT is $+0.042$ under transition independence, -0.125 under linear DiD, and -0.025 under logit DiD: the link pulls the counterfactual back inside the unit interval but still reports an effect of the wrong sign, because it cannot detect that the treated group's disproportionate weight in the low state mechanically raises its untreated mean.

Remark 13 (Nonlinear DiD and transition independence)**.** The nonlinear DiD of Wooldridge (2023) (building on Blundell et al., 2004; Blundell and Costa Dias, 2009; Puhani, 2012; Athey and Imbens, 2006) is the leading parametric alternative for discrete outcomes. Proposition 11 and Corollary 12 show that, under transition independence, it inherits the linear DiD composition bias and adds only an index-link curvature term, so a logit or probit transformation does not remedy the bias when baseline distributions differ across groups. Wooldridge (2023) also considers a conditional version that conditions on the full pre-treatment history; this is a parametric special case of the transition-based estimator under Assumption CPT, and hence under Assumption TI by Proposition 14. As noted in Remark 15, conditioning on a single binarized outcome is generally insufficient for consistency, while conditioning on the full history is not directly implementable because the conditioning set grows exponentially in the length of the pre-treatment period.

2.2.2 Conditional Parallel Trends with Pre-Treatment Outcomes

Another relevant notion in the DiD literature is the conditional parallel trends assumption, which allows for the parallel trends assumption to hold conditional on the lagged outcomes. Instead of the unconditional parallel trends assumption in Assumption PT, we may consider the following conditional parallel trends assumption given a sequence of pre-treatment outcomes.

Assumption CPT (Parallel trends, conditional on pre-treatment outcomes)**.** For all $\boldsymbol{x}_1^{T_0} \in \mathcal{X}^{T_0}$ and post-treatment period $t = T_0 + 1, ..., T$,

$$\mathbb{E}\left[\boldsymbol{X}_t(0) - \boldsymbol{X}_{T_0}(0) \mid \boldsymbol{X}_1^{T_0} = \boldsymbol{x}_1^{T_0}, D = 0\right] = \mathbb{E}\left[\boldsymbol{X}_t(0) - \boldsymbol{X}_{T_0}(0) \mid \boldsymbol{X}_1^{T_0} = \boldsymbol{x}_1^{T_0}, D = 1\right]. \tag{11}$$

The following proposition shows that the conditional parallel trends assumption (Assumption CPT) is equivalent to the transition independence assumption (Assumption TI).

Proposition 14. *Assumption TI holds if and only if Assumption CPT holds.*

Proposition 14 implies that our approach is equivalent to conditional DiD estimator that conditions on the full pre-treatment history, an equivalence result that clarifies identification but is not directly implementable in typical panel data because the conditioning set grows exponentially in the length of the pre-treatment period. Our contribution is to make this identification result operational. In the next subsection, we impose a Markov restriction to control the dimensionality of the conditioning set, and in the following section we introduce a latent-type structure to accommodate unobserved heterogeneity in transition dynamics.

Remark 15 (Conditional DiD with categorical outcomes)**.** Additional care is required when extending the conditional DiD representation implied by Assumption CPT to outcomes with more than two discrete categories. Suppose we aim to estimate the ATT for category k, defined as

$$\mu_t^{(k)} := \mathbb{E}[X_t^{(k)}(1) - X_t^{(k)}(0) \mid D = 1].$$

A natural approach is to construct a binary indicator for category k, $X_t^{(k)} = \mathbf{1}\{Y_t = \bar{y}^{(k)}\}$, and apply a conditional DiD estimator conditioning on the sequence $\{X_s^{(k)}\}_{s=1}^{T_0}$.

However, conditioning on a single binarized outcome captures only limited information about the unit's underlying categorical history. In particular, Assumption CPT need not hold even if the corresponding conditional parallel trends condition holds for each binarized outcome separately, i.e., even if

$$\mathbb{E}\Big[X_t^{(k)}(0) - X_{T_0}^{(k)}(0) \mid \{X_s^{(k)}(0)\}_{s=1}^{T_0}, D = 0\Big] = \mathbb{E}\Big[X_t^{(k)}(0) - X_{T_0}^{(k)}(0) \mid \{X_s^{(k)}(0)\}_{s=1}^{T_0}, D = 1\Big]$$

for all $k = 1, \ldots, K$.

Consequently, conditioning only on the pre-treatment history of a single binarized outcome is generally insufficient for consistency under Assumption CPT, unless the full collection of binary histories $\{\{X_s^{(k)}\}_{s=1}^{T_0}\}_{k=1}^{K}$ is included in the conditioning variables.

2.3 An Issue in Conditioning on Pre-Treatment Outcomes

Implementing our estimator under Assumption TI entails an exponential increase in the number of possible past outcome paths as the number of pre-treatment periods grows. This, in turn, can introduce non-negligible finite sample bias, since few observations fall into certain paths and overlap between treated and control units in the conditioning variables becomes weak.

To address this challenge, we first propose matching units on a limited history of pre-treatment outcomes, say $\boldsymbol{X}_{T_0-\ell+1}^{T_0}(0) := \{\boldsymbol{X}_{T_0-\ell+1}(0), ..., \boldsymbol{X}_{T_0}(0)\}$ for some small $\ell \geq 1$, rather than the full pre-treatment sequence $\boldsymbol{X}_1^{T_0}(0) := \{\boldsymbol{X}_1(0), ..., \boldsymbol{X}_{T_0}(0)\}$.

Assumption TI-LH (Transition Independence with Limited History)**.** For some $\ell \geq 1$ and for all $t = T_0 + 1, \ldots, T$,

$$\begin{aligned} &\Pr\Big(\boldsymbol{X}_t(0) = \boldsymbol{x}_t \,\Big|\, \boldsymbol{X}_{T_0-\ell+1}^{T_0}(0) = \boldsymbol{x}_{T_0-\ell+1}^{T_0},\, D = 0\Big) \\ &\quad = \Pr\Big(\boldsymbol{X}_t(0) = \boldsymbol{x}_t \,\Big|\, \boldsymbol{X}_{T_0-\ell+1}^{T_0}(0) = \boldsymbol{x}_{T_0-\ell+1}^{T_0},\, D = 1\Big) \end{aligned} \tag{12}$$

for all $\boldsymbol{x}_t \in \mathcal{X}$ and all histories $\boldsymbol{x}_{T_0-\ell+1}^{T_0} \in \mathcal{X}^\ell$.

In practice, we recommend reporting ATT estimates using different lengths of pre-treatment conditioning histories. The "pre-transition" assumption can also be tested, as described in Remark 4, by reporting test statistics for the null hypothesis $H_0 : \Pr(\boldsymbol{X}_{T_0}(0) = \boldsymbol{x}_{T_0} \mid \boldsymbol{X}_{T_0-\ell}^{T_0-1}(0) = \boldsymbol{x}_{T_0-\ell}^{T_0-1}(0), D = 0) = \Pr(\boldsymbol{X}_{T_0}(0) = \boldsymbol{x}_{T_0} \mid \boldsymbol{X}_{T_0-\ell}^{T_0-1}(0) = \boldsymbol{x}_{T_0-\ell}^{T_0-1}(0), D = 1)$ using pre-treatment data. Additionally, comparing BIC values from likelihood-based estimation of transition probabilities—either using pre-treatment data from both groups or using the full sample of control units—can be used to guide the choice of lag length.

Adopting Assumption TI-LH in place of Assumption TI provides a practical solution to finite sample bias as well as weak overlap between treated and control units due to exponentially increasing past outcome paths, yet at the cost of potential violation of Assumption TI-LH due to the presence of latent heterogeneity in transition dynamics. Section 3 considers the possibility of latent heterogeneity in transition dynamics.

3 Discrete Outcome Models with Latent Heterogeneity

This section introduces latent heterogeneity into the transition-based framework. By augmenting the model with a finite discrete latent type, we show that both latent-type-specific and aggregate ATTs remain identified from short panel data under type-specific transition independence and a Markov assumption (Propositions 16 and 17).

Assumption TI or TI-LH may be invalid in the presence of unobserved heterogeneity in transition dynamics that may simultaneously affect both the treatment allocations and the transition probabilities of potential outcomes under no treatment. To address such concerns, we introduce a finite discrete latent type to capture unobserved heterogeneity. We assume each unit i belongs to one of J latent types, where J is known. Let $Z_i \in \mathcal{J} := \{1, ..., J\}$ denote unit i's latent type, with $\pi^j := \Pr(Z_i = j)$ representing the population probability of type j for $j = 1, 2, ..., J$.

We consider assumptions analogous to Assumptions TI and CS, conditional on latent types.

Assumption TI-LT (Transition independence, conditional on latent types)**.**

$$\Pr(\boldsymbol{X}_t(0) = \boldsymbol{x}_t \mid \boldsymbol{X}_1^{T_0} = \boldsymbol{x}_1^{T_0}, D = 0, Z = j) = \Pr(\boldsymbol{X}_t(0) = \boldsymbol{x}_t \mid \boldsymbol{X}_1^{T_0} = \boldsymbol{x}_1^{T_0}, D = 1, Z = j) \quad (13)$$

for all $\boldsymbol{x}_1^{T_0} \in \mathcal{X}^{T_0}$ and $j = 1, \ldots, J$.

Assumption CS-LT (Common support conditional on latent types)**.** There exists a positive constant $\epsilon > 0$ such that $\epsilon \leq Pr(D_i = 1 | \boldsymbol{X}_1^{T_0} = \boldsymbol{x}_1^{T_0}, Z = j) \leq 1 - \epsilon$ for all $\boldsymbol{x}_1^{T_0} \in \mathcal{X}^{T_0}$ and $j = 1, \ldots, J$.

One can consider the latent-type average treatment effects on the treated, ATT for the latent type where transition independence holds within. We define a vector of the *average treatment effects on the treated for latent type j (LTATTs)* as

$$\boldsymbol{\mu}_t^{\mathrm{ATT},j} := \mathbb{E}\left[\mathbf{X}_t(1) - \mathbf{X}_t(0) \mid D = 1, Z = j\right] \quad \text{for } j \in \mathcal{J}. \quad (14)$$

Taking the average over latent-type-specific ATTs across types, we may estimate the average treatment effects on treated (ATTs) as

$$\boldsymbol{\mu}_t^{\mathrm{ATT}} := \mathbb{E}\left[\mathbf{X}_t(1) - \mathbf{X}_t(0) \mid D = 1\right] = \sum_{j=1}^{J} \Pr(Z = j \mid D = 1) \boldsymbol{\mu}_t^{\mathrm{ATT},j} \quad \text{for } t = T_0 + 1, ..., T. \quad (15)$$

We account for latent heterogeneity in transition dynamics using a finite mixture model under Assumptions NA, TI-LT, and CS-LT. Let $\boldsymbol{W}_i := (\boldsymbol{X}_{i1}^\top, \ldots, \boldsymbol{X}_{iT}^\top, D_i)^\top \in \boldsymbol{\mathcal{W}} := \boldsymbol{\mathcal{X}}^T \times \{0, 1\}$ denote the vector of binary outcomes over T periods and the treatment indicator for unit i. We assume that the data are generated from the mixture distribution whose probability mass function (PMF) is given by:

$$p_{\boldsymbol{W}}(\boldsymbol{w}) = \sum_{j=1}^{J} \pi^j p_{\boldsymbol{W}}^j(\boldsymbol{w}), \quad (16)$$

where $\pi^j := \Pr(Z_i = j)$ is the population proportion of latent type j, and

$$p_{\boldsymbol{W}}^j(\boldsymbol{w}) = \Pr(\boldsymbol{W} = \boldsymbol{w} | Z = j)$$

denotes the conditional PMF of $\boldsymbol{W} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_T, D)$ given $Z = j$. The true number of components, J, is defined as the smallest integer such that the data distribution admits the representation (16).

Assumption RSM. (Random sampling from a finite mixture distribution) (a) We observe a sample of n i.i.d. draws $\{\boldsymbol{W}_i\}_{i=1}^n$, where $\boldsymbol{W}_i \overset{i.i.d.}{\sim} p_{\boldsymbol{W}}(\boldsymbol{w})$ given in (16), where $p^j_{\boldsymbol{W}}(\boldsymbol{w})$ satisfies Assumptions NA, TI-LT, and CS-LT with the relationship between observed outcome and potential outcomes given in (2). (b) The true number of components J in (16) is known. (c) The population mixture weights are ordered such that $\pi^1 < \pi^2 < \cdots < \pi^J$.

Assumption RSM(a) consolidates the maintained structural restrictions—transition independence, the no-anticipation condition, and the overlap requirement—into a unified sampling framework by specifying that the observed data are drawn from a J-component finite mixture, where each component corresponds to a latent type with its own transition dynamics. Assumption RSM(b) specifies that the number of latent components J is known. In practice, this can be assessed using information criteria (e.g., BIC) or other procedures, which we discuss below. Assumption RSM(c) imposes an ordering restriction to ensure model identifiability, since finite mixture models are identified only up to permutation of their components.

For identification, we assume the potential outcome without treatment $\{\boldsymbol{X}_t(0)\}_{t=1}^T$ follows a first-order Markov process conditional on the initial outcome $\boldsymbol{X}_1(0)$ and latent type Z. This Markovian assumption also serves as a practical condition for identifying the LTATTs from observed data, avoiding the computational burden and finite sample issues that arise from conditioning on complete outcome histories. Recall $\boldsymbol{X}_1^{t-1}(d) := \{\boldsymbol{X}_s(d)\}_{s=1}^{t-1}$.

Assumption M (the first-order Markov)**.** For all $j = 1, 2, ..., J$, conditional on $Z = j$ and $D = d$, $\{\boldsymbol{X}_t(d) : t = 1, ..., T\}$ follows a (non-stationary) first-order Markov process, i.e., for $t = 2, ..., T$ and all $d \in \{0, 1\}$,

$$\Pr\left(\boldsymbol{X}_t(d) \mid \{\boldsymbol{X}_s(d)\}_{s=1}^{t-1}, D = d, Z = j\right) = \Pr\left(\boldsymbol{X}_t(d) \mid \boldsymbol{X}_{t-1}(d), D = d, Z = j\right).$$

Collect the unknown probability mass functions for Z and the type-specific transition probabilities of the potential outcomes $\boldsymbol{X}_t(d)$, $d \in \{0, 1\}$, into the parameter vector

$$\boldsymbol{\psi} = (\boldsymbol{\pi}, \boldsymbol{\varphi}^1, ..., \boldsymbol{\varphi}^J) \in \Theta_{\boldsymbol{\psi}},$$

where $\boldsymbol{\pi} = (\pi^1, ..., \pi^J)^\top$ satisfies $\epsilon \le \pi^j \le 1-\epsilon$ for some $\epsilon > 0$ and $\sum_{j=1}^J \pi^j = 1$, and each component $\boldsymbol{\varphi}^j$ is defined as a collection of type-specific initial distributions and transition probabilities as follows:

$$\boldsymbol{\varphi}^j = \left\{ p^j_{\boldsymbol{X}_1(0),D}(\cdot,\cdot), \left\{ p^j_{\boldsymbol{X}_t(0)|\boldsymbol{X}_{t-1}(0)}(\cdot|\cdot) \right\}_{t=2}^T, \left\{ p^j_{\boldsymbol{X}_t(1)|\boldsymbol{X}_{t-1}(1)}(\cdot|\cdot) \right\}_{t=T_0+1}^T \right\},$$

where

$$p^j_{\boldsymbol{X}_1(0),D}(\boldsymbol{x}_1, d) := \Pr\left(\boldsymbol{X}_1(0) = \boldsymbol{x}_1, D = d \mid Z = j\right),$$
$$p^j_{\boldsymbol{X}_t(d)|\boldsymbol{X}_{t-1}(d)}(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) := \Pr\left(\boldsymbol{X}_t(d) = \boldsymbol{x}_t \mid \boldsymbol{X}_{t-1}(d) = \boldsymbol{x}_{t-1}, Z = j\right)$$

for $d \in \{0,1\}$ and $t \in \{2,...,T\}$.

Under Assumption M, and noting from Assumption NA that $\boldsymbol{X}_t(0) = \boldsymbol{X}_t(1)$ for $t = 1, \dots, T_0$, by explicitly writing its dependence on the parameters $\boldsymbol{\psi}$ and $\boldsymbol{\varphi}^j$, the PMF of $\boldsymbol{W} = (\boldsymbol{X}_1, \dots, \boldsymbol{X}_T, D)$ in (16) can be written as

$$p_{\boldsymbol{W}}(\boldsymbol{w}; \boldsymbol{\psi}) := \sum_{j=1}^{J} \pi^j p^j_{\boldsymbol{W}}(\boldsymbol{w}; \boldsymbol{\varphi}^j) \tag{17}$$

where

$$\begin{aligned} &p^j_{\boldsymbol{W}}(\boldsymbol{w}; \boldsymbol{\varphi}) \\ &:= \begin{cases} p^j_{\boldsymbol{X}_1(0),D}(\boldsymbol{x}_1, 0) \prod_{t=2}^{T} p^j_{\boldsymbol{X}_t(0)|\boldsymbol{X}_{t-1}(0)}(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) & \text{if } D = 0 \\ p^j_{\boldsymbol{X}_1(0),D}(\boldsymbol{x}_1, 1) \prod_{t=2}^{T_0} p^j_{\boldsymbol{X}_t(0)|\boldsymbol{X}_{t-1}(0)}(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) \prod_{t=T_0+1}^{T} p^j_{\boldsymbol{X}_t(1)|\boldsymbol{X}_{t-1}(1)}(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) & \text{if } D = 1. \end{cases} \end{aligned}$$

Let $\boldsymbol{\psi}^*$ denote the true value of $\boldsymbol{\psi}$, so that the true distribution of $\boldsymbol{W}$ is given by $p_{\boldsymbol{W}}(\boldsymbol{w}; \boldsymbol{\psi}^*)$.

By extending the arguments in Anderson (1954), Madansky (1960), Hu (2008), Kasahara and Shimotsu (2009), Carroll et al. (2010), Hu and Shum (2012), and Ishimaru (2025), the following proposition establishes the identification of the mixture models for the potential outcome processes $\{\boldsymbol{X}_t(0)\}_{t=1}^T$ and $\{\boldsymbol{X}_t(1)\}_{t=T_0+1}^T$ in (17), even when the panel is relatively short.

Proposition 16. *Suppose that Assumptions RSM, M, and ID hold. If $T_0 \geq k+1$, $T \geq 2(k+1)$, and $J \leq |\mathcal{X}|^k$ for $k = 1, 2, 3, \dots$, then $\boldsymbol{\psi}$ is uniquely identified from $\{p_{\boldsymbol{W}}(\boldsymbol{w}; \boldsymbol{\psi}) : \boldsymbol{w} \in \boldsymbol{\mathcal{W}}\}$.*

Assumption ID, presented in the proof of Proposition 16, imposes a set of regularity conditions for identification. The regularity conditions ensure that certain matrices of transition probabilities are non-singular, guaranteeing that variations in past outcomes induce sufficiently heterogeneous changes in current outcome probabilities across latent types for identification.

Applying Proposition 16 with $k = 1$, we can identify $\boldsymbol{\psi}$ when $J \leq |\mathcal{X}|$, $T_0 = 2$, and $T = 4$ under these regularity conditions. This implies, for example, that when the discrete outcome takes three possible values (e.g., E, U, and O in the ADA example), we may identify three latent types ($J = 3$) from four-period panel data with two pre-treatment periods ($T = 4$, $T_0 = 2$) under the first-order Markov assumption. When the panel is longer, we may identify more latent types.

While the regularity conditions in Assumption ID are not directly verifiable from the data, they are generically satisfied: the set of parameter values violating any of the rank conditions has Lebesgue measure zero in the parameter space. In practice, failure of these conditions would manifest as observational equivalence between models with different numbers of latent types, which can be diagnosed through the model selection procedures described in Remark 22.

Under Assumption M, Assumption TI-LT implies that

$$\Pr\left(\boldsymbol{X}_t(0) = \boldsymbol{x}_t \mid \boldsymbol{X}_{T_0}(0) = \boldsymbol{x}_{T_0}, D = 0, Z = j\right) = \Pr\left(\boldsymbol{X}_t(0) = \boldsymbol{x}_t \mid \boldsymbol{X}_{T_0}(0) = \boldsymbol{x}_{T_0}, D = 1, Z = j\right).$$

Then, given the identification of $\boldsymbol{\psi}$ in Proposition 16, we may identify the LTATTs $\boldsymbol{\mu}_t^{ATT,j}$ (14) for each j as the following proposition states.

Proposition 17. *Suppose that Assumptions RSM, M, TI-LT, and ID hold. Then, for each post-treatment $t \geq T_0+1$ and for all $j \in \mathcal{J}$, we may uniquely identify $\boldsymbol{\mu}_t^{ATT,j}$ from $\{p_{\boldsymbol{W}}(\boldsymbol{w};\boldsymbol{\psi}) : \boldsymbol{w} \in \boldsymbol{\mathcal{W}}\}$ as*

$$\boldsymbol{\mu}_t^{ATT,j} = \sum_{\boldsymbol{x}_{T_0} \in \mathcal{X}} \Pr(\boldsymbol{X}_{T_0} = \boldsymbol{x}_{T_0} | D = 1, Z = j) \times \left(\mathbb{E}\left[\boldsymbol{X}_t \mid \boldsymbol{X}_{T_0} = \boldsymbol{x}_{T_0}, D = 1, Z = j\right] - \mathbb{E}\left[\boldsymbol{X}_t \mid \boldsymbol{X}_{T_0} = \boldsymbol{x}_{T_0}, D = 0, Z = j\right]\right). \tag{18}$$

As a corollary, we may identify the ATT $\boldsymbol{\mu}_t^{ATT}$ defined in (15) for each post-treatment t.

Corollary 18. *Under Assumptions TI-LT, RSM, M, and ID, for each post-treatment $t \geq T_0 + 1$, we may uniquely identify the ATT $\boldsymbol{\mu}_t^{ATT}$ from $\{p_{\boldsymbol{W}}(\boldsymbol{w};\boldsymbol{\psi}) : \boldsymbol{w} \in \boldsymbol{\mathcal{W}}\}$.*

Remark 19 (Latent heterogeneity and nonlinear DiD)**.** With latent types ($J > 1$), the nonlinear (logit) DiD of Remark 13 faces two further sources of bias beyond those in Corollary 12, even when index parallel trends holds *within* each latent type. First, because the latent type is unobserved, a nonlinear DiD that conditions only on observed covariates (Wooldridge, 2023, Assumption CIPT, eq. (2.25)) cannot account for differences in the latent-type composition of the treated and control groups. Second, the logit link does not aggregate linearly across types: the logit of a mixture mean is not the type-weighted average of the type-specific logits (Jensen's inequality), so type-wise index parallel trends does not imply index parallel trends for the pooled population. The transition-based estimator avoids both problems by identifying the latent types (Proposition 16) and aggregating the type-specific counterfactual distributions linearly in probabilities, as in (18). Table 1 illustrates both wedges in a two-type binary example in which index parallel trends holds within each type yet fails for the pooled population.

The first-order Markov assumption in Assumption M can be straightforwardly relaxed to a higher (but finite) order Markov process, provided that the length of T_0 and T is sufficiently large as the following proposition shows.

Assumption M$_\ell$ (the ℓth-order Markov)**.** For all $j = 1, 2, ..., J$, conditional on $Z = j$ and $D = d$, $\{\boldsymbol{X}_t(d) : t = 1, ..., T\}$ follows a (non-stationary) ℓth-order Markov process, i.e., for $t = \ell+1, ..., T$ and all $d \in \{0, 1\}$, $\Pr(\boldsymbol{X}_t(d) \mid \{\boldsymbol{X}_s(d)\}_{s=1}^{t-1}, D = d, Z = j) = \Pr(\boldsymbol{X}_t(d) \mid \{\boldsymbol{X}_s(d)\}_{s=t-\ell}^{t-1}, D = d, Z = j)$.

Under Assumptions M$_\ell$ and NA, with abuse of notation, we redefine the parameter vector as $\boldsymbol{\psi} = (\boldsymbol{\pi}, \boldsymbol{\varphi}^1, \dots, \boldsymbol{\varphi}^J) \in \Theta_{\boldsymbol{\psi}}$ with each component given by

$$\boldsymbol{\varphi}^j = \left\{ p^j_{\{\boldsymbol{X}_s(0)\}_{s=1}^{\ell}, D}(\cdot, \cdot), \left\{ p^j_{\boldsymbol{X}_t(0) | \{\boldsymbol{X}_s(0)\}_{s=t-\ell}^{t-1}}(\cdot|\cdot) \right\}_{t=\ell+1}^{T}, \left\{ p^j_{\boldsymbol{X}_t(1) | \{\boldsymbol{X}_s(1)\}_{s=t-\ell}^{t-1}}(\cdot|\cdot) \right\}_{t=T_0+1}^{T} \right\},$$

Table 1: Latent Heterogeneity: Transition Independence versus Pooled Nonlinear (Logit) DiD ($J = 2$).

	Equal type composition	Unequal type composition
Latent-mixture counterfactual (m^{TI})	0.50	0.44
Pooled logit DiD counterfactual (m^{W})	0.50	0.33
Discrepancy	0.00	0.11

Notes: Two latent types with type-specific (pre-treatment, post-treatment) untreated means $(0.9, 0.6)$ for type 1 and $(0.3, 0.4)$ for type 2; the same type-level dynamics hold for treated and control units, so transition independence holds within each type. In the "equal composition" column both groups place weight 1/2 on each type; in the "unequal composition" column the treated group places weights $(0.2, 0.8)$ on types 1 and 2 while the control group keeps $(0.5, 0.5)$. The latent-mixture counterfactual aggregates the type-specific counterfactuals linearly in probabilities, whereas the pooled logit DiD applies index parallel trends to the pooled marginal means. The two coincide under equal composition and diverge under unequal composition, reflecting the unobserved-composition and nonlinear-aggregation (Jensen) wedges.

where $p^j_{\{\boldsymbol{X}_s(0)\}_{s=1}^{\ell},D}(\cdot,\cdot)$ and $p^j_{\boldsymbol{X}_t(d)|\{\boldsymbol{X}_s(d)\}_{s=t-\ell}^{t-1}}(\cdot|\cdot)$ for $d \in \{0,1\}$ are defined similarly to $p^j_{X_t(0),D}(\cdot,\cdot)$ and $p^j_{\boldsymbol{X}_t(d)|\boldsymbol{X}_{t-1}(d)}(\cdot|\cdot)$ but using $\{\boldsymbol{X}_s(0)\}_{s=1}^{\ell}$ and $\{\boldsymbol{X}_s(d)\}_{s=t-\ell}^{t-1}$ in place of $X_t(0)$ and $\boldsymbol{X}_{t-1}(d)$, respectively.

Proposition 20. *Suppose that Assumptions RSM, M_ℓ, and ID_ℓ hold. If $T_0 \geq 2\ell$, $T \geq 4\ell$, and $J \leq |\mathcal{X}|^\ell$ for $\ell = 1, 2, \ldots$, then: (a) $\boldsymbol{\psi}$ is uniquely identified from $\{p_{\boldsymbol{W}}(\boldsymbol{w};\boldsymbol{\psi}) : \boldsymbol{w} \in \boldsymbol{\mathcal{W}}\}$; (b) for each post-treatment $t \geq T_0 + 1$ and for all $j \in \mathcal{J}$, we may uniquely identify $\boldsymbol{\mu}_t^{ATT,j}$ from $\{p_{\boldsymbol{W}}(\boldsymbol{w};\boldsymbol{\psi}) : \boldsymbol{w} \in \boldsymbol{\mathcal{W}}\}$ as*

$$
\begin{aligned}
\boldsymbol{\mu}_t^{ATT,j} = \sum_{\boldsymbol{x}_{T_0-\ell+1}^{T_0} \in \mathcal{X}^\ell} & \Pr\left(\boldsymbol{X}_{T_0-\ell+1}^{T_0} = \boldsymbol{x}_{T_0-\ell+1}^{T_0} | D = 1, Z = j\right) \\
& \times \Big\{ \mathbb{E}\big[\boldsymbol{X}_t \mid \boldsymbol{X}_{T_0-\ell+1}^{T_0} = \boldsymbol{x}_{T_0-\ell+1}^{T_0}, D = 1, Z = j\big] \\
& \quad - \mathbb{E}\big[\boldsymbol{X}_t \mid \boldsymbol{X}_{T_0-\ell+1}^{T_0} = \boldsymbol{x}_{T_0-\ell+1}^{T_0}, D = 0, Z = j\big] \Big\}.
\end{aligned} \tag{19}
$$

Under the second-order Markov assumption, for example, Proposition 20 with $\ell = 2$ implies that the LTATTs are identified when $J \leq |\mathcal{X}|^2$, $T_0 = 4$, and $T = 8$, provided that the stated regularity conditions are satisfied.

