

[*[Click here for the most recent version](#)*]

November 20, 2025

## COMPOSITIONAL DIFFERENCE-IN-DIFFERENCES FOR CATEGORICAL OUTCOMES

ONIL BOUSSIM

**ABSTRACT.** In DiD settings with categorical or compositional outcomes (e.g., occupation or major choice, voting outcomes, electricity mix), treatments can affect both total counts (extensive margin) and within-category shares (intensive margin). Standard linear DiD applied to shares or totals separately can generate invalid counterfactuals (e.g., negative values) and lack compatibility with discrete choice models. I propose Compositional DiD (CoDiD), which jointly identifies counterfactual totals and shares under a parallel growths assumption: absent treatment, each category's absolute size evolves proportionally at the same rate in treated and control groups. Under a standard random utility model, I show that this is equivalent to parallel trends in expected utilities, i.e., the change in the average latent value for each alternative is identical across groups. Consequently, relative preferences between any pair of categories (how one alternative is preferred to another) evolve in parallel, and counterfactual counterfactual distributions follow parallel trajectories in the probability simplex. I extend CoDiD to (i) partial-identification bounds under weaker assumptions, (ii) staggered treatment adoption, and (iii) a synthetic DiD analog. Empirically, I apply CoDiD to study early voting reforms on U.S. presidential elections and to study the impact of the Regional Greenhouse Gas Initiative on states' electricity generation mix.

*Keywords:* categorical outcomes, difference-in-differences, causal inference, compositional data, voting, RGGI.

*JEL codes:* C14, C23, C31

---

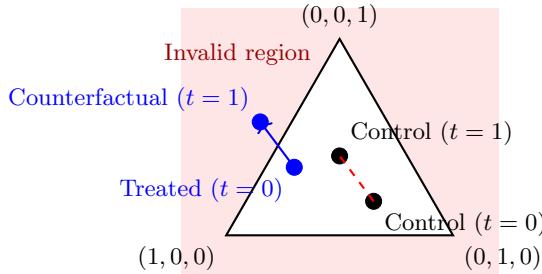
Corresponding address: oib5044@psu.edu, Department of Economics, The Pennsylvania State University, University Park, PA 16801. I am grateful to Marc Henry, Andres Aradillas-Lopez, and Sung Jae Jun for their guidance on this project. I also thank Keisuke Hirano, Ismael Mourifié, Désiré Kédagni, and Patrick Guggenberger, as well as all participants of the Penn State Econometrics Seminar, the Penn State Student Econometrics Workshop, the RESA Alumni Conference 2025, and the EconNect Africa JMC Symposium, for their valuable comments and suggestions. All remaining errors are my own.

## INTRODUCTION

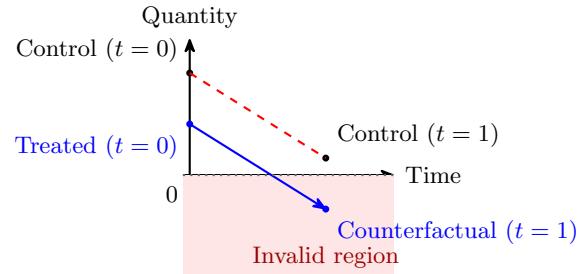
The classical difference-in-differences (DiD) framework identifies the average treatment effect on the treated under the parallel trends assumption for scalar outcomes. However, in many applications, the outcome of interest is categorical, and the notion of an average level is generally undefined. Examples include voting decision (democrat, republican, others), occupational sector (agriculture, manufacturing, and services), or field of study (STEM, humanities, business, or others). For categorical outcomes, the key quantitative objects are the category-specific quantities, which jointly determine both the overall total quantity and the corresponding probability mass function across categories. In this case, one would like to analyze both absolute changes (how many units are added or removed in absolute terms for each category, and overall) and relative changes (how the shares are reallocated across categories) in a consistent way. For example, a voting policy can differently affect both the total number of voters (turnout) and the reallocation of vote shares for each party. A minimum wage policy can change both the number of people across all employment sectors and the probability of being in each of them. An education reform may reallocate students between fields, but policymakers may also care about the total number of new graduates overall. The same structure arises with compositional outcomes, vectors of quantities that partition a well-defined whole, such as budget allocations, GDP composition across sectors, or the electricity generation mix. In these cases, the observed outcome is directly the vector of quantities itself, rather than individual-level categorical choices.

The classical linear parallel trends assumption, which imposes a linear evolution in outcome levels, may be inappropriate for categorical quantities for several reasons already noted in the literature (see [Lechner \(2011\)](#); [Athey and Imbens \(2006\)](#); [Wooldridge \(2023\)](#)). A naive strategy would be to apply linear parallel trends separately to the raw quantities of each category, obtain counterfactual quantities, and then normalize them to form a counterfactual distribution. However, this approach is problematic for several reasons. First, applying linear parallel trends to raw quantities is scale-dependent, as it ignores differences in group size. Transferring the same additive change from a large control group to a smaller treated group can distort counterfactual quantities, exaggerating or understating effects, and thus producing misleading totals and distributions. Second, this method can yield negative counterfactual quantities when a negative shift in the control group outweighs the pre-treatment

value in the treated group. On top of that, applying parallel trends directly to probabilities is theoretically possible but often yields invalid counterfactuals, with values below zero or exceeding one. The resulting counterfactual distribution can fall outside the probability simplex (the set of all vectors of positive numbers summing to one), rendering them invalid. We illustrate all this graphically in the figures 1 and 2,



**Figure 1.** Illustration of how PT in shares can lead to invalid counterfactuals in the 2-dimensional probability simplex.



**Figure 2.** Illustration of how PT in quantities can lead to invalid counterfactuals for quantities, as it can lead to negative values.

Moreover, imposing linear parallel trends on either quantities or shares is not consistent with the implications of standard discrete choice models. Such models typically generate nonlinear relationships between underlying utilities and observed market shares or quantities. Therefore, assuming linear trends would imply implausible dynamics in category choices, suggesting that agents adjust their selections across alternatives through time in a purely mechanical fashion. This approach lacks a behavioral or structural foundation in economic theory, which instead predicts that changes in shares or quantities arise from shifts in relative utilities, not just from linear extrapolations over time.

Most existing nonlinear DiD methods for discrete outcomes (e.g., CiC, copula-based methods) rely on a natural ordering of the categories. Such an ordering is crucial because these approaches typically operate through cumulative distribution functions (CDFs), which require a well-defined order of categories to accumulate probabilities meaningfully. When the outcome is unordered, such as choices among qualitatively distinct alternatives like fields of study, occupations, or parties, the CDF is no longer well-defined, and these methods cannot

be applied directly. In these settings, the probability mass function (PMF) is the only mathematically coherent representation of the outcome distribution. Additionally, these methods do not produce counterfactuals for quantities; they only focus on the shares. Ideally, one would like to formulate a nonlinear parallel trends assumption on categorical quantities that is interpretable in a natural sense, resolves the scale-dependence problem, does not require imposing an artificial ordering on categories, and ensures that the evolution of shares remains consistent with the framework of standard discrete choice theory.

This paper introduces Compositional difference-in-differences (CoDiD), a new framework for analyzing categorical outcomes in a DiD setting. The method is built on a nonlinear parallel trends assumption, which I call *parallel growths*, which allows us to point-identify the counterfactual categorical quantity, and thus the counterfactual total and probability mass functions within a DiD framework. I do so while addressing all the challenges outlined above. In addition, I define new treatment effect measures tailored to this context. The growth treatment effect on the treated (GTT) quantifies how each category's absolute size changes as a result of treatment, and also how the total quantity has evolved as a result of treatment. The second object is the compositional treatment effect on the treated (CTT), which measures how treatment proportionally reallocates probability mass across categories. Alternatively, one may focus on the simple difference in shares as the parameter of interest (ATT). This yields an absolute change in category shares and reflects how probability mass is reallocated in absolute terms in response to treatment.

I begin with the canonical  $2 \times 2$  DiD setting (two groups and two time periods). The parallel growths assumption is a parallel trends of log-transformed categorical quantities, meaning that the absolute size of each category would grow or shrink at the same proportional rate in both treated and control groups. Using a standard random utility model, I show that this assumption is equivalent to assuming parallel trends of latent utilities for each category. This means that in the absence of treatment, both groups would have experienced the same evolution in how desirable each option is on average. For example, if in the control group, the expected utility for voting democrat increase, the same movement will happen in the treated group in the absence of treatment. The implied evolution of the shares is a parallel trends in the log-odds of category shares. In the random utility model, log-odds capture utility differences, therefore relative preferences. Under this interpretation,

parallel growths mean that, absent treatment, the evolution of relative preferences between any pair of categories would be similar in both groups. In other words, the evolution of the way people chose between categories in the control group will be similar to what happens in the treated group. Under this interpretation, we can view the causal effect as the way the treatment shift the way categories are desired.

Furthermore, without relying on a random utility framework, I derive another economic implication of parallel growths. In the control group, category shares naturally evolve over time, with some categories gaining mass at the expense of others. The assumption implies that, in the absence of treatment, the treatment group would have experienced the same reallocation pattern observed in the control group. This interpretation allows us to view the causal effect as the extent to which the treatment changes the way probability mass is redistributed across categories. Finally, in geometric terms, each probability mass function (PMF) corresponds to a point in the simplex, and its evolution traces a path within this space. The parallel growths assumption means that, in the absence of treatment, the treated group's counterfactual path would be parallel to the control group's path under Aitchison geometry, the standard geometry for the simplex in compositional data analysis ([Aitchison \(1982\)](#); [Egozcue et al. \(2003\)](#)).

I also discuss the connections and key differences between my approach and the multinomial logit two-way fixed-effects model. In the simple  $2 \times 2$  case, I show that—when focusing solely on category shares, both methods produce the same counterfactual shares. The key distinction is that my approach also recovers counterfactual quantities, which cannot be obtained directly from the logistic specification. Moreover, my method naturally accommodates compositional data and aggregated quantities, whereas the multinomial logit model is designed for individual-level choice data and does not extend directly to aggregate settings. Moreover, my framework introduces treatment effect parameters that offer a more natural and interpretable measure than the change in log-odds implied by the logistic specification. Even when the analysis is limited to shares, the differences become more pronounced when extending the framework to multiple time periods (e.g., staggered adoption case) and relaxations of the identifying assumption. My approach provides a more flexible and easily extendable foundation for such generalizations. Another comparison is made with the DiD and the changes-in-changes (CiC) frameworks. While DiD targets expectations (points on

the real line) and CiC, cumulative distribution functions (points in a space of CDFs), my method also targets probability mass functions (points in the probability simplex). The underlying identification logic is consistent across all three methods: each characterizes the control group’s evolution via a translation mapping and applies that same mapping to the treated group’s initial state to recover its counterfactual. The fundamental distinction, therefore, lies not in the core logic but in the object being transported, expectations, CDFs, or PMFs, and the geometry that defines the translation. From this unified perspective, the CoDiD emerges naturally as the DiD/CiC analogue for probability mass functions.

In a more general framework with multiple time periods, the parallel growths assumption can be assessed visually by plotting the log-transformed quantities. If the trends appear non-parallel in pre-treatment periods, the assumption becomes less credible. In this case, I first propose a set of relaxations of the assumption in the spirit of [Ban and Kédagni \(2022\)](#). Specifically, I bound the counterfactual log differences between groups by restricting their change to lie within the convex hull of observed changes across pre-treatment periods. This approach produces bounds for both counterfactual quantities and counterfactual distributions. When multiple potential control groups are available, I develop a synthetic DiD (see [Arkhangelsky et al. \(2021\)](#)) analog, synthetic CoDiD, that merges the flexibility of synthetic control methods with the causal logic of difference-in-differences. This extension constructs a weighted combination of control units whose pre-treatment log-trajectory closely matches that of the treated unit, while preserving the parallel growths structure in the post-treatment period. The resulting control group not only improves pre-treatment fit but also maintains interpretability within the CoDiD framework. Finally, I generalize the methodology to settings with staggered treatment adoption, where units receive treatment at different points in time. I follow the same assumptions as in [Callaway and Sant’Anna \(2021\)](#), propose two types of parallel growths assumption in this case, and derive the corresponding counterfactual in each case.

As applications, I first examine the impact of early voting programs on the distribution of voter support across political parties in U.S. presidential elections. These programs, allowing voters to cast ballots before election day, either in person or by mail, have been the subject of intense partisan debate. I revisit this question by analyzing both total turnout and the compositional shift in voter preferences. I consider Maryland and New Jersey,

which introduced early voting before 2008 in the treated group, while Pennsylvania and New York serve as the control group. As expected, the treatment increased overall turnout ( increase of 4.4%) and the number of votes for all parties. The ATT shows that the Republican party lost 1.11 percentage points, 0.9 in favor of democrats, and 0.3 in favor of alternative parties. However, the compositional treatment effect on the Treated (CTT) reveals a nuanced reallocation: a previously indifferent voter (assigning equal initial weight to all parties) would, after the treatment, shift support more toward alternative parties (since they gain the most in relative terms). Secondly, I analyze the Regional Greenhouse Gas Initiative (RGGI) and its effects on the mix of electricity generation sources in participating states. RGGI, a cap-and-trade program designed to reduce carbon emissions from the power sector, is expected to shift the composition of electricity generation across coal, natural gas, nuclear, and renewable sources by promoting cleaner technologies and discouraging carbon-intensive ones. I assess how RGGI affected the size of each source in the electricity generation process over time. I find that RGGI sharply cut fossil fuel use, especially coal, by shifting generation to gas and nuclear rather than boosting renewables, underscoring the need for complementary policies to accelerate a truly renewable-driven energy transition.

## RELATED LITERATURE

This paper makes some methodological contributions at the intersection of difference-in-differences and compositional data analysis.

First, it contributes to the extensive DiD literature. Classical DiD focuses on scalar outcomes, estimating treatment effects on the mean, with surveys and applications summarized in [Lechner \(2011\)](#); [Cafri et al. \(2019\)](#); [De Chaisemartin and d'Haultfoeuille \(2023\)](#); [Roth et al. \(2023\)](#); [Baker et al. \(2025\)](#). Recent works have extended DiD to more complex settings: [Callaway and Sant'Anna \(2020\)](#) develops methods for staggered treatment adoption, [Arkhangelsky et al. \(2021\)](#) combines DiD with synthetic controls, and [Manski and Pepper \(2018\)](#); [Rambachan and Roth \(2020\)](#); [Ban and Kédagni \(2022\)](#) relax the parallel trends assumption. A second wave of research extends DiD to entire outcome distributions. Early contributions include quantile DiD [Meyer et al. \(1990\)](#) and the changes-in-changes (CiC) approach [Athey and Imbens \(2006\)](#), which identifies counterfactual distributions non-parametrically. Subsequent work has generalized these ideas to multivariate outcomes [Torous](#)

et al. (2021), settings where identifying assumptions apply to cumulative distribution functions Havnes and Mogstad (2015); Roth and Sant'Anna (2021), copula-based approaches Callaway et al. (2018); Callaway and Li (2019); Ghanem et al. (2023), and methods using characteristic functions Bonhomme and Sauder (2011). However, these methods do not directly address categorical outcome distributions, which are probability mass functions. Graves et al. (2022) proposes a DiD for categorical outcomes while relying on parallel trends for proportions, which I explained earlier, is not appropriate in general. Zhou et al. (2025) extend the difference-in-differences framework to non-Euclidean data, including probability mass functions, by employing Fréchet means and geodesic transport. While their approach is mathematically elegant, the identifying assumptions are primarily geometric and offer limited economic interpretability. In contrast, my framework delivers both counterfactual quantities and probability mass functions that retain economic intuition while incorporating a geometric perspective specifically designed for categorical outcomes. Another related paper is Tchetgen Tchetgen et al. (2024), who propose a general DiD framework applicable to count data in the binomial setting. My analysis differs by focusing on the multinomial case, allowing for richer categorical structures and more general forms of distributional change. Second, the paper contributes to the compositional data analysis (CoDA) literature. CoDA, pioneered by Aitchison (1982, 1990, 1992, 2002), provides tools for analyzing data constrained to the simplex. Subsequent developments Egozcue et al. (2003); Billheimer et al. (2001); Barceló-Vidal et al. (2001) refine transformations and models that respect the geometric structure of compositional data. More recently, Arnold et al. (2020) connected CoDA to causal inference. I also combine compositional data methods with econometric identification strategies to analyze causal effects when outcomes are probability mass functions, providing a bridge between DiD and compositional statistics.

