THE UNIVERSITY OF CHICAGO

Ethical Sourcing and Consumer Demand for Coffee

By

Adam Wu

June, 2024

A paper submitted in partial fulfillment of the requirements for
the Master of Arts degree in the Master of Arts in
Computational Social Science

Faculty Advisor: Kirill Ponomarev

Preceptor: Joseph Hardwick

I am grateful to Kirill Ponomarev and Joseph Hardwick for their continued guidance and support.

Contents

Abstract

We analyze the impact of product labels indicative of ethical sourcing on consumer demand for coffee beans. The paper complements an existing field experiment by re-analyzing the empirical setting under weak assumptions with a more granular level of study which is subject to selection bias. In contrast to standard assumptions on consumer behavior in discrete choice models, this paper takes a fully nonparametric difference-in-differences approach to flexibly capture heterogeneous treatment effects and substitution patterns using spillovers. We also propose plausible structural bounds for partial identification as treatments can switch off in a multi-period setting. We derive semiparametric estimators based on the moment conditions to avoid the negative weighting problems associated with standard fixed effects models under heterogeneity. However by including certain covariates, we show that average effects may be estimated by a simple linear model on specific subsets of the data that achieve balanced propensity scores. Overall we find similar results as existing work, and in particular, the average consumer is sensitive to prices and unwilling to pay a price premium for ethically sourced coffees.

JEL Classification: D12; C1

Keywords: ethical sourcing; heterogeneous treatment effects; difference-in-differences

1 Introduction

A rapidly growing trend in consumer markets are ethically sourced products, where consumers derive value from not just the good but the manner in which the good was produced. In particular, some consumers may derive utility from goods which are produced in an environmentally or socially responsible way that is sustainable in the long-run. To differentiate between how goods are produced, a number of certification organizations have emerged in the last few decades that provide credible signals through product labels. One of the most widely recognized initiatives is the Fair Trade certification, which aims to promote long-term sustainability by ensuring profitability for farmers in developing countries in addition to enacting various environmental and social standards such as the prohibition of exploitative child labor. The most well-established product traded on Fair Trade networks is coffee, which is the empirical focus of this paper.

In the coffee industry, the majority of production is done by small family-run farms in Latin America, Southeast Asia, and Africa. One of the most prized coffee varieties are those from Ethiopia where the coffee plant *Coffea arabica* originates, which is also one of the poorest

nations in the world. Due to the significantly higher purchasing power of wealthier nations, production and exports of coffee beans are a lucrative industry that plays a key role in sustaining the economy of these developing countries. However as a result, the global supply of coffee beans grows much faster than demand. Downward price pressures along with price shocks, uncertain weather patterns, and poor soil conditions due to over-exploitation can make coffee production unsustainable for basic living needs of the farmers and workers in poor countries. Méndez et al. (2010) surveys 469 coffee-producing households in Nicaragua, Guatemala, El Salvador, and Mexico, and found that 63% of those surveyed struggle to meet basic food needs every year. Bacon et al. (2008) found similar results, with 69% of the households surveyed in Nicaragua reporting food insecurity.

The primary mechanism of Fair Trade is the enactment of a guaranteed price floor that is intended to cover the average costs of production and a livable minimum wage for farmers. This also offers protection against volatile prices in the event of uncertain weather or supply shocks, and discourages poor environmental practices such as the usage of certain agricultural chemicals that may increase production in the short-run but have detrimental effects in the long-run. In addition, Fair Trade-certified products also carry a fixed price premium, where local communities then democratically allocate the funds to invest in projects and infrastructure such as the development of schools, housing, hospitals, and roads that provide positive externalities to those outside the coffee industry. Fair Trade producers also enter in long-term contracts with buyers and have access to financing for greater economic stability.

Since the rapid growth of modern Fair Trade networks in the last two decades, there is an extensive and debated theoretical literature on the long-run sustainability of Fair Trade markets. However, empirical evidence on the socioeconomic impacts of Fair Trade is limited and often face difficult identification challenges due to selection into Fair Trade certification by the producers. Existing studies indicate a strong positive relationship between Fair Trade certification and greater incomes, education, access to financing, and more environmentally friendly practices by the producers (Arnould, Plastina, and Ball 2009; Jaffee 2008; Bacon et al. 2008). The majority of empirical studies understandably focus on the socioeconomic effect of Fair Trade certification on the producers.

In this paper, we focus on the effect of Fair Trade certification on the consumers. Since the defining features of Fair Trade are its enacted price floor and social premiums, for the mechanism to function effectively consumers must be willing and able to pay more for Fair Trade-certified products. We investigate whether this is the case in practice.

Existing survey evidence suggest the majority of consumers prefer and are willing to pay more for ethically sourced products. For example, the National Bureau of Economic Re-

search found in 1999 that 80% of those surveyed stated they would be willing to pay more for goods made under ethical working conditions (Elliott and Freeman 2003). Hertel, Scruggs, and Heidkamp (2009) finds that over 75% of those surveyed would be willing to pay at least an additional 50 cents per pound for Fair Trade coffee, and over 50% would be willing to pay at least a dollar more per pound. There is an extensive literature in microeconomic theory and behavioral economics that debate whether such behavior may be attributed to pure altruism, social desirability bias, or that perhaps people simply derive a feeling of "warm glow" satisfaction (Andreoni 1990). However empirical studies on whether consumers would actually prefer and be willing to pay more for ethically sourced products is limited, and one major challenge is the endogeneity of prices.