Remark 21 (Testing for transition independence across pre-treatment periods, continued)**.** Under the Markovian assumption, testing transition independence in pre-treatment periods (Remark 4) can be applied recursively across pre-treatment periods and can be conducted using a graphical diagnostic comparing conditional means between treated and control units, analogous to the standard eyeball tests for pre-trends in event-study plots (Freyaldenhoven et al., 2019; Kahn-Lang and Lang, 2020; Roth, 2022; Rambachan and Roth, 2023). In particular, under Assumption M, the conditioning set in the null hypothesis (6) collapses from the entire sequence of past outcomes to the most recent outcome alone, yielding the following specification:

$$\begin{aligned} H_0 : &\Pr\left(\boldsymbol{X}_t = \boldsymbol{x}_t | \boldsymbol{X}_{t-1} = \boldsymbol{x}_{t-1}, D = 0\right) \\ &= \Pr\left(\boldsymbol{X}_t = \boldsymbol{x}_t | \boldsymbol{X}_{t-1} = \boldsymbol{x}_{t-1}, D = 1\right) \quad \text{for all } (\boldsymbol{x}_{t-1}, \boldsymbol{x}_t) \in \mathcal{X}^2, t = 2, \dots, T_0. \end{aligned} \tag{20}$$

Under the null hypothesis (20), the difference in conditional probabilities should be equal to zero for all pre-treatment periods $t = 2, \dots, T_0$. This implication yields a graphical diagnostic for transition independence analogous to pre-trends testing in difference-in-differences. Specifically, one may plot the difference in conditional means of $\boldsymbol{X}_t$ between treated and control units for each lagged outcome $\boldsymbol{x}_{t-1}$ across pre-treatment periods $t = 2, \dots, T_0$ and visually assess whether these differences are close to zero. The null hypothesis (20) can also be formally tested by constructing uniform confidence bands for the differences across all pre-treatment periods using bootstrap methods, extending the approach in Callaway and Sant'Anna (2021) to our setting. In the presence of latent types, type-specific conditional means for treated and control units for each latent type can also be consistently estimated using posterior type probabilities obtained via Bayes' rule. Implementation details are provided in Appendix D.

Remark 22 (Choosing the number of latent types)**.** The number of latent types J is a key parameter in the model and is assumed to be known to the econometrician. One may determine the number of latent types using the Bayesian Information Criterion (BIC), computed from the likelihood function in (17) with an additional penalty term as in Bonhomme and Manresa (2015), or through sequential hypothesis testing based on likelihood ratio tests (e.g., Kasahara and Shimotsu, 2015).

Alternatively, one may implement the following iterative procedure: beginning with $J = 1$, estimate the model and test transition independence using the procedure in Remark 21. If the null hypothesis is not rejected, select J as the number of latent types. Otherwise, increment J by one and repeat.

4 Estimation

We develop a two-stage estimator that is consistent and asymptotically normal (Proposition 23). The first stage estimates the mixture model via maximum likelihood using the EM algorithm; the second computes treatment effects using the estimated posterior type probabilities.

We propose the following two-stage estimation procedure. For brevity, we assume a first-order Markov process, but extension to higher-order Markov processes is straightforward.

In the first stage, we estimate $\boldsymbol{\psi}$ by the maximum likelihood estimator

$$\hat{\boldsymbol{\psi}} = \arg\max_{\boldsymbol{\psi} \in \Theta_{\boldsymbol{\psi}}} \sum_{i=1}^{n} \log p_{\boldsymbol{W}}(\boldsymbol{W}_i; \boldsymbol{\psi}), \tag{21}$$

where $p_{\boldsymbol{W}}(\boldsymbol{w};\boldsymbol{\psi})$ is the likelihood function defined in (17), and $\{\boldsymbol{W}_i\}_{i=1}^n$ is a random sample of n i.i.d. observations as described in Assumption RSM.

In the second stage, define the conditional type probabilities of $Z = j$ given $\boldsymbol{W} = \boldsymbol{w}$ as a function of $\boldsymbol{\psi}$ as

$$\tau^j(\boldsymbol{w};\boldsymbol{\psi}) := \frac{\pi^j p^j_{\boldsymbol{W}}(\boldsymbol{w};\boldsymbol{\varphi}^j)}{\sum_{k=1}^J \pi^k p^k_{\boldsymbol{W}}(\boldsymbol{w};\boldsymbol{\varphi}^k)}.$$

From the MLE $\hat{\boldsymbol{\psi}}$, we obtain the *estimated type probabilities* of $Z = j$ for each observation as

$$\hat{\tau}_i^j := \tau^j(\boldsymbol{W}_i;\hat{\boldsymbol{\psi}}) = \frac{\hat{\pi}^j p^j_{\boldsymbol{W}}(\boldsymbol{W}_i;\hat{\boldsymbol{\varphi}}^j)}{\sum_{k=1}^J \hat{\pi}^k p^k_{\boldsymbol{W}}(\boldsymbol{W}_i;\hat{\boldsymbol{\varphi}}^k)}. \tag{22}$$

Note that the posterior type probability $\hat{\tau}_i^j$ conditions on the full observation vector $\boldsymbol{W}_i$, which includes both pre- and post-treatment outcomes. This is valid because $\boldsymbol{W}_i$ is an observed (realized) quantity for each unit: the posterior simply uses all available data to classify units into latent types, and no counterfactual quantities enter the computation.

Then, we consistently estimate the LTATTs (14) by the sample analogue estimator of (18) using $\{\hat{\tau}_i^j\}_{i=1}^n$ as weights: for $t \geq T_0 + 1$,

$$\begin{aligned}\hat{\boldsymbol{\mu}}_t^{ATT,j} = \sum_{\boldsymbol{x}_{T_0}\in\mathcal{X}} & \widehat{\Pr}(\boldsymbol{X}_{T_0} = \boldsymbol{x}_{T_0}|D = 1, Z = j) \\ & \times \left\{\widehat{\mathbb{E}}\left[\boldsymbol{X}_t \mid \boldsymbol{X}_{T_0} = \boldsymbol{x}_{T_0}, D = 1, Z = j\right] - \widehat{\mathbb{E}}\left[\boldsymbol{X}_t \mid \boldsymbol{X}_{T_0} = \boldsymbol{x}_{T_0}, D = 0, Z = j\right]\right\}\end{aligned} \tag{23}$$

where

$$\widehat{\Pr}(\boldsymbol{X}_{T_0} = \boldsymbol{x}_{T_0}|D = 1, Z = j) = \frac{\sum_{i=1}^n \mathbf{1}\{\boldsymbol{X}_{iT_0} = \boldsymbol{x}_{T_0}, D_i = 1\}\hat{\tau}_i^j}{\sum_{i=1}^n \mathbf{1}\{D_i = 1\}\hat{\tau}_i^j} \tag{24}$$

and the estimated conditional expectation is

$$\widehat{\mathbb{E}}\left[\boldsymbol{X}_t \mid \boldsymbol{X}_{T_0} = \boldsymbol{x}_{T_0}, D = d, Z = j\right] = \frac{\sum_{i=1}^n \boldsymbol{X}_{it}\mathbf{1}\{\boldsymbol{X}_{iT_0} = \boldsymbol{x}_{T_0}, D_i = d\}\hat{\tau}_i^j}{\sum_{i=1}^n \mathbf{1}\{\boldsymbol{X}_{iT_0} = \boldsymbol{x}_{T_0}, D_i = d\}\hat{\tau}_i^j}. \tag{25}$$

We also propose the following consistent estimator for the ATT:

$$\hat{\boldsymbol{\mu}}_t^{ATT} = \sum_{j=1}^J \widehat{\Pr}(Z = j \mid D = 1)\, \hat{\boldsymbol{\mu}}_t^{ATT,j}, \quad \widehat{\Pr}(Z = j \mid D = 1) = \frac{\sum_{i=1}^n \mathbf{1}\{D_i = 1\}\hat{\tau}_i^j}{\sum_{i=1}^n \mathbf{1}\{D_i = 1\}}. \tag{26}$$

Let $\boldsymbol{I}(\boldsymbol{\psi}^*)$ be the Fisher information matrix for the MLE in (21). We assume the following regularity conditions for the MLE $\hat{\boldsymbol{\psi}}$.

Assumption MLE. (a) $\Theta_{\boldsymbol{\psi}}$ is compact. All PMF entries in $\boldsymbol{\varphi}^j$ and mixture weights π^j are bounded away from 0 and 1 by a common constant $\epsilon > 0$. (b) $\boldsymbol{\psi}^*$ lies in the interior of $\Theta_{\boldsymbol{\psi}}$. (c) the

Fisher information matrix $\boldsymbol{I}(\boldsymbol{\psi}^*)$ is nonsingular.

Let $\boldsymbol{\theta}_t := \big(\mathrm{vec}(\boldsymbol{\mu}_t^{ATT,1})^\top, \ldots, \mathrm{vec}(\boldsymbol{\mu}_t^{ATT,J})^\top,\ \mathrm{vec}(\boldsymbol{\mu}_t^{ATT})^\top\big)^\top$ and $\boldsymbol{\theta} := \big(\boldsymbol{\theta}_{T_0+1}^\top, \ldots, \boldsymbol{\theta}_T^\top\big)^\top$. Denote the true value of $\boldsymbol{\theta}$ by $\boldsymbol{\theta}^*$, and its two-step estimator, defined above, by $\hat{\boldsymbol{\theta}}$.

Proposition 23 (Consistency and asymptotic normality of LTATT and ATT estimators)**.** *Suppose that Assumptions RSM, M, ID, and MLE hold. Then: (a)* $\hat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}^*$. *(b)* $\sqrt{n}\,(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) \Rightarrow \mathcal{N}(0, \boldsymbol{V})$, *where* $\boldsymbol{V} = \mathbb{E}[\boldsymbol{\Psi}(\boldsymbol{W};\boldsymbol{\psi}^*)\boldsymbol{\Psi}(\boldsymbol{W};\boldsymbol{\psi}^*)^\top]$ *is the variance of the combined influence function* $\boldsymbol{\Psi} := \boldsymbol{\phi} + \boldsymbol{A}^*\boldsymbol{S}$*, accounting for both the second-stage sampling variability* ($\boldsymbol{\phi}$) *and the first-stage estimation error propagated through the score* ($\boldsymbol{A}^*\boldsymbol{S}$)*. Explicit expressions are given in the proof.*

To obtain a consistent estimator $\hat{\boldsymbol{V}}$, we employ a nonparametric weighted bootstrap that yields consistent standard errors for the estimated LTATTs and ATTs while accounting for the dependence structure induced by the two-stage estimation procedure. Further details on the EM algorithm and the weighted bootstrap are provided in Appendix C.

5 Applications

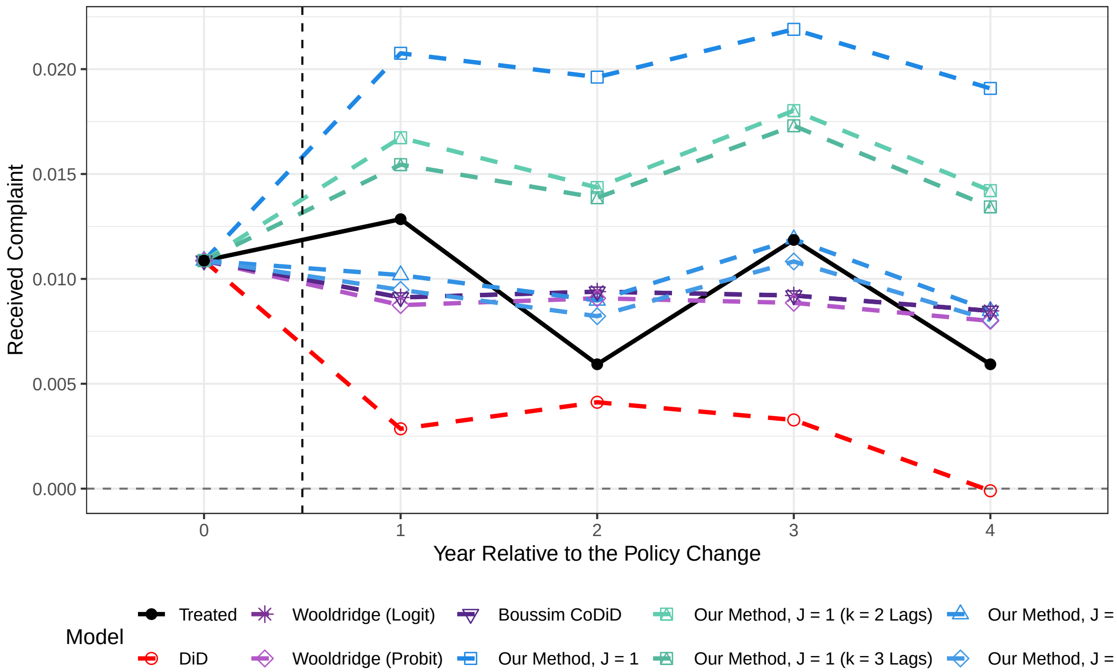
We illustrate the proposed methodology through three empirical applications, each highlighting a distinct limitation of conventional DiD with discrete outcomes. In Charoenwong et al. (2019)'s study of the Dodd-Frank Act, DiD counterfactuals fall below zero, producing out-of-bounds predictions that our transition-based approach avoids by construction. In Hvide and Jones (2018)'s analysis of a Norwegian patent reform, large baseline differences in patenting rates between treated and control groups induce mean-reversion bias, leading DiD to overstate the reform's negative effect. In the ADA application, DiD fails to detect statistically significant employment effects because pre-treatment level differences mask the underlying transition dynamics; our method reveals significant negative effects operating through specific labor-force exit channels.

5.1 Application to Charoenwong et al. (2019)

Charoenwong et al. (2019) investigate how regulatory jurisdiction affects the quality of investment advisor regulation by exploiting a unique policy shift under the Dodd-Frank Act implemented in 2012, which transferred oversight of midsize registered investment advisers (RIAs) from the SEC to state regulators. This 2012 reform left fiduciary standards unchanged but allowed the SEC to focus on other areas. Charoenwong et al. (2019) compare midsize RIAs transitioned to state oversight (treated) with the remaining RIAs under SEC oversight (control) and document an increase in client complaints among midsize advisers under state oversight, indicating a decrease in service quality following the reform.

We revisit Charoenwong et al. (2019) by comparing a canonical difference-in-differences estimator and our proposed estimators. The main outcome of interest is the annual rate of client complaints, i.e., whether a RIA received a complaint from customers, which indicates a decline

Figure 7: Counterfactual Complaint Rates from the Dodd-Frank Act.

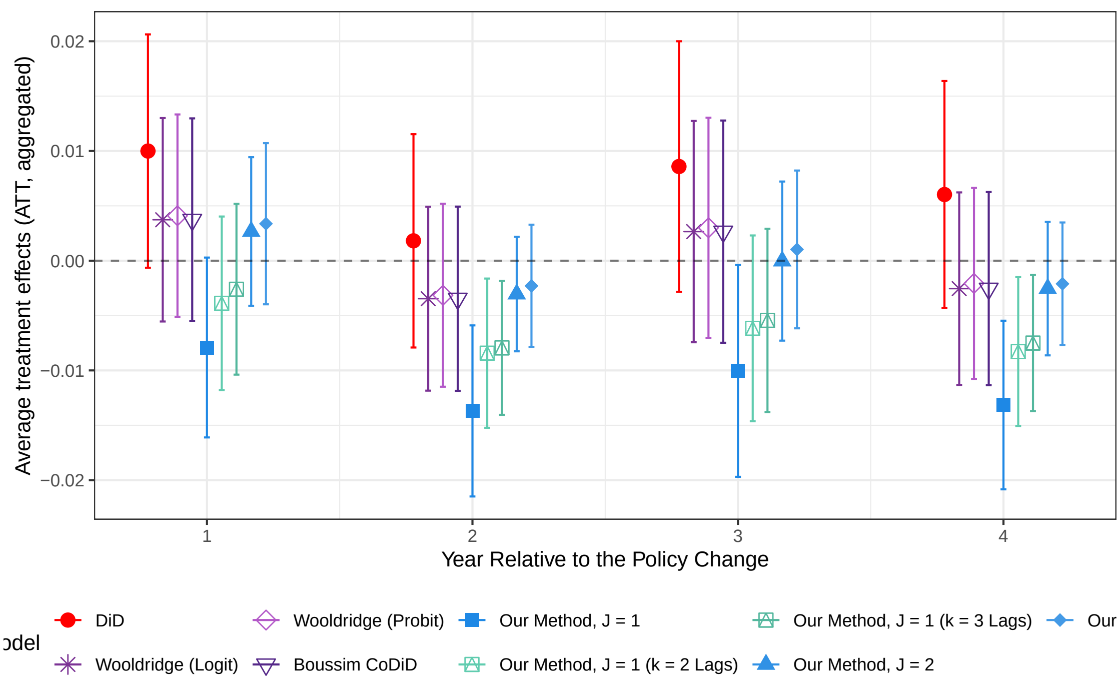


Notes: This figure reports average counterfactual complaint rates for treated units in the absence of treatment, using the data of Charoenwong et al. (2019). The black solid line represents the observed average complaint rates from treated units. The red dashed line represents the counterfactual complaint rates implied by the parallel trends assumption for difference-in-differences. Blue markers denote counterfactual complaint rates from the proposed transition-based method with $J \in \{1, 2, 3\}$ latent types. Green markers correspond to lag-augmented variants for $J = 1$ conditioning on $k = 2$ and $k = 3$ lags of the outcome variable.

in the service quality. We note that the original regression specification in Charoenwong et al. (2019) includes customer fixed effects and state-level time fixed effects as well, rather than having only RIA-level fixed effects and time fixed effects. This feature makes their reported estimates numerically different from difference-in-differences estimates under the parallel trends assumption. For consistency with previous examples, we implement a canonical difference-in-differences estimator with RIA-level fixed effects and time fixed effects only. Although this modification leads to numerically different estimates from Charoenwong et al. (2019), the qualitative conclusions remain unchanged: the difference-in-differences estimates still suggest an increase in complaint rates following the reform.

The counterfactual untreated average outcomes implied by the parallel trends assumption fall below zero as illustrated in Figure 1, indicating an out-of-bounds issue common in linear probability models. In contrast, our transition-based approach naturally restricts counterfactual untreated potential outcomes within the feasible range of probabilities by construction. This difference in the counterfactuals leads to the opposite signs of the ATTs across methods: we find a slight decrease in complaint rates following the reform using our transition-based approach. See Figure 7 for full

Figure 8: ATT Estimates of the Dodd-Frank Act on Complaint Rates.

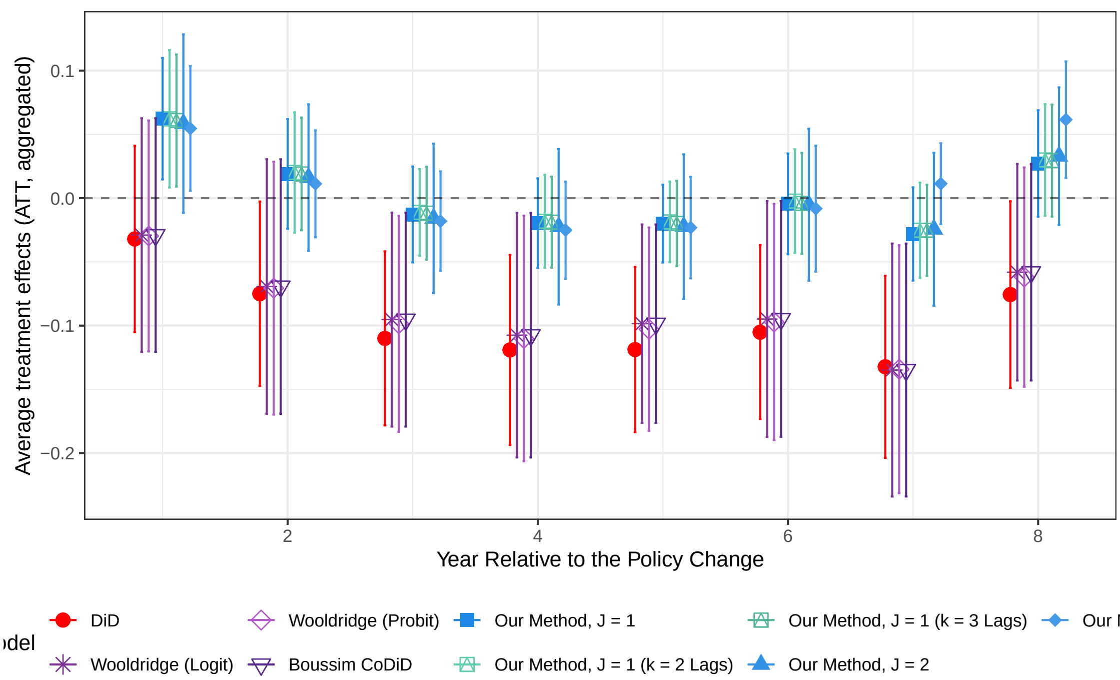


Notes: This figure summarizes the estimated average treatment effects on the treated (ATTs) of the Dodd-Frank Act's 2012 reform on investment adviser complaint rates. The x-axis measures years relative to the policy change, and the y-axis reports aggregated ATT estimates. Red circles represent conventional difference-in-differences (DiD) estimates. Blue markers denote estimates from the proposed transition-based method with varying numbers of latent types ($J \in \{1, 2, 3\}$). Green markers correspond to lag-augmented variants for $J = 1$ conditioning on $k = 2$ and $k = 3$ lags of the outcome variable. Vertical bars indicate 95% bootstrap uniform confidence intervals.

comparison of counterfactuals across different numbers of latent types and lag orders. Figure 8 summarizes the ATT estimates from both approaches across alternative model specifications. The magnitude of the estimated decline varies across specifications, reflecting the sensitivity of the counterfactual construction to the degree of allowed latent heterogeneity in transition dynamics and the number of conditioning lags, although they all indicate either improvement or null impacts on service quality, in contrast to the findings implied by DiD.

To formally assess transition independence in the pre-reform period, we compute differences in transition probabilities between treated and control units across all four possible transition pairs (from no complaint to no complaint, from no complaint to complaint, etc.) and report bootstrap uniform confidence intervals for $J = 1, 2, 3$ latent types (see Figure 15, Figure 16, and Figure 17 in the Appendix). The estimated differences are small and rarely statistically distinguishable from zero, providing support for transition independence prior to the Dodd-Frank Act.

Figure 9: ATT Estimates of the Norwegian Law Reform on Annual Patenting Rates.



Notes: This figure summarizes the estimated average treatment effects on the treated (ATTs) of the Norwegian patent reform on annual patenting rates of university inventors, using the data of Hvide and Jones (2018). The x-axis measures years relative to the reform, and the y-axis reports aggregated ATT estimates. Red circles represent conventional difference-in-differences (DiD) estimates. Blue markers denote estimates from the proposed transition-based method with varying numbers of latent types ($J \in \{1, 2, 3\}$). Green markers correspond to lag-augmented variants for $J = 1$ conditioning on $k = 2$ and $k = 3$ lags of the outcome variable. Vertical bars indicate 95% bootstrap uniform confidence intervals.

5.2 Application to Hvide and Jones (2018)

Hvide and Jones (2018) examine the impact of the 2003 Norwegian reform that transferred one-third of patent rights to universities from researchers, who previously held full ownership. Hvide and Jones (2018) use a difference-in-differences design for the period 1995-2010 comparing inventors who were university researchers employed from 2000-2002 (treated) and non-university inventors (control). Their main TWFE specification (Equation 1 in Hvide and Jones, 2018) uses a dummy variable for annual patent applications as the outcome. They report a statistically significant estimate of -0.045 (Table 9 in Hvide and Jones, 2018), corresponding to a 4.5 percentage point decline in patenting probability for university researchers. We replicate their analysis using the same data to implement both the standard DiD design and our transition-based method.

In contrast to the findings of Hvide and Jones (2018), our proposed method does not find a significant change in patenting rates following the reform. Figure 9 compares the ATT estimates across different methods and post-treatment periods. The results indicate that our transition-based estimates are closer to zero than the negative impacts implied by DiD. For robustness, we condition on additional lagged outcomes ($k = 2$ and $k = 3$) and introduce latent types ($J = 2$ and $J = 3$) to account for potential violations of the transition independence assumption. Even after incorporating additional lag terms and latent transition heterogeneity, the ATT estimates remain negligible across all post-reform periods. The aggregate ATT (the average impact on patenting rates over post-reform periods) shows a statistically insignificant decline of approximately 0.1 percentage points across all specifications ($J = 1, 2$, and 3).

The large discrepancy between the DiD and our transition-based estimates is attributable to substantial pre-treatment differences in patenting rates between university inventors (treated) and non-university inventors (control), as illustrated in Figure 2. In fact, in the year immediately preceding the reform, university inventors were nearly twice as likely to patent as their non-university counterparts. As discussed in subsection 2.2, this baseline disparity most likely biases DiD estimates: due to higher baseline rates, mean reversion would lead treated units to experience a steeper decline than controls in the absence of the reform, thereby violating the parallel trends assumption.

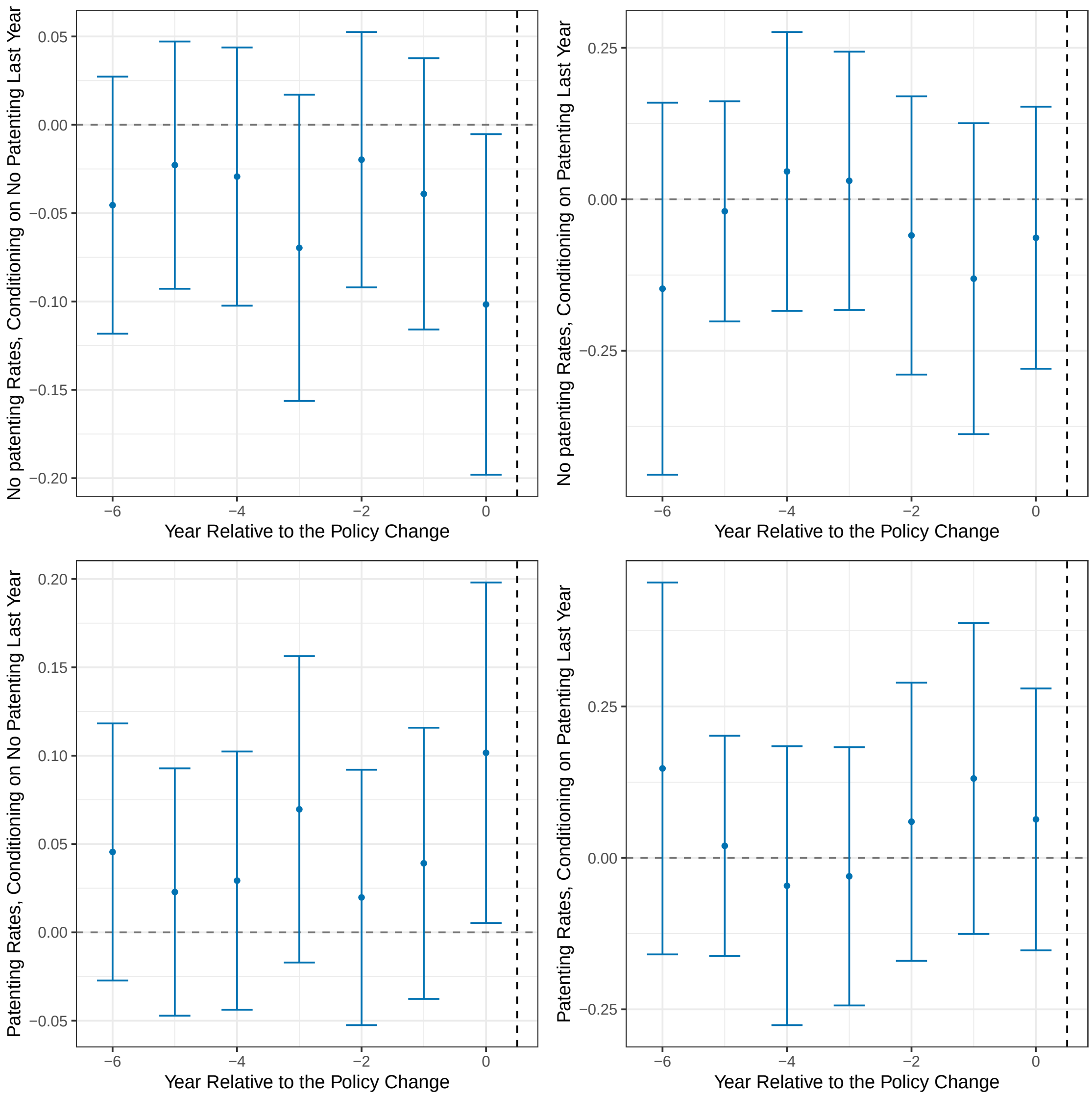
To assess the plausibility of the transition independence assumption, we examine pre-treatment differences in transition probabilities between treated and control units across all four possible transition pairs and report bootstrap uniform confidence intervals in Figure 10 for $J = 1$ (see Figure 18 and Figure 19 in the Appendix for $J = 2, 3$). Across all specifications, the estimated differences are consistently small and statistically indistinguishable from zero, providing empirical support for transition independence in this setting. Transitions conditioning on the more common no-patenting state exhibit tight confidence intervals centered near zero, while transitions conditioning on patenting show wider confidence intervals, reflecting the smaller number of inventors in this conditioning subset. While introducing additional latent types ($J = 2$ and $J = 3$) marginally improves the pre-treatment fit by isolating units with systematically different transition patterns, the estimated treatment effects remain stable across specifications.