The paper is organized as follows. Section 1 introduces the canonical  $2 \times 2$  case, presenting the main identifying assumption along with its economic and geometric interpretations. Section 2 generalizes the framework to settings with multiple time periods. Section 3 discusses the connection with existing approaches, and Section 4 illustrates the methodology through two empirical applications. Finally, Section 5 concludes.

## 1. CANONICAL $2 \times 2$ DiD CASE

In this section, I focus on the canonical, two-group, two-time period case in the DiD framework, and I clearly expose the assumptions and their implications. I will start with the analytical framework, later discuss the main identifying assumption, and all the economic and geometric implications.

### 1.1. Parallel growths and Identification.

1.1.1. *Analytical framework.* In this section, I introduce the basic framework of my method. A categorical outcome  $Y$  takes values in a finite set of  $p$  mutually exclusive categories:

$$Y \in \{c_1, c_2, \dots, c_p\}.$$

I focus primarily on the absolute quantities for each category, which can appear in several empirical contexts. They can arise when the outcome is observed at the individual level. For example, in evaluating a job training program, individuals may be classified as unemployed, part-time employed, or full-time employed. The categorical quantity is obtained by aggregating individuals into these outcome categories. Sometimes, the data consist directly of quantities across categories rather than individual-level observations. For example, the number of votes each political party receives in an election or the budget allocated to each category (for example, food, rent, healthcare, others). Unlike individual-level data, the identities or attributes of the units are not observed. In all these cases, the proposed method remains applicable, as it relies solely on knowledge of the categorical quantities.

I consider a setting with two groups,  $g \in \{0, 1\}$ , where  $g = 1$  is the treated group and  $g = 0$  is the control group. These groups are observed in two time periods,  $t \in \{0, 1\}$ , with  $t = 0$  being the pre-treatment period and  $t = 1$  the post-treatment period. For any category  $c_k$ , I define  $q_{g,t}^N(c_k)$ : the potential untreated quantity (the quantity in  $c_k$  if the group  $g$  were untreated at time  $t$ ). I also have  $q_{g,t}^I(c_k)$ : the potential treated quantity (the quantity in  $c_k$  if the group  $g$  were treated at time  $t$ ). For the untreated potential outcomes, I aggregate across categories to define the total quantity for a group-period:

$$S_{gt}^N = \sum_{k=1}^p q_{g,t}^N(c_k).$$

The share (or probability) for category  $c_k$  is then:

$$\pi_{g,t}^N(c_k) = \frac{q_{g,t}^N(c_k)}{S_{gt}^N}.$$

The vector of quantities and the entire probability mass function are represented as:

$$q_{g,t}^N = [q_{g,t}^N(c_1), \dots, q_{g,t}^N(c_p)], \quad \pi_{g,t}^N = [\pi_{g,t}^N(c_1), \dots, \pi_{g,t}^N(c_p)], \quad \text{where } \sum_{k=1}^p \pi_{g,t}^N(c_k) = 1.$$

All the same definitions apply to the treated potential quantities,  $q_{g,t}^I(c_k)$ , and their associated totals and shares. From the data, I can identify the vectors  $q_{0,0}^N$  for the control group and  $q_{1,0}^N$  for the treated group in the pre-treatment period. In the post-treatment period, I can identify from data the untreated vector of quantities for the control group,  $q_{0,1}^N$ , and the treated counterpart for the treated group,  $q_{1,1}^I$ . The counterfactual of interest is the vector of quantities that I would have observed for the treated group in the absence of treatment  $q_{1,1}^N$ . Identifying this missing object is the core challenge of my analysis. Once this is identified, I can consider the treatment effect parameters.

**Treatment effect parameters:** I introduce treatment effect parameters to capture the impact of the intervention. The first is the growth treatment effect on the treated (GTT), which quantifies the causal effect on the absolute size of a category. For each category  $c_k$  (where  $k = 1, \dots, q$ ), the GTT is defined as the proportional change in its quantity:

$$\text{GTT}(c_k) = \frac{q_{1,1}^I(c_k)}{q_{1,1}^N(c_k)} - 1, \quad GTT = \frac{S_{1,1}^I}{S_{1,1}^N} - 1. \quad (1)$$

This parameter has an intuitive interpretation:  $GTT = 0$  implies the treatment had no effect on the absolute.  $GTT > 0$  indicates that the treatment caused category  $c_k$  to grow in absolute terms. For example, a GTT of 0.15 means the category's size increased by 15% due to the treatment.  $GTT < 0$  signifies that the treatment caused the category to shrink. A GTT of  $-0.10$ , for instance, corresponds to a 10% decline.

When comparing the shares, there are several ways to do it, depending on the type of effects that we would like to emphasize. To quantify the absolute change in shares, we can focus on the Average treatment effect on the treated (ATT), defined, defined as the simple difference in category proportions (or shares) between the treated potential outcome distribution and its untreated benchmark.

$$ATT = \pi_{1,1}^I - \pi_{1,1}^N. \quad (2)$$

This representation is intuitive, easy to interpret, and directly reflects the absolute reallocation of probability mass across categories induced by the treatment.

Sometimes, it might be interesting to see which category has gained relatively the most as a result of treatment. The compositional treatment effect on the treated (CTT) captures how treatment redistributes weight across categories proportionally and also tells us which category gained relatively the most. To define it, I consider the compositional difference operator commonly used in compositional data analysis to compare probability mass functions. This operator quantifies how each category expands or contracts relative to all others. For intuition, fix a category  $c_k$  and consider its proportional change:

$$r_k = \frac{\pi_{1,1}^I(c_k)}{\pi_{1,1}^N(c_k)}, \quad \begin{cases} r_k > 1 & \text{category } k \text{ grew,} \\ r_k < 1 & \text{category } k \text{ shrunk,} \\ r_k = 1 & \text{no change.} \end{cases}$$

Dividing by the total change gives a vector showing each category's relative gain or loss. I obtain the compositional difference operation used in compositional data analysis for the control group.

$$\pi_{1,1}^N \ominus \pi_{1,1}^N = \frac{1}{\sum_{k=1}^p r_k} [r_1, r_2, \dots, r_p] = \frac{1}{\sum_{k=1}^p \frac{\pi_{1,1}^I(c_k)}{\pi_{0,0}^N(c_k)}} \left[ \frac{\pi_{1,1}^N(c_1)}{\pi_{1,1}^N(c_1)}, \dots, \frac{\pi_{1,1}^N(c_p)}{\pi_{1,1}^N(c_p)} \right]. \quad (3)$$

Each component then reflects how much a category's importance has shifted relative to the others. Components larger than  $1/p$  indicate categories that gain relative importance, while components smaller than  $1/p$  indicate categories that lose. If all ratios equal one and the operator yields the uniform composition  $(1/p, \dots, 1/p)$ , we therefore have:

$$\text{CTT} = \pi_{1,1}^I \ominus \pi_{1,1}^N. \quad (4)$$

A useful way to think about CTT is through a neutral individual who starts indifferent, assigning equal probability to all  $p$  categories ( $1/p$ ). After treatment, this person updates their weights according to CTT. Components greater than  $1/P$  indicate categories that gain relative importance, while smaller components indicate categories that lose relative importance. For example, in a three-party voting scenario, a neutral voter starts with  $(1/3, 1/3, 1/3)$ . After a campaign (treatment), CTT might imply  $(0.5, 0.3, 0.2)$ . This shows that support for the first party increased, while the other two declined, with the second party still stronger than the third.

One can also consider the treatment effects on some functionals of the distribution. Consider, for example, a policy evaluated using a scalar index that is a nonlinear functional of compositional shares, such as the Herfindahl–Hirschman Index (HHI):

$$\text{HHI} = \sum_{k=1}^K \pi(c_k)^2$$

A common approach would be to apply DiD directly to the index. However, this strategy extrapolates the index directly, ignoring that it is derived from shares that must satisfy some mathematical and economic constraints. A better approach is to get the counterfactual shares first and later recompute the counterfactual index. Consider a general function  $H$  on the probability mass function, a causal analysis on  $H(\pi)$  would be the functional treatment effect on the treated (FTT), defined for a given choice of  $H$

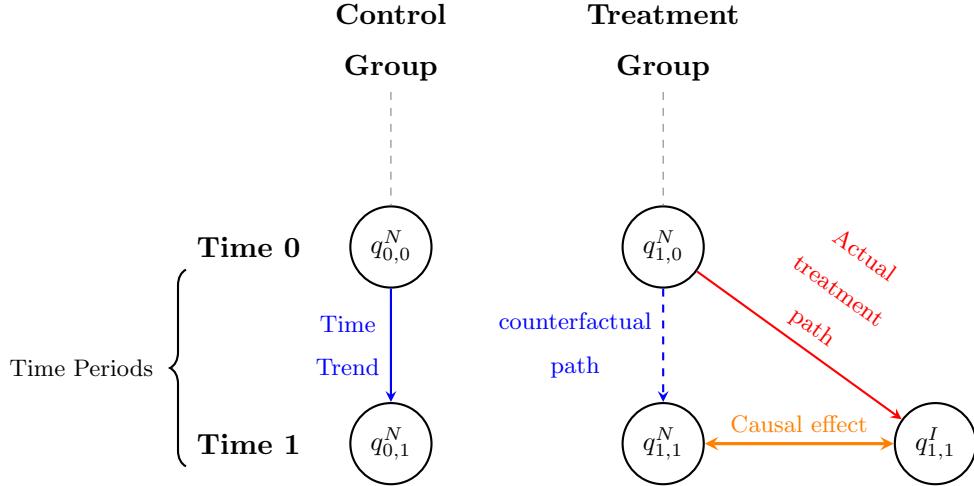
$$\text{FTT} = \pi_{1,1}^I - \pi_{1,1}^N. \quad (5)$$

In the next paragraph, I introduce the identifying assumption and its implications.

**1.1.2. Parallel growths.** This section develops the identification strategy and clarifies its implications for the probability mass functions. I start with the following assumption.

**Assumption 1** (Common Support). *The support, denoted by  $\bar{\mathcal{Y}} = \{c_1, \dots, c_p\}$ , is the same for all untreated potential categorical outcomes.*

This assumption ensures that all observed and counterfactual outcomes in the absence of treatment are defined over a common set of categories. It rules out situations where the support changes across groups or over time, which would make comparisons and counterfactual reasoning impossible. In practice, it implies that the set of possible outcomes is fixed. Such support conditions are standard in the literature. To introduce the main identifying assumption, I provide intuition for the DiD identification logic. The core idea is that, in the absence of treatment, the treated group would have followed the same trajectory as the control group. Figure 3 illustrates this DiD identification logic: the control group, unexposed to the treatment, serves as a counterfactual benchmark, mirroring what would have happened to the treated group had it not received the intervention. As I said before, in many empirical settings, the sizes of the groups defined in each category can vary dramatically in scale. Therefore, a standard parallel trends assumption on the raw quantities is problematic



**Figure 3.** Visual representation of the DiD identification strategy. The counterfactual outcome for the treated group in the post-treatment period ( $q_{1,1}^N$ ) is constructed by extrapolating the time trend observed in the control group (blue arrows) to the treated group's pre-treatment level. The causal effect is the difference between the observed treated outcome ( $q_{1,1}^I$ ) and this counterfactual ( $q_{1,1}^N$ ), shown by the orange double-headed arrow.

in this context, as it would impose the same absolute change on both large and small groups, which is often unrealistic. To address this issue of scale and to formulate a more plausible identifying assumption, I instead consider parallel trends on the log transformation of the quantities. This approach focuses on proportional, rather than absolute, changes. For a given vector, define the functions  $(\log)$  and  $(\exp)$  which apply respectively the logarithm and the exponential functions component-wise to the vectors.

**Assumption 2.** *(parallel growths)*

$$\log(q_{1,1}^N) - \log(q_{1,0}^N) = \log(q_{0,1}^N) - \log(q_{0,0}^N).$$

Assumption 2 is the core identifying condition for the model. It states that, in the absence of treatment, the proportional growth (or decline) of each category in the treated group would have been the same as that of the control group. This reflects the intuitive idea that underlying trends affect all categories proportionally to their size. This may appear as a more realistic and flexible assumption than one based on raw quantities in many economic contexts, especially when the sizes of categories differ significantly. It captures the idea that trends often affect all units proportionally to their size. It is also possible to use discrete

covariates to make the assumption more believable. I discuss this possibility in appendix B. From this assumption, the counterfactual quantity vector for the treated group can be easily recovered, and the corresponding counterfactual probability mass can be obtained by normalization, as we can see in Theorem 1. Taken together, assumptions 1 and 2 impose sufficient structure to point-identify the counterfactual distributions. The following theorem 1 formalizes this result

**Theorem 1.** *Under assumptions 1 and 2, the counterfactual quantities are identified as:*

$$q_{1,1}^N = \exp (\log (q_{1,0}^N) + \log (q_{0,1}^N) - \log (q_{0,0}^N)), \quad (6)$$

$$S_{1,1}^N = \sum_{k=1}^p \exp (\log (q_{1,0}^N(c_k)) + \log (q_{0,1}^N(c_k)) - \log (q_{0,0}^N(c_k))), \quad (7)$$

$$\pi_{1,1}^N = \frac{1}{S_{1,1}^N} q_{1,1}^N. \quad (8)$$

This identification result is constructive and provides a straightforward way to implement the method. In the next section, I will discuss further justifications for parallel growths, exploring its economic and geometric implications.

**1.2. Economic justifications of parallel growths.** In this section, I provide economic justifications for parallel growths using the standard random utility framework (McFadden, 1972, 1974, 1977).