To investigate the effect of Fair Trade labeling on actual consumer demand for coffee, Hainmueller, Hiscox, and Sequeira (2015) conducts a field experiment in partnership with a major U.S. grocery store chain in the Northeastern United States. They perform a set of two experiments across 26 stores over 8 weeks, where two bulk coffee types were labeled as Fair Trade certified while the others were not. In the second experiment, they also experimentally raised the price of the Fair Trade labeled coffees to investigate whether consumers would be willing to pay more for ethically sourced coffees. Treatments were assigned by a matched-pairs design, where stores were matched on key store characteristics and socioeconomic variables then randomly assigned to a treatment-control sequence that switches halfway through. Using a reduced-form version of a discrete choice model on aggregate market shares, they find that Fair Trade labels increased weekly sales by approximately 10% when prices were the same. When prices were experimentally raised, demand for the more expensive coffee type was inelastic and did not reduce sales. In fact, sales increased by 2%. However for the cheaper coffee type, sales decreased by approximately 30%.

We build directly on the work of Hainmueller, Hiscox, and Sequeira (2015) by focusing primarily on the nature and dynamics of heterogeneity in treatment effects across units and time. The central theme behind this paper is to complement their work by analyzing the empirical setting under weak assumptions.

In contrast to their work which begins with structural assumptions and parametric restrictions as is standard in the discrete choice literature, we use general potential outcomes as our primitives for nonparametric identification in a primarily difference-in-differences setting. Another key difference in our approach is that we focus on each coffee as the unit of study, where the interpretation of treatments are different and faces a problem of selection bias. We also allow for unrestricted heterogeneity across units and dynamic effects across time. However by imposing minimal economic structure, several complications arise including the possibility of spillovers and carryover effects as treatments can switch off. To deal

with this, we propose a plausible partial identification approach after treatments switch off where the bounds collapse to a point if there are no carryover effects.

Section 2 provides an overview of the empirical setting, treatment assignment mechanism, and data. Section 3 presents the main identification argument, where we condition on the set of high-dimensional matching variables to justify parallel trends between coffees of similar types. Section 4 discusses estimation and inference, where we follow the literature on flexible semiparametric estimation of the effects of interest. However by including certain covariates along with knowledge of the treatment assignment mechanism, the moment conditions simplify into a simple difference-in-differences between appropriate coffee types.

Overall, we find similar results as Hainmueller, Hiscox, and Sequeira (2015) where there is an approximately 9.5% increase in sales when Fair Trade labels were put on coffees. Similarly, there is a large drop in sales of the same magnitude as their findings when prices were raised in addition to Fair Trade labeling. Our results also show some interesting time-varying dynamics, where consumer preferences gradually adjust to the treatment. Conditional treatment effects have a heavy left tail, suggesting that some neighborhoods are particularly averse to price premiums associated with Fair Trade. This is complemented by a heavy right tail in conditional spillover effects on the other coffees in the stores, which may be interpreted as substitution effects when prices were raised. In contrast to survey responses, the results suggest that the average consumer is indeed sensitive to prices and unwilling to pay the price premium.

2 Empirical Setting

In 2009, Hainmueller, Hiscox, and Sequeira (2015) conducts a set of two field experiments to investigate consumer demand for Fair Trade coffees. These experiments were carried out in 26 stores of a major grocery store chain in the northeastern United States. A particular strength of their experimental design is that key product characteristics such as Fair Trade labeling and prices were experimentally set, thus avoiding the endogeneity issues typically involved with demand estimation.

Each store carries the same set of seven bulk coffee bean types. Two coffee types are cheaper options priced at \$10.99/pound, while the remaining are all priced identically at \$11.99/pound. Table 2 shows some basic summary statistics of the various coffee types across all stores.

2.1 Treatment Assignment Mechanism

The experimental design is based on a matched pairs design with a treatment crossover. First, the stores were matched into pairs based on a large set of store-level and neighborhood characteristics including historical average sales, sales growth, income, and demographics. For each pair of stores, one was randomly assigned to a treatment-control sequence while the other was assigned to a control-treatment sequence where the treatment status switches halfway through. Each of the two experiments were conducted for 8 weeks[1], and we will refer to the first half of each experiment (Weeks 1-4) as *Phase 1*, and the second half (Weeks 5-8) as *Phase 2*.

More precisely, and letting (S_1, S_2) denote the matched pairs, the treatment status of each pair of stores is shown in Table 1.

	Phase 1	Phase 2
S_1	Treated	Untreated
S_2	Untreated	Treated

Table 1: Treatment status of each pair of stores

However even though treatments were conditionally randomly assigned at the store-level, if we were to focus on each coffee as the unit of study then there could be some selection bias. In particular only two coffee types, French Roast and Coffee Blend, were ever subject to treatment. French Roast is historically one of the best-selling coffees in the stores, while Coffee Blend one of the cheapest. Following the terminology used in the experiments, we will refer to these two coffee types as *test coffees*.

The first experiment, hereafter called the *Label Experiment*, involved attaching a Fair Trade label to the bulk coffee bins of the test coffees in treated stores. To avoid the possibility that simply having a label could affect consumer choices unrelated to Fair Trade, the control group was also given a generic label. Figure 1 shows an example of the product labels used in the experiments.

The second experiment, hereafter called the *Price Experiment*, involved labeling the test coffees with Fair Trade as before but also raising the prices by \$1/pound. It should be noted that French Roast then becomes the most expensive coffee in the stores, while Coffee Blend is no longer one of the cheaper options. To avoid any carryover effects, the Price Experiment occurred several months following the end of the Label Experiment.

[1]In the Price Experiment, a few stores had delays in administering the treatment crossover so the two weeks immediately following the crossover were dropped by Hainmueller, Hiscox, and Sequeira (2015). The second phase was extended so that the total experimental period remains 8 weeks.

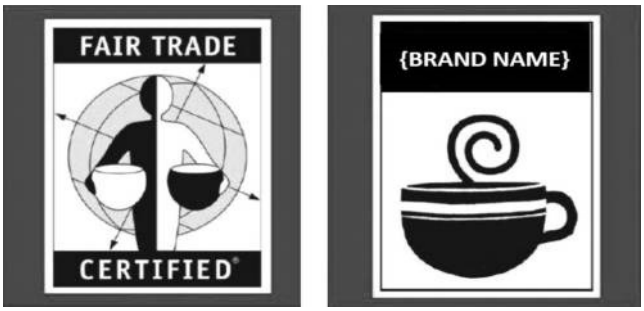


Figure 1: Fair Trade labels (left) and generic labels (right) during the experiments. *Source:* Hainmueller, Hiscox, and Sequeira (2015).