5.3 Employment Effects of the Americans with Disabilities Act of 1990

In this example, we examine the effects of the Americans with Disabilities Act (ADA) signed into law in 1990 on employment comparing working-age individuals with work-related disability (treated) and individuals without work-related disability (control). The ADA was enacted to prohibit discrimination against individuals with disabilities in employment, public services, and accommodations. Title I of the ADA specifically mandated that employers with 15 or more employees could not discriminate against qualified individuals with disabilities in hiring, advancement, or discharge decisions, and required that reasonable accommodations be provided in the workplace.

Previous studies on the ADA (Acemoglu and Angrist, 2001; Hotchkiss, 2003; Lise et al., 2023) have largely relied on before-and-after or difference-in-differences designs to evaluate employment impacts. These studies report declining employment rates among disabled individuals, suggesting that compliance costs and fear of litigation may have reduced employers' incentives to hire disabled workers.

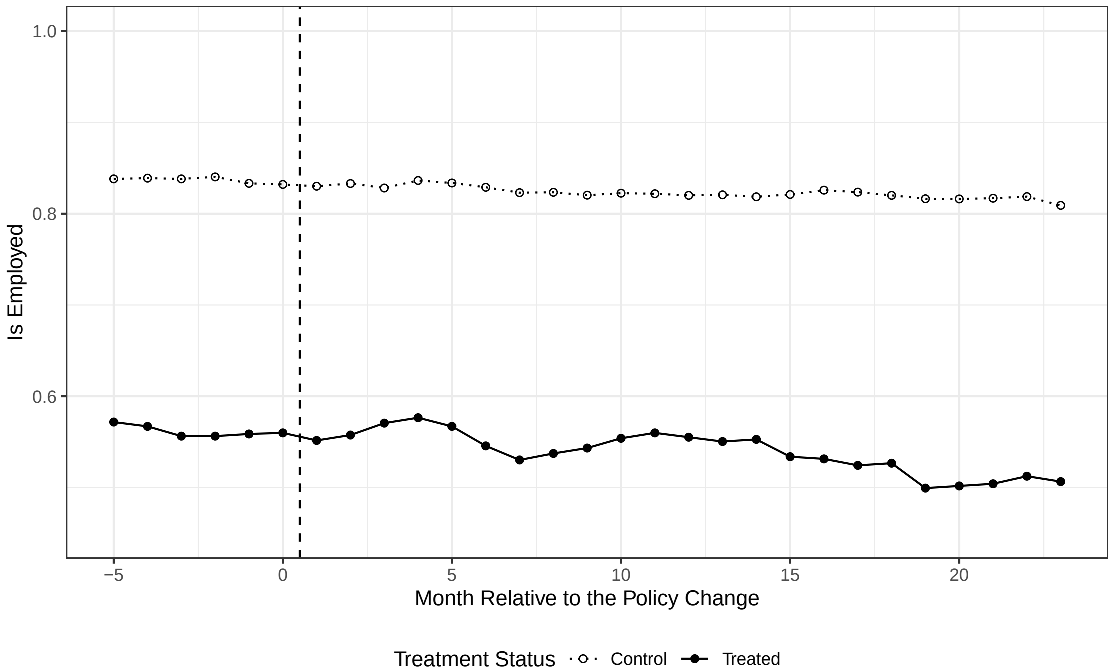
Figure 10: Differences in Patenting Transition Probabilities Before the Norwegian Law Reform.



Notes: This figure reports estimated differences in annual patenting transition probabilities between university researchers (treated) and non-university inventors (control) before the Norwegian patent reform. Each panel corresponds to a distinct patenting transition, where rates represent the annual probability of moving from one patenting status (No Patent or Patent) to another. The x-axis measures years relative to the introduction of the reform (with zero denoting the last pre-treatment period; treatment begins at period 1), and the y-axis shows the estimated difference in transition probabilities between treated and control groups. Vertical bars represent 95% bootstrap uniform confidence intervals across periods within each transition pair. The dashed vertical line marks the timing of the policy introduction.

We study monthly labor force status that takes the value of one of three mutually exclusive states: employed, unemployed, and out-of-labor-force (OLF). We use data from Rotation Group 1 of the 1990 Survey of Income and Program Participation (SIPP) panel (United States Census Bureau, 1990–1993), which provides monthly observations from January 1990 through October

Figure 11: Baseline Employment Rate Differences Before the ADA of 1990.

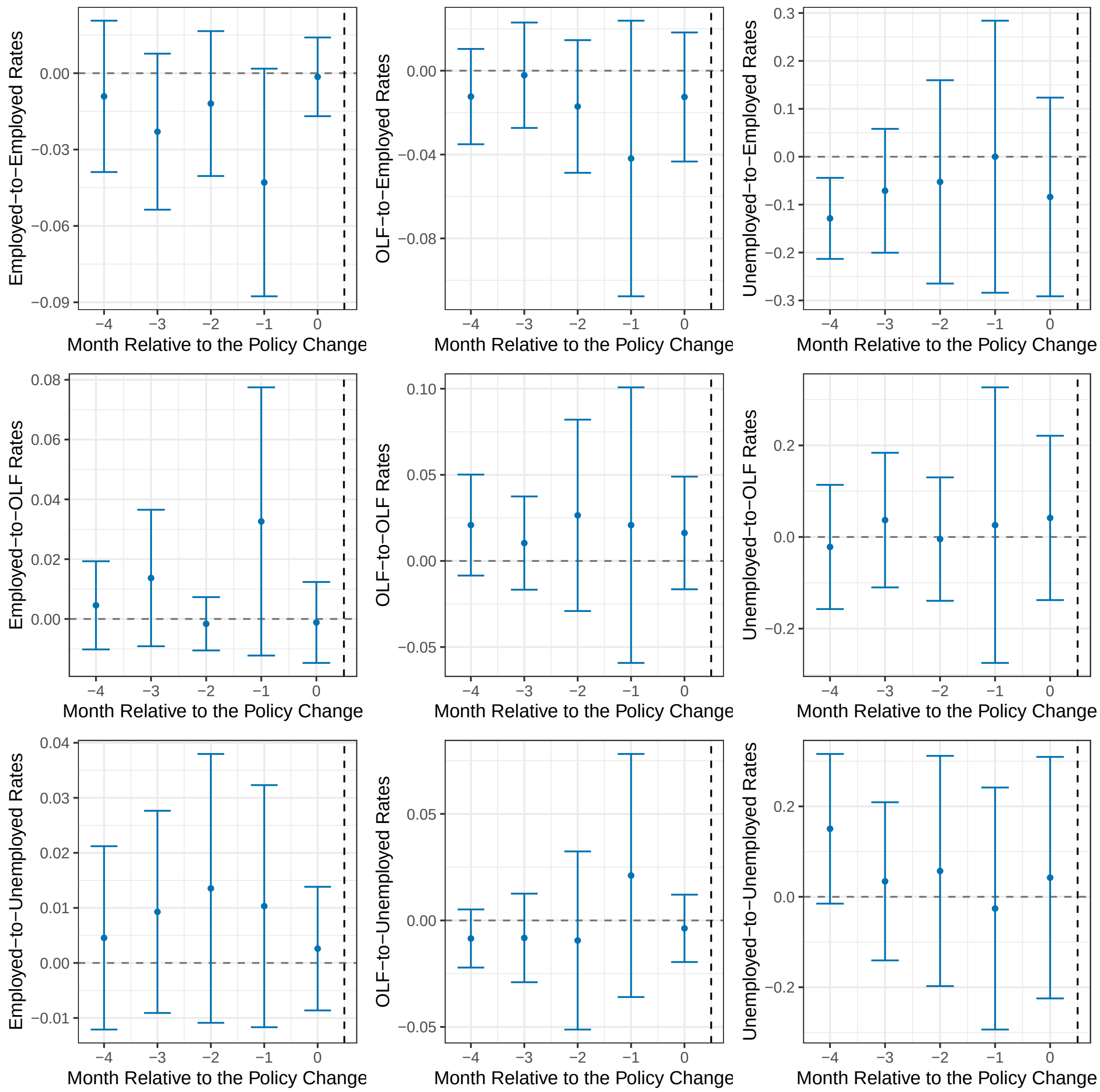


Notes: This figure shows average monthly employment rates for individuals with disabilities (treated group) and those without disabilities (control group) from 1989 to 1991, surrounding the enactment of the ADA. The x-axis reports calendar months, with 1990 marking the implementation year of the ADA. The y-axis shows the fraction of individuals employed. The dashed vertical line indicates the timing of the ADA's introduction.

1991. Taking the ADA signing in July 1990 as the treatment date, this yields 6 pre-treatment and up to 22 post-treatment monthly observations per individual, thereby capturing the full labor force transition dynamics before and after the ADA. Since the ADA was implemented nationwide without explicit state-level variation, we compare individuals with and without disabilities, as in difference-in-differences strategies used in prior research (Acemoglu and Angrist, 2001; Lise et al., 2023). Throughout the remainder of the paper, individuals with work-related disabilities are referred to as the treated group, and the others as the control group. We restrict the sample to adults aged 21-58 and classify individuals reporting a work-limiting disability as disabled, as in Lise et al. (2023).

We first note that the treated and control groups exhibited different baseline employment rates as well as diverging trends even before the introduction of the ADA, as illustrated in Figure 11. For example, in the pre-treatment period, the employment rate among individuals with disabilities was 27.5 percentage points lower than that of individuals without disabilities in our sample. This pre-existing disparity raises concerns about the validity of the parallel trends assumption due to potential mean reversion. Beyond employment status (employed), similar level differences are present across all possible labor force states, as displayed (see Figure 20 in the Appendix).

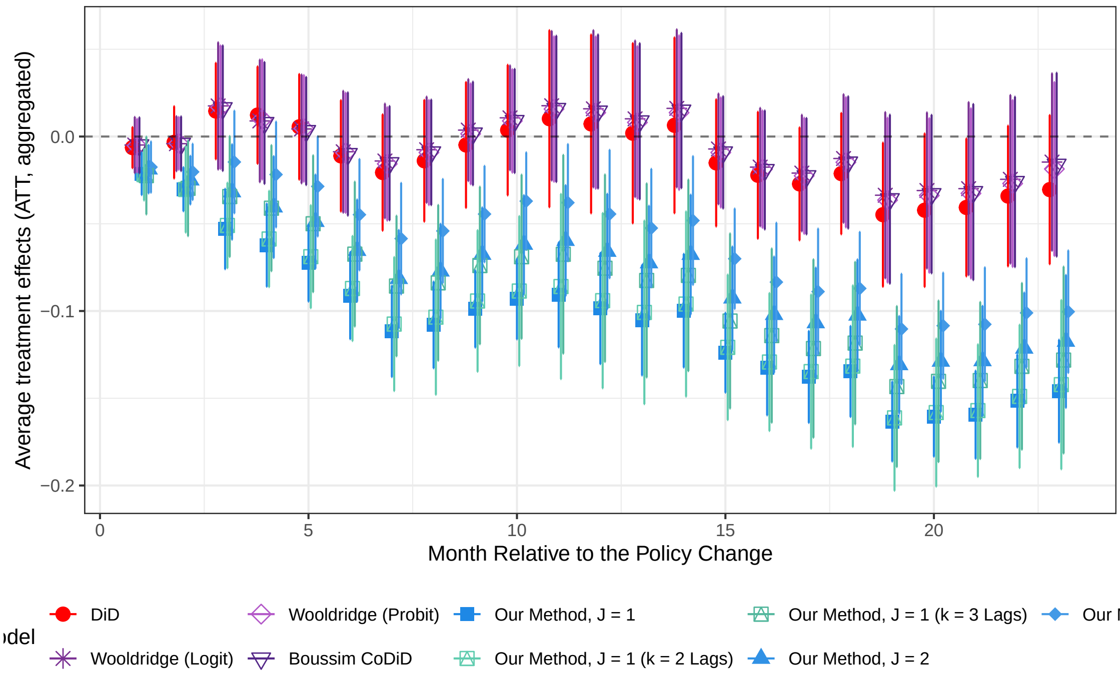
Figure 12: Difference in Labor Force Transition Probabilities Before the ADA.



Notes: This figure reports estimated differences in monthly labor force transition probabilities between individuals with and without disabilities before the enactment of the ADA. Each panel corresponds to a distinct labor force transition, where rates represent the monthly probability of moving from one labor force status (e.g., employed, unemployed, out of the labor force (OLF)) to another. The x-axis measures calendar months relative to the introduction of the ADA (with zero denoting the last pre-treatment period; treatment begins at period 1), and the y-axis shows the estimated difference in transition probabilities between treated (individuals with disabilities) and control (individuals without disabilities) groups. Vertical bars represent 95% bootstrap uniform confidence intervals across periods within each transition pair. The dashed vertical line marks the timing of the ADA introduction.

In contrast, the data suggest that treated and control units exhibit broadly similar transition dynamics across employment statuses, indicating transition independence. To formally assess transition independence in the pre-legislation period, we compute differences in transition probabilities between treated and control units across all nine possible transition pairs and report bootstrap

Figure 13: ATT Estimates of the ADA on Employment.

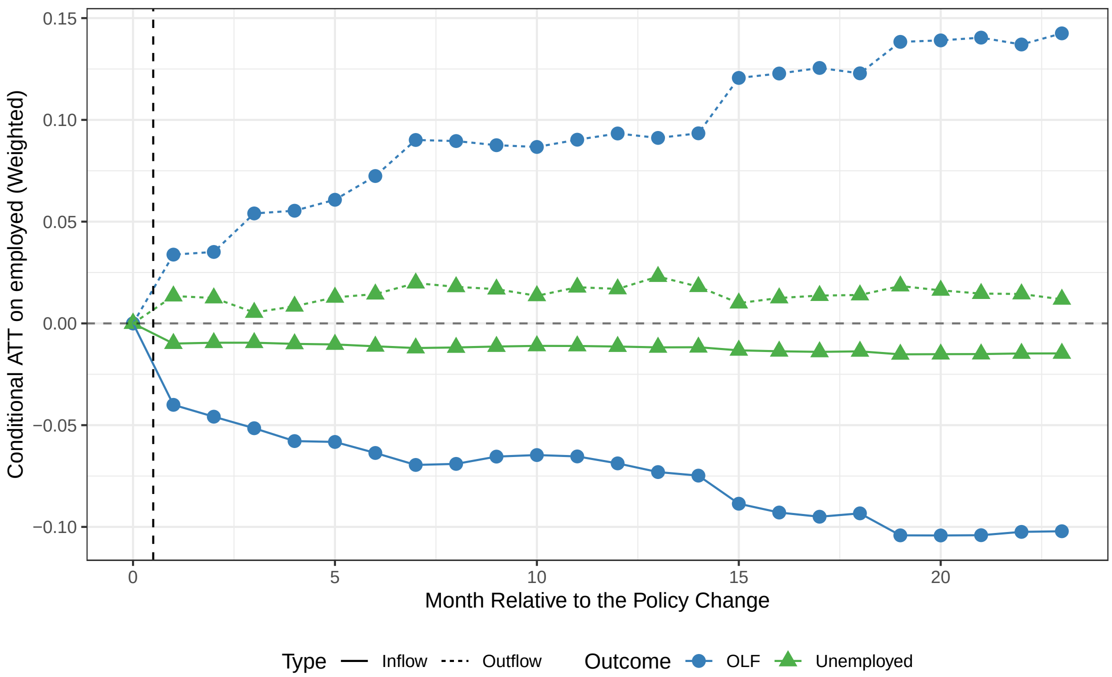


Notes: This figure summarizes the estimated average treatment effects on the treated (ATTs) of the Americans with Disabilities Act (ADA) on employment outcomes using several estimation strategies. The x-axis measures calendar months relative to the implementation of the ADA, and the y-axis reports aggregated ATT estimates across post-treatment periods. Red circles represent conventional difference-in-differences (DiD) estimates, while green markers correspond to lag-augmented DiD models conditioning on one to three lags of the outcome variable. Blue markers denote estimates from the proposed transition-based method with varying numbers of latent types ($J \in \{1, 2, 3\}$). Vertical bars indicate 95% bootstrap uniform confidence intervals.

uniform confidence intervals. Across all specifications ($J = 1, 2, 3$), the estimated pre-treatment differences are small and statistically insignificant, supporting the plausibility of transition independence (Figure 12 for $J = 1$; see Figure 21 and Figure 22 in the Appendix for $J = 2, 3$). The pattern remains qualitatively similar across latent types when allowing for heterogeneity in transition dynamics.

Our transition-based estimates indicate a statistically significant negative impact on employment across all specifications, whereas the standard DiD model, which is likely biased due to differences in pre-treatment employment levels, yields estimates that fail to reject the null hypothesis of no negative employment effects. The ATT estimates across different methods and specifications are provided in Figure 13. The aggregate ATT estimates from the transition-based estimators are robust across specifications: for $J = 1$, 2, and 3 as well as lag-augmented variants ($k = 2$ and $k = 3$ lags), all consistently indicate negative employment effects, suggesting that the conclusions are not driven by specific modeling choices. Overall, the results indicate that the ADA may have had an unintended short-run negative impact on employment for individuals with disabilities.

Figure 14: Decomposing ATTs of the ADA on Employment by Employment Channels.



Notes: This figure applies the flow decomposition in (8) to the ADA's effect on employment, with the focal state $\bar{y}^{(k)}$ = Employed. Solid lines plot inflow effects: the difference in transition probabilities into employment between treated and control groups, weighted by the treated group's pre-treatment distribution—corresponding to the first sum in (8). Specifically, the "Unemployed → Employed" series measures the inflow effect from y = Unemployed, and the "OLF → Employed" series measures the inflow effect from y = OLF. Dashed lines plot outflow effects: the difference in transition probabilities out of employment, weighted by the treated group's probability of being employed at T_0—corresponding to the subtracted sum in (8). All series are plotted as signed contributions to the employment ATT (i.e., outflow effects include the minus sign from (8)), so that negative values indicate channels that reduce employment. The "Employed → Unemployed" series measures the outflow effect to y = Unemployed, and the "Employed → OLF" series measures the outflow effect to y = OLF. The dominant channel is the outflow from employment to OLF. The x-axis measures calendar months relative to the ADA enactment, and the y-axis displays weighted conditional ATTs on employment.

A distinctive advantage of the transition-based framework is its ability to decompose treatment effects into specific transition channels via the flow decomposition in (8). Figure 14 applies this decomposition to identify how the employment declines arise, namely, which specific inflow and outflow transition channels account for the reduced employment. Figure 14 shows that the dominant channel in the decline in employment is an increase in transitions from employment directly into OLF among individuals with disabilities after the ADA, with a secondary contribution from reduced inflows from OLF into employment. In contrast, transitions involving unemployment (both Employment→Unemployment outflows and Unemployment→Employment inflows) play a limited role. This decomposition reveals a mechanism not apparent from the initial-status analysis: the ADA's adverse employment effect operates primarily through labor-force exits and, to a lesser extent, through reduced labor-force entry from OLF, rather than through changes in job-search outcomes. Such channel-specific insights are unavailable from standard DiD, which estimates only the net change in outcome levels.

A Proofs

A.1 Proof of Proposition 1

The result follows from the identity $\mathbb{E}\left[\mathbf{X}_t(1) \mid D=1\right] = \mathbb{E}\left[\mathbf{X}_t \mid D=1\right]$ and

$$\begin{aligned}
\mathbb{E}\left[\mathbf{X}_t(0) \mid D=1\right] &= \mathbb{E}\left[\mathbb{E}\left[\mathbf{X}_t(0) \mid \boldsymbol{X}_1^{T_0}(0), D=1\right] \mid D=1\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\mathbf{X}_t(0) \mid \boldsymbol{X}_1^{T_0}(0), D=0\right] \mid D=1\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\mathbf{X}_t \mid \boldsymbol{X}_1^{T_0}, D=0\right] \mid D=1\right],
\end{aligned}$$

where the first equality follows from the law of iterated expectations, where the outer expectation over $\boldsymbol{X}_1^{T_0}(0)$ is taken with respect to $\Pr(\boldsymbol{X}_1^{T_0}(0) \mid D=1)$; the second from Assumption TI together with the no-anticipation assumption (Assumption NA) and the common support assumption (Assumption CS); and the third from the fact that $\boldsymbol{X}_t(0) = \boldsymbol{X}_t$ for all t when $D=0$, which also uses Assumption NA to replace $\boldsymbol{X}_1^{T_0}(0)$ with the observed $\boldsymbol{X}_1^{T_0}$ in the conditioning set. Substituting these expressions into the definition of $\boldsymbol{\mu}_t^{\text{ATT}}$ in (3) yields equation (4). $\square$

A.2 Proof of Proposition 10

By definition,

$$\boldsymbol{\mu}_t^{\text{DiD}} = \mathbb{E}[\boldsymbol{X}_t - \boldsymbol{X}_{T_0} \mid D=1] - \mathbb{E}[\boldsymbol{X}_t - \boldsymbol{X}_{T_0} \mid D=0], \qquad \boldsymbol{\mu}_t^{\text{ATT}} = \mathbb{E}[\boldsymbol{X}_t(1) - \boldsymbol{X}_t(0) \mid D=1].$$

By Assumption NA, $\boldsymbol{X}_t = \boldsymbol{X}_t(0)$, $t \geq T_0+1$, for $D=0$ and $\boldsymbol{X}_{T_0} = \boldsymbol{X}_{T_0}(0)$ for all units, so

$$\boldsymbol{\mu}_t^{\text{DiD}} - \boldsymbol{\mu}_t^{\text{ATT}} = \mathbb{E}[\boldsymbol{X}_t(0) - \boldsymbol{X}_{T_0}(0) \mid D=1] - \mathbb{E}[\boldsymbol{X}_t(0) - \boldsymbol{X}_{T_0}(0) \mid D=0]. \tag{27}$$

Applying the Law of Iterated Expectations to the first term,

$$\mathbb{E}[\boldsymbol{X}_t(0) \mid D=1] = \mathbb{E}\Big[\mathbb{E}[\boldsymbol{X}_t(0) \mid \boldsymbol{X}_1^{T_0}, D=1] \mid D=1\Big].$$

By Assumption TI,

$$\mathbb{E}[\boldsymbol{X}_t(0) \mid \boldsymbol{X}_1^{T_0}, D=1] = \mathbb{E}[\boldsymbol{X}_t(0) \mid \boldsymbol{X}_1^{T_0}, D=0] = \mathbb{E}[\boldsymbol{X}_t \mid \boldsymbol{X}_1^{T_0}, D=0]. \tag{28}$$

Using (28) and the same argument for $D=0$, equation (27) becomes

$$\boldsymbol{\mu}_t^{\text{DiD}} - \boldsymbol{\mu}_t^{\text{ATT}} = \mathbb{E}[\mathbb{E}[\boldsymbol{X}_t \mid \boldsymbol{X}_1^{T_0}, D=0] - \boldsymbol{X}_{T_0} \mid D=1] - \mathbb{E}[\mathbb{E}[\boldsymbol{X}_t \mid \boldsymbol{X}_1^{T_0}, D=0] - \boldsymbol{X}_{T_0} \mid D=0].$$

Under Assumption CS, expand the above equation over the support $\mathcal{X}^{T_0}$:

$$\boldsymbol{\mu}_t^{\text{DiD}} - \boldsymbol{\mu}_t^{\text{ATT}} = \sum_{\boldsymbol{x}_1^{T_0} \in \mathcal{X}^{T_0}} \underbrace{\mathbb{E}[\boldsymbol{X}_t - \boldsymbol{x}_{T_0} \mid \boldsymbol{X}_1^{T_0} = \boldsymbol{x}_1^{T_0}, D = 0]}_{\text{control mean trend}} \times \left\{ \Pr\left(\boldsymbol{X}_1^{T_0} = \boldsymbol{x}_1^{T_0} \mid D = 1\right) - \Pr\left(\boldsymbol{X}_1^{T_0} = \boldsymbol{x}_1^{T_0} \mid D = 0\right)\right\}.$$

This proves the stated result. $\square$.