Consider a standard Random Utility Model, in which a decision-maker selects one alternative from a finite set of mutually exclusive choices  $\{c_1, c_2, \dots, c_p\}$ . The utility that the individual derives from alternative  $c_k$  in the absence of treatment is modeled as:

$$U_{gt}^N(c_k) = V_{gt}^N(c_k) + \varepsilon_{gt}(c_k),$$

where  $U_{gt}^N(c_k)$  denotes the latent utility of category  $k$  in group  $g$  at time  $t$  in the absence of treatment, and  $V_{gt}^N(c_k)$  is the systematic (deterministic) component of utility (can be seen as the measure of the total attractiveness of the category) and  $\varepsilon_{gt}(c_k)$  are the random unobserved shocks. The individual is assumed to choose the alternative that yields the highest latent utility:

$$\text{Choice} = \arg \max_{c_k} U_{gt}^N(c_k).$$

I decompose the total attractiveness into two parts:

$$V_{gt}^N(c_k) = \underbrace{\mu_{gt}^N(c_k) - \log \left( \sum_k e^{\mu_{gt}^N(c_k)} \right)}_{\text{relative attractiveness}} + \underbrace{\log(S_{gt}^N)}_{\text{absolute attractiveness}}.$$

This decomposition shows that  $V_{gt}^N(c_k)$  contains two additive components: the first is the relative desirability of the category (within-choice structure)  $\mu_{gt}^N(c_k) - \log \left( \sum_j e^{\mu_{gt}^N(c_j)} \right)$ . This term reflects how desirable option  $c_k$  is relative to all other alternatives at the same  $(g, t)$ . The second is absolute desirability (aggregate intensity)  $\log(S_{gt}^N)$ . This term captures the overall size or participation level (e.g., how many people are making a choice at all). It scales all utilities equally, affecting the total mass of choices but not their relative shares. When it rises (say, due to higher purchasing power, stronger participation, or favorable aggregate conditions), all utilities shift up equally, people are more likely to choose something, but the proportions among categories might remain the same unless the first component also change. Together, they form a utility concept that blends relative preferences and absolute participation. This decomposition is useful to see how both components can be affected. We finally have:

$$U_{gt}^N(c_k) = \mu_{gt}^N(c_k) - \log \left( \sum_k e^{\mu_{gt}^N(c_k)} \right) + \log(S_{gt}^N) + \varepsilon_{gt}(c_k).$$

In the traditional RUM, we assume that the random terms  $(\varepsilon_{gt}(c_1), \dots, \varepsilon_{gt}(c_p))$  all follow standard type-1 extreme value distribution and are independent. Because each marginal error term remains Gumbel-distributed with mean  $\gamma \approx 0.5772$  (the Euler–Mascheroni constant), the expected utility of alternative  $c_k$  is:

$$\mathbb{E}[U_{gt}^N(c_k)] = \mu_{gt}^N(c_k) - \log \left( \sum_k e^{\mu_{gt}^N(c_k)} \right) + \log(S_{gt}^N) + \gamma.$$

$\mathbb{E}[U_{gt}^N(c_k)]$  captures the average latent total desirability in the population for category  $c_k$ . The first implication of the RUM is given by the following proposition.

**Proposition 1** (Implication for expected utilities). *Under assumptions 1 and the RUM model assumptions, assumption 2 is equivalent to parallel trends of expected utilities:*

$$\mathbb{E}[U_{11}^N(c_k)] - \mathbb{E}[U_{10}^N(c_k)] = \mathbb{E}[U_{01}^N(c_k)] - \mathbb{E}[U_{00}^N(c_k)] \quad \forall c_k \in \bar{\mathcal{Y}}. \quad (9)$$

This equivalence provides a clear economic interpretation of the identifying assumption. It states that, in the absence of treatment, both groups would have experienced the same evolution in how valuable each option is on average. For example, in a voting context, suppose the control group experiences a rise in the expected utility of voting Democrat (e.g., due to national political trends). My assumption implies that the treated group would have experienced the same increase in expected utility had it not been exposed to the policy. The same logic applies to all categories. Thus, the parallel growths assumption is not merely a mathematical condition on outcomes; it can also be seen as an economic hypothesis about the common evolution of underlying category-specific attractiveness in the absence of treatment. This provides a micro-founded justification for CoDiD: it estimates how a policy shifts the total attractiveness of alternatives as measured by expected utility.

Now, we will see how it also restrict the evolution of the shares. Define the multinomial logistic transform. For  $\pi = [\pi(c_1), \dots, \pi(c_p)] \in \mathcal{S}^{p-1}$ . The log-odds (or multinomial logit) transformation  $\ell : \mathcal{S}^{p-1} \rightarrow \mathbb{R}^{p-1}$  maps probabilities to real-valued indices:

$$\ell(\pi) = \left[ \log\left(\frac{\pi(c_1)}{\pi(c_p)}\right), \dots, \log\left(\frac{\pi(c_{p-1})}{\pi(c_p)}\right) \right]. \quad (10)$$

Each component of this vector is the log-odds of category  $c_k$  compared to the baseline category  $c_p$ . In fact, the ratio  $\pi_{g,t}^N(c_k)/\pi_{g,t}^N(c_p)$  is the odds of category  $c_k$  relative to the reference category  $c_p$  and taking the logarithm gives the log-odds. The choice of the baseline category can be arbitrary. Its inverse is given by the softmax function: for any  $y = (y_1, \dots, y_{p-1}) \in \mathbb{R}^{p-1}$ ,

$$\ell^{-1}(y) = \frac{1}{1 + \sum_{i=1}^{p-1} e^{y_i}} (e^{y_1}, \dots, e^{y_{p-1}}, 1). \quad (11)$$

The parallel growths assumption implies a parallel trends assumption on the log-odds transformation of the probability distribution, as stated in this proposition.

**Proposition 2** (Implication for shares). *Assumptions 1 and 2 imply parallel trends in log-odds, i.e.,*

$$\ell(\pi_{1,1}) - \ell(\pi_{1,0}) = \ell(\pi_{0,1}) - \ell(\pi_{0,0}).$$

Moreover, parallel trends in log-odds is equivalent to parallel trends in utility differences for any pair of categories under the RUM:

$$\begin{aligned} \left\{ \mathbb{E}[U_{11}^N(c_k)] - \mathbb{E}[U_{11}^N(c_l)] \right\} - \left\{ \mathbb{E}[U_{10}^N(c_k)] - \mathbb{E}[U_{10}^N(c_l)] \right\} = \\ \left\{ \mathbb{E}[U_{01}^N(c_k)] - \mathbb{E}[U_{01}^N(c_l)] \right\} - \left\{ \mathbb{E}[U_{00}^N(c_k)] - \mathbb{E}[U_{00}^N(c_l)] \right\}, \quad \forall c_k, c_l \in \bar{\mathcal{Y}}. \end{aligned} \quad (12)$$

The log-odds of choosing one category over another provide a direct empirical analogue to differences in the underlying expected utilities. They summarize how much more desirable one option is relative to another, capturing the relative strength or simply the preferences in the population. For instance, in a voting context, the log-odds of choosing Democrat over Republican quantify the population's relative preference for Democrats. When the log-odds rise, it means that the expected utility, or average attractiveness, of voting Democrat has increased relative to that of voting Republican. The parallel growths assumption implies that, in the absence of treatment, these relative preferences would have evolved in parallel across groups for any pair of categories. That is, any secular change, such as a national political shift increasing the attractiveness of Democrats relative to Republicans, would have affected both treated and control groups identically had the treatment not occurred. The treatment effect is then identified as the deviation from this common trajectory. From that, I can also identify the counterfactual distribution with an alternative formula:

$$\pi_{1,1}^N = \ell^{-1} (\ell(\pi_{1,0}) + \ell(\pi_{0,1}) - \ell(\pi_{0,0})). \quad (13)$$

In a  $2 \times 2$  DiD with categorical outcomes, I have shown that assuming parallel trends in log-quantity ratios is not ad hoc, it is equivalent to assuming parallel trends in expected utilities, a natural behavioral assumption under the Random Utility Model. Because utility differences (not levels) drive choice, this assumption is economically meaningful. This foundation justifies the CoDiD approach as a method for estimating how a policy changes the relative attractiveness of each alternative compared to others. In Appendix D, I extend this result to more general Random Utility Models that allow for dependence among the error terms, thereby relaxing the IIA assumption implicitly assumed in the standard multinomial RUM.

**1.3. Geometric justifications of parallel growths.** In this section, I provide some geometric justifications for parallel growths. I start by interpreting the restriction on the shares

using the compositional difference operator.

**Interpretation using the compositional difference interpretation:** To further support the relevance of the assumption in this context, I provide an alternative interpretation that does not depend on the random utility framework or its underlying assumptions. Suppose our goal is to understand how the treatment redistributes probability mass across categories, that is, how each category gains or loses relative shares compared to the others as a consequence of treatment. An intuitive way of building the counterfactual would be to assume that the relative growth/decline in shares of each category in the control group reflects what would have happened in the treated group without treatment. The compositional difference operator ( $\ominus$ ) defined in equation 3 provides the precise mathematical language to formalize this intuition. I now show that the parallel growths assumption can be interpreted using that intuition. Specifically, assumption 2 implies that, in the absence of treatment, the relative growth or decline of category shares would have been similar across the treated and control groups. This interpretation is formally established in the following proposition.

**Proposition 3.** *Assumption 1, assumption 2 implies that the compositional difference for the untreated potential categorical distributions is identical for the treated and control groups:*

$$\pi_{1,1}^N \ominus \pi_{1,0}^N = \pi_{0,1}^N \ominus \pi_{0,0}^N.$$

The equality states that these two compositional changes are identical. In essence, the underlying forces that caused shares in the control group to be reallocated in a particular way would have produced the exact same pattern of reallocation within the treated group. This provides a clear, distributional interpretation of the parallel growths assumption.

**Geometric interpretation:** Parallel growths also imply parallel trajectories of probability mass functions in the probability simplex. Probability mass functions lie in the simplex  $\mathcal{S}^{p-1}$ , where movement can be meaningfully defined using the geometry introduced by [Aitchison \(1982\)](#). Unlike the Euclidean case, this geometry accounts for the simplex's constraints and forms the basis of compositional data analysis, redefining operations such as addition, scaling, and distance.

**Proposition 4** ([Aitchison \(2002\)](#), [Egozcue et al. \(2003\)](#)). . Let  $\mathcal{S}^{p-1}$  be the interior of the  $p - 1$  simplex, define the following operations: for all  $\pi_1, \pi_2 \in \mathcal{S}^{p-1}$ , we have:

$$\pi_1 \oplus \pi_2 = \left[ \frac{\pi_1(c_1)\pi_2(c_1)}{\sum_{k=1}^q \pi_1(c_k)\pi_2(c_k)}, \dots, \frac{\pi_1(c_p)\pi_2(c_p)}{\sum_{k=1}^q \pi_1(c_k)\pi_2(c_k)} \right],$$

and for all  $\pi_1 \in \mathcal{S}$  and  $\alpha \in \mathbb{R}$ ,

$$\alpha \odot \pi_1 = \left[ \frac{\pi_1^\alpha(c_1)}{\sum_{k=1}^q \pi_1^\alpha(c_k)}, \dots, \frac{\pi_1^\alpha(c_p)}{\sum_{k=1}^q \pi_1^\alpha(c_k)} \right], \quad \pi_1^\alpha \text{ is the usual powering.}$$

$(\mathcal{S}^{p-1}, \oplus, \odot)$  is a vector space of dimension  $p - 1$  where  $\oplus$  denotes the perturbation operator (analogous to vector addition),  $\odot$  denotes the powering operator (analogous to scalar multiplication), and the uniform distribution  $\left[\frac{1}{p}, \dots, \frac{1}{p}\right]$  acts as the zero element of the space.

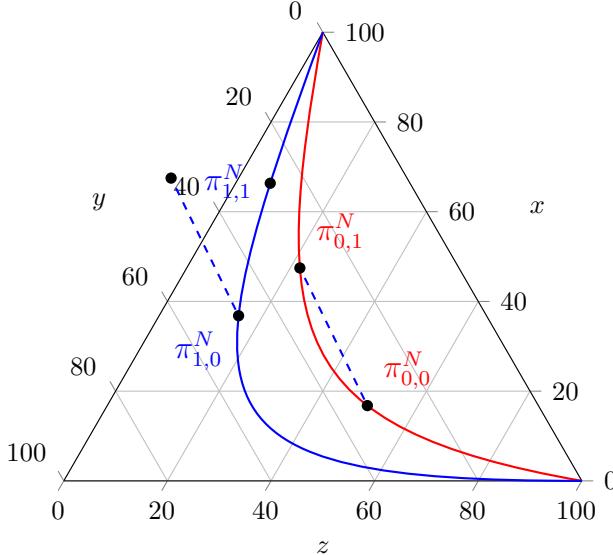
The compositional difference between two distributions  $\pi_1$  and  $\pi_2$  can be redefined as:

$$\pi_1 \ominus \pi_2 = \pi_1 \oplus (-1 \odot \pi_2).$$

This formulation highlights that  $\ominus$  works like a genuine vector difference in this vector space: it represents the perturbation needed to move from  $\pi_2$  to  $\pi_1$ . In other words, it captures the direction and magnitude of the adjustment between the two points. Viewing  $\ominus$  this way makes the operator's mathematical structure natural. Under this structure, parallel growths implies that the movements of the two PMFs correspond to trajectories that remain at a constant separation without intersecting, having the same direction. Figure 4 illustrates this idea: in the Aitchison geometry of the simplex, the trajectories of two distributions evolve in parallel within the 2-dimensional simplex  $\mathcal{S}^2$ , representing all valid probability distributions over three categories. In figure 4, each point in the triangle represents a distribution over three categories. As time evolves, distributions trace curves within the simplex. Although these paths may look non-linear in Euclidean space, in the simplex geometry, they are parallel, maintaining a consistent direction without intersecting. Here, I say that the implied evolution of the distributions is consistent with the nature of their space. Because of that, the compositional treatment effect on the treated can be written as:

$$CTT = \pi_{1,1} \ominus \pi_{1,0} \oplus \pi_{0,1} \ominus \pi_{0,0}. \tag{14}$$

This expression mirrors the structure of the standard average treatment effect on the treated (ATT) in difference-in-differences, but it operates in the probability simplex rather than in the real line. In the classical DiD, differences in means capture changes in average outcomes



**Figure 4.** Illustration of parallel growths in the simplex. The red curve represents the trajectory of the control group from pre-treatment ( $\pi_{0,0}^N$ ) to post-treatment ( $\pi_{0,1}^N$ ), while the blue curve shows the counterfactual trajectory of the treated group from pre-treatment ( $\pi_{1,0}^N$ ) to post-treatment ( $\pi_{1,1}^N$ ). Dashed lines indicate the linear translation (parallel growths).

across time and groups. In contrast, the compositional DiD captures differences in entire categorical distributions: each component of the treated group's pre-treatment shares is rescaled by the corresponding ratio of control group shares, and the resulting vector is normalized to yield a valid probability distribution. Thus, (CTT) extends the ATT from the scalar outcome space to the compositional space, preserving the relative structure of category shares while identifying how treatment shifts the allocation of probability mass across categories. This also helps formalize the geometric implication of the assumption.

All these perspectives reinforce the idea that imposing the parallel growths assumption on the quantities leads to desirable and interpretable restrictions on the evolution of category shares. It links the economic intuition of stable evolution of preferences with a geometric structure of parallel movements in the probability simplex.

## 2. EXTENSIONS WITH MULTIPLE TIME PERIODS

In this section, I propose several extensions of the method in the case where we have multiple time periods and possibly more potentially heterogeneous control groups.

**2.1. Relaxations of parallel growths.** When multiple pre-treatment periods are available, the plausibility of the parallel growths assumption can be evaluated by examining the pre-treatment trajectories of the log-quantities. If the estimated pre-treatment trends appear non-parallel, this raises concerns about the validity of the assumption. Rather than discarding the design altogether, one can relax the assumption in a way that delivers informative bounds on the treatment effect. In particular, I adopt a relaxation strategy closely related to the approach in [Ban and Kédagni \(2022\)](#). They relax the parallel trends assumption by allowing the counterfactual difference between treated and control units to lie within the convex hull of the differences observed in the two most recent pre-treatment periods ( $t = 0$  and  $t = -1$ ). By taking the convex hull of the last two differences, they are saying: the counterfactual evolution cannot jump outside the range spanned by recent dynamics. If the differences between groups have been changing slowly, then the most recent differences provide a plausible envelope for what would have happened in the absence of treatment. This is weaker than requiring parallelism and rules out wild deviations inconsistent with observed history. Analogously, in our setting, I extend this idea by applying the same type of convex-hull relaxation. I consider a more general case, where one can either only use the two most recent periods, all the past periods, or all the past periods with higher weights to the most recent ones. I keep the same notation as before, but now, we have that  $t \in \{-T_1, \dots, 0, 1\}$ .