2.2 Data

Data from the field experiments was obtained from Hainmueller (2017). The dataset consists of repeated observations of the weekly sales of the seven bulk coffee bean types across the 26 stores from 2007-2009.

We also observe the characteristics used to match the stores[2], including its past sales for the year, growth rate, historical sales of the various coffee types, as well as aggregate socioeconomic data from the 2000 U.S. Census for the zip code the stores are located in. The socioeconomic data includes average household incomes, education levels, age, race, and the proportion of those on social security or public assistance programs.

In Section 3 we describe the identification argument which is based primarily on a difference-in-differences design. Since it is a balanced panel, for each experiment there was a total of 1572 observations[3] in the experimental period and the week immediately prior. The Label Experiment occurred in early 2009, and the Price Experiment in late 2009. However we also observe a large set of historical sales data prior to the experiments, and we incorporate it to improve estimation and inference. With the additional historical data, there are then a total of 17674 observations for the Label Experiment and 5908 observations for the Price Experiment.

Tables 2 and 3 along with Figure 2 reports some basic descriptive statistics of the data in the full historical period since 2007. An interesting observation is the large socioeconomic differences across the stores' neighborhoods which may indicate potentially strong heterogeneity in treatment effects and valuation of ethical sourcing. For example, the average household income across all neighborhoods is $64,101, but can range from a relatively

[2]However, the exact matching procedure and which stores were matched is not entirely clear from the data.

[3]Across both experiments, there were a total of 145 entries with missing sales data and 21 missing entries in the pre-experimental period. As Hainmueller, Hiscox, and Sequeira (2015) points out that these are due to out-of-stocks or logistical issues, these observations were dropped.

poor neighborhood with median income $26,689 to a relatively wealthy neighborhood with median income $123,622. It is possible that wealthy people on average would be more supportive of Fair Trade initiatives and less sensitive to price premiums, and similarly young or highly educated neighborhoods could have different distributions of tastes. These socioeconomic differences may contribute to substantially different valuations in ethical sourcing, and induce strong heterogeneity in treatment effects.

	Price	Average Weekly Revenue	Average Weekly Sales	Share of Bulk Coffee Sales
Breakfast Blend	11.99	125.057	10.710	0.134
Coffee Blend	10.99	114.822	10.793	0.109
Colombian Supremo	10.99	159.573	15.407	0.168
French Roast Extra Dark	11.99	125.898	10.730	0.132
French Roast Regular	11.99	206.164	17.619	0.222
Mexican	11.99	93.475	8.032	0.097
Regional Blend	11.99	130.752	10.979	0.140

Table 2: Average historical sales and share of bulk sales for each coffee type across stores. The boldfaced test coffees are subject to treatment in the experiments, while the remaining are never treated. Note that the price column indicates the regular prices which are held fixed or experimentally raised during the experimental period, so average revenues and sales do not exactly correspond in the historical period due to occasional promotions and price changes.

	mean	std	min	25%	50%	75%	max
2008 Sales (1M$)	26.15	11.48	4.66	16.08	27.58	35.28	54.57
2008-2009 Sales Growth (%)	3.85	1.69	0.69	2.36	4.06	5.19	8.03
Average 4-week Sales for Bulk Coffees ($)	77.55	45.40	24.86	46.19	61.23	94.44	216.40
Average 4-week Sales for Instant Coffees ($)	38.87	12.34	19.72	29.74	39.42	45.06	78.06
Average 4-week Sales for Packaged Coffees ($)	40.08	12.00	22.11	31.29	37.08	47.03	69.64
Average 52-week Sales for Bulk Coffees ($)	76.79	46.25	21.52	42.80	62.24	92.53	212.44
Average 52-week Sales for Instant Coffees ($)	34.83	10.42	21.91	24.72	35.56	41.35	60.77
Average 52-week Sales for Packaged Coffees ($)	38.12	10.75	20.60	30.98	35.36	42.83	64.76
Population	24303.65	12611.86	4793.00	14199.00	22542.00	33790.25	56185.00
African-American Population (%)	4.13	4.57	0.37	0.91	2.43	6.16	16.68
Foreign-born Population (%)	13.70	7.79	3.21	8.91	10.13	18.78	30.65
Median Household Income	64104.81	24140.58	26689.00	44382.00	60111.00	78083.25	123622.00
Social Security (%)	9.78	2.79	5.51	7.79	9.68	11.78	15.23
Public Assistance (%)	36.93	24.94	11.75	16.79	30.40	45.93	101.39
Family Households (%)	59.22	18.61	21.70	41.14	64.97	74.36	80.33
Head of Household Aged 15-34 (%)	23.21	15.04	6.28	11.71	17.60	38.18	54.94
Head of Household Aged 65+ (%)	22.48	6.87	11.98	17.18	22.51	27.01	37.69
Attending Highschool (%)	12.88	5.83	0.68	8.01	14.07	16.70	21.12
Attending College (%)	12.54	13.12	3.51	4.66	6.29	18.17	60.05
Highschool Dropouts (%)	8.36	6.28	1.26	4.63	6.13	9.38	27.48
Highschool Degrees (%)	17.46	8.72	5.98	10.23	15.57	23.63	34.04
College Degrees (%)	28.22	7.09	12.20	24.38	30.60	32.83	36.26
Graduate Degrees (%)	27.10	13.40	6.22	17.56	27.43	37.49	51.23

Table 3: Summary statistics of historical sales and socioeconomic characteristics across the stores.