A.3 Proof of Proposition 11

Proof of Proposition 11. By construction, each estimand is the observed treated post-treatment mean minus an imputed untreated counterfactual, $\boldsymbol{\mu}_t^{\bullet} = \mathbb{E}[\boldsymbol{X}_t \mid D = 1] - m_t^{\bullet}$. Hence

$$\boldsymbol{\mu}_t^{W} - \boldsymbol{\mu}_t^{\text{ATT}} = \left(\mathbb{E}[\boldsymbol{X}_t \mid D = 1] - m_t^{W}\right) - \left(\mathbb{E}[\boldsymbol{X}_t \mid D = 1] - m_t^{\text{TI}}\right) = m_t^{\text{TI}} - m_t^{W},$$

where $m_t^{\text{TI}} = \mathbb{E}\left[\mathbb{E}[\boldsymbol{X}_t \mid \boldsymbol{X}_1^{T_0}, D = 0] \mid D = 1\right]$ is the transition-independence counterfactual that identifies $\boldsymbol{\mu}_t^{\text{ATT}}$ under Assumptions TI, NA, and CS (Proposition 1). Adding and subtracting the linear DiD counterfactual $m_t^{\text{DiD}} = \mathbb{E}[\boldsymbol{X}_{T_0} \mid D = 1] + \mathbb{E}[\boldsymbol{X}_t - \boldsymbol{X}_{T_0} \mid D = 0]$ yields the two terms of (10),

$$\boldsymbol{\mu}_t^{W} - \boldsymbol{\mu}_t^{\text{ATT}} = \left(m_t^{\text{TI}} - m_t^{\text{DiD}}\right) + \left(m_t^{\text{DiD}} - m_t^{W}\right).$$

For the first term, substitute $\mathbb{E}[\boldsymbol{X}_{T_0} \mid D = d] = \sum_{\boldsymbol{x}_1^{T_0}} \boldsymbol{x}_{T_0} \Pr(\boldsymbol{X}_1^{T_0} = \boldsymbol{x}_1^{T_0} \mid D = d)$ and $\mathbb{E}[\boldsymbol{X}_t \mid D = 0] = \sum_{\boldsymbol{x}_1^{T_0}} \mathbb{E}[\boldsymbol{X}_t \mid \boldsymbol{X}_1^{T_0} = \boldsymbol{x}_1^{T_0}, D = 0] \Pr(\boldsymbol{X}_1^{T_0} = \boldsymbol{x}_1^{T_0} \mid D = 0)$:

$$\begin{aligned}
m_t^{\text{TI}} - m_t^{\text{DiD}} &= \sum_{\boldsymbol{x}_1^{T_0}} \mathbb{E}[\boldsymbol{X}_t \mid \boldsymbol{X}_1^{T_0} = \boldsymbol{x}_1^{T_0}, D = 0] \, \Pr(\boldsymbol{X}_1^{T_0} = \boldsymbol{x}_1^{T_0} \mid D = 1) \\
&\quad - \sum_{\boldsymbol{x}_1^{T_0}} \boldsymbol{x}_{T_0} \, \Pr(\boldsymbol{X}_1^{T_0} = \boldsymbol{x}_1^{T_0} \mid D = 1) \\
&\quad - \sum_{\boldsymbol{x}_1^{T_0}} \mathbb{E}[\boldsymbol{X}_t \mid \boldsymbol{X}_1^{T_0} = \boldsymbol{x}_1^{T_0}, D = 0] \, \Pr(\boldsymbol{X}_1^{T_0} = \boldsymbol{x}_1^{T_0} \mid D = 0) \\
&\quad + \sum_{\boldsymbol{x}_1^{T_0}} \boldsymbol{x}_{T_0} \, \Pr(\boldsymbol{X}_1^{T_0} = \boldsymbol{x}_1^{T_0} \mid D = 0) \\
&= \sum_{\boldsymbol{x}_1^{T_0} \in \mathcal{X}^{T_0}} \mathbb{E}\left[\boldsymbol{X}_t - \boldsymbol{x}_{T_0} \mid \boldsymbol{X}_1^{T_0} = \boldsymbol{x}_1^{T_0}, D = 0\right] \left\{ \Pr(\boldsymbol{X}_1^{T_0} = \boldsymbol{x}_1^{T_0} \mid D = 1) - \Pr(\boldsymbol{X}_1^{T_0} = \boldsymbol{x}_1^{T_0} \mid D = 0)\right\},
\end{aligned}$$

which is the right-hand side of (9), equal to $\boldsymbol{\mu}_t^{\text{DiD}} - \boldsymbol{\mu}_t^{\text{ATT}}$. Each term is identified because the control conditional means, the pre-treatment history distributions, and the marginal means are all functions of the observed data distribution. $\square$

A.4 Proof of Proposition 14

Suppose that Assumption TI holds. Let $\bar{\boldsymbol{x}}^{(k)} \in \mathcal{X}$ denote the value of $\boldsymbol{X}_t(0)$ corresponding to $Y_t(0) = \bar{y}^{(k)}$, where the k-th element equals 1 and all other elements equal 0. Then, $\Pr(\boldsymbol{X}_t(0) = \bar{\boldsymbol{x}}^{(k)} \mid \boldsymbol{X}_1^{T_0}(0) = \boldsymbol{x}_1^{T_0}, D = d] = \Pr(X_t^{(k)}(0) = 1 \mid \boldsymbol{X}_1^{T_0}(0) = \boldsymbol{x}_1^{T_0}, D = d] = \mathbb{E}[X_t^{(k)}(0) \mid \boldsymbol{X}_1^{T_0}(0) = \boldsymbol{x}_1^{T_0}, D = d]$. Therefore, Assumption TI implies that $\mathbb{E}[\boldsymbol{X}_t(0) \mid \boldsymbol{X}_1^{T_0}(0) = \boldsymbol{x}_1^{T_0}, D = 0] = \mathbb{E}[\boldsymbol{X}_t(0) \mid \boldsymbol{X}_1^{T_0}(0) = \boldsymbol{x}_1^{T_0}, D = 1]$ for all $t \geq T_0 + 1$, and Assumption CPT is satisfied.

Conversely, suppose that Assumption CPT holds. Then, for each $k = 1, \ldots, K$, $\mathbb{E}[X_t^{(k)}(0) \mid \boldsymbol{X}_1^{T_0}(0) = \boldsymbol{x}_1^{T_0}, D = 0] = \mathbb{E}[X_t^{(k)}(0) \mid \boldsymbol{X}_1^{T_0}(0) = \boldsymbol{x}_1^{T_0}, D = 1]$. Substituting $\mathbb{E}[X_t^{(k)}(0) \mid \boldsymbol{X}_1^{T_0}(0) = \boldsymbol{x}_1^{T_0}, D = d] = \Pr(\boldsymbol{X}_t(0) = \bar{\boldsymbol{x}}^{(k)} \mid \boldsymbol{X}_1^{T_0}(0) = \boldsymbol{x}_1^{T_0}, D = d)$ into both sides yields

$$\Pr\left(\boldsymbol{X}_t(0) = \bar{\boldsymbol{x}}^{(k)} \mid \boldsymbol{X}_1^{T_0}(0) = \boldsymbol{x}_1^{T_0}, D = 0\right) = \Pr\left(\boldsymbol{X}_t(0) = \bar{\boldsymbol{x}}^{(k)} \mid \boldsymbol{X}_1^{T_0}(0) = \boldsymbol{x}_1^{T_0}, D = 1\right).$$

for all $k = 1, \ldots, K$. Since $\mathcal{X} = \{\bar{\boldsymbol{x}}^{(1)}, \ldots, \bar{\boldsymbol{x}}^{(K)}\}$, it follows that Assumption TI holds. $\square$

A.5 Proof of Proposition 16

Suppose that $T_0 = k + 1$, $T = 2(k + 1)$, and $J = |\mathcal{X}|^k$. Partition $(\boldsymbol{X}_1, ..., \boldsymbol{X}_T)$ as $(\boldsymbol{X}_1, ..., \boldsymbol{X}_T) = (\xi_1, \xi_2, \xi_3, \xi_4)$, where

$$\xi_1 := \boldsymbol{X}_1^k \in \mathcal{X}^k, \quad \xi_2 := \boldsymbol{X}_{k+1} \in \mathcal{X}, \quad \xi_3 = \boldsymbol{X}_{k+2} \in \mathcal{X}, \text{ and } \xi_4 := \boldsymbol{X}_{k+3}^T \in \mathcal{X}^{T-(k+2)}. \tag{29}$$

Note that, when $k = 1$, we have $\xi_t = \boldsymbol{X}_t$ for $t = 1, ..., 4$. For $d \in \{0, 1\}$, define the corresponding potential outcomes $\xi_1(d) := \boldsymbol{X}_1^k(d)$, $\xi_2(d) := \boldsymbol{X}_{k+1}(d)$, $\xi_3(d) = \boldsymbol{X}_{k+2}(d)$, and $\xi_4(d) := \boldsymbol{X}_{k+3}^T(d)$.

Because $\boldsymbol{X}_t = \boldsymbol{X}_t(0)$ for $t = 1, 2, ..., T$ for all control units with $D = 0$, we have $\xi_j = \xi_j(0)$ for $j = 1, 2, 3, 4$ for all units with $D = 0$. Then, the probability mass function of (ξ_2, ξ_3, ξ_4) given $(\xi_1, D) = (\xi_1, 0)$ is written as

$$\begin{aligned} p_{\xi_2,\xi_3,\xi_4|\xi_1,D}(\xi_2, \xi_3, \xi_4|\xi_1, 0) &= \sum_{j=1}^{J} \tau^j_{\xi_1(0),D}(\xi_1, 0) p^j_{\xi_2(0)|\xi_1(0)}(\xi_2|\xi_1) p^j_{\xi_3(0)|\xi_2(0)}(\xi_3|\xi_2) p^j_{\xi_4(0)|\xi_3(0)}(\xi_4|\xi_3) \\ &= \sum_{j=1}^{J} \lambda_1^j(\xi_2|\xi_1) \lambda_2^j(\xi_3|\xi_2) \lambda_3^j(\xi_4|\xi_3), \end{aligned} \tag{30}$$

where

$$\tau^j_{\xi_1(0),D}(\xi_1, d) := \frac{\pi^j p^j_{\xi_1,D}(\xi_1, d)}{\sum_{k=1}^{J} \pi^k p^k_{\xi_1,D}(\xi_1, d)}, \quad \lambda_1^j(\xi_2|\xi_1) := \tau^j_{\xi_1(0),D}(\xi_1, 0) p^j_{\xi_2(0)|\xi_1(0)}(\xi_2|\xi_1),$$
$$\lambda_2^j(\xi_3|\xi_2) := p^j_{\xi_3(0)|\xi_2(0)}(\xi_3|\xi_2), \text{ and } \lambda_3^j(\xi_4|\xi_3) := p^j_{\xi_4(0)|\xi_3(0)}(\xi_4|\xi_3).$$

Here, because ξ_4 is a block of periods, $p^j_{\xi_4(0)|\xi_3(0)}(\xi_4 \mid \xi_3)$ represents the product of first-order transition probabilities within that block.

Evaluating (30) at $(\xi_1, \xi_2, \xi_3, \xi_4) = (a, \xi_2, \xi_3, b)$ gives

$$p_{\xi_2,\xi_3,\xi_4|\xi_1,D}(\xi_2, \xi_3, b|a, 0) = \sum_{j=1}^{J} \lambda_1^j(\xi_2|a)\lambda_2^j(\xi_3|\xi_2)\lambda_3^j(b|\xi_3). \tag{31}$$

Similarly, evaluating $p_{\xi_2,\xi_3|\xi_1,D}(\xi_2, \xi_3|\xi_1, 0) = \sum_{j=1}^{J} \tau^j_{\xi_1(0),D}(\xi_1, 0)p^j_{\xi_2|\xi_1}(\xi_2|\xi_1)p^j_{\xi_3(0)|\xi_2(0)}(\xi_3|\xi_2)$ at $(\xi_1, \xi_2, \xi_3) = (a, \xi_2, \xi_3)$ gives

$$p_{\xi_2,\xi_3|\xi_1,D}(\xi_2, \xi_3|a, 0) = \sum_{j=1}^{J} \lambda_1^j(\xi_2|a)\lambda_2^j(\xi_3|\xi_2). \tag{32}$$

Denote

$$q_{\xi_2,\xi_3}(a, b) := p_{\xi_2,\xi_3,\xi_4|\xi_1,D}(\xi_2, \xi_3, b|a, 0) \quad \text{and} \quad \bar{q}_{\xi_2,\xi_3}(a) := p_{\xi_2,\xi_3|\xi_1,D}(\xi_2, \xi_3|a, 0). \tag{33}$$

Evaluating (31) at $a = a_1, ..., a_J$ and $b = b_1, ..., b_{J-1}$ gives $J(J-1)$ equations while evaluating (32) at $a = a_1, ..., a_J$ gives J equations.

Using matrix notation, we collect these $J(J-1) + J = J^2$ equations as

$$\boldsymbol{Q}_{\xi_2,\xi_3} = \boldsymbol{L}_{\xi_3}\boldsymbol{D}_{\xi_3|\xi_2}\bar{\boldsymbol{L}}_{\xi_2}^{\top}, \tag{34}$$

where

$$\boldsymbol{Q}_{\xi_2,\xi_3} := \begin{bmatrix} \bar{q}_{\xi_2,\xi_3}(a_1) & \bar{q}_{\xi_2,\xi_3}(a_2) & \cdots & \bar{q}_{\xi_2,\xi_3}(a_J) \\ q_{\xi_2,\xi_3}(a_1, b_1) & q_{\xi_2,\xi_3}(a_2, b_1) & \cdots & q_{\xi_2,\xi_3}(a_J, b_1) \\ \vdots & \vdots & \ddots & \vdots \\ q_{\xi_2,\xi_3}(a_1, b_{J-1}) & q_{\xi_2,\xi_3}(a_2, b_{J-1}) & \cdots & q_{\xi_2,\xi_3}(a_J, b_{J-1}) \end{bmatrix}, \tag{35}$$

$$\bar{\boldsymbol{L}}_{\xi_2} := \begin{bmatrix} \lambda_1^1(\xi_2|a_1) & \cdots & \lambda_1^J(\xi_2|a_1) \\ \vdots & \ddots & \cdots \\ \lambda_1^1(\xi_2|a_J) & \cdots & \lambda_1^J(\xi_2|a_J) \end{bmatrix}, \quad \boldsymbol{L}_{\xi_3} := \begin{bmatrix} 1 & \cdots & 1 \\ \lambda_3^1(b_1|\xi_3) & \cdots & \lambda_3^J(b_1|\xi_3) \\ \vdots & \ddots & \vdots \\ \lambda_3^1(b_{J-1}|\xi_3) & \cdots & \lambda_3^J(b_{J-1}|\xi_3) \end{bmatrix}, \tag{36}$$

and $\boldsymbol{D}_{\xi_3|\xi_2} := \text{diag}\left(\lambda_2^1(\xi_3|\xi_2), ..., \lambda_2^J(\xi_3|\xi_2)\right)$.

Define

$$\tau^j_{\xi_1(0),\xi_2(0),D}(\xi_1, \xi_2, 1) := \tau^j_{\xi_1(0),D}(\xi_1, 1)p^j_{\xi_2(0)|\xi_1(0)}(\xi_2|\xi_1).$$

We make the following assumption.

Assumption ID. (a) There exists a value ξ_3^* that satisfies the following condition: For every $\xi_3 \in \mathcal{X}$, we can find $(\check{\xi}_2, \bar{\xi}_2, \bar{\xi}_3) \in \mathcal{X}^3$, $(a_1, ..., a_J) \in \mathcal{X}^{kJ}$ and $(b_1, ..., b_{J-1}) \in \mathcal{X}^{(T-(k+2))(J-1)}$ such that (i) $\bar{\boldsymbol{L}}_{\check{\xi}_2}$, $\bar{\boldsymbol{L}}_{\bar{\xi}_2}$, $\boldsymbol{L}_{\xi_3}$, $\boldsymbol{L}_{\xi_3^*}$, and $\boldsymbol{L}_{\bar{\xi}_3}$ are non-singular; (ii) all the diagonal elements of $\boldsymbol{D}_{\xi_3,\xi_3}$ defined in (38) with $\xi_3^* = \xi_3$ take distinct values. (b) For every $(\xi_2, \xi_3, \xi_4) \in \mathcal{X}^{T-k}$, $p^j_{\xi_3(0)|\xi_2(0)}(\xi_3|\xi_2) > 0$

and $p^j_{\xi_4(0)|\xi_3(0)}(\xi_4|\xi_3) > 0$ for $j = 1, ..., J$. (c) For every value of $\xi_1 \in \mathcal{X}^k$, there exists $\{c_j\}_{j=1}^J \in \mathcal{X}^J$ such that

$$\begin{pmatrix} p^1_{\xi_2(0)|\xi_1(0)}(c_1|\xi_1) & p^2_{\xi_2(0)|\xi_1(0)}(c_1|\xi_1) & \cdots & p^J_{\xi_2(0)|\xi_1(0)}(c_1|\xi_1) \\ p^1_{\xi_2(0)|\xi_1(0)}(c_2|\xi_1) & p^2_{\xi_2(0)|\xi_1(0)}(c_2|\xi_1) & \cdots & p^J_{\xi_2(0)|\xi_1(0)}(c_2|\xi_1) \\ \vdots & \vdots & \ddots & \vdots \\ p^1_{\xi_2(0)|\xi_1(0)}(c_J|\xi_1) & p^2_{\xi_2(0)|\xi_1(0)}(c_J|\xi_1) & \cdots & p^J_{\xi_2(0)|\xi_1(0)}(c_J|\xi_1) \end{pmatrix}$$

is non-singular. (d) For each value of $\xi_2 \in \mathcal{X}$, there exists $\{e_j\}_{j=1}^J \in \mathcal{X}^{kJ}$ such that

$$\begin{pmatrix} \tau^1_{\xi_1(0),\xi_2(0),D}(e_1, \xi_2, 1) & \tau^2_{\xi_1(0),\xi_2(0),D}(e_1, \xi_2, 1) & \cdots & \tau^J_{\xi_1(0),\xi_2(0),D}(e_1, \xi_2, 1) \\ \tau^1_{\xi_1(0),\xi_2(0),D}(e_2, \xi_2, 1) & \tau^2_{\xi_1(0),\xi_2(0),D}(e_2, \xi_2, 1) & \cdots & \tau^J_{\xi_1(0),\xi_2(0),D}(e_2, \xi_2, 1) \\ \vdots & \vdots & \ddots & \vdots \\ \tau^1_{\xi_1(0),\xi_2(0),D}(e_J, \xi_2, 1) & \tau^2_{\xi_1(0),\xi_2(0),D}(e_J, \xi_2, 1) & \cdots & \tau^J_{\xi_1(0),\xi_2(0),D}(e_J, \xi_2, 1) \end{pmatrix}$$

is non-singular.

We fix the value of ξ_3^* and, for each ξ_3, we choose $(\check{\xi}_2, \bar{\xi}_2, \bar{\xi}_3)$, $(a_1, ..., a_J)$, and $(b_1, ..., b_{J-1})$ as stated in Assumption ID(a). Evaluating (34) at four different points, $(\xi_3, \check{\xi}_2)$, $(\bar{\xi}_3, \check{\xi}_2)$, $(\bar{\xi}_3, \bar{\xi}_2)$, and $(\xi_3^*, \bar{\xi}_2)$ gives

$$\boldsymbol{Q}_{\check{\xi}_2,\xi_3} = \boldsymbol{L}_{\xi_3}\boldsymbol{D}_{\xi_3|\check{\xi}_2}\bar{\boldsymbol{L}}^\top_{\check{\xi}_2}, \quad \boldsymbol{Q}_{\check{\xi}_2,\bar{\xi}_3} = \boldsymbol{L}_{\bar{\xi}_3}\boldsymbol{D}_{\bar{\xi}_3|\check{\xi}_2}\bar{\boldsymbol{L}}^\top_{\check{\xi}_2},$$
$$\boldsymbol{Q}_{\bar{\xi}_2,\bar{\xi}_3} = \boldsymbol{L}_{\bar{\xi}_3}\boldsymbol{D}_{\bar{\xi}_3|\bar{\xi}_2}\bar{\boldsymbol{L}}^\top_{\bar{\xi}_2}, \quad \boldsymbol{Q}_{\bar{\xi}_2,\xi_3^*} = \boldsymbol{L}_{\xi_3^*}\boldsymbol{D}_{\xi_3^*|\bar{\xi}_2}\bar{\boldsymbol{L}}^\top_{\bar{\xi}_2}.$$

Then, following the identification argument in Carroll et al. (2010), under Assumption ID(a)(b), we have

$$A_{\xi_3,\xi_3^*} := \boldsymbol{Q}_{\check{\xi}_2,\xi_3}(\boldsymbol{Q}_{\check{\xi}_2,\bar{\xi}_3})^{-1}\boldsymbol{Q}_{\bar{\xi}_2,\bar{\xi}_3}(\boldsymbol{Q}_{\bar{\xi}_2,\xi_3^*})^{-1} = \boldsymbol{L}_{\xi_3}\boldsymbol{D}_{\xi_3,\xi_3^*}\boldsymbol{L}^{-1}_{\xi_3^*}, \tag{37}$$

where

$$\boldsymbol{D}_{\xi_3,\xi_3^*} := \boldsymbol{D}_{\xi_3|\check{\xi}_2}(\boldsymbol{D}_{\bar{\xi}_3|\check{\xi}_2})^{-1}\boldsymbol{D}_{\bar{\xi}_3|\bar{\xi}_2}(\boldsymbol{D}_{\xi_3^*|\bar{\xi}_2})^{-1}. \tag{38}$$

We first identify $\boldsymbol{L}_{\xi_3}$ for all $\xi_3 \in \mathcal{X}$ up to an unknown permutation matrix. Evaluating (37) and (38) at $\xi_3^* = \xi_3$, we have

$$\boldsymbol{A}_{\xi_3,\xi_3}\boldsymbol{L}_{\xi_3} = \boldsymbol{L}_{\xi_3}\boldsymbol{D}_{\xi_3,\xi_3}.$$

Because $\boldsymbol{A}_{\xi_3,\xi_3}$ has J distinct eigenvalues under Assumption ID(a), the eigenvalues of $\boldsymbol{A}_{\xi_3,\xi_3}$ determine the diagonal elements of $\boldsymbol{D}_{\xi_3,\xi_3}$ while the right eigenvectors of $\boldsymbol{A}_{\xi_3,\xi_3}$ determine the columns of $\boldsymbol{L}_{\xi_3}$ up to multiplicative constant and the ordering of its columns. Namely, collecting the right eigenvectors of $\boldsymbol{A}_{\xi_3,\xi_3}$ into a matrix in descending order of their eigenvalues, we identify

$$\boldsymbol{B} := \boldsymbol{L}_{\xi_3}\Delta_{\xi_3}\boldsymbol{C},$$

where $\boldsymbol{B}$ satisfies $\boldsymbol{A}_{\xi_3,\xi_3}\boldsymbol{B} = \boldsymbol{B}\boldsymbol{D}_{\xi_3,\xi_3}$, Δ_{ξ_3} is an unknown permutation matrix, and $\boldsymbol{C}$ is some diagonal matrix with non-zero diagonal elements.

We can determine the diagonal matrix $\boldsymbol{CD}_{\xi_3,\xi_3}$ from the first row of $\boldsymbol{A}_{\xi_3,\xi_3}\boldsymbol{B} = \boldsymbol{BD}_{\xi_3,\xi_3} = \boldsymbol{L}_{\xi_3}\Delta_{\xi_3}\boldsymbol{CD}_{\xi_3,\xi_3}$ because the first row of $\boldsymbol{L}_{\xi_3}\Delta_{\xi_3}$ is a vector of ones. Then, $\boldsymbol{L}_{\xi_3}\Delta_{\xi_3}$ is determined from $\boldsymbol{A}_{\xi_3,\xi_3}\boldsymbol{B}$ and $\boldsymbol{CD}_{\xi_3,\xi_3}$ as $\boldsymbol{L}_{\xi_3}\Delta_{\xi_3} = \boldsymbol{A}_{\xi_3,\xi_3}\boldsymbol{B}(\boldsymbol{CD}_{\xi_3,\xi_3})^{-1}$ in view of $\boldsymbol{A}_{\xi_3,\xi_3}\boldsymbol{B} = \boldsymbol{L}_{\xi_3}\Delta_{\xi_3}\boldsymbol{CD}_{\xi_3,\xi_3}$. Repeating the above argument for all values of $\xi_3 \in \mathcal{X}$, the eigenvalue decomposition algorithm identifies the matrices

$$\tilde{\boldsymbol{L}}_{\xi_3} := \boldsymbol{L}_{\xi_3}\Delta_{\xi_3} \quad \text{for all } \xi_3 \in \mathcal{X}, \tag{39}$$

where Δ_{ξ_3} is an unknown permutation matrix that depends on ξ_3.

Next, we identify permutation matrices that re-arrange $\boldsymbol{L}_{\xi_3}\Delta_{\xi_3}$ into a common order of latent types across different values of ξ_3, following the identification argument in Higgins and Jochmans (2023). Pre- and post-multiplying (37) by $\tilde{\boldsymbol{L}}_{\xi_3}^{-1}$ and $\tilde{\boldsymbol{L}}_{\xi_3^*}$, respectively, we obtain

$$\tilde{\boldsymbol{D}}_{\xi_3,\xi_3^*} := \tilde{\boldsymbol{L}}_{\xi_3}^{-1}\boldsymbol{A}_{\xi_3,\xi_3^*}\tilde{\boldsymbol{L}}_{\xi_3^*} = \Delta_{\xi_3}^{-1}\boldsymbol{D}_{\xi_3,\xi_3^*}\Delta_{\xi_3^*} = \left(\Delta_{\xi_3}^{-1}\Delta_{\xi_3^*}\right)\left(\Delta_{\xi_3^*}^{-1}\boldsymbol{D}_{\xi_3,\xi_3^*}\Delta_{\xi_3^*}\right),$$

where the last equality inserts the identity matrix $\Delta_{\xi_3^*}\Delta_{\xi_3^*}^{-1}$. Define the relative permutation matrix $\boldsymbol{P}_{\xi_3} := \Delta_{\xi_3}^{-1}\Delta_{\xi_3^*}$ and the diagonal matrix $\boldsymbol{D}^*_{\xi_3,\xi_3^*} := \Delta_{\xi_3^*}^{-1}\boldsymbol{D}_{\xi_3,\xi_3^*}\Delta_{\xi_3^*}$. We observe that $\tilde{\boldsymbol{D}}_{\xi_3,\xi_3^*} = \boldsymbol{P}_{\xi_3}\boldsymbol{D}^*_{\xi_3,\xi_3^*}$. Because $\boldsymbol{P}_{\xi_3}$ is a permutation matrix, $\tilde{\boldsymbol{D}}_{\xi_3,\xi_3^*}$ is a column-permuted diagonal matrix. Specifically, the j-th column of $\tilde{\boldsymbol{D}}_{\xi_3,\xi_3^*}$ contains exactly one non-zero element, which corresponds to a diagonal element of $\boldsymbol{D}^*_{\xi_3,\xi_3^*}$. Therefore, each diagonal element of $\boldsymbol{D}^*_{\xi_3,\xi_3^*}$ is identified by the sum of the elements in the corresponding column of $\tilde{\boldsymbol{D}}_{\xi_3,\xi_3^*}$. Consequently, the matrix $\boldsymbol{D}^*_{\xi_3,\xi_3^*} = \Delta_{\xi_3^*}^{-1}\boldsymbol{D}_{\xi_3,\xi_3^*}\Delta_{\xi_3^*}$ is identified.