**Assumption 3.** *For every category  $c_k$  where  $k = 1, \dots, p$ , we have:*

$$\log(q_{1,1}^N(c_k)) - \log(q_{0,1}^N(c_k)) \left[ \min_{t \leq 0} \omega_t \log \left( \frac{q_{1,t}^N(c_k)}{q_{0,t}^N(c_k)} \right), \quad \max_{t \leq 0} \omega_t \log \left( \frac{q_{1,t}^N(c_k)}{q_{0,t}^N(c_k)} \right) \right].$$

Assumption 3 allows the counterfactual log-difference of each category in the post-treatment period to lie within the convex hull of the weighted pre-treatment log-growth ratios. The weights  $\omega_t$  determine how much each pre-treatment period contributes to this range. This assumption ensures that deviations from parallelism in the pre-treatment periods are accounted for, while still preserving enough structure to obtain informative bounds on the counterfactual and therefore treatment effects. In practice, the researcher can consider different scenarios with the weights. The first case considers uniform weights for the  $p$  most recent pre-treatment periods, defined as  $\omega_t = 1$  for  $t = 0, -1, \dots, -p + 1$ , and  $\omega_t = 0$  for all other pre-treatment periods. In this specification, all pre-treatment periods are treated symmetrically. A second case is when more recent pre-treatment periods receive greater

importance. This specification is often empirically justified when recent dynamics are considered more informative about the counterfactual trajectory. It adjust the bounds by giving greater relevance to near-period observations and diminishing the influence of older periods. In all these cases, Assumption 3 ensures that deviations from parallelism observed before treatment are explicitly incorporated in the identification region. The identified set is the collection of all quantities that are consistent with the data and the model assumptions.

**Theorem 2** (Partial Identification).

*Under assumptions 1, and 3, for all  $\{c_k, k = 1, \dots, p\}$ , we have:*

$$\begin{aligned} q_{1,1}^N &\in \left[ b_{min}, b_{max} \right], \\ S_{1,1}^N &\in \left[ \sum_{k=1}^p b_{max}(c_k), \sum_{k=1}^p b_{max}(c_k) \right] \\ \pi_{1,1}^N &\in \left\{ \pi \in \mathcal{S}^{p-1} : \pi_k = \frac{q_k}{\sum_j q_j}, q_k \in [b_{min}^k, b_{max}^k] \text{ for all } k \right\}. \end{aligned}$$

where we have:

$$\begin{aligned} b_{min}^k &= \exp \left( \min_{t \leq 0} \omega_t \log \left( \frac{q_{1,t}^N(c_k)}{q_{0,t}^N(c_k)} \right) + \log(q_{0,1}^N(c_k)) \right), \\ b_{max}^k &= \exp \left( \max_{t \leq 0} \omega_t \log \left( \frac{q_{1,t}^N(c_k)}{q_{0,t}^N(c_k)} \right) + \log(q_{0,1}^N(c_k)) \right). \end{aligned}$$

Theorem 2 establishes sharp bounds on both the growth treatment effect on the treated and the compositional treatment effect on the treated by allowing for deviations from parallelism between treated and control groups. The CTT bounds describe a feasible set within the simplex of compositions, capturing all possible redistributions consistent with those counterfactuals. The choice of  $\omega_t$  governs the degree of relaxation: uniform weights yield a fully agnostic specification based solely on pre-treatment extremes, while time-increasing weights change the bounds by prioritizing near-treatment dynamics.

**2.2. Staggered adoption.** In this section, I extend the staggered treatment adoption framework of Callaway and Sant'Anna (2021), originally developed for scalar outcomes, to the setting of categorical outcomes. Such designs are increasingly common in applied policy

evaluation, where reforms or programs are rolled out gradually across units rather than implemented simultaneously. I consider a panel of  $n$  units observed over periods  $t = 0, 1, \dots, T$ . Let  $D_{i,t} \in \{0, 1\}$  indicate whether unit  $i$  is treated in period  $t$ . I impose the standard irreversibility assumption:

**Assumption 4** (Irreversibility of Treatment). *No unit is treated at baseline:  $D_{i,0} = 0$  almost surely. Moreover, treatment is permanent: if  $D_{i,t-1} = 1$ , then  $D_{i,t} = 1$  for all  $t = 1, \dots, T$ .*

Under this assumption, each unit that ever receives treatment has a well-defined first treatment time, denoted  $G_i = \min\{t : D_{i,t} = 1\}$ . Units never treated are assigned  $G_i = \infty$ . This partitions the sample into cohorts: for each  $g \in \{1, \dots, T\}$ , group  $G_g$  consists of units first treated in period  $g$ ; the never-treated group is  $G_\infty$ . The outcome of interest is categorical: for each group and time, I observe a vector of positive quantities  $q_{G_g,t}$ . I also work with the associated share vector  $\pi_{G_g,t} = q_{G_g,t} / \sum_{k=1}^p q_{G_g,t}(c_k)$ .

For each cohort  $G_g$ , I observe pre-treatment (untreated) outcomes for all  $t < g$ , and post-treatment (treated) outcomes for  $t \geq g$ . The goal is to identify the counterfactual untreated quantity  $q_{G_g,t}^N$  for  $t \geq g$ . Let  $\bar{g} = \max\{G_i : G_i < \infty\}$  denote the last period in which any unit is first treated, and define  $\mathcal{G} = \text{Supp}(G) \setminus \{\bar{g}\}$  as the set of all treated cohorts except the last one. (The last cohort cannot be used as a control for itself under the second identification strategy, as explained below). I propose two identification strategies, both generalizing the parallel growths assumption (assumption 2) to the staggered setting:

**Assumption 5** (Parallel Growths : Never-Treated Control). *For every  $g \in \mathcal{G}$  and every  $t \geq g$ ,*

$$\log(q_{G_g,t}^N) - \log(q_{G_\infty,t}^N) = \log(q_{G_g,t-1}^N) - \log(q_{G_\infty,t-1}^N).$$

**Assumption 6** (Parallel Growths : Not-Yet-Treated Controls). *For every  $g \in \mathcal{G}$ , and for all  $s, t$  such that  $t \geq g$  and  $t \leq s < \bar{g}$ ,*

$$\log(q_{G_g,t}^N) - \log(q_{G_s,t}^N) = \log(q_{G_g,t-1}^N) - \log(q_{G_s,t-1}^N).$$

Assumption 5 states that, in the absence of treatment, the log-quantity paths of cohort  $G_g$  and the never-treated group  $G_\infty$  evolve in parallel. Assumption 6 replaces the never-treated group with not-yet-treated cohorts  $G_s$  (those first treated after time  $t$ ), which serve as internal controls. These conditions enable us to identify counterfactuals by extrapolating

pre-treatment log-gaps into post-treatment periods. The resulting identification formulas are constructive, interpretable, and directly estimable from observed data, as formalized in the following theorem.

**Theorem 3** (Identification). *Under Assumption 5, for every  $g \in \{1, \dots, T\}$  and every  $t \geq g$ ,*

$$q_{G_g,t}^N = \exp \left( \log(q_{G_g,g-1}^N) + \log(q_{G_\infty,t}^N) - \log(q_{G_\infty,g-1}^N) \right).$$

*Under Assumption 6, for every  $g \in \mathcal{G}$  and every  $t \geq g$ , and for any  $s$  such that  $t < s \leq \bar{g}$  (i.e.,  $G_s$  is not treated by time  $t$ ),*

$$q_{G_g,t}^N = \exp \left( \log(q_{G_g,g-1}^N) + \log(q_{G_s,t}^N) - \log(q_{G_s,g-1}^N) \right).$$

The corresponding distributions are obtained by normalizing the counterfactual quantities and the total by summing them.

**2.3. Synthetic CoDiD.** Now suppose we have a setting with  $G > 2$  groups observed over  $T > 2$  time periods. Only the first group ( $g = 1$ ) receives treatment, beginning after period  $T_0$ ; all other groups ( $g = 2, \dots, G$ ) serve as never-treated controls. In this context, Arkhangelsky et al. (2021) propose the Synthetic difference-in-differences (SDID) method, an extension of the classic DiD framework that constructs a data-driven synthetic control by optimally reweighting the control units to best reproduce the pre-treatment trajectory of the treated unit. This improves the credibility of post-treatment comparisons by enforcing a stronger pre-treatment fit than standard DiD. In this section, I adapt SDID to settings where the outcome is a categorical quantity vector, for example, expenditures across categories, counts of events by type, or any positive-valued composition. For each group  $g$  and time  $t$ , let  $q_{g,t}$  denote the vector of quantities across  $p$  categories, and let  $T_0$  be the last pre-treatment period. The objective is to estimate treatment effects on both total quantity and shares. The procedure is applied category by category; for clarity of exposition, I suppress the category index ( $c_k$ ) in what follows, noting that all steps are repeated for each component of the quantity vector.

- **Find Unit Weights( $\omega^*$ ):** I find weights for the control groups so that their weighted pre-treatment path best matches the pre-treatment path of the treated group.

$$\omega^* = \arg \min_{\omega \in \mathcal{S}^{J-1}} \sum_{t=1}^{T_0} \left\| \log(\mathbf{q}_{1,t}) - \sum_{g=2}^G \omega_j \log(\mathbf{q}_{g,t}) \right\|^2.$$

- **Find Time Weights ( $\lambda^*$ ):** I find weights for the pre-treatment time periods so that the pre-treatment average of the controls best matches their post-treatment average.

$$\boldsymbol{\lambda}^* = \arg \min_{\boldsymbol{\lambda} \in \mathcal{S}^{T_0-1}} \sum_{g=2}^G \left\| \frac{1}{T - T_0} \sum_{t=T_0+1}^T \log(\mathbf{q}_{g,t}) - \sum_{t=1}^{T_0} \lambda_t \log(\mathbf{q}_{g,t}) \right\|^2.$$

The average categorical quantities are obtained by:

$$(\text{Treated}, T_0) \quad \mathbf{q}_{\text{treated}, T_0} = \exp(\lambda_{T_0}^* \log \mathbf{q}_{1, T_0}).$$

$$(\text{Control}, T_0) \quad \mathbf{q}_{\text{control}, T_0} = \exp \left( \sum_{g=2}^G \omega_g^* \lambda_{T_0}^* \log \mathbf{q}_{g, T_0} \right).$$

$$(\text{Control, After}, t > T_0) \quad \mathbf{q}_{\text{control}, t} = \exp \left( \sum_{g=2}^G \omega_g^* \log \mathbf{q}_{g, t} \right).$$

Finally, the counterfactuals for the treated group for all periods  $t > T_0$  are given by:

$$\begin{aligned} \mathbf{q}_{\text{treated}, t}^N &= \exp (\log(\mathbf{q}_{\text{treated}, T_0}) + \log(\mathbf{q}_{\text{control}, t}) - \log(\mathbf{q}_{\text{control}, T_0})) \\ \mathbf{S}_{\text{treated}, t}^N &= \sum_k \mathbf{q}_{\text{treated}, t}^N (c_k) \\ \boldsymbol{\pi}_{\text{treated}, t}^N &= \mathbf{q}_{\text{treated}, t}^N / \mathbf{S}_{\text{treated}, t}^N \end{aligned}$$

### 3. CONNECTIONS WITH EXISTING APPROACHES

In this section, I discuss the link between my method and existing methods in the literature.

**3.1. Connection with Logistic model.** Recall:  $Y_{it} \in \{c_1, c_2, \dots, c_p\}$ : outcome for unit  $i$  at time  $t$ ,  $G_i \in \{0, 1\}$ : group indicator (1 = treated),  $T_t \in \{0, 1\}$ : time indicator (1 = post), Interaction:  $D_{it} = G_i \cdot T_t$  (standard DiD treatment indicator). Pick category  $c_p$  as the baseline. The two-way fixed effects multinomial logit model without covariates is:

$$\mathbb{P}(Y_{it} = c_k \mid G_i, T_t) = \frac{\exp(\alpha_k + \delta_k G_i + \gamma_k T_t + \beta_k D_{it})}{1 + \sum_{j=1}^{p-1} \exp(\alpha_j + \delta_j G_i + \gamma_j T_t + \beta_j D_{it})}, \quad k = 1, \dots, p-1.$$

This specification is fully saturated: for each non-baseline outcome category  $c_k$ , the model includes four parameters, an intercept  $\alpha_k$ , a group effect  $\delta_k$ , a time effect  $\gamma_k$ , and an interaction term  $\beta_k$ , which together exactly span the log-odds in all four group-time cells.

The interaction coefficient  $\beta_k$  captures the difference-in-differences contrast in log-odds and admits a causal interpretation under the assumption of parallel trends on the log-odds. The parallel trends in log-odds are the same restriction that my method imposes on the shares. The compositional treatment effect parameter (CTT) can be linked to the coefficients of a multinomial logit model as follows:

$$CTT = \ell^{-1}(\beta_1, \beta_2, \dots, \beta_{p-1}),$$

where  $\ell^{-1}(\cdot)$  denotes the inverse compositional log-ratio transformation, which maps the vector of log-odds coefficients back into category shares. Although using the multinomial logit representation coincides with my approach to modeling category shares in the simple  $(2 \times 2)$  setup, several conceptual and practical differences are worth emphasizing.

The first key difference is that the multinomial logit model does not directly and jointly deliver counterfactual category quantities. In contrast, my approach constructs these counterfactuals explicitly, allowing recovery of treatment effects on both quantities and shares within a unified framework. Second, when only aggregate data are available, as in many policy applications, my method applies directly, whereas multinomial logit models are designed for individual-level choice data and do not naturally extend to aggregate counts without additional modeling layers. Third, in multi-period settings, the multinomial logit specification offers no transparent way to relax identifying assumptions once they appear violated. My framework, however, accommodates partial-identification relaxations seamlessly through Assumption 3. Fourth, under staggered treatment timing, multinomial logit requires estimating an entire vector of time-varying treatment coefficients. My method instead estimates each post-treatment effect separately, avoiding joint optimization and substantially reducing computational burden. Finally, the ATT and CTT parameters produced by my framework are directly interpretable as effects on category shares, unlike contrasts in log-odds ratios from multinomial logit models, which can be difficult to interpret economically.

**3.2. Unified view with DiD and CiC.** My method identifies the counterfactual probability mass function for unordered categorical outcomes. I show how the identifying assumption relates to parallel trends (DiD) and rank invariance (CiC). DiD targets objects like expectations, CiC targets the CDF for ordered outcomes, and CoDiD extends the same logic to the PMF for unordered outcomes. I proceed in two steps. First, I show that DiD and CiC,

although targeting different objects (means versus CDFs), share the same geometric identification logic. Second, I show how CoDiD applies that logic to probability mass functions. A recall of the setup: I consider two groups  $g \in 0, 1$  (treatment and control), observed in two periods  $t \in 0, 1$ . Let  $m_{g,t}^N$  denote the potential parameter of interest for unit  $g$  at time  $t$  in the absence of treatment.