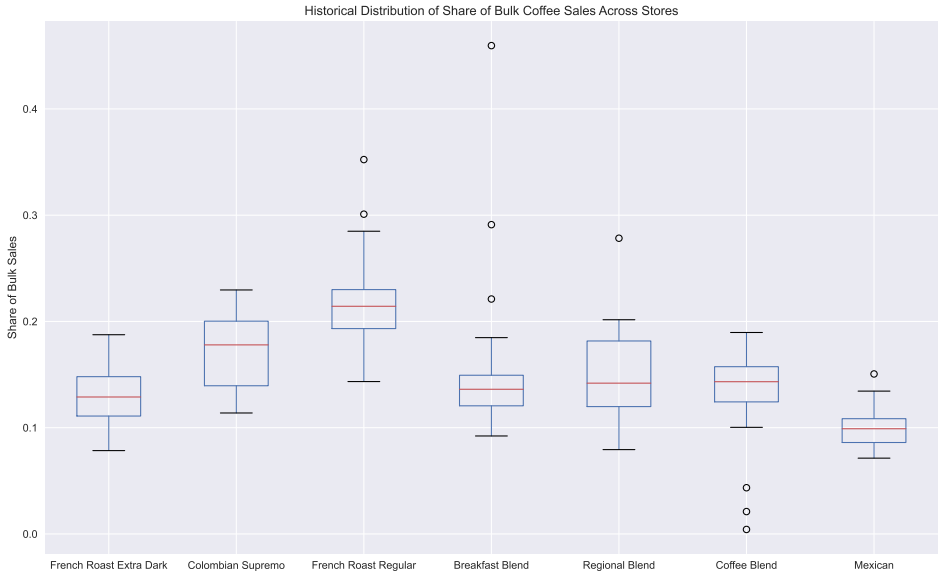


Figure 2: Historical distribution of the share of sales across stores for each coffee type.

3 Identification

In this section we introduce the general framework using potential outcomes as our primitives, define the effects of interest, and discuss the main assumptions for identification. Given the identical design of the two experiments, the discussion here will mainly be in the context of the Label Experiment to avoid confusion but these assumptions are taken to also hold for the Price Experiment.

Notation

Let $S = 26$ denote the total number of stores, and $T = 8$ denote the total number of weeks during each experiment. The time period $t = 0$ refers to the pre-treatment period. Recall that Phase 1 refers to Weeks 1-4, and Phase 2 refers to Weeks 5-8 after the treatment crossover. The set of all coffee types in each store is indexed by $J = \{1, \ldots, 7\}$, and we set the first two elements as the test coffees[4] denoted by $J^{(tc)} = \{1, 2\}$.

Store-level outcomes are denoted by the vector $\mathbf{Y}_{st} = (Y_{1st}, Y_{2st}, \ldots, Y_{|J|st})$, where the jth element is the revenue of coffee type j in store s at time t. The pre-treatment covariates are denoted by $\mathbf{X}_s = (X_{1s}, X_{2s}, \ldots, X_{|J|s})$, and Z_{st} is a binary treatment indicator at the store-level. To ease notation, subscripts are dropped when they can be inferred from the context.

Pre-Treatment Covariates

The set of pre-treatment covariates includes a rich collection of store-level and neighborhood characteristics such as income, education, age, and demographics that may be relevant for characterizing preferences for Fair Trade. Importantly, this is the same set of covariates used to match stores as described in Section 2 which we use to strongly justify the plausibility of the assumptions. We further include a binary indicator $X_j^{(tc)}$ for whether the coffee is a test coffee. At the coffee-level, we may then view $X_{js} = (X_j^{(tc)}, X_s)$ as a high-dimensional vector[5] consisting of the single coffee-specific characteristic $X_j^{(tc)}$ and the store-level matching characteristics X_s.

3.1 Potential Outcomes

Assumption 1. (Random Cluster Sampling)

$$\{\mathbf{Y}_{s0}, \mathbf{Y}_{s1}, \ldots, \mathbf{Y}_{sT}, Z_{s0}, \ldots, Z_{sT}, \mathbf{X}_s\}_{s=1}^{S} \text{ is i.i.d.}$$

[4] Recall that of the 7 coffee types, test coffees refer to French Roast and Coffee Blend which are the only types subject to treatment. The remaining are never treated.

[5] There are a total of 28 covariates.

This assumption states that we observe an i.i.d. panel, where each store is drawn from a superpopulation. In particular, it also implies that the sales of each coffee type j, $(Y_{jst})_{s=1}^{S}$ are i.i.d. across stores.

We then focus on each coffee as the unit of study, and aim to estimate treatment effects at the coffee-level to capture heterogeneity and dynamic effects. However one complication of analyses at the coffee-level is the possibility of spillover effects from other coffees within the same store being treated. For example even if a coffee was untreated, if another coffee in the same store was labeled with Fair Trade this might induce consumers to substitute away from the untreated coffee. Not accounting for this would likely result in an upward bias[6] in estimated treatment effects.

To accommodate spillover effects, for each unit potential outcomes $Y_{jst}(d)$ are then given by its own treatment status as well as whether other units in the store were treated. We then observe

$$Y_{jst} = \sum_{d=0}^{2} \mathbf{1}\{D_{jst} = d\} Y_{jst}(d)$$

where $Y_{jst}(2)$ refers to the potential outcome had they been a treated coffee in a treated store, $Y_{jst}(1)$ refers to being an untreated coffee in a treated store, and $Y_{jst}(0)$ refers to being an untreated coffee in an untreated store. This notation implicitly relaxes the interference condition of SUTVA to hold only at the store-level. In particular, we assume the potential outcomes of coffees in store s are not affected by the treatment status of those in store $s' \neq s$. Since the stores are spread out across several states, this assumption is likely reasonable.