With $\tilde{\boldsymbol{D}}_{\xi_3,\xi_3^*}$ and $\boldsymbol{D}^*_{\xi_3,\xi_3^*}$ identified, we can identify the relative permutation matrix as $\boldsymbol{P}_{\xi_3} = \tilde{\boldsymbol{D}}_{\xi_3,\xi_3^*}(\boldsymbol{D}^*_{\xi_3,\xi_3^*})^{-1}$. Substituting the definitions back into this expression, we have identified $\Delta_{\xi_3}^{-1}\Delta_{\xi_3^*} = \tilde{\boldsymbol{D}}_{\xi_3,\xi_3^*}(\Delta_{\xi_3^*}^{-1}\boldsymbol{D}_{\xi_3,\xi_3^*}\Delta_{\xi_3^*})^{-1}$. Finally, we identify $\boldsymbol{L}_{\xi_3}$ up to a common permutation matrix $\Delta_{\xi_3^*}$ (which does not depend on ξ_3 under Assumption ID(a)) by aligning the columns of $\tilde{\boldsymbol{L}}_{\xi_3}$ using the identified relative permutation as

$$\boldsymbol{L}^*_{\xi_3} := \boldsymbol{L}_{\xi_3}\Delta_{\xi_3^*} = \tilde{\boldsymbol{L}}_{\xi_3}(\Delta_{\xi_3}^{-1}\Delta_{\xi_3^*}) = \tilde{\boldsymbol{L}}_{\xi_3}\tilde{\boldsymbol{D}}_{\xi_3,\xi_3^*}\left(\Delta_{\xi_3^*}^{-1}\boldsymbol{D}_{\xi_3,\xi_3^*}\Delta_{\xi_3^*}\right)^{-1}. \tag{40}$$

In the next step, we identify $\{\tau^j_{\xi_1(0),D}(\cdot,0), p^j_{\xi_2(0)|\xi_1(0)}(\cdot|\cdot), p^j_{\xi_3(0)|\xi_2(0)}(\cdot|\cdot)\}_{j=1}^J$ up to a permutation matrix $\Delta_{\xi_3^*}$. Evaluating

$$p_{\xi_2,\xi_3,\xi_4|\xi_1,D}(\xi_2,\xi_3,\xi_4|\xi_1,0) = \sum_{j=1}^J \tau^j_{\xi_1(0),D}(\xi_1,0)p^j_{\xi_2(0)|\xi_1(0)}(\xi_2|\xi_1)p^j_{\xi_3(0)|\xi_2(0)}(\xi_3|\xi_2)p^j_{\xi_4(0)|\xi_3(0)}(\xi_4|\xi_3)$$

at $\xi_4 = b_1, b_2, ..., b_{J-1}$ and collecting them into a vector together with $p_{\xi_2,\xi_3|\xi_1,D}(\xi_2,\xi_3|\xi_1,0) = \sum_{j=1}^J \tau^j_{\xi_1(0),D}(\xi_1,0)p^j_{\xi_2(0)|\xi_1(0)}(\xi_2|\xi_1)p^j_{\xi_3(0)|\xi_2(0)}(\xi_3|\xi_2)$ gives

$$\boldsymbol{q}_{\xi_2,\xi_3|\xi_1} = \boldsymbol{L}_{\xi_3}\boldsymbol{\ell}_{\xi_2,\xi_3|\xi_1} = \boldsymbol{L}^*_{\xi_3}\Delta_{\xi_3^*}^{-1}\boldsymbol{\ell}_{\xi_2,\xi_3|\xi_1} \tag{41}$$

where

$$\boldsymbol{q}_{\xi_2,\xi_3|\xi_1} := \begin{pmatrix} p_{\xi_2,\xi_3|\xi_1,D}(\xi_2,\xi_3|\xi_1,0) \\ p_{\xi_2,\xi_3,\xi_4|\xi_1,D}(\xi_2,\xi_3,b_1|\xi_1,0) \\ \vdots \\ p_{\xi_2,\xi_3,\xi_4|\xi_1,D}(\xi_2,\xi_3,b_{J-1}|\xi_1,0) \end{pmatrix} \quad \text{and}$$

$$\boldsymbol{\ell}_{\xi_2,\xi_3|\xi_1} := \begin{pmatrix} \ell^1(\xi_2,\xi_3|\xi_1) \\ \ell^2(\xi_2,\xi_3|\xi_1) \\ \vdots \\ \ell^J(\xi_2,\xi_3|\xi_1) \end{pmatrix} := \begin{pmatrix} \tau^1_{\xi_1(0),D}(\xi_1,0)p^1_{\xi_2(0)|\xi_1(0)}(\xi_2|\xi_1)p^1_{\xi_3(0)|\xi_2(0)}(\xi_3|\xi_2) \\ \tau^2_{\xi_1(0),D}(\xi_1,0)p^2_{\xi_2(0)|\xi_1(0)}(\xi_2|\xi_1)p^2_{\xi_3(0)|\xi_2(0)}(\xi_3|\xi_2) \\ \vdots \\ \tau^J_{\xi_1(0),D}(\xi_1,0)p^J_{\xi_2(0)|\xi_1(0)}(\xi_2|\xi_1)p^J_{\xi_3(0)|\xi_2(0)}(\xi_3|\xi_2) \end{pmatrix}.$$

From (40) and (41), we identify $\boldsymbol{\ell}_{\xi_1,\xi_2,\xi_3}$ up to $\Delta_{\xi_3^*}$ for all $(\xi_1,\xi_2,\xi_3) \in |\mathcal{X}|^{k+2}$ as

$$\Delta^{-1}_{\xi_3^*}\boldsymbol{\ell}_{\xi_2,\xi_3|\xi_1} = \begin{pmatrix} \ell^{\delta(1)}(\xi_2,\xi_3|\xi_1) \\ \ell^{\delta(2)}(\xi_2,\xi_3|\xi_1) \\ \vdots \\ \ell^{\delta(J)}(\xi_2,\xi_3|\xi_1) \end{pmatrix} = \left(\boldsymbol{L}^*_{\xi_3}\right)^{-1} \boldsymbol{q}_{\xi_2,\xi_3|\xi_1}, \tag{42}$$

where

$$\delta : \{1,2,...,J\} \to \{1,2,...,J\}$$

is a permutation implied by $\Delta_{\xi_3^*}$. Then, $\tau^{\delta(j)}_{\xi_1(0),D}(\xi_1,0)$ is identified from $\ell^{\delta(j)}(\xi_2,\xi_3|\xi_1)$ in (42) as

$$\begin{aligned} \tau^{\delta(j)}_{\xi_1(0),D}(\xi_1,0) &= \sum_{(\xi_2,\xi_3)\in\mathcal{X}^2} \ell^{\delta(j)}(\xi_2,\xi_3|\xi_1) \\ &= \sum_{(\xi_2,\xi_3)\in\mathcal{X}^2} \tau^{\delta(j)}_{\xi_1(0),D}(\xi_1,0)\, p^{\delta(j)}_{\xi_2(0)|\xi_1(0)}(\xi_2|\xi_1)\, p^{\delta(j)}_{\xi_3(0)|\xi_2(0)}(\xi_3|\xi_2). \end{aligned}$$

Given that $\tau^{\delta(j)}_{\xi_1(0),D}(\xi_1,0)$ is identified, $p^{\delta(j)}_{\xi_2(0)|\xi_1(0)}(\xi_2|\xi_1)$ and $p^{\delta(j)}_{\xi_3(0)|\xi_2(0)}(\xi_3|\xi_2)$ are also identified as

$$p^{\delta(j)}_{\xi_2(0)|\xi_1(0)}(\xi_2|\xi_1) = \sum_{\xi_3\in\mathcal{X}} \ell^{\delta(j)}(\xi_2,\xi_3|\xi_1)/\tau^{\delta(j)}_{\xi_1(0),D}(\xi_1,0)$$

and

$$p^{\delta(j)}_{\xi_3(0)|\xi_2(0)}(\xi_3|\xi_2) = \ell^{\delta(j)}(\xi_2,\xi_3|\xi_1)/(\tau^{\delta(j)}_{\xi_1(0),D}(\xi_1,0)p^{\delta(j)}_{\xi_2(0)|\xi_1(0)}(\xi_2|\xi_1)).$$

This proves the identification of $\{\tau^{\delta(j)}_{\xi_1(0),D}(\cdot,0), p^{\delta(j)}_{\xi_2(0)|\xi_1(0)}(\cdot|\cdot), p^{\delta(j)}_{\xi_3(0)|\xi_2(0)}(\cdot|\cdot)\}_{j=1}^J$ up to $\Delta_{\xi_3^*}$.

To identify $\{p^j_{\xi_4(0)|\xi_3(0)}(\cdot|\cdot)\}_{j=1}^J$ up to $\Delta_{\xi_3^*}$, define

$$\bar{\boldsymbol{q}}_{\check{\xi}_2,\xi_3,\xi_4} := \begin{pmatrix} p_{\xi_2,\xi_3,\xi_4|\xi_1,D}(\check{\xi}_2,\xi_3,\xi_4|a_1,0) \\ \vdots \\ p_{\xi_2,\xi_3,\xi_4|\xi_1,D}(\check{\xi}_2,\xi_3,\xi_4|a_J,0) \end{pmatrix} \quad \text{and} \quad \bar{\boldsymbol{\ell}}_{\xi_4|\xi_3} := \begin{pmatrix} p^1_{\xi_4(0)|\xi_3(0)}(\xi_4|\xi_3) \\ \vdots \\ p^J_{\xi_4(0)|\xi_3(0)}(\xi_4|\xi_3) \end{pmatrix}.$$

For each (ξ_3,ξ_4), we have the linear system:

$$\begin{aligned} \bar{\boldsymbol{q}}_{\check{\xi}_2,\xi_3,\xi_4} &= \bar{\boldsymbol{L}}_{\check{\xi}_2}\boldsymbol{D}_{\xi_3|\check{\xi}_2}\bar{\boldsymbol{\ell}}_{\xi_4|\xi_3} \\ &= (\bar{\boldsymbol{L}}_{\check{\xi}_2}\Delta_{\xi_3^*})(\Delta_{\xi_3^*}^{-1}\boldsymbol{D}_{\xi_3|\check{\xi}_2}\Delta_{\xi_3^*})(\Delta_{\xi_3^*}^{-1}\bar{\boldsymbol{\ell}}_{\xi_4|\xi_3}). \end{aligned}$$

The matrices $\bar{\boldsymbol{L}}_{\check{\xi}_2}\Delta_{\xi_3^*}$ and $\Delta_{\xi_3^*}^{-1}\boldsymbol{D}_{\xi_3|\check{\xi}_2}\Delta_{\xi_3^*}$ are now identified from

$$\{\tau^{\delta(j)}_{\xi_1(0),D}(\cdot,0),\, p^{\delta(j)}_{\xi_2(0)|\xi_1(0)}(\cdot|\cdot),\, p^{\delta(j)}_{\xi_3(0)|\xi_2(0)}(\cdot|\cdot)\}_{j=1}^J$$

up to $\Delta_{\xi_3^*}$. Both are invertible by assumption. We can therefore solve for the permuted component vector as

$$\Delta_{\xi_3^*}^{-1}\bar{\boldsymbol{\ell}}_{\xi_4|\xi_3} = \begin{pmatrix} p^{\delta(1)}_{\xi_4(0)|\xi_3(0)}(\xi_4|\xi_3) \\ \vdots \\ p^{\delta(J)}_{\xi_4(0)|\xi_3(0)}(\xi_4|\xi_3) \end{pmatrix} = (\Delta_{\xi_3^*}^{-1}\boldsymbol{D}_{\xi_3|\check{\xi}_2}\Delta_{\xi_3^*})^{-1}(\bar{\boldsymbol{L}}_{\check{\xi}_2}\Delta_{\xi_3^*})^{-1}\bar{\boldsymbol{q}}_{\check{\xi}_2,\xi_3,\xi_4}.$$

Therefore, $\{p^{\delta(j)}_{\xi_4(0)|\xi_3(0)}(\cdot|\cdot)\}_{j=1}^J$ is identified.

We proceed to prove the identification of $\{\tau^j_{\xi_1(0),D}(\cdot,1), p^j_{\xi_3(1)|\xi_2(1)}(\cdot|\cdot), p^j_{\xi_4(1)|\xi_3(1)}(\cdot|\cdot)\}_{j=1}^J$ up to the permutation $\delta(\cdot)$ implied by $\Delta_{\xi_3^*}$.

First, to identify $\{\tau^{\delta(j)}_{\xi_1(0),D}(\cdot,1)\}_{j=1}^J$, we note that $\xi_j(0)$ for $j=1,2$ is observed for both $D=0$ and $D=1$. For each $\xi_1 \in \mathcal{X}^{kJ}$, we choose $\{c_j\}_{j=1}^J$ as in Assumption ID(c). We then evaluate the PMF $p_{\xi_2|\xi_1,D}(\xi_2|\xi_1,d) = \sum_{j=1}^J \tau^{\delta(j)}_{\xi_1(0),D}(\xi_1,d)p^{\delta(j)}_{\xi_2(0)|\xi_1(0)}(\xi_2|\xi_1)$ at $d=0,1$ and $\xi_2 = c_1,\ldots,c_J$. This yields the matrix system:

$$\begin{aligned} &\begin{pmatrix} p_{\xi_2|\xi_1,D}(c_1|\xi_1,0) & \cdots & p_{\xi_2|\xi_1,D}(c_J|\xi_1,0) \\ p_{\xi_2|\xi_1,D}(c_1|\xi_1,1) & \cdots & p_{\xi_2|\xi_1,D}(c_J|\xi_1,1) \end{pmatrix} \\ &= \begin{pmatrix} \tau^{\delta(1)}_{\xi_1(0),D}(\xi_1,0) & \cdots & \tau^{\delta(J)}_{\xi_1(0),D}(\xi_1,0) \\ \tau^{\delta(1)}_{\xi_1(0),D}(\xi_1,1) & \cdots & \tau^{\delta(J)}_{\xi_1(0),D}(\xi_1,1) \end{pmatrix} \begin{pmatrix} p^{\delta(1)}_{\xi_2(0)|\xi_1(0)}(c_1|\xi_1) & \cdots & p^{\delta(1)}_{\xi_2(0)|\xi_1(0)}(c_J|\xi_1) \\ \vdots & \ddots & \vdots \\ p^{\delta(J)}_{\xi_2(0)|\xi_1(0)}(c_1|\xi_1) & \cdots & p^{\delta(J)}_{\xi_2(0)|\xi_1(0)}(c_J|\xi_1) \end{pmatrix}. \end{aligned}$$

The $2\times J$ matrix on the left is observed. The second $J\times J$ matrix on the right is identified from the $D=0$ case and is invertible by Assumption ID(c). Thus, the $2\times J$ matrix of permuted τ

components is identified by inversion:

$$\begin{pmatrix} \tau^{\delta(1)}_{\xi_1(0),D}(\xi_1,0) & \cdots & \tau^{\delta(J)}_{\xi_1(0),D}(\xi_1,0) \\ \tau^{\delta(1)}_{\xi_1(0),D}(\xi_1,1) & \cdots & \tau^{\delta(J)}_{\xi_1(0),D}(\xi_1,1) \end{pmatrix}$$
$$= \begin{pmatrix} p_{\xi_2|\xi_1,D}(c_1|\xi_1,0) & \cdots & p_{\xi_2|\xi_1,D}(c_J|\xi_1,0) \\ p_{\xi_2|\xi_1,D}(c_1|\xi_1,1) & \cdots & p_{\xi_2|\xi_1,D}(c_J|\xi_1,1) \end{pmatrix} \begin{pmatrix} p^{\delta(1)}_{\xi_2(0)|\xi_1(0)}(c_1|\xi_1) & \cdots & p^{\delta(1)}_{\xi_2(0)|\xi_1(0)}(c_J|\xi_1) \\ \vdots & \ddots & \vdots \\ p^{\delta(J)}_{\xi_2(0)|\xi_1(0)}(c_1|\xi_1) & \cdots & p^{\delta(J)}_{\xi_2(0)|\xi_1(0)}(c_J|\xi_1) \end{pmatrix}^{-1}.$$

This identifies the second row, $\{\tau^{\delta(j)}_{\xi_1(0),D}(\cdot,1)\}_{j=1}^J$.

Next, for the identification of $\{p^{\delta(j)}_{\xi_3(1)|\xi_2(1)}(\cdot|\cdot), p^{\delta(j)}_{\xi_4(1)|\xi_3(1)}(\cdot|\cdot)\}_{j=1}^J$, we define:

$$\tau^{\delta(j)}_{\xi_1,\xi_2,D}(\xi_1,\xi_2,1) := \tau^{\delta(j)}_{\xi_1(0),D}(\xi_1,1)p^{\delta(j)}_{\xi_2(0)|\xi_1(0)}(\xi_2|\xi_1) \quad \text{and}$$
$$p^{\delta(j)}_{\xi_3,\xi_4|\xi_2}(\xi_3,\xi_4|\xi_2) := p^{\delta(j)}_{\xi_3(1)|\xi_2(1)}(\xi_3|\xi_2)p^{\delta(j)}_{\xi_4(1)|\xi_3(1)}(\xi_4|\xi_3).$$

Note that $\tau^{\delta(j)}_{\xi_1,\xi_2,D}$ is identified from the preceding steps. The conditional PMF for $D=1$ can be expressed as:

$$p_{\xi_2,\xi_3,\xi_4|\xi_1,D}(\xi_2,\xi_3,\xi_4|\xi_1,1) = \sum_{j=1}^J \tau^{\delta(j)}_{\xi_1,\xi_2,D}(\xi_1,\xi_2,1)p^{\delta(j)}_{\xi_3,\xi_4|\xi_2}(\xi_3,\xi_4|\xi_2).$$

For each $\xi_2 \in \mathcal{X}$, choose $\{e_j\}_{j=1}^J$ as in Assumption ID(d). Evaluating the PMF at $\xi_1 = e_1,\ldots,e_J$ yields:

$$\begin{pmatrix} p_{\xi_2,\xi_3,\xi_4|\xi_1,D}(\xi_2,\xi_3,\xi_4|e_1,1) \\ \vdots \\ p_{\xi_2,\xi_3,\xi_4|\xi_1,D}(\xi_2,\xi_3,\xi_4|e_J,1) \end{pmatrix} = \begin{pmatrix} \tau^{\delta(1)}_{\xi_1,\xi_2,D}(e_1,\xi_2,1) & \cdots & \tau^{\delta(J)}_{\xi_1,\xi_2,D}(e_1,\xi_2,1) \\ \vdots & \ddots & \vdots \\ \tau^{\delta(1)}_{\xi_1,\xi_2,D}(e_J,\xi_2,1) & \cdots & \tau^{\delta(J)}_{\xi_1,\xi_2,D}(e_J,\xi_2,1) \end{pmatrix} \begin{pmatrix} p^{\delta(1)}_{\xi_3,\xi_4|\xi_2}(\xi_3,\xi_4|\xi_2) \\ \vdots \\ p^{\delta(J)}_{\xi_3,\xi_4|\xi_2}(\xi_3,\xi_4|\xi_2) \end{pmatrix}.$$

The $J \times J$ matrix of $\tau^{\delta(j)}_{\xi_1,\xi_2,D}$ components is identified and is invertible by Assumption ID(d). We can therefore solve for the vector of permuted product probabilities:

$$\begin{pmatrix} p^{\delta(1)}_{\xi_3,\xi_4|\xi_2}(\xi_3,\xi_4|\xi_2) \\ \vdots \\ p^{\delta(J)}_{\xi_3,\xi_4|\xi_2}(\xi_3,\xi_4|\xi_2) \end{pmatrix} = \begin{pmatrix} \tau^{\delta(1)}_{\xi_1,\xi_2,D}(e_1,\xi_2,1) & \cdots & \tau^{\delta(J)}_{\xi_1,\xi_2,D}(e_1,\xi_2,1) \\ \vdots & \ddots & \vdots \\ \tau^{\delta(1)}_{\xi_1,\xi_2,D}(e_J,\xi_2,1) & \cdots & \tau^{\delta(J)}_{\xi_1,\xi_2,D}(e_J,\xi_2,1) \end{pmatrix}^{-1} \begin{pmatrix} p_{\ldots}(\xi_2,\xi_3,\xi_4|e_1,1) \\ \vdots \\ p_{\ldots}(\xi_2,\xi_3,\xi_4|e_J,1) \end{pmatrix}.$$

The individual components $\{p^{\delta(j)}_{\xi_3(1)|\xi_2(1)}(\cdot|\cdot), p^{\delta(j)}_{\xi_4(1)|\xi_3(1)}(\cdot|\cdot)\}_{j=1}^J$ are then identified by marginalizing $p^{\delta(j)}_{\xi_3,\xi_4|\xi_2}$ (e.g., summing over ξ_4 to find $p^{\delta(j)}_{\xi_3(1)|\xi_2(1)}$).

Finally, we identify the base probabilities $\{\pi^{\delta(j)}, p^{\delta(j)}_{\xi_1(0),D}(\cdot,\cdot)\}_{j=1}^J$ using the identified

$\{\tau^{\delta(j)}_{\xi_1(0),D}(\cdot,\cdot)\}_{j=1}^J$ and the observable PMF $p_{\xi_1,D}(\xi_1,d)$:

$$\pi^{\delta(j)} = \sum_{(\xi_1,d)} \tau^{\delta(j)}_{\xi_1(0),D}(\xi_1,d) p_{\xi_1,D}(\xi_1,d),$$
$$p^{\delta(j)}_{\xi_1(0),D}(\xi_1,d) = \frac{\tau^{\delta(j)}_{\xi_1(0),D}(\xi_1,d) p_{\xi_1,D}(\xi_1,d)}{\pi^{\delta(j)}}.$$

Under Assumption RSM(c), re-ordering the components such that $\pi^1 < \pi^2 < \cdots < \pi^J$ fixes the permutation $\delta(\cdot)$ and uniquely identifies all components, which completes the proof. □

A.6 Proof of Proposition 17

The expectation of untreated potential outcomes for treated units can be expressed as

$$\begin{aligned}
\mathbb{E}\left[\boldsymbol{X}_t(0) \mid D=1, Z_i=j\right] &= \mathbb{E}\left[\mathbb{E}\left[\boldsymbol{X}_t(0) \mid \boldsymbol{X}_1^{T_0}(0), D=1, Z=j\right] \mid D=1, Z=j\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\boldsymbol{X}_t(0) \mid \boldsymbol{X}_1^{T_0}(0), D=0, Z=j\right] \mid D=1, Z=j\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\boldsymbol{X}_t(0) \mid \boldsymbol{X}_{T_0}(0), D=0, Z=j\right] \mid D=1, Z=j\right] \\
&= \sum_{\boldsymbol{x}_{T_0}\in\mathcal{X}} \Pr(\boldsymbol{X}_{T_0}=\boldsymbol{x}_{T_0} \mid D=1, Z=j)\mathbb{E}\left[\boldsymbol{X}_t \mid \boldsymbol{X}_{T_0}=\boldsymbol{x}_{T_0}, D=0, Z=j\right],
\end{aligned}$$

where the first equality follows from the law of iterated expectations, the second from transition independence, the third from the Markov property of $\boldsymbol{X}_t(0)$, and the fourth from the definition of conditional expectations.

Similarly,

$$\mathbb{E}\left[\boldsymbol{X}_t(1) \mid D=1, Z_i=j\right] = \sum_{\boldsymbol{x}_{T_0}} \Pr(\boldsymbol{X}_{T_0}=\boldsymbol{x}_{T_0} \mid D=1, Z=j)\mathbb{E}\left[\boldsymbol{X}_t \mid \boldsymbol{X}_{T_0}=\boldsymbol{x}_{T_0}, D=1, Z=j\right].$$

Therefore, (18) follows from subtracting the two objects given $\boldsymbol{\mu}_t^{ATT,j} = \mathbb{E}\left[\boldsymbol{X}_t(1) \mid D=1, Z_i=j\right] - \mathbb{E}\left[\boldsymbol{X}_t(0) \mid D=1, Z_i=j\right]$.

Moreover, since Proposition 16 implies that the jth type-specific probabilities and expectations appearing in (18) are identified from $\{p_{\boldsymbol{W}}(\boldsymbol{w};\boldsymbol{\psi}) : \boldsymbol{w} \in \boldsymbol{\mathcal{W}}\}$, it follows directly from (18) that $\boldsymbol{\mu}_t^{ATT,j}$ is identified. □

A.7 Proof of Proposition 20

Suppose that $T_0 = 2\ell$, $T = 4\ell$, and $J = |\mathcal{X}|^\ell$. Partition $(\boldsymbol{X}_1,\ldots,\boldsymbol{X}_T)$ as $(\boldsymbol{X}_1,\ldots,\boldsymbol{X}_T) = (\xi_1,\xi_2,\xi_3,\xi_4)$, where

$$\xi_1 := \boldsymbol{X}_1^\ell \in \mathcal{X}^\ell,\ \xi_2 := \boldsymbol{X}_{\ell+1}^{2\ell} \in \mathcal{X}^\ell,\ \xi_3 := \boldsymbol{X}_{2\ell+1}^{3\ell} \in \mathcal{X}^\ell,\ \text{and}\ \xi_4 := \boldsymbol{X}_{3\ell+1}^{T} \in \mathcal{X}^\ell. \tag{43}$$

For $d \in \{0,1\}$, define the corresponding potential outcomes $\xi_1(d)$, $\xi_2(d)$, $\xi_3(d)$, and $\xi_4(d)$ analogously, by replacing $\boldsymbol{X}_t$ with $\boldsymbol{X}_t(d)$ in the above definitions.

Assumption ID$_\ell$. Assumption ID holds with the ξ's defined in (43) in place of those in (29).

Under Assumption ID$_\ell$, we can replicate the proof of Proposition 16 using the block definitions in (43) and the corresponding potential outcomes. This establishes identification of

$$\begin{gathered}\{\pi^j,\, p^j_{\xi_1(0),D}(\cdot,\cdot),\, p^j_{\xi_2(0)|\xi_1(0)}(\cdot \mid \cdot),\, p^j_{\xi_3(0)|\xi_2(0)}(\cdot \mid \cdot),\\ p^j_{\xi_4(0)|\xi_3(0)}(\cdot \mid \cdot),\, p^j_{\xi_3(1)|\xi_2(1)}(\cdot \mid \cdot),\, p^j_{\xi_4(1)|\xi_3(1)}(\cdot \mid \cdot)\}_{j=1}^J,\end{gathered}$$

thereby proving part (a).

Part (b) follows similarly by repeating the proof of Proposition 17 under Assumption M$_\ell$ in place of Assumption M. ☐

A.8 Proof of Proposition 23

The consistency of the MLE $\hat{\boldsymbol{\psi}}$ follows directly from Theorem 2.5 of Newey and McFadden (1994). Specifically, condition (i) of that theorem is satisfied by Proposition 16, while conditions (ii)–(iv) are implied by Assumption MLE(a). Asymptotic normality then follows from Theorem 3.3 of Newey and McFadden (1994), whose conditions (i)–(v) are readily verified under Assumption MLE. Hence,

$$\hat{\boldsymbol{\psi}} \xrightarrow{p} \boldsymbol{\psi}^*, \tag{44}$$

$$\sqrt{n}(\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}^*) = \boldsymbol{I}(\boldsymbol{\psi}^*)^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^n \boldsymbol{S}(\boldsymbol{W}_i; \boldsymbol{\psi}^*) + o_p(1) \tag{45}$$

$$\Rightarrow N(0, \boldsymbol{I}(\boldsymbol{\psi}^*)^{-1}), \tag{46}$$

where $\boldsymbol{S}(\boldsymbol{W}_i; \boldsymbol{\psi}^*)$ is the individual score function evaluated at $\boldsymbol{\psi}^*$ for the MLE in (21).

By continuity of the posterior mapping, the consistency of $\hat{\boldsymbol{\psi}}$ implies that

$$\sup_i \left|\hat{\tau}_i^j - \tau^j(\boldsymbol{W}_i; \boldsymbol{\psi}^*)\right| \xrightarrow{p} 0, \qquad \tau^j(\boldsymbol{W}; \boldsymbol{\psi}^*) := \Pr(Z = j \mid \boldsymbol{W}; \boldsymbol{\psi}^*). \tag{47}$$

Furthermore, Assumption MLE(a) ensures that $\mathbb{E}[\tau^j(\boldsymbol{W}; \boldsymbol{\psi})]$ and $\mathbb{E}[\tau^j(\boldsymbol{W}; \boldsymbol{\psi})\mathbf{1}\{\boldsymbol{X}_{T_0} = \boldsymbol{x}_{T_0}, D = d\}]$ are strictly positive in a neighborhood of $\boldsymbol{\psi}^*$.

Then, in (24),

$$\begin{aligned}\widehat{\Pr}(\boldsymbol{X}_{T_0} = \boldsymbol{x}_{T_0} \mid D = d, Z = j) &= \frac{\frac{1}{n}\sum_{i=1}^n \mathbf{1}\{\boldsymbol{X}_{iT_0} = \boldsymbol{x}_{T_0}, D_i = d\}\, \tau^j(\boldsymbol{W}_i; \boldsymbol{\psi}^*) + o_p(1)}{\frac{1}{n}\sum_{i=1}^n \tau^j(\boldsymbol{W}_i; \boldsymbol{\psi}^*) + o_p(1)}\\ &\xrightarrow{p} \frac{\mathbb{E}\big[\mathbf{1}\{\boldsymbol{X}_{T_0} = \boldsymbol{x}_{T_0}, D = d\}\, \tau^j(\boldsymbol{W}; \boldsymbol{\psi}^*)\big]}{\mathbb{E}\big[\tau^j(\boldsymbol{W}; \boldsymbol{\psi}^*)\big]}\\ &= \Pr(\boldsymbol{X}_{T_0} = \boldsymbol{x}_{T_0} \mid D = d, Z = j),\end{aligned}$$

where the first equality follows from (47), the second from the Law of Large Numbers and the Continuous Mapping Theorem, and the last from $\mathbb{E}[\tau^j(\boldsymbol{W};\boldsymbol{\psi}^*)g(\boldsymbol{W})] = \mathbb{E}[\mathbf{1}\{Z=j\}g(\boldsymbol{W})]$ for any measurable function g.

Identical arguments apply to $\widehat{\mathbb{E}}[\boldsymbol{X}_t \mid \boldsymbol{X}_{T_0} = \boldsymbol{x}_{T_0}, D=d, Z=j]$ and $\widehat{\Pr}(Z=j \mid D=1)$ in (25) and (26). Hence,

$$\widehat{\mathbb{E}}[\boldsymbol{X}_t \mid \boldsymbol{X}_{T_0} = \boldsymbol{x}_{T_0}, D=d, Z=j] \xrightarrow{p} \mathbb{E}[\boldsymbol{X}_t \mid \boldsymbol{X}_{T_0} = \boldsymbol{x}_{T_0}, D=d, Z=j]\,,$$
$$\widehat{\Pr}(Z=j \mid D=1) \xrightarrow{p} \Pr(Z=j \mid D=1).$$

Both $\hat{\boldsymbol{\mu}}_t^{ATT,j}$ and $\hat{\boldsymbol{\mu}}_t^{ATT}$ in (23) and (26) are a continuously differentiable function of these probabilities and expectations. Therefore, by the Continuous Mapping Theorem,

$$\hat{\boldsymbol{\mu}}_t^{ATT,j} \xrightarrow{p} \boldsymbol{\mu}_t^{ATT,j}, \qquad \hat{\boldsymbol{\mu}}_t^{ATT} \xrightarrow{p} \boldsymbol{\mu}_t^{ATT},$$

which establishes part (a).