In DiD, the parameter of interest  $m_{g,t}^N$  is the expectation of some individual-level outcome,  $m_{g,t}^N = E(Y_{g,t}^N)$ . The construction of the counterfactual is given by:

$$E(Y_{1,1}^N) = E(Y_{1,0}^N) \underbrace{-E(Y_{0,0}^N) + E(Y_{0,1}^N)}_{\text{control group change over time}}.$$

In CiC, the parameter of interest  $m_{g,t}^N$  is the cumulative distribution function of some individual-level outcome,  $m_{g,t}^N = F_{Y_{g,t}^N}$ . The construction of the counterfactual is given by:

$$F_{Y_{1,1}^N} = F_{Y_{1,0}^N} \circ \underbrace{F_{Y_{0,0}^N}^{-1} \circ F_{Y_{0,1}^N}}_{\text{control group change over time}}.$$

DiD and CiC share a simple underlying idea: the object of interest exists within a space where the notion of movement can be well-defined, and there is a transformation that characterizes how the control group evolves within this space. To obtain the counterfactual for the treated group, we then apply the same transformation to the treated group. Mathematically, if the control group moves from  $m_{0,0}^N$  to  $m_{0,1}^N$ , I write:

$$m_{0,1}^N = \varphi^c(m_{0,0}^N),$$

where  $\varphi^c$  captures the change or movement in the control group. The counterfactual for the treated group is then obtained by applying the same transformation to the initial state of the treated group  $m_{1,0}^N$ :

$$m_{1,1}^N = \varphi^c(m_{1,0}^N).$$

In DiD:

$$\forall m \in \mathbb{R}, \quad \varphi^c(m) = m - E_{Y_{0,0}^N} + E_{Y_{0,1}^N}. \quad (15)$$

In CiC:

$$\forall F \in \mathcal{F}, \quad \varphi^c(F) = F \circ F_{Y_{0,0}^N}^{-1} \circ F_{Y_{0,1}^N}. \quad (16)$$

where  $\mathcal{F}$  is the space of cumulative distribution functions. The map  $\varphi^c$  has the same interpretation in both methods and can be understood as a translation map. If I want to use a map to define the idea of movement within a space. A natural way to formalize that is through the concept of translation. A translation operator is defined as an operator that shifts points by a certain amount in a certain direction. I consider a general definition of this operator. Intuitively, suppose the outcome space forms a group, meaning it is a set of objects equipped with a binary operation that can help define the idea of movement (moves that can be combined, reversed, and have a neutral element that does nothing). These moves can be addition (for numeric outcomes) or function composition (for distributional transformations). Let's start with the formal definition of a group.

**Definition 1** (Group). *A group  $(G, \wedge)$  consists of a set  $G$  and a binary operation*

$$\wedge : G \times G \rightarrow G, \quad (a, b) \mapsto a \wedge b,$$

*satisfying the following axioms: i) Associativity:  $(a \wedge b) \wedge c = a \wedge (b \wedge c)$  for all  $a, b, c \in G$ . ii) Identity: There exists  $e \in G$  such that  $e \wedge a = a \wedge e = a$  for all  $a \in G$ . iii) Inverses: For each  $a \in G$ , there exists  $a^{-1} \in G$  such that  $a^{-1} \wedge a = a \wedge a^{-1} = e$ .*

For example, in DiD, the outcome space is  $\mathbb{R}$  with addition  $+$  as the group operation, identity  $e = 0$ , and inverse given by negation  $-a$ . In CiC, the outcome space is the set of cumulative distribution functions, the operation is composition  $\circ$ , the identity is the identity function  $Id(x) = x$ , and the inverse is the quantile function. Once the group structure is defined, I can formalize the idea of a movement using translations.

**Definition 2** (Right Translation Operator ([Hamilton, 2017](#))). *Let  $(G, \wedge)$  be a group with identity element  $e$ . For a fixed  $h \in G$ , the left translation by  $h$  is the map:*

$$R_h : G \rightarrow G, \quad R_h(m) := m \wedge h, \quad \forall h \in G.$$

If you start at  $m$ , you then apply  $h$  on the right, ending at  $m \wedge h$ . Think of shifting relative to where you already are, like taking the current position and then performing the move locally. This formalizes the notion of a shift or transformation acting consistently across the entire outcome space. From this perspective, the map  $\varphi^c$  can be identified as the unique right translation that captures how the control group evolves over time.

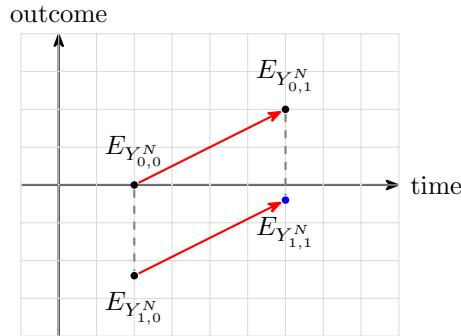
In DiD,  $\varphi^c$  is the right translation operator defining how the control group has moved from its initial state. we have that:

$$\varphi^c(m) = R_h(m) = m + h.$$

Since  $\varphi^c(E(Y_{0,0}^N)) = E(Y_{0,1}^N)$ , it implies that  $h = E(Y_{0,1}^N) - E(Y_{0,0}^N)$ . We therefore obtain that:

$$\varphi^c(m) = m + E(Y_{0,1}^N) - E(Y_{0,0}^N).$$

Graphically, it can be represented in figure 5



**Figure 5.** Geometric Illustration of Parallel Trends

In CiC,  $\varphi^c$  is also the left translation operator defining how the control group has moved from its initial state. We have that:

$$\varphi^c(F) = R_h(F) = F \circ h.$$

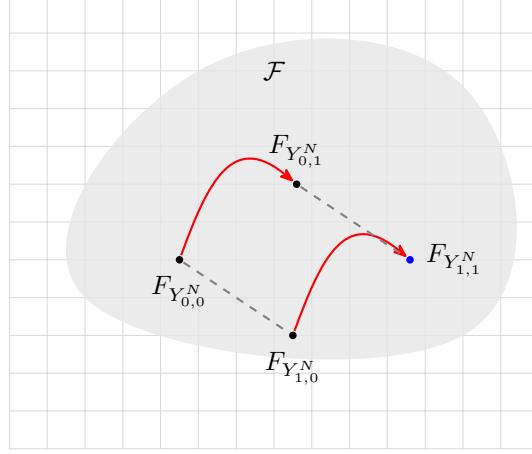
Since  $\varphi^c(F_{Y_{0,0}^N}) = F_{Y_{0,1}^N}$ , it implies that  $h = F_{Y_{0,0}^N}^{-1} \circ F_{Y_{0,1}^N}$ . We therefore obtain that:

$$\varphi^c(F) = F \circ F_{Y_{0,0}^N}^{-1} \circ F_{Y_{0,1}^N}.$$

Graphically, I can illustrate through figure 6. The uniqueness of  $\varphi^c$  is what guarantees point identification. If multiple transformations could map the control group across periods, the counterfactual is no longer unique. In that case, we have identified a set of feasible outcomes, as in the case of Changes-in-Changes (CiC) with a discrete ordered outcome.

CoDiD shares this same logic, but for probability mass functions, where the space is the simplex. Here, the parameter of interest  $m_{g,t}^N$  is the probability mass function,  $m_{g,t}^N = \pi_{g,t}^N$ . The construction of the counterfactual is given by:

$$\pi_{1,1}^N = \pi_{1,0}^N \quad \underbrace{\ominus \pi_{0,0}^N \oplus \pi_{0,1}^N}_{\text{control group change over time}}.$$



**Figure 6.** Geometric illustration of Rank invariance

The control group map is therefore defined in this way:

$$\forall \pi \in \mathcal{S}^{p-1}, \quad \varphi^c(\pi) = \pi \ominus \pi_{0,0}^N \oplus \pi_{0,1}^N. \quad (17)$$

In CoDiD, the group is the simplex equipped with the perturbation operator ( $\oplus$ ). In that method,  $\varphi^c$  is also a right translator operator defined as:

$$\varphi^c(\pi) = R_h(\pi) = \pi \oplus h.$$

Since  $\varphi^c(\pi_{0,0}^N) = \pi_{0,1}^N$ , it implies that  $h = \ominus \pi_{0,0}^N \oplus \pi_{0,1}^N$ . We therefore obtain that:

$$\varphi^c(\pi) = \pi \ominus \pi_{0,0}^N \oplus \pi_{0,1}^N$$

That unified logic breaks down if I use the Euclidean group structure for the simplex (doing parallel trends on proportions). In fact, when I restrict to the simplex, the structure no longer forms a group since the neutral element would have to be the zero vector (which is not part of the simplex). In contrast, under my formulation, the group structure is preserved, ensuring consistency with DiD and CiC. Table 1 summarizes this comparison. Let  $\mathcal{F}$ , be the space of non-decreasing real functions.

**Table 1.** Comparison of Identification Strategies

Method	Outcome Type	Geometric Space	Control group map
parallel trends (DiD)	Expectation	$\mathbb{R}$ (Euclidean)	$m \mapsto m - E_{Y_{0,0}^N} + E_{Y_{0,1}^N}$
Rank Invariance (CiC)	CDF	$\mathcal{F}$	$F \mapsto F \circ F_{Y_{0,0}^N}^{-1} \circ F_{Y_{0,1}^N}$
parallel growths (CoDiD)	PMF	Aitchison simplex	$\pi \mapsto \pi \ominus \pi_{0,0}^N \oplus \pi_{0,1}^N$

#### 4. APPLICATIONS

This section presents two empirical applications. The first examines the effect of early voting policies on voter support decisions in U.S. presidential elections. The second evaluates the impact of the regional greenhouse gas initiative (RGGI) on the composition of electricity generation across energy sources.

**4.1. Impact of early voting law on distributions of voters' support across parties.** Early voting allows registered voters to cast their ballots in person or by mail before Election Day. I evaluate the effect of this policy on voter decisions using the compositional difference-in-differences (CoDiD) framework. A credible analysis requires comparing states with similar political, social, and demographic characteristics so that observed differences in voter participation can be attributed to early voting rather than to preexisting trends. The treatment group includes Maryland and New Jersey, which introduced early voting before 2008, while Pennsylvania and New York serve as the control group. These states are geographically close, share comparable demographic and political profiles, and consistently supported Democratic presidential candidates from 1992 to 2004, indicating similar partisan alignment and voting behavior during the pre-treatment period. Table 2 reports their racial and ethnic composition around 2007 to further contextualize the comparison. This

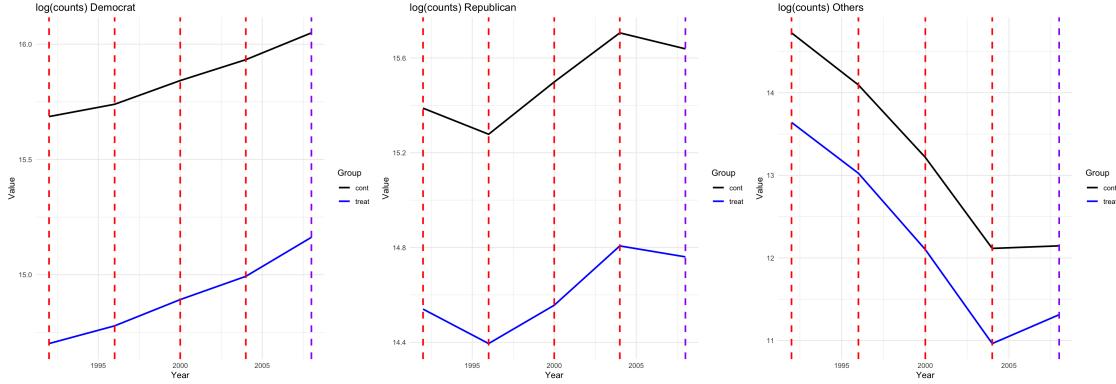
**Table 2.** Population Composition in 2007

State/Group	Total Pop. (2007)	White Afr. Am.	Black/ Afr. Am.	Am. Ind./ Alaska Nat.	Asian	Two or More Races	Hispanic Latino	White, not Hisp.
Maryland	5,618,344	63.6%	29.5%	0.3%	5.0%	1.6%	6.3%	58.1%
New Jersey	8,685,920	76.3%	14.5%	0.3%	7.5%	1.3%	15.9%	62.2%
<b>MD+NJ</b>	<b>14,304,264</b>	<b>71.3%</b>	<b>20.8%</b>	<b>0.3%</b>	<b>6.5%</b>	<b>1.4%</b>	<b>12.1%</b>	<b>60.5%</b>
New York	19,297,729	73.6%	17.3%	0.5%	6.9%	1.5%	16.4%	60.3%
Pennsylvania	12,432,792	85.6%	10.8%	0.2%	2.4%	1.0%	4.5%	81.8%
<b>NY+PA</b>	<b>31,730,521</b>	<b>78.0%</b>	<b>14.5%</b>	<b>0.4%</b>	<b>5.0%</b>	<b>1.3%</b>	<b>11.7%</b>	<b>69.9%</b>

*Source: U.S. Census Bureau, Population Division, released May 1, 2008*

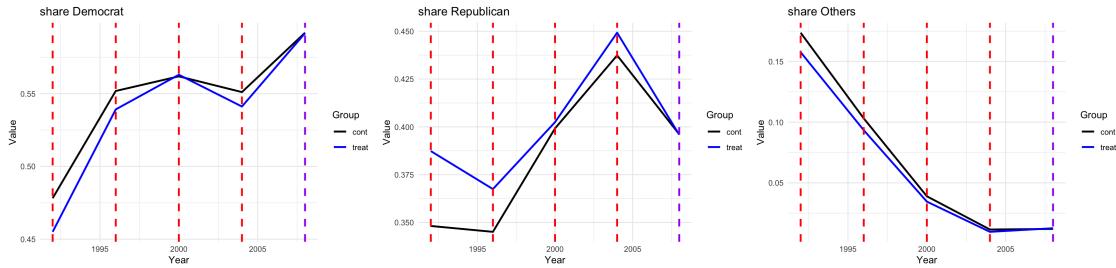
table shows that the two groups have almost similar racial compositions. Combined with their geographic proximity and political alignment over the study period, this makes them

well-suited for causal analysis. The outcome of interest is the voter's support across three categories: Democrats, Republicans, and Others. I consider that from the 1992 elections to 2008. These data are publicly available and consistently reported across states. I start by plotting the log-counts to assess the plausibility of parallel growths. Figure 7 displays the



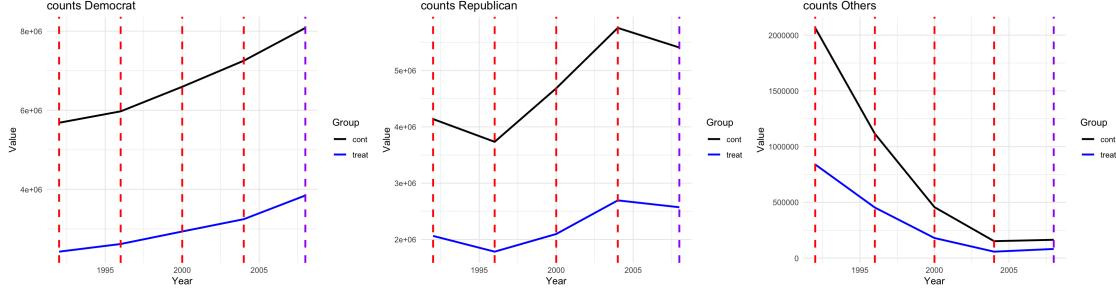
**Figure 7.** log-ratios evolution 1992-2008 (treated vs. control)

evolution of log transformations of each party vote counts across 5 presidential elections. The dotted red line denotes the pre-treatment elections, while the purple line indicates the post-treatment election. In the pre-treatment period, both states exhibit broadly parallel trajectories in log-counts, suggesting comparable underlying voter dynamics before the policy change. This visual similarity reinforces the plausibility of the identification strategy. In this empirical illustration, neither parallel trends on raw proportions nor parallel trends on quantity data provide an appropriate comparison, as evidenced by figure 8 and figure 9



**Figure 8.** shares evolution 1992-2008 (treated vs. control)

My method appears to be the most appropriate in this context, as it relies on the visual assessment of parallel trends between the treatment and control groups. The observed similarity in pre-treatment trajectories provides reassuring evidence that the identifying assumptions are reasonable, thereby strengthening the credibility of the causal analysis.