Analogously, the coffee-level treatment is defined as

$$D_{jst} = \begin{cases} 2 & \text{if } Z_{st} = 1, j \in J^{(tc)} \\ 1 & \text{if } Z_{st} = 1, j \notin J^{(tc)} \\ 0 & \text{else} \end{cases}$$

We then view the data as given by the coffee-level panel $\{Y_{js0}, \ldots, Y_{jsT}, D_{js0}, \ldots, D_{jsT}, X_{js}\}_{s=1;j\in J}^{S}$. Note that the coffee-level panel is i.i.d. across stores by Assumption 1, but may be highly dependent within stores due to the shared covariates across coffees within the same store. We discuss this further in Section 4 for estimation and inference.

3.2 Parameters of Interest

With the potential outcomes framework previously introduced, we may now define the main parameters of interest at the coffee-level.

[6] For the Price Experiment, this would likely be a downward bias.

First, the *individual treatment effect* is defined as

$$\theta_{jst}^{treat} = Y_{jst}(2) - Y_{jst}(0)$$

which represents the effect of going from no treatment at all to a treatment on coffee j.

Closely related are *individual spillover effects*, which are defined as

$$\theta_{jst}^{spill} = Y_{jst}(1) - Y_{jst}(0)$$

and represents the spillover effects on untreated coffees when other coffees in the same store were treated.

While not directly useful as it is difficult to interpret[7], we may also define *individual total effects* by

$$\theta_{jst}^{tot} = Y_{jst}(2) - Y_{jst}(1)$$

which can be thought of as the effect of going from being an untreated coffee in a treated store, to a treated coffee. These parameters are all related by

$$\underbrace{Y_{jst}(2) - Y_{jst}(1)}_{\text{total effect}} = \underbrace{\Big(Y_{jst}(2) - Y_{jst}(0)\Big)}_{\text{treatment effect}} - \underbrace{\Big(Y_{jst}(1) - Y_{jst}(0)\Big)}_{\text{spillover on untreated}}$$

Dropping the unit-level subscripts in the notation, the average effects of interest in the population are then defined by

$$\theta_t^{treat} = \mathbb{E}\Big[Y_t(2) - Y_t(0)\Big|D = 2\Big]$$
$$\theta_t^{spill} = \mathbb{E}\Big[Y_t(1) - Y_t(0)\Big|D = 1\Big]$$

where we allow for arbitrary heterogeneity across time and units. It should be noted that within each phase, treatments are absorbing in the sense that $D_{t-1} = d \implies D_t = d$ a.s., but this does not hold across phases due to the treatment switching.

3.3 Point Identifying Assumptions for Phase 1

In this section, consider the first sub-experiment corresponding to Weeks 1-4. Denote $T^{(1)} = \{1, \ldots, 4\}$ as the set of event time indices for Phase 1.

[7] For this reason, we do not consider total effects but mention it here as it highlights the potential bias from disregarding the possibility of spillovers.

Assumption 2. (No Anticipation) For all $d \in \{0, 1, 2\}$, in the pre-treatment period

$$Y_0 = Y_0(d) \text{ a.s.}$$

This assumption states that in the pre-treatment period, there are no anticipation effects. This holds by the experimental design, where all references to Fair Trade were removed from the stores several weeks prior to the experiments.

Assumption 3. (Overlap) For all $d \in \{0, 1, 2\}, t \in T^{(1)}$,

$$\mathbb{P}(D_t = d) > 0, \mathbb{P}(D_t = d|X) < 1 \text{ a.s.}$$

This assumption states that for each possible level of treatment, there exists some proportion of the population which was treated. Furthermore for any sub-population, there exists some untreated units who may be used as controls. These support conditions are naturally implied by the design of the experiments.

For identification of the average effects on the treated, the fundamental problem is that we never observe the counterfactual outcome $Y(0)$ in the treated group. As is standard in the difference-in-differences literature, we impose variants on parallel trends as a way to impute what the counterfactual outcome would have been had they not been treated.

To identify average treatment effects θ_t^{treat}, we make the following parallel trends assumption.

Assumption 4. (Conditional Parallel Trends) For all $t \in T^{(1)}$,

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

This assumption states that for the treated coffees in treated stores, the counterfactual trend in sales had the coffee and the store not been treated is the same as the untreated coffees in untreated stores after conditioning on covariates. Since the covariates X includes both the type of coffee as well as the set of store-level and neighborhood characteristics used in matching, this strongly justifies the validity of this assumption.

Remark. We may think of the set of all coffee types as partitioned into $J = (J^{tc}, J^{rest})$, where J^{tc} are the test coffees and J^{rest} are the alternatives. By the matched design, the natural counterfactual for J^{tc} in treated stores is J^{tc} in untreated stores since the stores were matched. Similarly, the natural counterfactual for J^{rest} in treated stores (which are subject to spillovers) is J^{rest} in untreated stores. This choice of covariates gives the strongest identification argument. On the other hand, removing the test coffee indicator $X^{(tc)}$ from

the covariates and incorporating further coffee-level covariates could potentially allow for the usage of more data and improve estimation and inference, but weakens the identification argument. This trade-off is discussed further in Section 4.

Proposition 1. Under Assumptions 2-4, θ_t^{treat} is non-parametrically point-identified. Furthermore, conditional treatment effects are given by

$$\theta_t^{treat}(X) = \mathbb{E}[Y_t - Y_0|X, D = 2] - \mathbb{E}[Y_t - Y_0|X, D = 0]$$

which is a difference-in-differences between the treated coffees in treated stores and untreated coffees in untreated stores. The proof is in Appendix A.1.

Next for identification of average spillover effects θ_t^{spill}, we impose a different parallel trends assumption.

Assumption 5. (Conditional Parallel Trends of Indirect Effects) For all $t \in T^{(1)}$,

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

This assumption states that for the untreated coffees in treated stores, the counterfactual trend had the store not been treated is the same as untreated coffees in untreated stores after conditioning on covariates.