Let $\hat{\boldsymbol{m}}_n(\hat{\boldsymbol{\psi}})$ denote the vector collecting the estimators for the probabilities and expectations defined in (24)-(26). Specifically, for each $j = 1, 2, \ldots, J$ and $(\boldsymbol{x}_{T_0}, d) \in \mathcal{X} \times \{0,1\}$,

$$\hat{m}_{1,n}^{j,\boldsymbol{x}_{T_0}}(\hat{\boldsymbol{\psi}}) = \frac{\frac{1}{n}\sum_{i=1}^n \mathbf{1}\{\boldsymbol{X}_{iT_0} = \boldsymbol{x}_{T_0}, D_i = 1\}\,\tau^j(\boldsymbol{W}_i;\hat{\boldsymbol{\psi}})}{\frac{1}{n}\sum_{i=1}^n \mathbf{1}\{D_i=1\}\tau^j(\boldsymbol{W}_i;\hat{\boldsymbol{\psi}})},$$
$$\hat{m}_{2,n}^{j,\boldsymbol{x}_{T_0},d}(\hat{\boldsymbol{\psi}}) = \frac{\frac{1}{n}\sum_{i=1}^n \boldsymbol{X}_{it}\,\mathbf{1}\{\boldsymbol{X}_{iT_0} = \boldsymbol{x}_{T_0}, D_i = d\}\,\tau^j(\boldsymbol{W}_i;\hat{\boldsymbol{\psi}})}{\frac{1}{n}\sum_{i=1}^n \mathbf{1}\{\boldsymbol{X}_{iT_0} = \boldsymbol{x}_{T_0}, D_i = d\}\,\tau^j(\boldsymbol{W}_i;\hat{\boldsymbol{\psi}})}, \quad \hat{m}_{3,n}^{j}(\hat{\boldsymbol{\psi}}) = \frac{\frac{1}{n}\sum_{i=1}^n \mathbf{1}\{D_i=1\}\hat{\tau}_i^j}{\frac{1}{n}\sum_{i=1}^n \mathbf{1}\{D_i=1\}},$$

and $\hat{\boldsymbol{m}}_n(\hat{\boldsymbol{\psi}})$ stacks all such elements into a vector.

Let $\boldsymbol{m}(\boldsymbol{\psi})$ denote the corresponding vector of population ratios under the parameter value $\boldsymbol{\psi}$, with elements given by

$$m_1^{j,\boldsymbol{x}_{T_0}}(\boldsymbol{\psi}) = \frac{\mathbb{E}\big[\mathbf{1}\{\boldsymbol{X}_{T_0} = \boldsymbol{x}_{T_0}, D=1\}\,\tau^j(\boldsymbol{W};\boldsymbol{\psi})\big]}{\mathbb{E}[\mathbf{1}\{D_i=1\}\tau^j(\boldsymbol{W};\boldsymbol{\psi})]},$$
$$m_2^{j,\boldsymbol{x}_{T_0},d}(\boldsymbol{\psi}) = \frac{\mathbb{E}\big[\boldsymbol{X}_t\,\mathbf{1}\{\boldsymbol{X}_{T_0} = \boldsymbol{x}_{T_0}, D=d\}\,\tau^j(\boldsymbol{W};\boldsymbol{\psi})\big]}{\mathbb{E}[\mathbf{1}\{\boldsymbol{X}_{T_0} = \boldsymbol{x}_{T_0}, D=d\}\,\tau^j(\boldsymbol{W};\boldsymbol{\psi})]}, \quad m_3^j(\boldsymbol{\psi}) = \frac{\mathbb{E}[\mathbf{1}\{D_i=1\}\tau^j(\boldsymbol{W};\boldsymbol{\psi})]}{\mathbb{E}[\mathbf{1}\{D_i=1\}]}.$$

By definition, $\boldsymbol{m}(\boldsymbol{\psi}^*)$ collects the true population probabilities and expectations evaluated at the true parameter $\boldsymbol{\psi}^*$.

Note that

$$\sqrt{n}\big(\hat{\boldsymbol{m}}(\hat{\boldsymbol{\psi}}) - \boldsymbol{m}(\boldsymbol{\psi}^*)\big) = \underbrace{\sqrt{n}(\hat{\boldsymbol{m}}_n(\boldsymbol{\psi}^*) - \boldsymbol{m}(\boldsymbol{\psi}^*))}_{(I)} + \underbrace{\sqrt{n}\Big(\hat{\boldsymbol{m}}_n(\hat{\boldsymbol{\psi}}) - \hat{\boldsymbol{m}}_n(\boldsymbol{\psi}^*)\Big)}_{(II)}. \tag{48}$$

For term (II), apply a mean-value (componentwise) expansion of the map $\boldsymbol{\psi} \mapsto \hat{\boldsymbol{m}}_n(\boldsymbol{\psi})$: there exists $\tilde{\boldsymbol{\psi}}$ on the line segment between $\hat{\boldsymbol{\psi}}$ and $\boldsymbol{\psi}^*$ such that

$$(II) = \left(\nabla_{\boldsymbol{\psi}} \hat{\boldsymbol{m}}_n(\tilde{\boldsymbol{\psi}})\right)^\top \sqrt{n}\,(\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}^*),$$

where $\nabla_{\boldsymbol{\psi}} \hat{\boldsymbol{m}}_n(\tilde{\boldsymbol{\psi}}) = \partial \hat{\boldsymbol{m}}_n(\boldsymbol{\psi})/\partial \boldsymbol{\psi}|_{\boldsymbol{\psi}=\tilde{\boldsymbol{\psi}}}$. Since $\hat{\boldsymbol{\psi}} \xrightarrow{p} \boldsymbol{\psi}^*$, we have $\tilde{\boldsymbol{\psi}} \xrightarrow{p} \boldsymbol{\psi}^*$. Under the Uniform Law of Large Numbers for the class $\{\nabla_{\boldsymbol{\psi}} \hat{\boldsymbol{m}}_n(\boldsymbol{\psi}) : \boldsymbol{\psi} \in \mathcal{N}(\boldsymbol{\psi}^*)\}$—which holds because $\mathcal{W}$ is finite (so the supremum over $\boldsymbol{w}$ reduces to a finite maximum and the envelope condition is trivially satisfied), combined with continuity of $\tau^j(\cdot;\boldsymbol{\psi})$ and its derivatives in a neighborhood $\mathcal{N}(\boldsymbol{\psi}^*)$—it follows that

$$\nabla_{\boldsymbol{\psi}} \hat{\boldsymbol{m}}_n(\tilde{\boldsymbol{\psi}}) \ \xrightarrow{p}\ \nabla_{\boldsymbol{\psi}} \boldsymbol{m}(\boldsymbol{\psi}^*).$$

Combining this with the asymptotic linear representation of the MLE in (45), we obtain

$$(II) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \boldsymbol{A}^* \boldsymbol{S}(\boldsymbol{W}_i; \boldsymbol{\psi}^*) \ + \ o_p(1). \tag{49}$$

where $\boldsymbol{A}^* := \left(\nabla_{\boldsymbol{\psi}} \boldsymbol{m}(\boldsymbol{\psi}^*)\right)^\top \boldsymbol{I}(\boldsymbol{\psi}^*)^{-1}$.

For term (I), each component of $\hat{\boldsymbol{m}}_n(\boldsymbol{\psi}^*)$ is a ratio of empirical means, e.g.

$$\hat{m}_{1,n}^{j,\boldsymbol{x}_{T_0}}(\boldsymbol{\psi}^*) = \frac{n^{-1}\sum_{i=1}^n a_{\boldsymbol{\psi}^*}}{n^{-1}\sum_{i=1}^n b_{\boldsymbol{\psi}^*}}, \quad \text{where}$$
$$a_{\boldsymbol{\psi}^*}(\boldsymbol{W}) := \tau^j(\boldsymbol{W};\boldsymbol{\psi}^*)\,\mathbf{1}\{\boldsymbol{X}_{T_0} = \boldsymbol{x}_{T_0}, D=1\},\ b_{\boldsymbol{\psi}^*}(\boldsymbol{W}) := \tau^j(\boldsymbol{W};\boldsymbol{\psi}^*)\mathbf{1}\{D_i = 1\}.$$

Let $A_i := a_{\boldsymbol{\psi}^*}(\boldsymbol{W}_i)$ and $B_i := b_{\boldsymbol{\psi}^*}(\boldsymbol{W}_i)$ with $\mu_A := \mathbb{E}[A_i]$ and $\mu_B := \mathbb{E}[B_i] > 0$. Write $\hat{m}_{1,n}^{j,\boldsymbol{x}_{T_0}}(\boldsymbol{\psi}^*) = g(\bar{A}_n, \bar{B}_n)$ with $g(a,b) = a/b$ and $\bar{A}_n = n^{-1}\sum_{i=1}^n A_i$, $\bar{B}_n = n^{-1}\sum_{i=1}^n B_i$. By the multivariate CLT and a first-order delta method,

$$\sqrt{n}\big(\hat{m}_{1,n}^{j,\boldsymbol{x}_{T_0}}(\boldsymbol{\psi}^*) - m_1^{j,\boldsymbol{x}_{T_0}}(\boldsymbol{\psi}^*)\big) = \nabla g(\mu_A,\mu_B)^\top \sqrt{n}\begin{pmatrix} \bar{A}_n - \mu_A \\ \bar{B}_n - \mu_B \end{pmatrix} + o_p(1).$$

Since $\nabla g(\mu_A,\mu_B) = (1/\mu_B,\ -\mu_A/\mu_B^2)^\top$, this equals

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n \left[\frac{A_i - \mu_A}{\mu_B} - \frac{\mu_A}{\mu_B^2}(B_i - \mu_B)\right] + o_p(1) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \phi_1^{j,\boldsymbol{x}_{T_0}}(\boldsymbol{W}_i;\boldsymbol{\psi}^*) + o_p(1),$$

where, using $m_1^{j,\boldsymbol{x}_{T_0}}(\boldsymbol{\psi}^*) = \mu_A/\mu_B$,

$$\phi_1^{j,\boldsymbol{x}_{T_0}}(\boldsymbol{W};\boldsymbol{\psi}^*) = \frac{\tau^j(\boldsymbol{W};\boldsymbol{\psi}^*)}{\mathbb{E}[\mathbf{1}\{D=1\}\tau^j(\boldsymbol{W};\boldsymbol{\psi}^*)]}\Big[\mathbf{1}\{\boldsymbol{X}_{T_0} = \boldsymbol{x}_{T_0}, D=1\} - \mathbf{1}\{D=1\}m_1^{j,\boldsymbol{x}_{T_0}}(\boldsymbol{\psi}^*)\Big].$$

By analogous arguments,

$$\phi_2^{j,\boldsymbol{x}_{T_0},d}(\boldsymbol{W};\boldsymbol{\psi}^*) = \frac{\tau^j(\boldsymbol{W};\boldsymbol{\psi}^*)\mathbf{1}\{\boldsymbol{X}_{T_0}=\boldsymbol{x}_{T_0}, D=d\}}{\mathbb{E}[\tau^j(\boldsymbol{W};\boldsymbol{\psi}^*)\mathbf{1}\{\boldsymbol{X}_{T_0}=\boldsymbol{x}_{T_0}, D=d\}]}\big[\boldsymbol{X}_t - m_2^{j,\boldsymbol{x}_{T_0},d}(\boldsymbol{\psi}^*)\big],$$
$$\phi_3^j(\boldsymbol{W};\boldsymbol{\psi}^*) = \frac{\mathbf{1}\{D=1\}}{\Pr(D=1)}\Big(\tau^j(\boldsymbol{W};\boldsymbol{\psi}^*) - \mathbb{E}[\tau^j(\boldsymbol{W};\boldsymbol{\psi}^*) \mid D=1]\Big).$$

Stacking all such elements into a vector $\boldsymbol{\phi}(\boldsymbol{W};\boldsymbol{\psi}^*)$ gives

$$(I) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \boldsymbol{\phi}(\boldsymbol{W}_i;\boldsymbol{\psi}^*) + o_p(1), \quad \mathbb{E}[\boldsymbol{\phi}(\boldsymbol{W};\boldsymbol{\psi}^*)] = \mathbf{0}. \tag{50}$$

Therefore, combining (48)–(50), we obtain the asymptotic linear representation:

$$\sqrt{n}\big(\hat{\boldsymbol{m}}(\hat{\boldsymbol{\psi}}) - \boldsymbol{m}(\boldsymbol{\psi}^*)\big) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \Big(\boldsymbol{\phi}(\boldsymbol{W}_i;\boldsymbol{\psi}^*) + \boldsymbol{A}^*\,\boldsymbol{S}(\boldsymbol{W}_i;\boldsymbol{\psi}^*)\Big) + o_p(1).$$

Let $\boldsymbol{h}(\cdot)$ denote the continuously differentiable function that maps the moment vector $\boldsymbol{m}$ to the vector of LTATT and ATT parameters, $\boldsymbol{\theta}^* = \boldsymbol{h}(\boldsymbol{m}(\boldsymbol{\psi}^*))$, as defined by (23)–(26). Denote its Jacobian by $\nabla_{\boldsymbol{m}}\boldsymbol{h}(\boldsymbol{m}) := \partial \boldsymbol{h}(\boldsymbol{m})/\partial \boldsymbol{m}$. Then, by the multivariate delta method,

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \boldsymbol{\eta}(\boldsymbol{W}_i) + o_p(1) \;\Rightarrow\; \mathcal{N}(0, \boldsymbol{V}),$$

where the influence function and asymptotic covariance matrix are given by

$$\boldsymbol{\eta}(\boldsymbol{W}_i) := (\nabla_{\boldsymbol{m}}\boldsymbol{h}(\boldsymbol{m}(\boldsymbol{\psi}^*)))^\top\,\big[\boldsymbol{\phi}(\boldsymbol{W}_i;\boldsymbol{\psi}^*) + \boldsymbol{A}^*\,\boldsymbol{S}(\boldsymbol{W}_i;\boldsymbol{\psi}^*)\big], \qquad \boldsymbol{V} = \mathbb{E}[\boldsymbol{\eta}(\boldsymbol{W}_i)\boldsymbol{\eta}(\boldsymbol{W}_i)^\top].$$

This establishes part (b). □

B Extension to Staggered Treatment Adoption

Our identification strategy can be extended to staggered treatment adoption. First, we replace a treatment group D_i with a treatment *cohort*: the period in which treatment is given, $G_i \in \mathcal{G}$ with $\mathcal{G} \subset \mathcal{T}$, defined by $G_i = \min\{t \mid D_{it} = 1\}$ for treated units. We set $G_i = 0$ if the unit is never treated and we assume there is no unit who is treated in the initial period of 0.

We adopt the dynamic potential outcome framework with multi-stage treatment adoption as in Callaway and Sant'Anna (2021) and Sun and Abraham (2021). Let $Y_{it}(0)$ denote the untreated potential outcome of unit i in period t if she is never treated throughout $\mathcal{T}$. For treated potential outcomes for staggered adoption in period $g \in \mathcal{G}\setminus\{0\}$, we use $Y_{it}(g)$ to denote the potential outcome of unit i at period t if she is treated for the first time at period g.

For brevity, we focus on the binary outcome case where $Y_{it}(g) \in \mathcal{Y} := \{0, 1\}$. By extending

Callaway and Sant'Anna (2021) to our transition independence setup, we suggest never-treated or (and) not-yet-treated units as controls, as follows:

Assumption 1 (Transition independence with a never-treated group)**.** For each $g \in \mathcal{G}$ and $t \geq g$, the following hold:

$$\begin{aligned}&\Pr(Y_{it}(0) = y_t \mid Y_{i,g-1}(0) = y_{g-1}, ..., Y_{i0}(0) = y_0, G_i = g) \\ &= \Pr(Y_{it}(0) = y_t \mid Y_{i,g-1}(0) = y_{g-1}, ..., Y_{i0}(0) = y_0, G_i = 0),\end{aligned} \tag{51}$$

for all $\{y_{t'-1}\}_{t'=1}^{g} \in \mathcal{Y}^g$.

Assumption 2 (Transition independence with not-yet-treated groups)**.** For each $g \in \mathcal{G}$ and $t \geq g$, the following hold:

$$\begin{aligned}&\Pr(Y_{it}(0) = y_t \mid Y_{i,g-1}(0) = y_{g-1}, ..., Y_{i0}(0) = y_0, G_i = g) \\ &= \Pr(Y_{it}(0) = y_t \mid Y_{i,g-1}(0) = y_{g-1}, ..., Y_{i0}(0) = y_0, G_i > g, G_i \neq 0),\end{aligned} \tag{52}$$

for all $\{y_{t'-1}\}_{t'=1}^{g} \in \mathcal{Y}^g$.

Note that these two assumptions are distinct, and one can use either or both of them depending on the research question to choose proper controls. We can also extend the no anticipatory effects assumption (Assumption NA) and the overlap assumption (Assumption CS) to the staggered treatment adoption setting as follows:

Assumption 3 (No anticipatory effects for staggered treatment adoption)**.** For all $g \in \mathcal{G}$, $Y_{it}(g) = Y_{it}(0)$ holds for all units with $G_i = g$ and $t < g$.

Assumption 4 (Overlap for staggered treatment adoption)**.** There exists a positive constant $\epsilon > 0$ such that $\epsilon < \Pr(G_i = g | Y_{i,g-1} = y_{g-1}, \ldots, Y_{i0} = y_1) < 1 - \epsilon$ for all $\{y_{t'-1}\}_{t'=1}^{g} \in \mathcal{Y}^g$ and $g \in \mathcal{G}$.

Let $\overline{G}$ denote the latest treatment period in which control units are available, i.e.,

$$\overline{G} = \begin{cases} T & \text{if } 0 \in \mathcal{G} \\ \max\{g - 1 \mid g \in \mathcal{G}\} & \text{if } 0 \notin \mathcal{G}. \end{cases}$$

Then one can define the following notion of cohort-specific ATT for the staggered treatment adoption setting for all g and $g \leq t \leq \overline{G}$:

$$ATT_{g,t} = \mathbb{E}\left[Y_{it}(g) - Y_{it}(0) \mid G_i = g\right],$$

One can then extend Proposition 1 to the staggered treatment adoption setting by replacing $T_0 + 1$ with G_i and T with $\overline{G}$, as follows:

Proposition 24. *Suppose that either (or both) of Assumptions 1 or 2 holds, and Assumptions 3-4 hold. Let $\mathcal{G}(g,t)$ denote the set of treatment cohorts for controls against a treatment cohort g at post-treatment period t for the cohort g, i.e.,*

$$\mathcal{G}(g,t) \subset \underbrace{\{0\}}_{\textit{never-treated group}} \cup \underbrace{\{g' \in \mathcal{G} \mid g' > t\}}_{\textit{not-yet-treated groups}},$$

such that $\mathcal{G}(g,t)$ is set to be the former group $\{0\}$ if Assumption 1 holds and the latter group $\{g' \in \mathcal{G} \mid g' > t\}$ if Assumption 2 holds, and the union of the two groups if both assumptions hold. Then, for all $g \in \mathcal{G} \setminus \{0\}$ and $g \leq t \leq \overline{G}$, $ATT_{g,t}$ is identified by

$$ATT_{g,t} = \mathbb{E}\left[Y_{it}(g) - \mathbb{E}\left[Y_{it} \mid \{Y_{is}\}_{s=0}^{g-1}, G \in \mathcal{G}(g,t)\right] \middle| G_i = g\right],$$

where

$$\begin{aligned}
&\mathbb{E}\left[Y_{it} \mid \{Y_{is}\}_{s=0}^{g-1}, G \in \mathcal{G}(g,t)\right] \\
&= \sum_{y_t \in \mathcal{Y}} y_t \sum_{g' \in \mathcal{G}(g,t)} \Pr(G_i = g' \mid G \in \mathcal{G}(g,t)) p_{Y_t|\{Y_s\}_{s=0}^{g-1}, G_i = g'}(y_t|\{Y_{is}\}_{s=0}^{g-1}),
\end{aligned}$$

with $p_{Y_t|\{Y_s\}_{s=0}^{g-1}, G_i = g'}(y_t|\{Y_{is}\}_{s=0}^{g-1}) = \Pr(Y_{it} = y_t|Y_{ig-1}(0) = y_{g-1}, ..., Y_{i0}(0) = y_0, G_i = g')$ for $g' \in \mathcal{G}(g,t)$.

There are several ways to aggregate the cohort-specific ATTs to derive the overall ATT. One approach is to use the average of the cohort-specific ATTs, weighted by the cohort distribution, as follows:

$$ATT_t = \sum_{g \in \mathcal{G} \setminus \{0\}; g \leq t} \Pr(G = g \mid G \neq 0, g \leq t) ATT_{g,t},$$

where the weights are given by the cohort distribution $\Pr(G = g \mid G \neq 0, g \leq t)$. The choice of aggregation method depends on the research question and the context of the data. For further examples of alternative aggregation of cohort-specific ATTs and discussions on their corresponding weights for estimation, see Callaway and Sant'Anna (2021).

C Estimation Procedure and Weighted Bootstrap

Estimation Procedure. We provide below a detailed description of the two-stage estimator presented in the main text. In the first stage, we estimate the latent transition probabilities using the EM algorithm; in the second stage, we compute the counterfactual expected untreated potential outcome for each latent type to derive the average treatment effect on the treated for post-treatment observations as follows:

0. Initiate the process with a given starting value, $\hat{\boldsymbol{\psi}}^{(0)} = (\hat{\boldsymbol{\pi}}^{(0)}, \hat{\boldsymbol{\varphi}}^{1(0)}, ..., \hat{\boldsymbol{\varphi}}^{J(0)}) \in \Theta_{\boldsymbol{\psi}}$.

1. Estimate ψ by employing the following EM algorithm, starting with $s = 1$:

   (a) (E-step) Compute the posterior probabilities from $\hat{\tau}_i^{j(s)}$ for each i and j:

$$\hat{\tau}_i^{j(s)} = \frac{\hat{\pi}^{j(s-1)} p^j_{\boldsymbol{W}}(\boldsymbol{W}_i; \hat{\boldsymbol{\varphi}}^{j(s-1)})}{\sum_{k=1}^J \hat{\pi}^{k(s-1)} p^k_{\boldsymbol{W}}(\boldsymbol{W}_i; \hat{\boldsymbol{\varphi}}^{k(s-1)})}.$$

   (b) (M-step) Update the mixture weights $\hat{\boldsymbol{\pi}}^{(s)} = (\pi^{1(s)}, \dots, \pi^{J(s)})^\top$ and type-specific transition probabilities

$$\hat{\boldsymbol{\varphi}}^{j(s)} = \big\{\{\hat{p}^{j(s)}_{\boldsymbol{X}_1(0),D}(\cdot,\cdot),\ \hat{p}^{j(s)}_{\boldsymbol{X}_t(0)|\boldsymbol{X}_{t-1}(0)}(\cdot|\cdot)\}_{t=2}^T,\ \{\hat{p}^{j(s)}_{\boldsymbol{X}_t(1)|\boldsymbol{X}_{t-1}(1)}(\cdot|\cdot)\}_{t=T_0+1}^T\big\}$$

   for each latent type j as

$$\hat{\pi}^{j(s)} = \frac{1}{n}\sum_{i=1}^n \hat{\tau}_i^{j(s)},$$

$$\hat{p}^{j(s)}_{\boldsymbol{X}_1(0),D}(\boldsymbol{x}_1, d) = \frac{\sum_{i=1}^n \mathbf{1}\{\boldsymbol{X}_{i1} = \boldsymbol{x}_1, D_i = d\}\hat{\tau}_i^{j(s)}}{\sum_{i=1}^n \hat{\tau}_i^{j(s)}},$$

$$\hat{p}^{j(s)}_{\boldsymbol{X}_t(0)|\boldsymbol{X}_{t-1}(0)}(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) = \begin{cases} \frac{\sum_{i=1}^n \mathbf{1}\{\boldsymbol{X}_{it}=\boldsymbol{x}_t, \boldsymbol{X}_{it-1}=\boldsymbol{x}_{t-1}\}\hat{\tau}_i^{j(s)}}{\sum_{i=1}^n \mathbf{1}\{\boldsymbol{X}_{it-1}=\boldsymbol{x}_{t-1}\}\hat{\tau}_i^{j(s)}} & \text{for } t = 2, ..., T_0, \\ \frac{\sum_{i=1}^n \mathbf{1}\{\boldsymbol{X}_{it}=\boldsymbol{x}_t, \boldsymbol{X}_{it-1}=\boldsymbol{x}_{t-1}, D_i=0\}\hat{\tau}_i^{j(s)}}{\sum_{i=1}^n \mathbf{1}\{\boldsymbol{X}_{it-1}=\boldsymbol{x}_{t-1}, D_i=0\}\hat{\tau}_i^{j(s)}} & \text{for } t = T_0+1, ..., T, \end{cases}$$

$$\hat{p}^{j(s)}_{\boldsymbol{X}_t(1)|\boldsymbol{X}_{t-1}(1)}(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) = \frac{\sum_{i=1}^n \mathbf{1}\{\boldsymbol{X}_{it} = \boldsymbol{x}_t, \boldsymbol{X}_{it-1} = \boldsymbol{x}_{t-1}, D_i = 1\}\hat{\tau}_i^{j(s)}}{\sum_{i=1}^n \mathbf{1}\{\boldsymbol{X}_{it-1} = \boldsymbol{x}_{t-1}, D_i = 1\}\hat{\tau}_i^{j(s)}} \quad \text{for } t = T_0+1, \dots, T.$$

   (c) Set $s = s + 1$ and repeat until convergence.

2. Given the first-step estimate, we estimate LTATTs and ATTs as in (23) and (26), respectively.

Weighted Bootstrap. To consistently estimate the asymptotic covariance matrix $\boldsymbol{V}$, we employ a nonparametric weighted bootstrap. Specifically, we generate B bootstrap replications $\{\hat{\boldsymbol{\theta}}^{(b)}\}_{b=1}^B$ by reweighting each observation's contribution in the estimation process as follows:

1. *Generate weights.* For each bootstrap replication $b = 1, \dots, B$, draw n i.i.d. positive weights $\{\zeta_i^{(b)}\}_{i=1}^n$ from the standard exponential distribution, Exp(1).

2. *Weighted MLE.* Re-estimate the parameter vector $\boldsymbol{\psi}$ by maximizing the weighted log-likelihood function

$$\sum_{i=1}^n \zeta_i^{(b)} \log p_{\boldsymbol{W}}(\boldsymbol{W}_i; \boldsymbol{\psi}),$$

implementing the weights $\zeta_i^{(b)}$ within the Expectation–Maximization (EM) algorithm.

3. *Labeling correction.* After each weighted MLE, re-order the components of $\hat{\boldsymbol{\psi}}^{(b)}$ according to the identification constraint in Assumption RSM(c), ensuring consistent labeling of latent types across bootstrap replications.

4. *Weighted sample analogue.* Compute the bootstrap estimate $\hat{\boldsymbol{\theta}}^{(b)}$ by replacing all summations in the sample-analogue estimators (23), (24), and (25) with their weighted counterparts, incorporating both the random weights $\zeta_i^{(b)}$ and the estimated posterior probabilities $\hat{\tau}_i^{j(b)}$ obtained from $\hat{\boldsymbol{\psi}}^{(b)}$.