**Figure 9.** quantities evolution 1992-2008 (treated vs. control)

**Empirical findings** I begin by examining how the treatment affected the total number of votes cast for each party. Table 3 reports the estimated Growth Treatment Effects (GTT), that is, the proportional increase in votes due to the treatment, along with 95% confidence intervals obtained via the bootstrap procedure described in appendix C. Overall, the policy

**Table 3.** Estimated growth treatment effects

Party	Estimate [95% CI]
Democrat	0.0545 [0.0533, 0.0557]
Republican	0.0214 [0.0199, 0.0230]
Other	0.3747 [0.3578, 0.3915]
Total	0.0441 [0.0441, 0.0443]

increased turnout by around 4.4%, demonstrating that early voting can mobilize more voters. This finding is consistent with the existing literature on the effects of early voting. All three categories experienced a statistically significant increase in votes, although the magnitude of the effect varies considerably. The Democratic Party saw a modest 5.5% increase, while the Republican Party experienced a smaller rise of about 2.1%. In contrast, support for Other parties, which includes third-party and independent candidates, grew by nearly 37.5%, representing a much larger relative gain. Importantly, because “Other” parties started from a much smaller base, even a modest absolute increase translates into a large percentage change. Focusing on the two major parties, the Democratic Party benefited more from the policy than the Republican Party. This aligns with the results in Berry et al. (2025), which also found that such policies tend to favor Democrats over Republicans. However, I exercise

caution in interpreting these results, as the states I selected were already Democratic-leaning during this period. The next table 4 gives us what happened with the shares.

**Table 4.** Estimated categorical distributions

Party	Estimate [CI]	Counterfactual[CI]	ATT [CI]
Democrat	0.5915 [0.5911, 0.5915]	0.5823 [0.5818, 0.5829]	0.0092 [0.0086, 0.0099]
Republican	0.3959 [0.3954, 0.3962]	0.4076 [0.4071, 0.4082]	-0.0117 [-0.0124, -0.0111]
Other	0.0126 [0.0125, 0.0127]	0.0101 [0.0100, 0.0102]	0.0025 [0.0024, 0.0027]

Table 4 reports the observed distribution of voter support in the 2008 elections and the estimated counterfactual distribution implied by Theorem 1. The results suggest that the early voting policy slightly increased support for the Democratic Party relative to what we would expect in the absence of treatment: the share of Democratic support is estimated at 0.5915 but would have been 0.5823 without treatment. However, Republican support would have been higher in the absence of treatment (0.4076 counterfactual vs. 0.3959 observed), indicating a shift away from Republicans toward Democrats, suggesting that the main absolute redistribution of probability mass occurred between the two major parties. This is consistent with some of the results obtained in Berry et al. (2025). Interestingly, support for third parties also increased, suggesting that the treatment may have mobilized some voters outside the two major parties as well. Compared to the major parties, the outside option experienced the largest increase relative to its initial share. The next table (5) presents the results for the compositional treatment effects (CTT). Each entry represents

**Table 5.** Estimated CTT for 2008 elections

Party	Estimate [95% CI]
Democrat	0.3136 [0.3119, 0.3151]
Republican	0.2998 [0.2982, 0.3012]
Other	0.3867 [0.3838, 0.3898]

the post-treatment probability of selecting a given party, under a counterfactual scenario where, before treatment, all parties were equally likely (i.e., each had a baseline probability

of 1/3). After the treatment: the probability of voting for “Other” rises to 38.7%, up from the baseline of 33.3%. The probability of voting Republican falls to 30.0%. The probability of voting Democratic also declines slightly, to 31.4%, but remains higher than that of Republicans. Thus, while both major parties lose relative appeal, Republicans lose more ground than Democrats, and third-party candidates gain the most in relative terms. This finding is interesting because most prior work focuses exclusively on the two major parties. My analysis reveals that policy interventions can meaningfully affect the electoral relevance of minor parties, a dimension often overlooked in the literature. It is worth emphasizing that this shift is purely relative: the overall ranking of parties by actual vote share in the treated states remains unchanged (Democrats first, Republicans second, Others third). Hence, the treatment does not overturn the existing political hierarchy, but it does narrow the gap between major and minor parties in terms of voter consideration.

**4.2. Impact of RGGI on electricity generation composition.** In this section, I study how the Regional Greenhouse Gas Initiative (RGGI) affected the mix of electricity generation in participating states. RGGI, launched in 2009, is a cooperative, market-based program among Connecticut, Delaware, Maine, Maryland, Massachusetts, New Hampshire, New York, Rhode Island, and Vermont. Its goal is to cap and reduce carbon emissions from the power sector. The program works by setting a regional cap on total carbon dioxide emissions from power plants and issuing a limited number of tradable allowances. Each allowance gives a power plant the right to emit one short ton of carbon dioxide. States distribute these allowances through quarterly auctions. The revenue from these auctions is reinvested in energy efficiency, renewable energy, and other programs that benefit consumers. Evaluating the impact of the Regional Greenhouse Gas Initiative (RGGI) on electricity generation is an interesting causal question because it reveals whether the policy effectively shifts energy production toward cleaner sources and reduces carbon emissions. I analyze U.S. electricity generation data, which is publicly available on the U.S. Energy Information Administration’s website (<https://www.eia.gov/electricity/data/state/>). This dataset provides annual electricity generation figures in megawatts by energy source for each state. I consider the following four categories of energy sources: gas (combining natural gas and other gases), coal and oil (combining coal and petroleum), nuclear (corresponding solely to the nuclear

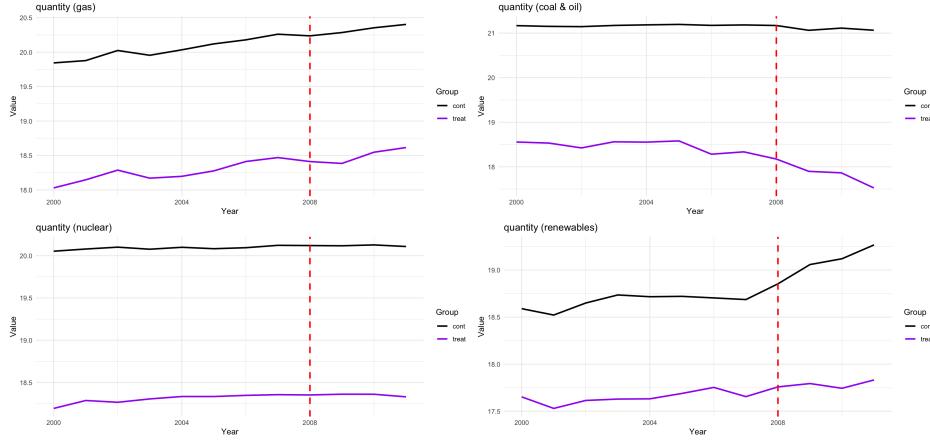
category), and renewables (combining hydroelectric, conventional, solar thermal and photovoltaic, geothermal, wind, and wood and wood-derived fuel). For each state and year, I observe the number of megawatts generated by different sources, and I aggregate these into two groups: the treated group and the control group. In constructing the control group, I exclude states that either joined RGGI, are strongly affected by leakage through electricity market integration, or implemented their own carbon pricing policies. In particular, Pennsylvania, West Virginia, Virginia, and Ohio are excluded due to their geographic proximity and electricity market linkages with RGGI states, which expose them to leakage and anticipatory effects. In fact, electricity markets in the U.S. Northeast and Mid-Atlantic are highly integrated (PJM, NYISO, ISO-NE), including these states in the control group risks contamination. California, Washington, and Oregon are also excluded, since those states adopted cap-and-trade or cap-and-invest programs that constitute carbon pricing mechanisms similar in nature to RGGI. Including them in the control pool would compromise the validity of the counterfactual by mixing treated and untreated units. The program was



**Figure 10.** Map of the United States showing the nine RGGI states in purple and six excluded control states (due to market leakage or carbon pricing policies) in dark gray, and the control group consists of the gray states.

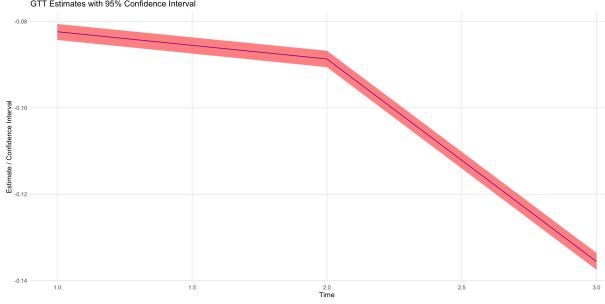
organized into three-year compliance periods, with the first spanning January 1, 2009, to December 31, 2011. I focus my analysis on this initial period because the parallel growth assumption becomes progressively less plausible over longer time horizons. Here, I also assess the validity of the parallel growths assumption by plotting both the raw quantities and their logarithmic transformations over the study period (2000-2011). As shown in Figure 11, the log-quantity trajectories for the treated and control groups move approximately in parallel

during the pre-treatment period, supporting the plausibility of the assumption. I examine



**Figure 11.** log-quantity evolution 2000-2011 (treated vs control group)

the impact of RGGI over the first three years of implementation (2009–2011), corresponding to the program’s initial compliance period. As shown in the dynamic estimates below, the policy induced a statistically significant and growing reduction in total electricity generation (see figure 12). These results show an 8.2% decline in total electricity generation in the first



**Figure 12.** GTT on total electricity production

year, increasing to 8.9% in the second year, and reaching 13.6% by the end of the initial compliance period in 2011. The growing effect suggests that RGGI not only reduced generation in the short term but also triggered accelerating adjustments over time, potentially through coal plant retirements, fuel switching, or improvements in demand-side efficiency. While total output fell, the composition of generation shifted, as I can see in table 6.

The table compares observed and counterfactual shares of electricity generation by source in RGGI states between 2009 and 2011, providing insight into how the policy reshaped the energy mix. Gas shares increased from 31.8% to 39.3%, consistently exceeding counterfactual values, suggesting that RGGI accelerated the substitution from coal to gas due

**Table 6.** Observed and Counterfactual Shares Across Three Years

Year	Category	Observed	Counterfactual
2009	Gas	0.3179 [0.3178, 0.3179]	0.3146 [0.3130, 0.3161]
	Coal & Oil	0.1954 [0.1953, 0.1954]	0.2123 [0.2113, 0.2132]
	Nuclear	0.3107 [0.3106, 0.3107]	0.2818 [0.2802, 0.2833]
	Renewables	0.1761 [0.1761, 0.1762]	0.1914 [0.1896, 0.1932]
2010	Gas	0.3594 [0.3593, 0.3594]	0.3215 [0.3199, 0.3231]
	Coal & Oil	0.1815 [0.1815, 0.1816]	0.2127 [0.2117, 0.2137]
	Nuclear	0.2982 [0.2982, 0.2983]	0.2716 [0.2701, 0.2731]
	Renewables	0.1609 [0.1608, 0.1609]	0.1942 [0.1924, 0.1960]
2011	Gas	0.3927 [0.3927, 0.3928]	0.3272 [0.3257, 0.3289]
	Coal & Oil	0.1322 [0.1322, 0.1322]	0.1965 [0.1956, 0.1974]
	Nuclear	0.2955 [0.2954, 0.2955]	0.2583 [0.2568, 0.2597]
	Renewables	0.1796 [0.1795, 0.1796]	0.2180 [0.2161, 0.2199]

to its carbon pricing incentives. In contrast, coal and oil shares dropped sharply, from 19.5% to 13.2%, compared to counterfactuals near 20%, indicating that, absent RGGI, these high-emission sources would have retained greater importance. Nuclear generation declined modestly but remained more stable than expected under the counterfactual, implying resilience as a low-carbon base source. Renewable shares increased slightly, though less than in the counterfactual scenario, suggesting that RGGI’s immediate effect on renewables was limited, likely due to longer investment lags. Overall, the results reflect a compositional reallocation within the generation simplex: probability mass shifted mainly from coal and oil toward gas, with minor adjustments across nuclear and renewables. These patterns indicate that the early RGGI-driven transition was led by natural gas rather than renewables.

Overall, RGGI effectively reduced coal and oil dependence, achieving substantial emissions reductions, as reflected in a 13.6% decline in total generation, but primarily through fuel switching to gas and increased nuclear output, rather than new renewable deployment. This highlights the role of existing infrastructure in shaping the composition of the clean energy

transition. It also reveals the importance of complementary policies, such as renewable subsidies, to accelerate the shift toward truly clean energy sources. The results are consistent with several other studies (see [Yan \(2021\)](#)).

## 5. CONCLUSION

This paper introduces compositional difference-in-differences (CoDiD), a causal inference framework for analyzing how treatments and policies affect entire categorical outcome quantities and distributions. It is especially suited to settings with discrete, unordered outcomes, such as employment status (employed, unemployed, out of the labor force), voting choice, or health categories, where the policy-relevant question is not whether a single average has changed, but how the composition of all categories has been reconfigured. CoDiD addresses a dual objective: it quantifies both absolute changes in total counts and relative shifts in shares (i.e., the probability mass function). This allows researchers and policymakers to identify which categories expanded or contracted, by how much, and in what combination of scale and composition. The framework is built on a parallel growths assumption, a multiplicative analog of standard parallel trends, which posits that, in the absence of treatment, the ratios of category sizes evolve proportionally over time. This assumption has a natural interpretation in random utility models: it corresponds to stable relative utility differences across categories, so that estimated effects map directly to shifts in underlying preferences. CoDiD is designed for practical use and naturally extends to richer settings, including relaxations of the identifying assumption, staggered treatment adoption, and a synthetic DiD analog for richer data contexts. A key direction for future work is the incorporation of continuous covariates, enabling adjustment for confounders and the estimation of heterogeneous treatment effects across subpopulations.

## REFERENCES

- J. Aitchison. The statistical analysis of compositional data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 44(2):139–160, 1982.
- J. Aitchison. Relative variation diagrams for describing patterns of compositional variability. *Mathematical Geology*, 22(4):487–511, 1990.
- J. Aitchison. On criteria for measures of compositional difference. *Mathematical Geology*, 24(4):365–379, 1992.
- J. Aitchison. Simplicial inference. In M. A. G. Viana and D. S. P. Richards, editors, *Algebraic Methods in Statistics and Probability*, volume 287 of *Contemporary Mathematics*, pages 1–22. American Mathematical Society, Providence, Rhode Island, 2002.
- D. Arkhangelsky, S. Athey, D. A. Hirshberg, G. W. Imbens, and S. Wager. Synthetic difference-in-differences. *American Economic Review*, 111(12):4088–4118, 2021.
- K. F. Arnold, L. Berrie, P. W. Tennant, and M. S. Gilthorpe. A causal inference perspective on the analysis of compositional data. *International journal of epidemiology*, 49(4):1307–1313, 2020.
- S. Athey and G. W. Imbens. Identification and inference in nonlinear difference-in-differences models. *Econometrica*, 74(2):431–497, 2006.
- A. Baker, B. Callaway, S. Cunningham, A. Goodman-Bacon, and P. H. Sant’Anna. Difference-in-differences designs: A practitioner’s guide. *arXiv preprint arXiv:2503.13323*, 2025.
- K. Ban and D. Kédagni. Robust difference-in-differences models. *arXiv preprint arXiv:2211.06710*, 2022.
- C. Barceló-Vidal, J. A. Martín-Fernández, and V. Pawlowsky-Glahn. Mathematical foundations of compositional data analysis. In G. Ross, editor, *Proceedings of the Sixth Annual Conference of the International Association for Mathematical Geology*, Cancun, Mexico, 2001. CD-ROM.
- S. T. Berry, C. Cox, and P. Haile. Selective turnout, voting policy, and partisan bias: Evidence from multi-level data. Technical report, National Bureau of Economic Research, 2025.
- D. Billheimer, P. Guttorp, and W. F. Fagan. Statistical interpretation of species composition. *Journal of the American Statistical Association*, 96(456):1205–1214, 2001.