Proposition 2. Under Assumptions 2, 3, and 5, θ_t^{spill} is non-parametrically point-identified. Furthermore, conditional spillover effects are given by

$$\theta_t^{spill}(X) = \mathbb{E}[Y_t - Y_0|X, D = 1] - \mathbb{E}[Y_t - Y_0|X, D = 0]$$

which is a difference-in-differences between the untreated coffees in treated stores and untreated coffees in untreated stores. The proof is very similar to Proposition 1 and omitted.

3.4 Partially Identifying Assumptions for Phase 2

In this section, consider the second sub-experiment corresponding to Weeks 5-8. To keep the notation clean, denote $T^{(2)} = \{1, 2, 3, 4\}$ as the set of event time indices for Phase 2, where we reindex the times so that $t = 0$ represents the pre-treatment period corresponding to Week 4 in calendar time.

Recall that at the start of Phase 2, there is a treatment crossover. This means that for the previously treated test coffees in Phase 1, they are now untreated starting at $t = 1$, while test coffees in a different set of stores are now treated.

If we were to assume no carryover effects from the previous phase, then we may view the two phases as two independent sub-experiments where the parameters of interest are point-identified using the same set of assumptions. However this is likely unrealistic, since the previous treatments might have induced a permanent change in the distribution of tastes. For example, some consumers might have initially preferred Breakfast Blend but switched to French Roast after the Fair Trade labels in Phase 1. Even if the labels were now removed, they may continue purchasing French Roast due to habit, ambiguity about whether it is still Fair Trade but mislabeled, or simply because they enjoy the taste. Assuming no carryover effects would then be analogous to assuming that the distribution of tastes in the treated group immediately revert back to the state had the entire experiment never happened. In terms of potential outcomes, this means there would likely be a violation of the previous parallel trends assumptions.

We do not impose any substantial restrictions on possible carryover effects, and instead take a partial identification approach where the bounds collapse to a point if these two phases can indeed be viewed as independent sub-experiments. The main challenge for identification in Phase 2 is due to the fact that there are no clean controls: untreated coffees are either affected by spillovers, or treatments switching off from the previous phase. To accommodate this, we will first enrich the potential outcomes framework to also include past treatment status.

3.4.1 Potential Outcomes and Parameters of Interest

Let $D^{(1)}$ denote the previous treatment status prior to Phase 2, so that $D^{(1)}_{jst} = D_{js0}$ for all units. Denote the actual treatment status during Phase 2 by $D^{(2)}$, where $D^{(1)}, D^{(2)} \in \{0, 1, 2\}$ are coffee-level treatments as before representing untreated coffees in untreated stores, untreated coffees in treated stores, and treated coffees in treated stores.

Potential outcomes are now defined by $Y_{jst}(d_1, d_2)$, where d_1 is the treatment status prior to Phase 2, and d_2 is the treatment status during Phase 2. We then observe

$$Y_{jst} = \sum_{d_2=0}^{2} \sum_{d_1=0}^{2} \mathbf{1}\{D^{(1)}_{jst} = d_1, D^{(2)}_{jst} = d_2\} Y_{jst}(d_1, d_2)$$

Note that the potential outcomes now refer to a path of treatments. For example, $Y_{jst}(0, 2)$ refers to the outcome had the coffee not been treated in Phase 1 but was treated in Phase 2. Analogous to Section 3.2, the individual effects in Phase 2 are defined by

$$\begin{aligned} \theta^{treat}_{jst} &= Y_{jst}(0, 2) - Y_{jst}(0, 0) \\ \theta^{spill}_{jst} &= Y_{jst}(0, 1) - Y_{jst}(0, 0) \end{aligned}$$

This notation emphasizes that treatment effects should be the difference between being treated in Phase 2, and never treated during the entire experiment[8]. Dropping the unit-level subscripts, average effects in the population are then defined similarly by

$$\theta_t^{treat} = \mathbb{E}[Y_t(0,2) - Y_t(0,0)|D^{(1)} = 0, D^{(2)} = 2]$$
$$\theta_t^{spill} = \mathbb{E}[Y_t(0,1) - Y_t(0,0)|D^{(1)} = 0, D^{(2)} = 1]$$

From these definitions, the main identification problem in Phase 2 becomes clear. We never observe $Y_t(0,0)$ since there are no never-treated coffees in never-treated stores. In other words, every coffee is either directly subject to a treatment, spillover effects from other coffees, or from past treatments switching off.

3.4.2 Structural Bounds on Counterfactuals

Assumption 6. (No Anticipation of Treatment Switching) For all $(d_1, d_2) \in \{0,1,2\}^2$, in the pre-treatment period

$$Y_{js0}(d_1, 0) = Y_{js0}(d_1, d_2) \text{ a.s.}$$

This assumption states that at time $t = 0$, there are no anticipation effects that the treatment status will change in the post-treatment period. This holds by the experimental design, where store employees were instructed to switch the labels during Phase 2. For the previously untreated coffees, this is the analogue of Assumption 2. For the previously treated coffees, there was no indication on the Fair Trade labels that it was a limited event which might abruptly change.

Assumption 7. (Support for Treated Groups) For all $(d_1, d_2) \neq (0,0), t \in T^{(2)}$,

$$\mathbb{P}(D_t^{(1)} = d_1, D_t^{(2)} = d_2) > 0, \mathbb{P}(D_t^{(1)} = d_1, D_t^{(2)} = d_2|X) < 1 \text{ a.s.}$$

A typical point-identifying argument for θ_t^{treat} would then proceed by assuming that $\mathbb{E}[Y_t(0,0) - Y_{t-1}(0,0)|X, D^{(1)} = 0, D^{(2)} = 2]$ is parallel to the observed outcomes of some valid control group. However the problem here is that there are no valid control groups, or more precisely, we have a violation of the typical support conditions since $\mathbb{P}(D^{(1)} = 0, D^{(2)} = 0) = 0$. We then make the following assumption that the counterfactual trends of all groups evolve in parallel.