5. *Covariance estimation.* Estimate the asymptotic covariance matrix $\boldsymbol{V}$ as

$$\hat{\boldsymbol{V}} = \frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{\boldsymbol{\theta}}^{(b)} - \bar{\boldsymbol{\theta}}\right)\left(\hat{\boldsymbol{\theta}}^{(b)} - \bar{\boldsymbol{\theta}}\right)^{\top}, \qquad \bar{\boldsymbol{\theta}} = \frac{1}{B}\sum_{b=1}^{B}\hat{\boldsymbol{\theta}}^{(b)}.$$

Cluster-Level Weighting. When the data exhibit cluster-level dependence (e.g., individuals within states, or firms within industries), the bootstrap weights are drawn at the cluster level rather than the individual level. Let C denote the number of clusters. For each bootstrap replication b, we draw independent exponential weights $\{w_c^{(b)}\}_{c=1}^{C}$ with $w_c^{(b)} \sim \text{Exp}(1)$, normalize them to $\tilde{w}_c^{(b)} = w_c^{(b)} / \sum_{c'=1}^{C} w_{c'}^{(b)}$, and assign each unit i in cluster c the weight $\zeta_i^{(b)} = \tilde{w}_c^{(b)}$. This weighted bootstrap procedure at the cluster level follows Cameron et al. (2008) and MacKinnon and Webb (2017), preserving the within-cluster correlation structure while allowing for inference that is robust to cluster-level heterogeneity.

Uniform Confidence Intervals. We construct uniform (simultaneous) confidence intervals for sequences of treatment effect estimates $\{\hat{\theta}_t\}_{t=1}^{T}$ using a studentized bootstrap procedure. Unlike pointwise confidence intervals that control coverage at each individual time period, uniform confidence intervals provide simultaneous coverage across all periods:

$$\Pr\left(\theta_t \in \text{CI}_t \text{ for all } t = 1, \ldots, T\right) \geq 1 - \alpha.$$

Let $\hat{\theta}_t$ denote the point estimate for period t. Given B bootstrap estimates $\{\hat{\theta}_t^{(b)}\}_{b=1}^{B}$ from the weighted bootstrap procedure:

1. Compute the bootstrap standard error for each period:

$$\hat{\sigma}_t = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{\theta}_t^{(b)} - \bar{\theta}_t\right)^2},$$

where $\bar{\theta}_t = B^{-1}\sum_{b=1}^{B}\hat{\theta}_t^{(b)}$.

2. For each bootstrap sample b and period t, compute the studentized statistic:

$$t_t^{(b)} = \frac{\hat{\theta}_t^{(b)} - \hat{\theta}_t}{\hat{\sigma}_t}.$$

3. For each bootstrap sample, compute the supremum of the absolute studentized statistics:

$$M^{(b)} = \max_{t \in \{1,\ldots,T\}} |t_t^{(b)}|.$$

4. Obtain the critical value $c_{1-\alpha}$ as the $(1-\alpha)$-quantile of $\{M^{(1)}, \ldots, M^{(B)}\}$.

5. Construct the uniform confidence interval for each period t:

$$\left[\hat{\theta}_t - c_{1-\alpha} \cdot \hat{\sigma}_t, \quad \hat{\theta}_t + c_{1-\alpha} \cdot \hat{\sigma}_t\right].$$

The key distinction from pointwise intervals is that uniform intervals use a single critical value $c_{1-\alpha}$ derived from the joint distribution of the supremum statistic, rather than period-specific quantiles. This accounts for the multiple testing problem inherent in examining effects across many time periods. When the uniform confidence band lies entirely above (or below) zero for a range of periods, we can conclude that the treatment effect is significantly positive (or negative) for all those periods simultaneously at the $(1-\alpha)$ confidence level.

Computational details. In the empirical applications, the EM algorithm is initialized using a two-stage multistart procedure: a large number of random starting values are evaluated for a small number of iterations, and the top candidates are then run to convergence. Specifically, we use 6,000 short-run initializations narrowed to 20 long-run candidates, with a convergence tolerance of 10^{-3} and a maximum of 100 EM iterations. Standard errors are obtained via the weighted bootstrap with $B = 500$ replications. For the ADA application, bootstrap weights are drawn at the state level to account for within-state correlation; for the remaining applications, weights are drawn at the unit level.

Since the EM algorithm is guaranteed to converge only to a local maximum of the likelihood, the multistart procedure is essential: by evaluating many initial points and selecting the best candidates for full convergence, it provides a practical safeguard against convergence to inferior local optima. While global optimality cannot be guaranteed, the large number of initializations (6,000) makes it unlikely that the global maximum is missed in practice.

D Construction of Type-Specific Transition Probabilities

Testing for transition independence across pre-treatment periods (Theorem 21) under the presence of latent heterogeneity in transition dynamics requires estimating type-specific transition probabilities for both treated and control units during pre-treatment periods.

In this section, we outline the procedure for constructing the type-specific transition probabilities across pre-treatment periods for each treatment status and latent type. First, we estimate the parameters of the latent models using the observed outcomes and treatment status via the EM algorithm described in Appendix C. Next, we compute the type-specific probabilities for each unit using the estimated parameters. Lastly, we estimate the type-specific transition probabilities for each latent type by weighting pre-treatment observations using the estimated type-specific probabilities. Specifically, the procedure is as follows:

1. Estimate the model parameters for each latent type $\hat{\boldsymbol{\psi}}$ via the EM algorithm described in Appendix C.

2. Obtain the estimated type-specific probabilities $\hat{\tau}_i^j$ for each unit i and latent type j from (22) from the estimated parameters $\hat{\boldsymbol{\psi}}$.

3. For each latent type j and treatment status $d \in \{0, 1\}$, estimate the type-specific transition probabilities across pre-treatment periods $t = 2, \ldots, T_0$ as follows:

$$\hat{p}^j_{\boldsymbol{X}_t(0)|\boldsymbol{X}_{t-1}(0),D}(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}, d) = \frac{\sum_{i=1}^n \mathbf{1}\{\boldsymbol{X}_{it} = \boldsymbol{x}_t, \boldsymbol{X}_{it-1} = \boldsymbol{x}_{t-1}, D_i = d\}\hat{\tau}_{ij}}{\sum_{i=1}^n \mathbf{1}\{\boldsymbol{X}_{it-1} = \mathbf{x}_{t-1}, D_i = d\}\hat{\tau}_{ij}}.$$

E Type-Specific Flow Decomposition

Corollary 25 (Type-specific flow decomposition)**.** *Under Assumption TI-LH with one lag, the LTATT on the kth outcome for latent type j decomposes as*

$$\begin{aligned}
&\mathbb{E}\big[X_t^{(k)}(1) - X_t^{(k)}(0) \mid D = 1, Z = j\big] \\
&= \sum_{y \neq \bar{y}^{(k)}} \underbrace{\Big\{\Pr(Y_t = \bar{y}^{(k)} \mid Y_{T_0} = y, D = 1, Z = j) - \Pr(Y_t = \bar{y}^{(k)} \mid Y_{T_0} = y, D = 0, Z = j)\Big\}}_{\textit{inflow CATT from } y \textit{ for type } j} \\
&\qquad\qquad \times\ \Pr(Y_{T_0} = y \mid D = 1, Z = j) \qquad (53)\\
&\quad - \sum_{y \neq \bar{y}^{(k)}} \underbrace{\Big\{\Pr(Y_t = y \mid Y_{T_0} = \bar{y}^{(k)}, D = 1, Z = j) - \Pr(Y_t = y \mid Y_{T_0} = \bar{y}^{(k)}, D = 0, Z = j)\Big\}}_{\textit{outflow CATT to } y \textit{ for type } j} \\
&\qquad\qquad \times\ \Pr(Y_{T_0} = \bar{y}^{(k)} \mid D = 1, Z = j).
\end{aligned}$$

The aggregate ATT on the kth outcome is recovered by aggregating over types:

$$\mathbb{E}\big[X_t^{(k)}(1) - X_t^{(k)}(0) \mid D = 1\big] = \sum_{j=1}^{J} \Pr(Z = j \mid D = 1)\ \mathbb{E}\big[X_t^{(k)}(1) - X_t^{(k)}(0) \mid D = 1, Z = j\big]. \quad (54)$$

Proof. Fix type j and outcome category k. By Proposition 17, the type-specific ATT on the kth outcome is

$$\mathbb{E}\big[X_t^{(k)}(1) - X_t^{(k)}(0) \mid D=1, Z=j\big] = \sum_{\boldsymbol{x}_{T_0}} \Pr(\boldsymbol{X}_{T_0} = \boldsymbol{x}_{T_0} \mid D=1, Z=j) \cdot \Delta_t^{j,k}(\boldsymbol{x}_{T_0}),$$

where $\Delta_t^{j,k}(\boldsymbol{x}_{T_0}) := \mathbb{E}[X_t^{(k)} \mid \boldsymbol{X}_{T_0} = \boldsymbol{x}_{T_0}, D=1, Z=j] - \mathbb{E}[X_t^{(k)} \mid \boldsymbol{X}_{T_0} = \boldsymbol{x}_{T_0}, D=0, Z=j]$. Under the first-order Markov assumption, conditioning on $\boldsymbol{X}_{T_0}$ reduces to Y_{T_0}. Partitioning the summation over Y_{T_0} into $\{Y_{T_0} = \bar{y}^{(k)}\}$ (outflows) and $\{Y_{T_0} \neq \bar{y}^{(k)}\}$ (inflows) yields the stated decomposition. Equation (54) follows from the law of iterated expectations applied to $\mathbb{E}[\cdot \mid D = 1] = \sum_j \Pr(Z=j \mid D=1)\,\mathbb{E}[\cdot \mid D=1, Z=j]$. □

F Reference Estimators Used in the Empirical Comparisons

The empirical figures compare our transition-based estimator with three reference estimators: Wooldridge nonlinear DiD with logit and probit links, and Boussim's compositional DiD (CoDiD). This section summarizes the estimators as implemented in the empirical exercises. Throughout, $D=1$ denotes the treated group, $D=0$ denotes the control group, and T_0 denotes the last pre-treatment period for the single treated cohort in each application.

F.1 Wooldridge Nonlinear DiD

Following Wooldridge (2023), the Wooldridge comparisons impose parallel trends after transforming a marginal conditional mean by a known link function. In our applications, the target outcome is the binary margin

$$R_t(\mathcal{A}) = \mathbb{1}\{Y_t \in \mathcal{A}\},$$

where $\mathcal{A}$ is the empirical outcome set of interest, such as receiving a complaint, filing a patent, or being employed. Let $\hat{p}_{d,t}$ denote the sample mean of $R_t(\mathcal{A})$ for group $D=d$ at period t, after applying the same sample restrictions used for the empirical analysis. For a link transformation $h \in \{\Lambda^{-1}, \Phi^{-1}\}$, corresponding to the logit and probit links, the implemented untreated counterfactual mean for treated units in post-treatment period $t > T_0$ is

$$\hat{m}_t^{W,h} = h^{-1}\{h(\hat{p}_{1,T_0}) + h(\hat{p}_{0,t}) - h(\hat{p}_{0,T_0})\}. \tag{55}$$

The corresponding ATT estimate is

$$\widehat{ATT}_t^{W,h} = \hat{p}_{1,t} - \hat{m}_t^{W,h}.$$

The implementation uses exactly one treated cohort and sets T_0 to the period immediately before treatment begins. The logit and probit versions require the treated pre-treatment, control pre-treatment, and control post-treatment cell means entering (55) to lie strictly inside $(0,1)$; no continuity correction is applied. Bootstrap confidence intervals are computed by resampling the empirical panel using the package implementation. The two plotted labels are "Wooldridge (Logit)" and "Wooldridge (Probit)."

F.2 Boussim CoDiD

The Boussim comparison follows the point-identified compositional DiD estimator in Boussim (2025). Unlike the Wooldridge comparisons, it uses all observed atomic outcome categories before aggregating to the target outcome set shown in the figures. Let $k = 1, \dots, K$ index the atomic categories in the restricted empirical sample, and let $\hat{q}_{d,k,t}$ denote the sample quantity for category k in group $D = d$ at period t. For each post-treatment period $t > T_0$, the implemented category-level counterfactual quantity is

$$\hat{q}^B_{k,t} = \frac{\hat{q}_{1,k,T_0}\hat{q}_{0,k,t}}{\hat{q}_{0,k,T_0}}. \tag{56}$$

These quantities are then normalized to obtain counterfactual shares,

$$\hat{s}^B_{k,t} = \frac{\hat{q}^B_{k,t}}{\sum_{\ell=1}^{K}\hat{q}^B_{\ell,t}}. \tag{57}$$

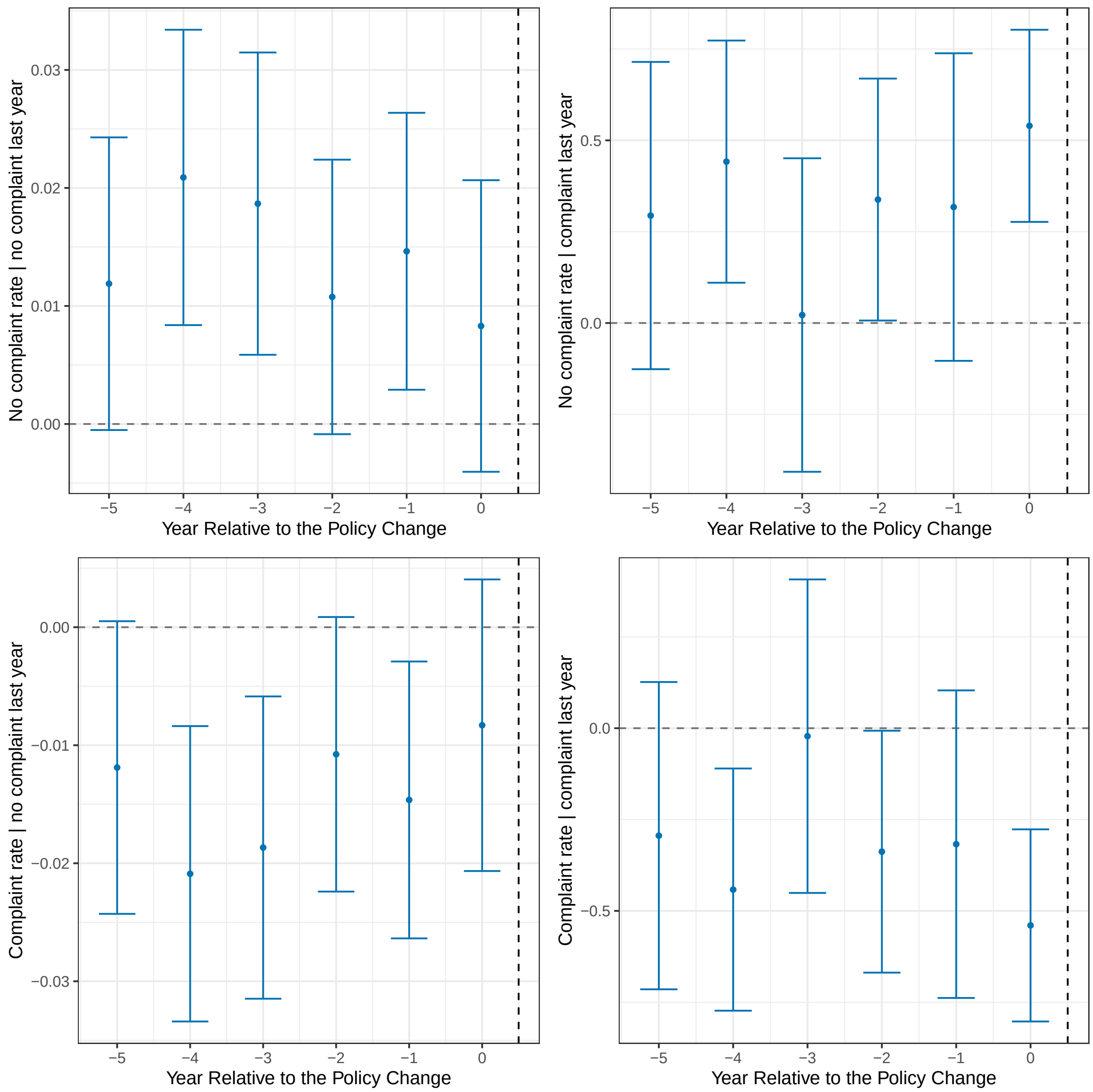
If $\mathcal{A}$ denotes the target set of categories plotted in the empirical application, the reported ATT is

$$\widehat{ATT}^B_t = \sum_{k\in\mathcal{A}}\hat{s}_{1,k,t} - \sum_{k\in\mathcal{A}}\hat{s}^B_{k,t}, \tag{58}$$

where $\hat{s}_{1,k,t}$ is the observed treated share of category k.

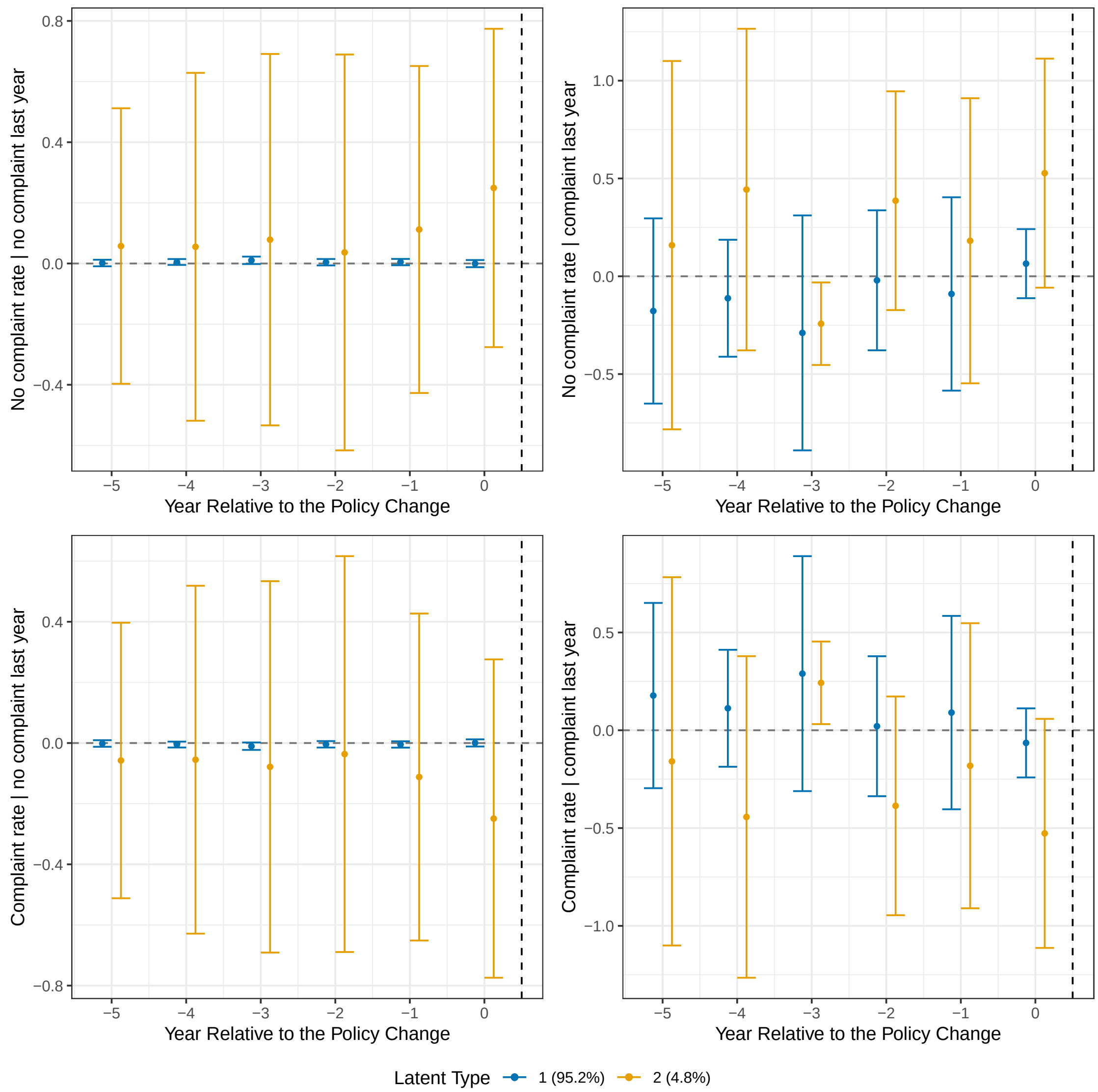
G Supplementary Figures

Figure 15: Differences in Complaint Transition Probabilities Before the Dodd-Frank Act ($J = 1$).



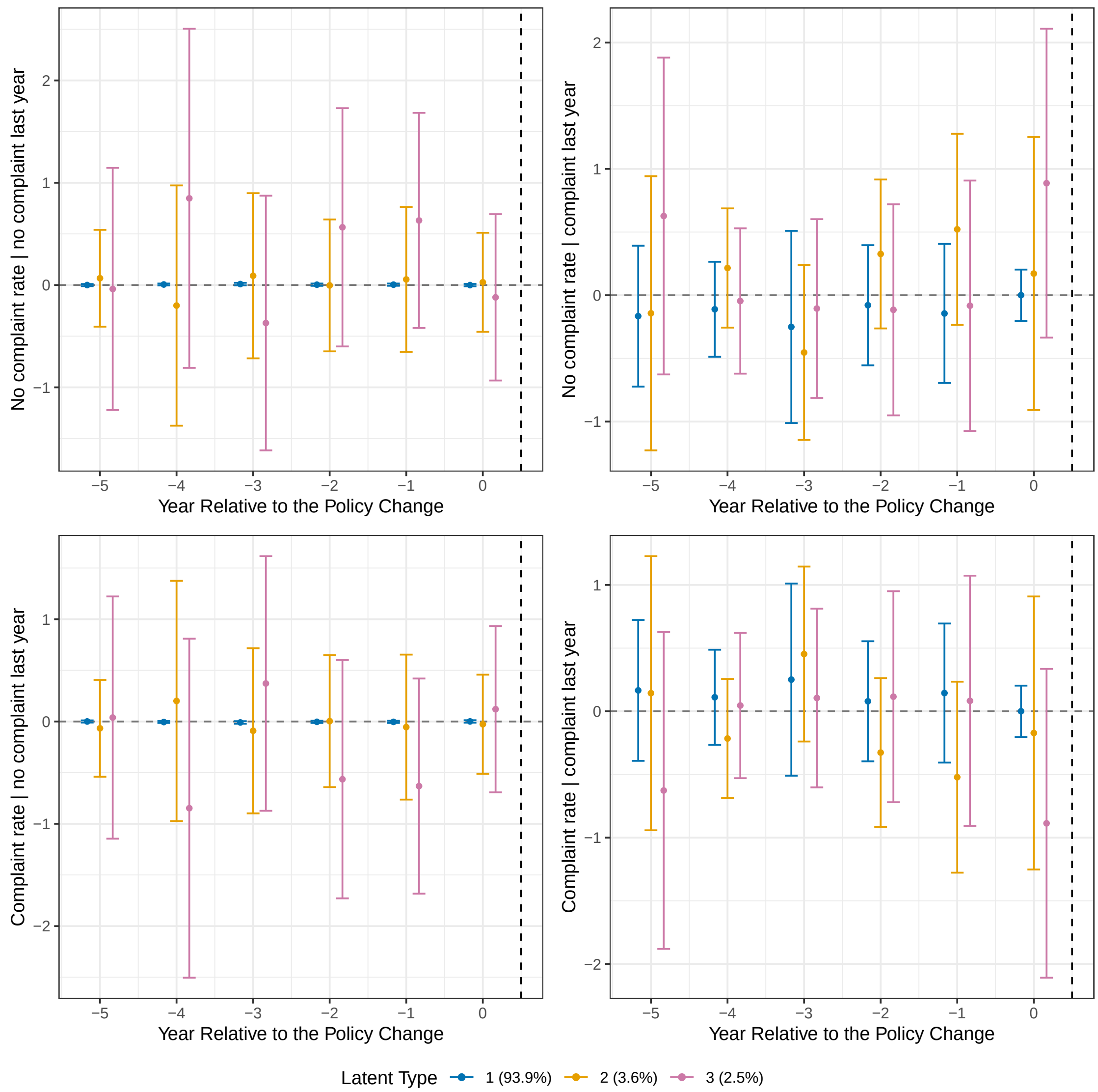
Notes: This figure reports estimated differences in annual complaint transition probabilities between midsize RIAs (treated) and other RIAs (control) before the enactment of the Dodd-Frank Act. Each panel corresponds to a distinct complaint transition, where rates represent the annual probability of moving from one complaint status (No Complaint or Complaint Received) to another. The x-axis measures years relative to the introduction of the Dodd-Frank Act (with zero denoting the last pre-treatment period; treatment begins at period 1), and the y-axis shows the estimated difference in transition probabilities between treated and control groups. Vertical bars represent 95% bootstrap uniform confidence intervals across periods within each transition pair. The dashed vertical line marks the timing of the policy introduction.

Figure 16: Differences in Complaint Transition Probabilities Before the Dodd-Frank Act ($J = 2$).



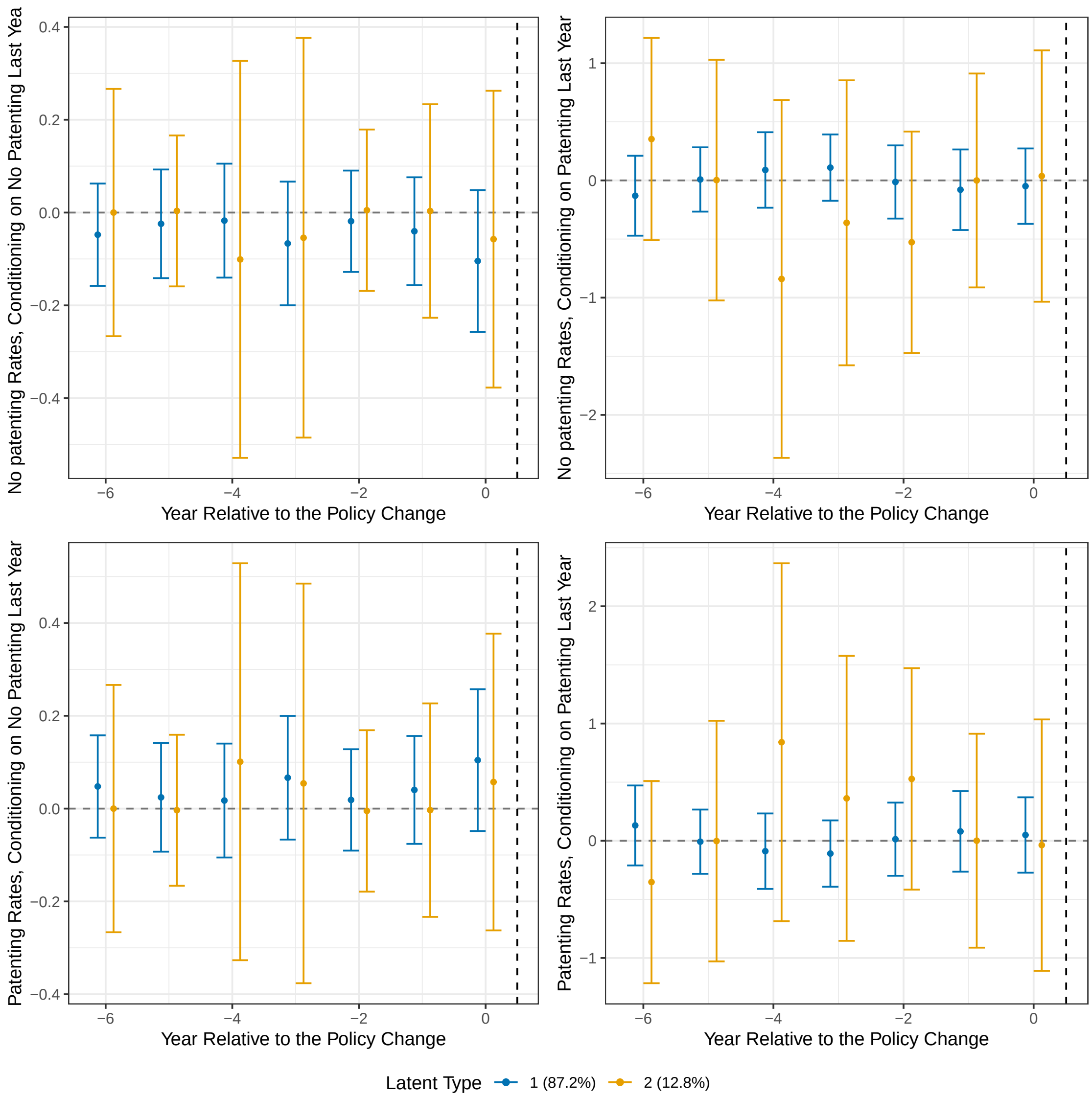
Notes: This figure reports estimated differences in annual complaint transition probabilities between midsize RIAs (treated) and other RIAs (control) before the enactment of the Dodd-Frank Act, for each latent type (shown in different colors). Each panel corresponds to a distinct complaint transition. The x-axis measures years relative to the introduction of the Dodd-Frank Act (with zero denoting the last pre-treatment period; treatment begins at period 1), and the y-axis shows the estimated difference in transition probabilities between treated and control groups. Percentages in the legend indicate the estimated population share of each latent type. Vertical bars represent 95% bootstrap uniform confidence intervals across periods within each transition pair and latent type, $j \in \{1, 2\}$. The dashed vertical line marks the timing of the policy introduction.