- S. Bonhomme and U. Sauder. Recovering distributions in difference-in-differences models: A comparison of selective and comprehensive schooling. *The Review of Economics and Statistics*, 93(2):479–494, 2011. ISSN 00346535, 15309142.
- G. Cafri, W. Wang, P. H. Chan, and P. C. Austin. A review and empirical comparison of causal inference methods for clustered observational data with application to the evaluation of the effectiveness of medical devices. *Statistical Methods in Medical Research*, 28(10-11):3142–3162, 2019.
- B. Callaway and T. Li. Quantile treatment effects in difference in differences models with panel data. *Quantitative Economics*, 10(4):1579–1618, 2019.
- B. Callaway and P. H. Sant’Anna. Difference-in-differences with multiple time periods. *Journal of Econometrics*, 2020. ISSN 0304-4076.
- B. Callaway and P. H. Sant’Anna. Difference-in-differences with multiple time periods. *Journal of econometrics*, 225(2):200–230, 2021.
- B. Callaway, T. Li, and T. Oka. Quantile treatment effects in difference in differences models under dependence restrictions and with only two time periods. *Journal of Econometrics*, 206(2):395–413, 2018.
- C. De Chaisemartin and X. d’Haultfoeuille. Two-way fixed effects and differences-in-differences with heterogeneous treatment effects: A survey. *The econometrics journal*, 26(3):C1–C30, 2023.
- J. J. Egozcue, V. Pawlowsky-Glahn, G. Mateu-Figueras, and C. Barcelo-Vidal. Isometric logratio transformations for compositional data analysis. *Mathematical geology*, 35(3):279–300, 2003.
- D. Ghanem, D. Kédagni, and I. Mourifié. Evaluating the impact of regulatory policies on social welfare in difference-in-difference settings. *arXiv preprint arXiv:2306.04494*, 2023.
- J. A. Graves, C. Fry, J. M. McWilliams, and L. A. Hatfield. Difference-in-differences for categorical outcomes. *Health Services Research*, 57(3):681–692, 2022.
- M. J. Hamilton. Chapter 1 lie groups and lie algebras: Basic concepts. In *Mathematical Gauge Theory: With Applications to the Standard Model of Particle Physics*, pages 3–82. Springer, 2017.
- T. Havnes and M. Mogstad. Is universal child care leveling the playing field? *Journal of Public Economics*, 127:100–114, 2015. ISSN 0047-2727. The Nordic Model.

- J. L. Horowitz. Bootstrap methods in econometrics. *Annual Review of Economics*, 11(1):193–224, 2019.
- M. Lechner. The estimation of causal effects by difference-in-difference methods. *Foundations and Trends® in Econometrics*, 4(3):165–224, 2011. ISSN 1551-3076.
- C. F. Manski and J. V. Pepper. How do right-to-carry laws affect crime rates? coping with ambiguity using bounded-variation assumptions. *Review of Economics and Statistics*, 100(2):232–244, 2018.
- D. McFadden. Conditional logit analysis of qualitative choice behavior. 1972.
- D. McFadden. The measurement of urban travel demand. *Journal of public economics*, 3(4):303–328, 1974.
- D. McFadden. Modelling the choice of residential location. 1977.
- B. D. Meyer, W. K. Viscusi, and D. Durbin. Workers’ compensation and injury duration: evidence from a natural experiment, 1990.
- A. Rambachan and J. Roth. An honest approach to parallel trends. Working Paper, 2020.
- J. Roth and P. Sant’Anna. When is parallel trends sensitive to functional form? Unpublished Manuscript, 2021.
- J. Roth, P. H. Sant’Anna, A. Bilinski, and J. Poe. What’s trending in difference-in-differences? a synthesis of the recent econometrics literature. *Journal of Econometrics*, 235(2):2218–2244, 2023.
- E. J. Tchetgen Tchetgen, C. Park, and D. B. Richardson. Universal difference-in-differences for causal inference in epidemiology. *Epidemiology*, 35(1), 2024.
- W. Torous, F. Gunsilius, and P. Rigollet. An optimal transport approach to causal inference. *arXiv preprint arXiv:2108.05858*, 2021.
- K. E. Train. *Discrete choice methods with simulation*. Cambridge university press, 2009.
- J. M. Wooldridge. Simple approaches to nonlinear difference-in-differences with panel data. *The Econometrics Journal*, 26(3):C31–C66, 2023.
- J. Yan. The impact of climate policy on fossil fuel consumption: Evidence from the regional greenhouse gas initiative (rggi). *Energy Economics*, 100:105333, 2021.
- Y. Zhou, D. Kurisu, T. Otsu, and H.-G. Müller. Geodesic difference-in-differences. *arXiv preprint arXiv:2501.17436*, 2025.

## APPENDIX A. PROOFS OF THE RESULTS IN THE MAIN TEXT

**A.1. Proof of Theorem 1.** Start from assumption 2:

$$\log(q_{1,1}^N) - \log(q_{1,0}^N) = \log(q_{0,1}^N) - \log(q_{0,0}^N).$$

Add  $\log(q_{1,0}^N)$  to both sides:

$$\log(q_{1,1}^N) = \log(q_{1,0}^N) + \log(q_{0,1}^N) - \log(q_{0,0}^N). \quad (1)$$

Since all vectors are strictly positive (by Assumption 1), the logarithm is defined component-wise, and I can apply the component-wise exponential function  $\exp$  to both sides of (1). Because  $\exp$  and  $\log$  are inverses, I obtain:

$$q_{1,1}^N = \exp(\log(q_{1,0}^N) + \log(q_{0,1}^N) - \log(q_{0,0}^N)).$$

By definition (implied by the structure of the model and assumption 1, the scalar  $S_{1,1}^N$  is the sum of the components of the vector  $q_{1,1}^N$ . Let the components be indexed by  $c_k$ , for  $k = 1, \dots, p$ . Then:

$$S_{1,1}^N = \sum_{k=1}^p q_{1,1}^N(c_k) = \sum_{k=1}^p \exp(\log q_{1,0}^N(c_k) + \log q_{0,1}^N(c_k) - \log q_{0,0}^N(c_k)),$$

By definition (again, standard in compositional or share-based models),

$$\pi_{1,1}^N = \frac{q_{1,1}^N}{S_{1,1}^N},$$

where the division is component-wise (i.e., normalizing the vector  $q_{1,1}^N$  to sum to 1). Since both numerator and denominator are identified in Steps 1 and 2,  $\pi_{1,1}^N$  is identified as stated.

**A.2. Proof of proposition 1.** we have:

$$U_{gt}^N(c_k) = \mu_{gt}^N(c_k) - \log\left(\sum_k e^{\mu_{gt}^N(c_k)}\right) + \log(S_{gt}^N) + \varepsilon_{gt}(c_k)$$

In the standard multinomial logit RUM:

$$\pi_{gt}(c_k) = \Pr\left(U_{gt}^N(c_k) \geq U_{gt}^N(c_j) \text{ for all } j \neq k\right).$$

Since  $\varepsilon_{gt}(c_k) \sim \text{i.i.d. Gumbel}$ , the choice probability has the closed-form logit formula:

$$\pi_{gt}(c_k) = \frac{\exp\left(\mu_{gt}^N(c_k) + \log(S_{gt}^N) - \log(\sum_j e^{\mu_{gt}^N(c_j)})\right)}{\sum_j \exp\left(\mu_{gt}^N(c_j) + \log(S_{gt}^N) - \log(\sum_l e^{\mu_{gt}^N(c_l)})\right)}.$$

This simplifies to:

$$\pi_{gt}(c_k) = \frac{e^{\mu_{gt}^N(c_k)} \cdot \frac{S_{gt}^N}{\sum_l e^{\mu_{gt}^N(c_l)}}}{\sum_j e^{\mu_{gt}^N(c_j)} \cdot \frac{S_{gt}^N}{\sum_l e^{\mu_{gt}^N(c_l)}}} = \frac{e^{\mu_{gt}^N(c_k)}}{\sum_j e^{\mu_{gt}^N(c_j)}}.$$

The  $S_{gt}^N$  term cancels out in the probability, which shows that the relative shares depend only on the relative utilities, while  $S_{gt}^N$  scales the absolute number of choices. The number of individuals choosing  $c_k$  is:

$$q_{gt}^N(c_k) = \pi_{gt}(c_k) \cdot S_{gt}^N = \frac{e^{\mu_{gt}^N(c_k)}}{\sum_j e^{\mu_{gt}^N(c_j)}} \cdot S_{gt}^N.$$

This leads to:

$$\log q_{gt}^N(c_k) = \mu_{gt}^N(c_k) - \log \left( \sum_{k=1}^p e^{\mu_{gt}^N(c_k)} \right) + \log(S_{gt}^N).$$

Now, we have :

$$\begin{aligned} \mathbb{E}[U_{g1}(c_k)] - \mathbb{E}[U_{g0}(c_k)] &= \mu_{g1}^N(c_k) - \log \left( \sum_k e^{\mu_{g1}^N(c_k)} \right) + \log(S_{g1}^N) + \gamma \\ &\quad - \mu_{g0}^N(c_k) - \log \left( \sum_k e^{\mu_{g0}^N(c_k)} \right) + \log(S_{g0}^N) - \gamma \\ &= \log q_{g1}^N(c_k) - \log q_{g0}^N(c_k) \end{aligned}$$

From there, we have that assumption 2 is equivalent to parallel trends in expected utilities:

$$\mathbb{E}[U_{11}^N(c_k)] - \mathbb{E}[U_{10}^N(c_k)] = \mathbb{E}[U_{01}^N(c_k)] - \mathbb{E}[U_{00}^N(c_k)], \quad \forall c_k \in \bar{\mathcal{Y}}.$$

**A.3. Proof of Proposition 2.** Pick a baseline category  $c_p$ . The log-odds for each category can be written as :

$$\log \left( \frac{\pi_{g,t}^N(c_k)}{\pi_{g,t}^N(c_p)} \right) = \log \left( \frac{q_{g,t}^N(c_k)}{S_{g,t}^N} \times \frac{S_{g,t}^N}{q_{g,t}^N(c_p)} \right) = \log(q_{g,t}^N(c_k)) - \log(q_{g,t}^N(c_p)).$$

Parallel growths implies hold for each category, so we have :

$$\log(q_{1,1}^N(c_k)) - \log(q_{1,0}^N(c_k)) = \log(q_{0,1}^N(c_k)) - \log(q_{0,0}^N(c_k)),$$

and also

$$\log(q_{1,1}^N(c_p)) - \log(q_{1,0}^N(c_p)) = \log(q_{0,1}^N(c_p)) - \log(q_{0,0}^N(c_p)).$$

Taking the difference between the two equalities gives:

$$\begin{aligned} \{ \log(q_{1,1}^N(c_k)) - \log(q_{1,1}^N(c_p)) \} - \{ \log(q_{1,0}^N(c_k)) - \log(q_{1,0}^N(c_p)) \} &= \\ \{ \log(q_{0,1}^N(c_k)) - \log(q_{0,1}^N(c_p)) \} - \{ \log(q_{0,0}^N(c_k)) - \log(q_{0,0}^N(c_p)) \} &= \end{aligned}$$

This is exactly parallel trends in log-odds :

$$\log \left( \frac{\pi_{1,1}^N(c_k)}{\pi_{1,1}^N(c_p)} \right) - \log \left( \frac{\pi_{0,1}^N(c_k)}{\pi_{0,1}^N(c_p)} \right) = \log \left( \frac{\pi_{1,0}^N(c_k)}{\pi_{1,0}^N(c_p)} \right) - \log \left( \frac{\pi_{0,0}^N(c_k)}{\pi_{0,0}^N(c_p)} \right).$$

From part 1, I also can derive that:

$$\begin{aligned} \{ \log(q_{1,1}^N(c_k)) - \log(q_{1,1}^N(c_l)) \} - \{ \log(q_{1,0}^N(c_k)) - \log(q_{1,0}^N(c_l)) \} = \\ \{ \log(q_{0,1}^N(c_l)) - \log(q_{0,1}^N(c_l)) \} - \{ \log(q_{0,0}^N(c_k)) - \log(q_{0,0}^N(c_l)) \} \end{aligned}$$

is equivalent to

$$\begin{aligned} \{ \mathbb{E}[U_{11}^N(c_k)] - \mathbb{E}[U_{11}^N(c_l)] \} - \{ \mathbb{E}[U_{10}^N(c_k)] - \mathbb{E}[U_{10}^N(c_l)] \} = \\ \{ \mathbb{E}[U_{01}^N(c_k)] - \mathbb{E}[U_{01}^N(c_l)] \} - \{ \mathbb{E}[U_{00}^N(c_k)] - \mathbb{E}[U_{00}^N(c_l)] \}, \quad \forall c_k, c_l \in \bar{\mathcal{Y}}. \end{aligned} \quad (18)$$

**A.4. Proof of Proposition 3.** I aim to show that the parallel growths assumption in compositional form is equivalent to a parallel trends condition in log-odds. Specifically, I prove the equivalence of the following two statements:

$$(i) \quad \pi_{1,1}^N \ominus \pi_{0,1}^N = \pi_{1,0}^N \ominus \pi_{0,0}^N$$

$$(ii) \quad \text{for all } k = 1, \dots, p-1, \log \left( \frac{\pi_{1,1}^N(c_k)}{\pi_{1,1}^N(c_p)} \right) - \log \left( \frac{\pi_{0,1}^N(c_k)}{\pi_{0,1}^N(c_p)} \right) = \log \left( \frac{\pi_{1,0}^N(c_k)}{\pi_{1,0}^N(c_p)} \right) - \log \left( \frac{\pi_{0,0}^N(c_k)}{\pi_{0,0}^N(c_p)} \right).$$

Here, all share vectors  $\pi$  belong to the interior of the  $(p-1)$ -dimensional simplex  $\mathcal{S}^{p-1}$ , and the operation  $\ominus$  denotes compositional subtraction (perturbation inverse), defined for any two compositions  $\pi_1, \pi_2 \in \mathcal{S}^{p-1}$  as:

$$\pi_1 \ominus \pi_2 := \frac{1}{\sum_{k=1}^p \frac{\pi_1(c_k)}{\pi_2(c_k)}} \left( \frac{\pi_1(c_1)}{\pi_2(c_1)}, \dots, \frac{\pi_1(c_p)}{\pi_2(c_p)} \right).$$

I know that  $\ell : \text{int}(\mathcal{S}^{p-1}) \rightarrow \mathbb{R}^{p-1}$  defined by:

$$\ell(\pi) = \left[ \log \left( \frac{\pi(c_1)}{\pi(c_p)} \right), \dots, \log \left( \frac{\pi(c_{p-1})}{\pi(c_p)} \right) \right].$$

I already know that this map is a bijection. A key property of that transformation is that it linearizes the Aitchison geometry of the simplex. In particular, for any  $\pi_1, \pi_2 \in \text{int}(\mathcal{S}^{p-1})$ ,

$$\ell(\pi_1 \ominus \pi_2) = \ell(\pi_1) - \ell(\pi_2).$$

By definition of  $\ominus$ , the  $k$ -th component of  $\pi_1 \ominus \pi_2$  is proportional to  $\pi_1(c_k)/\pi_2(c_k)$ .

$$\begin{aligned}\ell_k(\pi_1 \ominus \pi_2) &= \log \left( \frac{(\pi_1 \ominus \pi_2)(c_k)}{(\pi_1 \ominus \pi_2)(c_p)} \right) \\ &= \log \left( \frac{\pi_1(c_k)/\pi_2(c_k)}{\pi_1(c_p)/\pi_2(c_p)} \right) \\ &= \log \left( \frac{\pi_1(c_k)}{\pi_1(c_p)} \right) - \log \left( \frac{\pi_2(c_k)}{\pi_2(c_p)} \right) \\ &= \ell_k(\pi_1) - \ell_k(\pi_2).\end{aligned}$$

Because  $\ell$  is bijective, we have:

$$\pi_{1,1}^N \ominus \pi_{0,1}^N = \pi_{1,0}^N \ominus \pi_{0,0}^N \iff \ell(\pi_{1,1}^N \ominus \pi_{0,1}^N) = \ell(\pi_{1,0}^N \ominus \pi_{0,0}^N).$$

Using the linearity property, this becomes:

$$\ell(\pi_{1,1}^N) - \ell(\pi_{0,1}^N) = \ell(\pi_{1,0}^N) - \ell(\pi_{0,0}^N),$$

which holds if and only if the equality in (ii) holds component-wise for all  $k = 1, \dots, p-1$ .