Assumption 8. (Common Counterfactual Trends) There exists some sequence $(\delta_t(X))_{t\in\mathbb{N}}$

[8] In the special case of no carryover effects, one way is to assume that $Y_{jst}(2,0) = Y_{jst}(0,0) = Y_{jst}(1,0)$ almost surely for all $t \in T^{(2)}$ in Phase 2. In which case we may drop the previous treatments in the potential outcomes notation and the setting becomes identical to Phase 1.

s.t. for all $t \in T^{(2)}, (d_1, d_2) \neq (0,0)$,

$$\mathbb{E}[Y_t(0,0) - Y_{t-1}(0,0)|X, D^{(1)} = d_1, D^{(2)} = d_2] = \delta_t(X) \text{ a.s.}$$

The intuition behind this assumption is that we may partition the set of all coffees into four groups: treated coffees which became untreated ($D^{(1)} = 2, D^{(2)} = 0$), untreated coffees affected by spillovers but no longer ($D^{(1)} = 1, D^{(2)} = 0$), untreated coffees which became treated ($D^{(1)} = 0, D^{(2)} = 2$), and untreated coffees which were later affected by spillovers ($D^{(1)} = 0, D^{(2)} = 1$). This assumption then simply states that the counterfactual trends for all groups if the entire experiment never happened are all the same.

Under Assumptions 6 and 8, the general difference-in-differences decomposition for conditional effects can then be written as[9]

$$\theta_t^{treat}(X) = \underbrace{\mathbb{E}[Y_t - Y_0|X, D^{(1)} = 0, D^{(2)} = 2]}_{\text{trend in outcomes}} - \underbrace{\sum_{k=1}^{t} \delta_k(X)}_{\text{counterfactual trend}}$$

$$\theta_t^{spill}(X) = \mathbb{E}[Y_t - Y_0|X, D^{(1)} = 0, D^{(2)} = 1] - \sum_{k=1}^{t} \delta_k(X)$$

However these assumptions are of course insufficient for identification, since we do not actually know what $\delta_t(X)$ is. If we had observed never treated units, then setting $\delta_t(X)$ as the observed trend for these units would give point-identification for both θ_t^{treat} and θ_t^{spill}. Alternatively if we assumed no carryover effects from Phase 1, then we could also set $\delta_t(X)$ as the observed trend for the $(D^{(1)} = 2, D^{(2)} = 0)$ or $(D^{(1)} = 1, D^{(2)} = 0)$ group and obtain point identification. Instead we use a partial identification approach to obtain plausible structural bounds on $\delta_t(X)$.

Consider first the group of coffees that were initially treated in Phase 1. Suppose for now that Fair Trade labeling had a positive average effect on sales, either directly by incentivizing new purchases from consumers who otherwise would not have purchased any coffees, or indirectly by consumers substituting towards Fair Trade coffees. After the treatment switches off in Phase 2, there are no longer any direct effects since the coffees are no longer labeled.

For the indirect effects, some consumers might continue purchasing this group of coffees due to habit, ambiguity about whether it is actually Fair Trade but mislabeled, or simply because they have an acquired taste for these coffees. On the other hand, it is also possible that some consumers immediately revert back to their initial preferences had the experiment

[9] See Appendix A.2

never occurred. In addition, any late entrants to the market would make choices as if Phase 1 never occurred. On average, this then motivates the following upper bound

$$\delta_t(X) \leq \mathbb{E}[Y_t - Y_{t-1}|X, D^{(1)} = 2, D^{(2)} = 0]$$

For a lower bound, consider now the group of coffees that were untreated but subject to spillovers in Phase 1. If the treatment had induced consumers to substitute away from this group towards Fair Trade coffees, then we should generally expect the sign of treatment effects and spillover effects to be opposite. Thus if Fair Trade labeling had a positive average effect, then we should expect spillover effects to be negative (or at least non-positive[10]). The scenario where both effects are of the same sign would imply that people on average prefer Fair Trade, but at the same time are somehow incentivized to also purchase more non-Fair Trade coffees which is possible[11] but unlikely. Since treatment effects and spillover effects are closely related, by the previous argument this then motivates the following lower bound

$$\mathbb{E}[Y_t - Y_{t-1}|X, D^{(1)} = 1, D^{(2)} = 0] \leq \delta_t(X)$$

It may be helpful to instead consider the following. For the group of untreated coffees subject to spillovers in Phase 1, we might imagine a world where consumers are substituting away from these coffees towards Fair Trade, and a world where the experiment never occurred. During Phase 2, the best-case outcome for the spillover group is that there are no longer any more substitution effects, in which case the observed outcome is $\delta_t(X)$. The worst-case outcome is that everyone continues switching to Fair Trade coffees as if they were still treated.

Figure 3 shows a stylized visualization of the identification argument, and Figure 4 shows a plot based on the observed trend in sales averaged across stores for the four subgroups.

In the argument above, we assumed that treatment effects were positive for simplicity. It is certainly possible that treatment effects are negative, which is especially relevant for the Price Experiment where the treatment consisted of both Fair Trade labeling and raising prices. In the case of negative treatment effects, then by an analogous argument, the observed outcomes for the treated group becomes a lower bound on $\delta_t(X)$, and the observed outcomes for the spillover group becomes an upper bound.

[10] Even though average treatment effects and spillover effects are likely to be opposite signs, the effects are not necessarily symmetric. For example Fair Trade labeling might incentivize new purchases from consumers who would otherwise not have purchased any coffees in the stores, which results in no spillover on the untreated coffees.

[11] For example some consumers could be incentivized by the treatment to purchase both a Fair Trade and non-Fair Trade coffee to compare the quality, but for this to hold on average in the population is unlikely.

We then formalize the partial identification argument in general without sign restrictions by the following assumption.