Figure 17: Differences in Complaint Transition Probabilities Before the Dodd-Frank Act ($J = 3$).



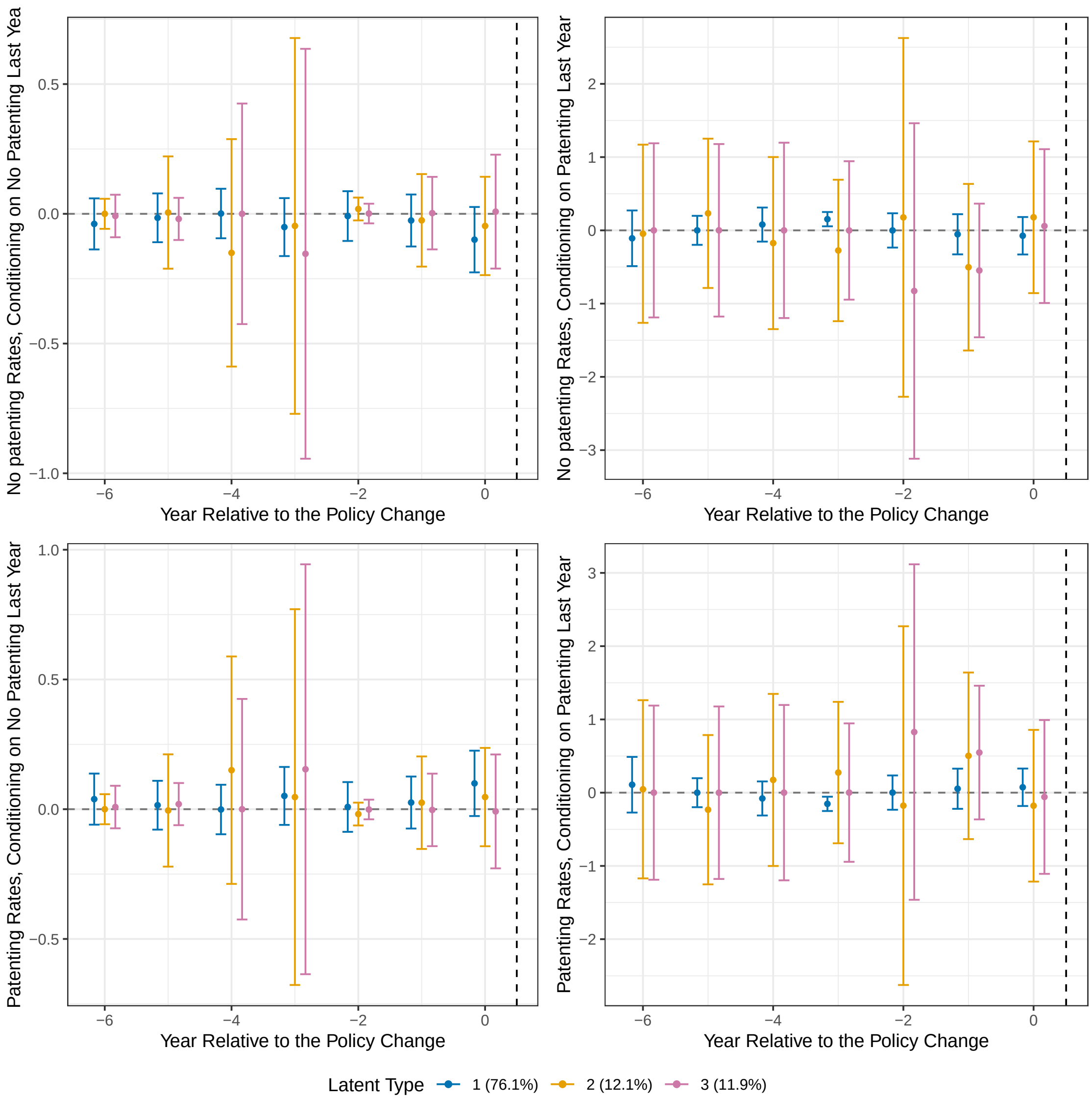
Notes: This figure reports estimated differences in annual complaint transition probabilities between midsize RIAs (treated) and other RIAs (control) before the enactment of the Dodd-Frank Act, for each latent type (shown in different colors). Each panel corresponds to a distinct complaint transition. The x-axis measures years relative to the introduction of the Dodd-Frank Act (with zero denoting the last pre-treatment period; treatment begins at period 1), and the y-axis shows the estimated difference in transition probabilities between treated and control groups. Percentages in the legend indicate the estimated population share of each latent type. Vertical bars represent 95% bootstrap uniform confidence intervals across periods within each transition pair and latent type, $j \in \{1, 2, 3\}$. The dashed vertical line marks the timing of the policy introduction.

Figure 18: Differences in Patenting Transition Probabilities Before the Norwegian Reform ($J = 2$).



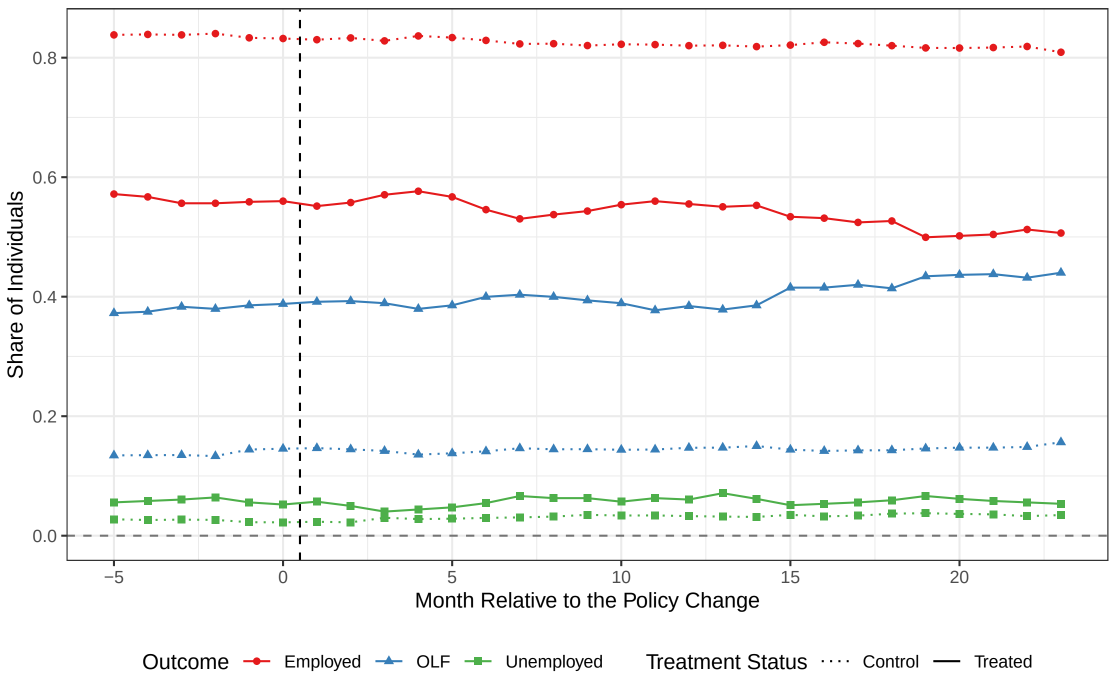
Notes: This figure reports estimated differences in annual patenting transition probabilities between university researchers (treated) and non-university inventors (control) before the Norwegian patent reform, for each latent type (shown in different colors). Each panel corresponds to a distinct patenting transition. The x-axis measures years relative to the introduction of the reform (with zero denoting the last pre-treatment period; treatment begins at period 1), and the y-axis shows the estimated difference in transition probabilities between treated and control groups. Percentages in the legend indicate the estimated population share of each latent type. Vertical bars represent 95% bootstrap uniform confidence intervals across periods within each transition pair and latent type, $j \in \{1, 2\}$. The dashed vertical line marks the timing of the policy introduction.

Figure 19: Differences in Patenting Transition Probabilities Before the Norwegian Reform ($J = 3$).



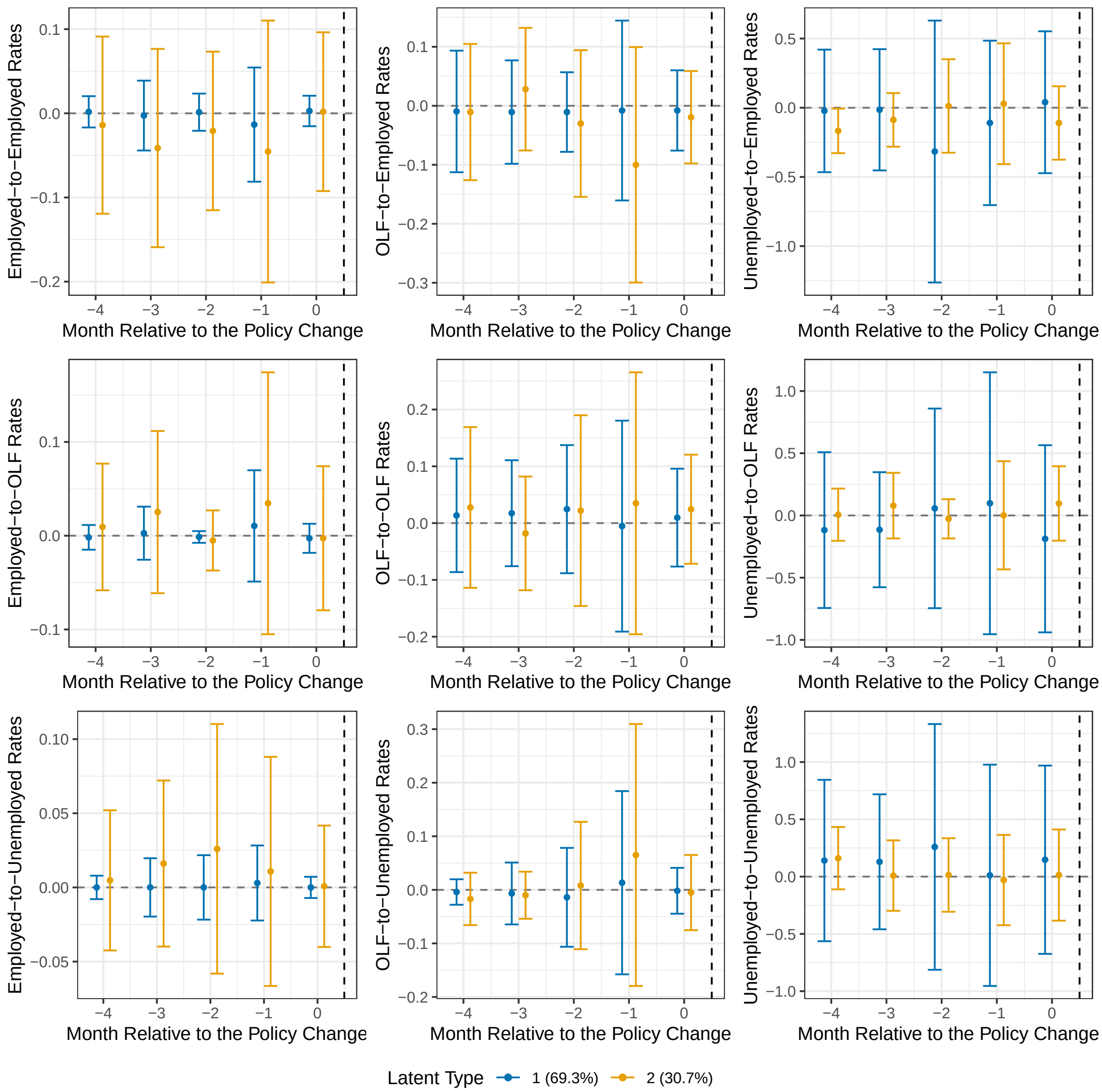
Notes: This figure reports estimated differences in annual patenting transition probabilities between university researchers (treated) and non-university inventors (control) before the Norwegian patent reform, for each latent type (shown in different colors). Each panel corresponds to a distinct patenting transition. The x-axis measures years relative to the introduction of the reform (with zero denoting the last pre-treatment period; treatment begins at period 1), and the y-axis shows the estimated difference in transition probabilities between treated and control groups. Percentages in the legend indicate the estimated population share of each latent type. Vertical bars represent 95% bootstrap uniform confidence intervals across periods within each transition pair and latent type, $j \in \{1, 2, 3\}$. The dashed vertical line marks the timing of the policy introduction.

Figure 20: Labor Force Status Trends Before and After the ADA of 1990.



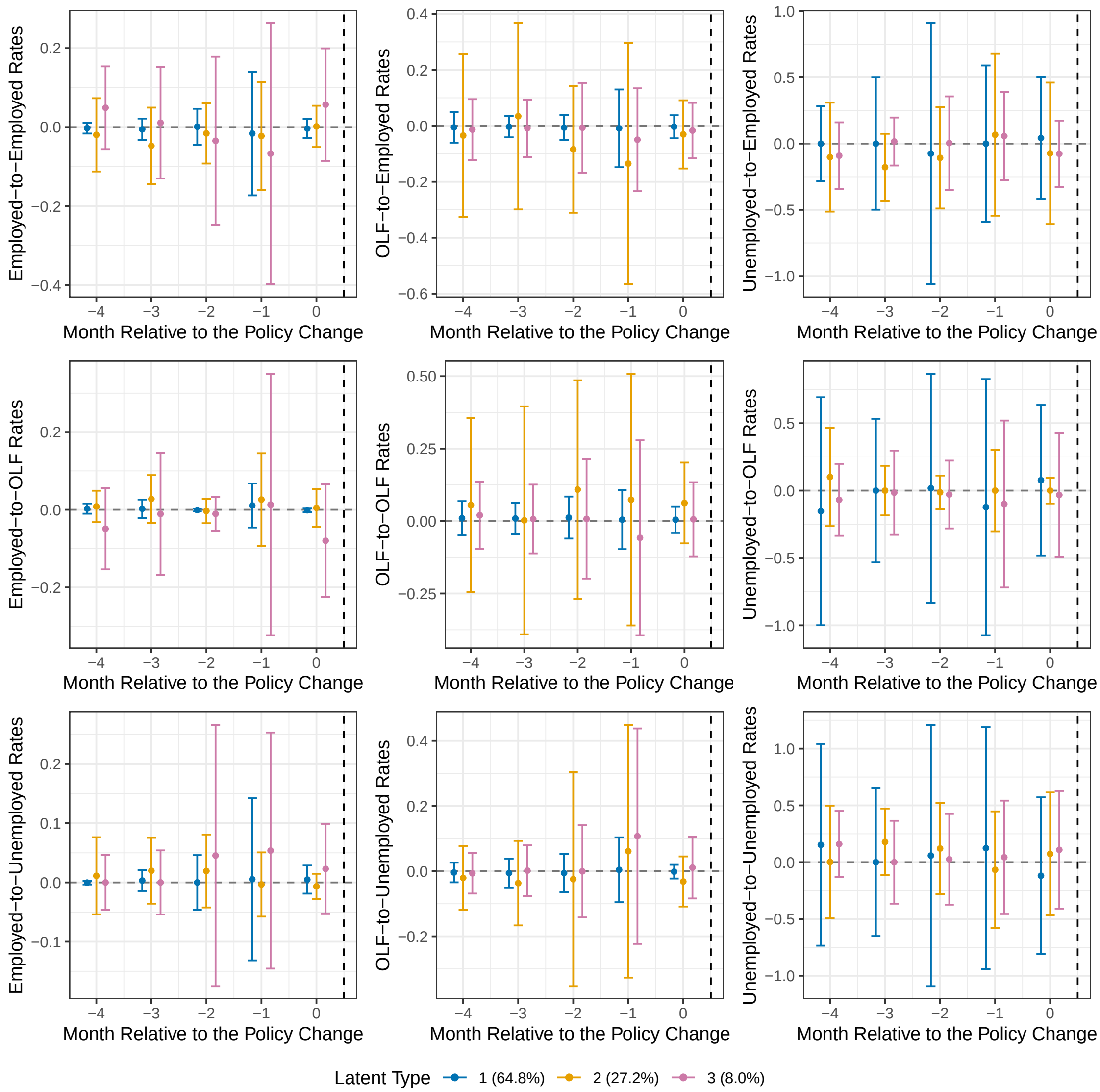
Notes: This figure shows average monthly employment rates for individuals with disabilities (treated group) and those without disabilities (control group) from 1989 to 1991, surrounding the enactment of the ADA. The x-axis reports calendar months, with 1990 marking the implementation year of the ADA. The y-axis shows the fraction of individuals employed. The dashed vertical line indicates the timing of the ADA's introduction.

Figure 21: Differences in Labor Force Transition Probabilities Before the ADA ($J = 2$).



Notes: This figure reports estimated differences in monthly labor force transition probabilities between individuals with and without disabilities before the enactment of the ADA, for each latent type (shown in different colors). Each panel corresponds to a distinct labor force transition, where rates represent the monthly probability of moving from one labor force status (e.g., employed, unemployed, or out of the labor force (OLF)) to another. The x-axis measures calendar months relative to the introduction of the ADA (with zero denoting the policy's introduction), and the y-axis shows the estimated difference in transition probabilities between treated (individuals with disabilities) and control (individuals without disabilities) groups. Percentages in the legend indicate the estimated population share of each latent type. Vertical bars represent 95% bootstrap uniform confidence intervals across periods within each transition pair and latent type, $j \in \{1, 2\}$. The dashed vertical line marks the timing of the ADA's introduction.

Figure 22: Differences in Labor Force Transition Probabilities Before the ADA ($J = 3$).



Notes: This figure reports estimated differences in monthly labor force transition probabilities between individuals with and without disabilities before the enactment of the ADA, for each latent type (shown in different colors). Each panel corresponds to a distinct labor force transition, where rates represent the monthly probability of moving from one labor force status (e.g., employed, unemployed, or out of the labor force (OLF)) to another. The x-axis measures calendar months relative to the introduction of the ADA (with zero denoting the policy's introduction), and the y-axis shows the estimated difference in transition probabilities between treated (individuals with disabilities) and control (individuals without disabilities) groups. Percentages in the legend indicate the estimated population share of each latent type. Vertical bars represent 95% bootstrap uniform confidence intervals across periods within each transition pair and latent type, $j \in \{1, 2, 3\}$. The dashed vertical line marks the timing of the ADA's introduction.

References

Acemoglu, Daron and Joshua D Angrist, "Consequences of employment protection? The case of the Americans with Disabilities Act," *Journal of Political Economy*, 2001, *109* (5), 915–957.

Anderson, Soren T. and James M. Sallee, "Using Loopholes to Reveal the Marginal Cost of Regulation: The Case of Fuel-Economy Standards," *American Economic Review*, June 2011, *101* (4), 1375–1409.

Anderson, T. W., "On Estimation of Parameters in Latent Structure Analysis," *Psychometrika*, 1954, *19* (1), 1–10.

Angrist, Joshua D and Jörn-Steffen Pischke, *Mostly harmless econometrics: An empiricist's companion*, Princeton University Press, 2009.

Arkhangelsky, Dmitry, Susan Athey, David A. Hirshberg, Guido W. Imbens, and Stefan Wager, "Synthetic Difference-in-Differences," *American Economic Review*, 2021, *111* (12), 4088–4118.

Athey, Susan and Guido W Imbens, "Identification and inference in nonlinear difference-in-differences models," *Econometrica*, 2006, *74* (2), 431–497.

Blundell, Richard and Monica Costa Dias, "Alternative Approaches to Evaluation in Empirical Microeconomics," *Journal of Human Resources*, 2009, *44* (3), 565–640.

_ , _ , Costas Meghir, and John Van Reenen, "Evaluating the Employment Impact of a Mandatory Job Search Program," *Journal of the European Economic Association*, 2004, *2* (4), 569–606.

Bonhomme, Stéphane and Elena Manresa, "Grouped patterns of heterogeneity in panel data," *Econometrica*, 2015, *83* (3), 1147–1184.

_ , Thibaut Lamadon, and Elena Manresa, "Discretizing Unobserved Heterogeneity," *Econometrica*, 2022, *90* (2), 625–643.

Boussim, Onil, "Compositional Difference-in-Differences for Categorical Outcomes," 2025.

Bustos, Paula, "Trade Liberalization, Exports, and Technology Upgrading: Evidence on the Impact of MERCOSUR on Argentinian Firms," *The American Economic Review*, 2011, *101* (1), 304–340.

Callaway, Brantly and Pedro H.C. Sant'Anna, "Difference-in-Differences with multiple time periods," *Journal of Econometrics*, 2021, *225* (2), 200–230. Themed Issue: Treatment Effect 1.

_ and Tong Li, "Quantile treatment effects in difference in differences models with panel data," *Quantitative Economics*, 2019, *10* (4), 1579–1618.

__ , __ , **and Tatsushi Oka**, "Quantile treatment effects in difference in differences models under dependence restrictions and with only two time periods," *Journal of Econometrics*, 2018, *206* (2), 395–413.

Cameron, A. Colin, Jonah B. Gelbach, and Douglas L. Miller, "Bootstrap-Based Improvements for Inference with Clustered Errors," *The Review of Economics and Statistics*, 2008, *90* (3), 414–427.

Card, David and Daniel Sullivan, "Measuring the Effect of Subsidized Training Programs on Movements In and Out of Employment," *Econometrica*, 1988, *56* (3), 497–530.

Carroll, Raymond J., Xiaohong Chen, and Yingyao Hu, "Identification and Estimation of Nonlinear Models Using Two Samples with Nonclassical Measurement Errors," *Journal of Nonparametric Statististics*, 2010, *22* (4), 379–399.

Charoenwong, Ben, Alan Kwan, and Tarik Umar, "Does regulatory jurisdiction affect the quality of investment-adviser regulation?," *American Economic Review*, 2019, *109* (10), 3681–3712.

Chen, Jiafeng and Jonathan Roth, "Logs with Zeros? Some Problems and Solutions," *The Quarterly Journal of Economics*, 2024, *139* (2), 891–936.

Di Francesco, Riccardo and Giovanni Mellace, "Causal inference for qualitative outcomes," *Economics Letters*, 2025, *256*, 112626.

Ding, Peng and Fan Li, "A bracketing relationship between difference-in-differences and lagged-dependent-variable adjustment," *Political Analysis*, 2019, *27* (4), 605–615.

Freyaldenhoven, Simon, Christian Hansen, and Jesse M Shapiro, "Pre-event trends in the panel event-study design," *American Economic Review*, 2019, *109* (9), 3307–3338.

Ghanem, Dalia, Désiré Kédagni, and Ismael Mourifié, "Evaluating the Impact of Regulatory Policies on Social Welfare in Difference-in-Difference Settings," *arXiv preprint arXiv:2306.04494*, 2023.

Graves, John A., Carrie Fry, J. Michael McWilliams, and Laura A. Hatfield, "Difference-in-Differences for Categorical Outcomes," *Health Services Research*, 2022, *57* (3), 681–692.

Heckman, James, Hidehiko Ichimura, and Petra Todd, "Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme," *The Review of Economic Studies*, 1997, *64* (4), 605–654.

__ , __ , **Jeffrey Smith, and Petra Todd**, "Characterizing Selection Bias Using Experimental Data," *Econometrica*, September 1998, *66* (5), 1017–1098.

Hernán, Miguel A. and James M. Robins, *Causal Inference: What If* 2025. Unpublished manuscript (Chapman & Hall/CRC, 2025).

Higgins, Ayden and Koen Jochmans, "Identification of mixtures of dynamic discrete choices," *Journal of Econometrics*, 2023, *237* (1), 105462.

Hotchkiss, Julie L, *The labor market experience of workers with disabilities: The ADA and beyond*, W.E. Upjohn Institute, 2003.

Hu, Yingyao, "Identification and estimation of nonlinear models with misclassification error using instrumental variables: A general solution," *Journal of Econometrics*, 2008, *144* (1), 27–61.

_ **and Matthew Shum**, "Nonparametric identification of dynamic models with unobserved state variables," *Journal of Econometrics*, 2012, *171* (1), 32–44.

Hvide, Hans K and Benjamin F Jones, "University innovation and the professor's privilege," *American Economic Review*, 2018, *108* (7), 1860–1898.

Ishimaru, Shoya, "Estimating Treatment Effects in Panel Data Without Parallel Trends," April 2025. Hitotsubashi University, Working Paper.

Kahn-Lang, Ariella and Kevin Lang, "The promise and pitfalls of differences-in-differences: Reflections on 16 and pregnant and other applications," *Journal of Business & Economic Statistics*, 2020, *38* (3), 613–620.

Kasahara, Hiroyuki and Katsumi Shimotsu, "Nonparametric Identification of Finite Mixture Models of Dynamic Discrete Choices," *Econometrica*, 2009, *77* (1), 135–175.

_ **and** _ , "Testing the Number of Components in Normal Mixture Regression Models," *Journal of the American Statistical Association (Theory and Methods)*, 2015, *110* (512), 1632–1645.

Lise, Jeremy, Elena Pastorino, and Luigi Pistaferri, "Revisiting the Employment Effects of the Americans with Disabilities Act," 2023.

MacKinnon, James G. and Matthew D. Webb, "Wild Bootstrap Inference for Wildly Different Cluster Sizes," *Journal of Applied Econometrics*, 2017, *32* (2), 233–254.

Madansky, Albert, "Determinantal methods in latent class analysis," *Psychometrika*, June 1960, *25* (2), 183–198.

Marx, Philip, Elie Tamer, and Xun Tang, "Heterogeneous Treatment Effects via Linear Dynamic Panel Data Models," 2025.

Newey, Whitney K. and Daniel McFadden, "Large sample estimation and hypothesis testing," in "Handbook of Econometrics," Vol. 4, Elsevier, jan 1994, pp. 2111–2245.

Puhani, Patrick A, “The treatment effect, the cross difference, and the interaction term in nonlinear “difference-in-differences” models,” *Economics Letters*, 2012, *115* (1), 85–87.

Rambachan, Ashesh and Jonathan Roth, “A More Credible Approach to Parallel Trends,” *The Review of Economic Studies*, 02 2023. rdad018.

Robins, J. M., “A New Approach To Causal Inference in Mortality Studies With a Sustained Exposure Period - Application To Control of the Healthy Worker Survivor Effect,” *Mathematical Modelling*, 1986, *7*, 1393–1512.

Roth, Jonathan, “Pretest with Caution: Event-Study Estimates after Testing for Parallel Trends,” *American Economic Review: Insights*, September 2022, *4* (3), 305–22.

_ **and Pedro H. C. Sant'Anna**, “When Is Parallel Trends Sensitive to Functional Form?,” *Econometrica*, 2023, *91* (2), 737–747.

_ **, Pedro H.C. Sant'Anna, Alyssa Bilinski, and John Poe**, “What's trending in difference-in-differences? A synthesis of the recent econometrics literature,” *Journal of Econometrics*, 2023.

Sun, Liyang and Sarah Abraham, “Estimating dynamic treatment effects in event studies with heterogeneous treatment effects,” *Journal of Econometrics*, 2021, *225* (2), 175–199.

United States Census Bureau, “Survey of Income and Program Participation, 1990 Panel,” 1990–1993.

Wooldridge, Jeffrey M, “Simple approaches to nonlinear difference-in-differences with panel data,” *The Econometrics Journal*, 2023, *26* (3), C31–C66.