**A.5. Proof of Theorem 2.** Define the log cross-group ratio at time  $t$  for category  $c_k$ :

$$\Delta_t^k := \log \left( \frac{q_{1,t}^N(c_k)}{q_{0,t}^N(c_k)} \right) = \log(q_{1,t}^N(c_k)) - \log(q_{0,t}^N(c_k)).$$

The unobserved quantity of interest is  $q_{1,1}^N(c_k)$ , or equivalently  $\Delta_1^k$ , since:

$$q_{1,1}^N(c_k) = \exp(\Delta_1^k + \log(q_{0,1}^N(c_k))). \quad (1)$$

Thus, bounding  $\Delta_1^k$  yields bounds on  $q_{1,1}^N(c_k)$ . Assumption 3 states:

$$\Delta_1^k \in \text{Conv} \left\{ \min_{t<0} \omega_t \Delta_t^k, \max_{t<0} \omega_t \Delta_t^k \right\} = [\underline{\Delta}_\omega^k, \bar{\Delta}_\omega^k],$$

where

$$\underline{\Delta}_\omega^k := \min_{t<0} \omega_t \Delta_t^k, \quad \bar{\Delta}_\omega^k := \max_{t<0} \omega_t \Delta_t^k.$$

Hence:

$$b_{\min}^k = \exp(\underline{\Delta}_\omega^k + \log(q_{0,1}^N(c_k))) = \exp \left( \min_{t<0} \omega_t \log \left( \frac{q_{1,t}^N(c_k)}{q_{0,t}^N(c_k)} \right) + \log(q_{0,1}^N(c_k)) \right),$$

$$b_{\max}^k = \exp(\bar{\Delta}_\omega^k + \log(q_{0,1}^N(c_k))) = \exp \left( \max_{t<0} \omega_t \log \left( \frac{q_{1,t}^N(c_k)}{q_{0,t}^N(c_k)} \right) + \log(q_{0,1}^N(c_k)) \right).$$

We have that:  $q_{1,1}^N(c_k) \in [b_{\min}^k, b_{\max}^k]$ . The counterfactual total is  $S_{1,1}^N = \sum_{k=1}^p q_{1,1}^N(c_k) \in [\sum_k b_{\min}^k, \sum_k b_{\max}^k]$ . Let  $\pi_{1,1}^I = q_{1,1}^I / S_{1,1}^I$  be the observed share vector (in the simplex  $\mathcal{S}^{p-1}$ ). The counterfactual share is  $\pi_{1,1}^N = q_{1,1}^N / S_{1,1}^N$ . The set of possible  $\pi_{1,1}^N$  is:

$$\left\{ \pi \in \mathcal{S}^{p-1} : \pi_k = \frac{q_k}{\sum_j q_j}, q_k \in [b_{\min}^k, b_{\max}^k] \text{ for all } k \right\}.$$

**A.6. Proof of Theorem 3.** Strategy 1: Use the never-treated group ( $G_\infty$ ) as control , it requires assumption 5: parallel growth (in logs) between each cohort  $G_g$  and the never-treated group  $G_\infty$ .

Strategy 2: Use not-yet-treated groups as controls, it requires assumption 6: parallel growth (in logs) between cohort  $G_g$  and any cohort  $G_s$  that is not yet treated at time  $t$  (i.e.,  $s > t$ ).

Because  $\mathcal{G}$  excludes  $\bar{g}$ , for every  $g \in \mathcal{G}$  and  $t \geq g$ , there exists at least one  $s > t$  such that  $G_s$  is not yet treated at time  $t$  (since  $s \leq \bar{g}$  and  $\bar{g} > g$ ). Assumption 5 states: for all  $g$  and  $t \geq g$ ,

$$\log q_{G_g,t}^N - \log q_{G_\infty,t}^N = \log q_{G_g,t-1}^N - \log q_{G_\infty,t-1}^N.$$

This implies the difference is constant over time for  $t \geq g$ . In particular, it equals its value at  $t = g - 1$  (the last pre-treatment period):

$$\log q_{G_g,t}^N - \log q_{G_\infty,t}^N = \log q_{G_g,g-1}^N - \log q_{G_\infty,g-1}^N.$$

Rearranging:

$$\log q_{G_g,t}^N = \log q_{G_g,g-1}^N + \log q_{G_\infty,t}^N - \log q_{G_\infty,g-1}^N.$$

Exponentiating both sides gives the result. Assumption 6 states: for all  $g \in \mathcal{G}$ , and for all  $s, t$  such that  $t \geq g$  and  $t < s \leq \bar{g}$ ,

$$\log q_{G_g,t}^N - \log q_{G_s,t}^N = \log q_{G_g,t-1}^N - \log q_{G_s,t-1}^N.$$

Again, this implies the log-difference is time-invariant from period  $g - 1$  onward (since  $G_s$  is untreated at all times  $\leq t < s$ , so  $q_{G_s,\tau}^N = q_{G_s,\tau}^I$  is observed for  $\tau \leq t$ ). Thus, for any  $t \geq g$  and any valid  $s > t$ ,

$$\log q_{G_g,t}^N - \log q_{G_s,t}^N = \log q_{G_g,g-1}^N - \log q_{G_s,g-1}^N.$$

Rearranging and exponentiating gives the result.

## APPENDIX B. ADDING COVARIATES

This section extends the main identification results to incorporate covariates. Let  $X \in \mathcal{X}$  denote a vector of covariates, which may be endogenous. For this analysis, I restrict my attention to the case where  $X$  is discrete. In this setting, I can redefine all relevant quantities by stratifying on  $X$ . This yields a set of new estimands, each conditional on a specific covariate profile  $x$ . Provided that the untreated potential quantities  $q_{g,t}^N(x)$  are strictly positive for all groups, time periods, and values  $x \in \mathcal{X}$ , I can formulate the following conditional assumption:

**Assumption 7.** (*Conditional parallel growths*)

$$\log(q_{1,1,x}^N) - \log(q_{1,0,x}^N) = \log(q_{0,1,x}^N) - \log(q_{0,0,x}^N).$$

The conditional parallel growths assumption is often more plausible in practice. The corresponding counterfactuals are derived in the same way as in the unconditional case.

$$q_{1,1,x}^N = \exp(\log(q_{1,0,x}^N) + \log(q_{0,1,x}^N) - \log(q_{0,0,x}^N)) \quad (19)$$

$$q_{1,1}^N = \sum_{x \in \mathcal{X}} \mathbb{P}(X = x) q_{1,1,x}^N \quad (20)$$

And the corresponding counterfactual probability mass can be obtained by: for each category  $c_k, k = 1, \dots, p$

$$\pi_{1,1,x}^N(c_k) = \frac{\exp(\log(q_{1,0,x}^N(c_k)) + \log(q_{0,1,x}^N(c_k)) - \log(q_{0,0,x}^N(c_k)))}{\sum_{j=1}^p \exp(\log(q_{1,0,x}^N(c_j)) + \log(q_{0,1,x}^N(c_j)) - \log(q_{0,0,x}^N(c_j)))} \quad (21)$$

$$\pi_{1,1}^N(c_k) = \sum_{x \in \mathcal{X}} \mathbb{P}(X = x) \pi_{1,1,x}^N(c_k) \quad (22)$$

The main challenge arises when  $X$  includes continuous variables, as it becomes necessary to model or approximate the conditional distribution in a flexible way, rather than relying on simple stratification. I do not explicitly discuss this case here.

## APPENDIX C. IMPLEMENTATION OF CoDiD AND INFERENCE PROCEDURE

For each group  $g$  and each time period  $t$ , we have a random sample of  $q_{g,t}$  individuals. Define  $\hat{q}_{g,t}(c_k)$  as the number of individuals in group  $g$  at time  $t$  in category  $c_k$ . The estimated counterfactual is given by:

$$\hat{q}_{1,1}^N = \exp \left( \log \left( \hat{q}_{1,0}^N \right) + \log \left( \hat{q}_{0,1}^N \right) - \log \left( \hat{q}_{0,0}^N \right) \right)$$

The total and shares are obtained using the estimated counterfactual quantities. The estimated treatment effects are given by:

$$G\hat{T}T(c_k) = \frac{\hat{q}_{1,1}^I(c_k)}{\hat{q}_{1,1}^N(c_k)} - 1$$

$$G\hat{T}T = \frac{\hat{S}_{1,1}^I}{\hat{S}_{1,1}^N} - 1$$

$$C\hat{T}T = \hat{\pi}_{1,1}^I \ominus \hat{\pi}_{1,1}^N$$

For inference, I use the parametric bootstrap to approximate the sampling distribution of the parameters. The parametric bootstrap is a resampling technique designed for situations where the data-generating process can be modeled parametrically. Instead of drawing resamples directly from the observed data, the procedure assumes a parametric model for the data and uses estimated parameters to generate new synthetic samples. For a detailed treatment of the method, see [Horowitz \(2019\)](#). In my case, the procedure proceeds as follows. For each bootstrap iteration  $b = 1, \dots, B$ , where  $B$  is the total number of iterations, we have:

- Resample quantities for each group-time pair  $(g, t)$  using a multinomial distribution:

$$q_{g,t}^{(b)}(c_1, \dots, c_p) \sim \text{Multinomial} (S_{gt}, \hat{\pi}_{g,t}(c_1), \dots, \hat{\pi}_{g,t}(c_p)) .$$

- Compute counterfactual total quantities and category shares based on the bootstrap samples.
- Calculate the bootstrap treatment effects:  $G\hat{T}T^b$  and  $C\hat{T}T^b$ .

The 95% confidence interval for each category  $c_k$  is then obtained from the 2.5% and 97.5% quantiles of the bootstrap distribution in each case.

## APPENDIX D. GENERAL RANDOM UTILITY MODEL WITHOUT IIA

Now, assume that the error terms jointly follow a Generalized Extreme Value (GEV) distribution,

$$F(\varepsilon_1, \dots, \varepsilon_p) = \exp \left[ -G(e^{-\varepsilon_1}, \dots, e^{-\varepsilon_p}) \right],$$

where  $G(\cdot)$  is a homogeneous generating function that determines the correlation structure among the unobserved utility shocks. When  $G(y_1, \dots, y_p) = \sum_{k=1}^p y_k$ , the shocks are independent Gumbel, and the model collapses to the multinomial logit case. More general choices of  $G$  (such as those defining the nested or cross-nested logit models) allow correlation across alternatives, thereby relaxing the independence of irrelevant alternatives (IIA) assumption while preserving closed-form choice probabilities. Train (2009) gives us different possible examples for  $G$  as shown in table 7. Consider the same model (modified to account for correlation between error terms) :

$$U_{gt}^N(c_k) = V_{gt}^N(c_k) + \varepsilon_{gt}(c_k) = \mu_{gt}^N(c_k) - \log(G(\mu_{gt}^N)) + \log(S_{gt}^N) + \varepsilon_{gt}(c_k)$$

**Table 7.** Examples of GEV Models

Model	$G(\cdot)$	$\pi^N(c_k)$
Logit	$\sum_{j=1}^p e^{V^N(c_j)}$	$\frac{e^{V^N(c_k)}}{\sum_j e^{V^N(c_j)}}$
Nested Logit	$\sum_{m=1}^M \left( \sum_{j \in \mathcal{B}_m} e^{V^N(c_j)/\lambda_m} \right)^{\lambda_m}$	$\frac{e^{V^N(c_k)/\lambda_m} \left( \sum_{j \in \mathcal{B}_m} e^{V^N(c_j)/\lambda_m} \right)^{\lambda_m-1}}{\sum_r \left( \sum_{j \in \mathcal{B}_r} e^{V^N(c_j)/\lambda_r} \right)^{\lambda_r}}$
PCL	$\sum_{r < s} \left( e^{V^N(c_r)/\lambda_{rs}} + e^{V^N(c_s)/\lambda_{rs}} \right)^{\lambda_{rs}}$	$\frac{\sum_{j \neq k} e^{V^N(c_k)/\lambda_{kj}} \left( e^{V^N(c_k)/\lambda_{kj}} + e^{V^N(c_j)/\lambda_{kj}} \right)^{\lambda_{kj}-1}}{\sum_{r < s} \left( e^{V^N(c_r)/\lambda_{rs}} + e^{V^N(c_s)/\lambda_{rs}} \right)^{\lambda_{rs}}}$
GNL	$\sum_{m=1}^M \left( \sum_{j \in \mathcal{B}_m} (\alpha_{jm} e^{V^N(c_j)})^{1/\lambda_m} \right)^{\lambda_m}$	$\frac{(\alpha_{km} e^{V^N(c_k)})^{1/\lambda_m} \left( \sum_{j \in \mathcal{B}_m} (\alpha_{jm} e^{V^N(c_j)})^{1/\lambda_m} \right)^{\lambda_m-1}}{\sum_r \left( \sum_{j \in \mathcal{B}_r} (\alpha_{jr} e^{V^N(c_j)})^{1/\lambda_r} \right)^{\lambda_r}}$

Let  $G_k(\mathbf{y}) = \frac{\partial G(\mathbf{y})}{\partial y_k}$ , with  $\mathbf{y} = (y_1, \dots, y_p)$ . This derivative is crucial in determining how likely an individual is to choose alternative  $c_k$  under the GEV model. If we evaluate it at  $\mathbf{y} = \mathbf{e}^{\boldsymbol{\mu}^N} = (e^{V^N(c_1)}, \dots, e^{V^N(c_p)})$ , the exponentiated systematic utilities, the term  $G_k(\mathbf{e}^{\boldsymbol{\mu}^N})$  tells us how the attractiveness of option  $c_k$  depends on the attractiveness of all the other

available options. It adjusts for the fact that some choices are more closely related than others. For example, if two alternatives are close substitutes, a high utility for one will reduce the probability of choosing the other. Conversely, if options are weakly related, changing one has little effect on the probability of picking the other. Now, consider the following assumption: from the RUM ([McFadden \(1977\)](#)), we have:

$$\pi_{gt}(c_k) = \frac{\exp(V_{gt}^N(c_k))G_k(\mathbf{V}_{\text{gt}}^N)}{G(\mathbf{V}_{\text{gt}}^N)} = \frac{\exp(\mu_{gt}^N(c_k))G_k(\mu_{\text{gt}}^N)}{G(\mu_{\text{gt}}^N)}$$

$$q_{gt}^N(c_k) = \pi_{gt}(c_k) \cdot S_{gt}^N = \frac{\exp(\mu_{gt}^N(c_k))G_k(\mu_{\text{gt}}^N)}{G(\mu_{\text{gt}}^N)} \cdot S_{gt}^N.$$

This leads to:

$$\log q_{gt}^N(c_k) = \mu_{gt}^N(c_k) - \log(G(\mu_{\text{gt}}^N)) + \log(S_{gt}^N) + \log(G_k(\mu_{\text{gt}}^N))$$

Here, we need an extra assumption which is :

$$\log(G_k(\mu_{11}^N)) - \log(G_k(\mu_{10}^N)) = \log(G_k(\mu_{01}^N)) - \log(G_k(\mu_{00}^N))$$

This assumption is trivially satisfied in the multinomial logit case. With this additional assumption, we obtain that that assumption 2 is equivalent to parallel trends in expected utilities. The result for utilities differences also follow.