Assumption 9. (Bounds on Counterfactual Trends)

$$\min\{\theta_t^{(2,0)}(X), \theta_t^{(1,0)}(X)\} \leq \sum_{k=1}^{t} \delta_k(X) \leq \max\{\theta_t^{(2,0)}(X), \theta_t^{(1,0)}(X)\}$$

where

$$\theta_t^{(2,0)}(X) = \mathbb{E}[Y_t - Y_0 | X, D^{(1)} = 2, D^{(2)} = 0]$$
$$\theta_t^{(1,0)}(X) = \mathbb{E}[Y_t - Y_0 | X, D^{(1)} = 1, D^{(2)} = 0]$$

Proposition 3. Under Assumptions 6-9, θ_t^{treat} is partially identified by the region

$$\theta_t^{treat} \in \Theta_t^{treat} = \left[L_t^{treat}, U_t^{treat}\right]$$

where

$$L_t^{treat} = \mathbb{E}[Y_t - Y_0 | D^{(1)} = 0, D^{(2)} = 2] - \int \max\{\theta_t^{(2,0)}(X), \theta_t^{(1,0)}(X)\} \, dP_{X|\{D^{(1)}=0, D^{(2)}=2\}}$$
$$U_t^{treat} = \mathbb{E}[Y_t - Y_0 | D^{(1)} = 0, D^{(2)} = 2] - \int \min\{\theta_t^{(2,0)}(X), \theta_t^{(1,0)}(X)\} \, dP_{X|\{D^{(1)}=0, D^{(2)}=2\}}$$

The proof is in Appendix A.2.

Since we have bounded the counterfactual trends $\delta_t(X)$ directly, we then also obtain partial identification for average spillover effects.

Proposition 4. Under Assumptions 6-9, θ_t^{spill} is partially identified by the region

$$\theta_t^{spill} \in \Theta_t^{spill} = \left[L_t^{spill}, U_t^{spill}\right]$$

where

$$L_t^{spill} = \mathbb{E}[Y_t - Y_0 | D^{(1)} = 0, D^{(2)} = 1] - \int \max\{\theta_t^{(2,0)}(X), \theta_t^{(1,0)}(X)\} \, dP_{X|\{D^{(1)}=0, D^{(2)}=1\}}$$
$$U_t^{spill} = \mathbb{E}[Y_t - Y_0 | D^{(1)} = 0, D^{(2)} = 1] - \int \min\{\theta_t^{(2,0)}(X), \theta_t^{(1,0)}(X)\} \, dP_{X|\{D^{(1)}=0, D^{(2)}=1\}}$$

The proof is similar to Proposition 3 and omitted.

Remark. Suppose there is no carryover effect from Phase 1, in the sense that $Y_t(2,0) =$

$Y_t(0,0) = Y_t(1,0)$ a.s. for all $t \in \mathcal{T}^{(2)}$, or under the weaker condition of

$$\mathbb{E}[Y_t(d,0) - Y_0(d,0)|X, D^{(1)} = d, D^{(2)} = 0] = \mathbb{E}[Y_t(0,0) - Y_0(0,0)|X, D^{(1)} = d, D^{(2)} = 0]$$

for both $d \in \{1,2\}$. Then the bounds collapse to a point and we obtain point identification of $\theta_t^{treat}, \theta_t^{spill}$. In other words, for the bounds to collapse we require both the treated and untreated coffees in treated stores to revert back to the counterfactual state where Phase 1 never happened.

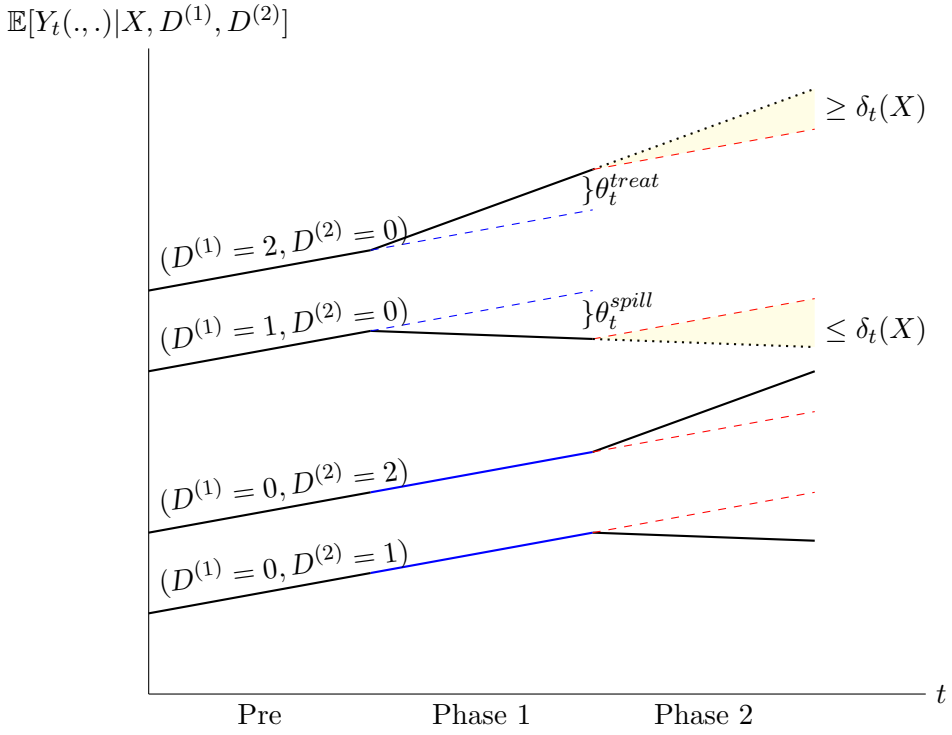


Figure 3: Stylized visualization of the identification argument. Solid lines indicate observable outcomes, while the blue and red dashed lines are the counterfactual trends. In Phase 1, we assume that the blue counterfactual for the treated groups are parallel to the observed outcomes of the untreated for point-identification. In Phase 2, we assume that the red counterfactual for the treated groups are bounded by the observable outcomes (in the shaded yellow region) of the no longer treated groups for partial identification.

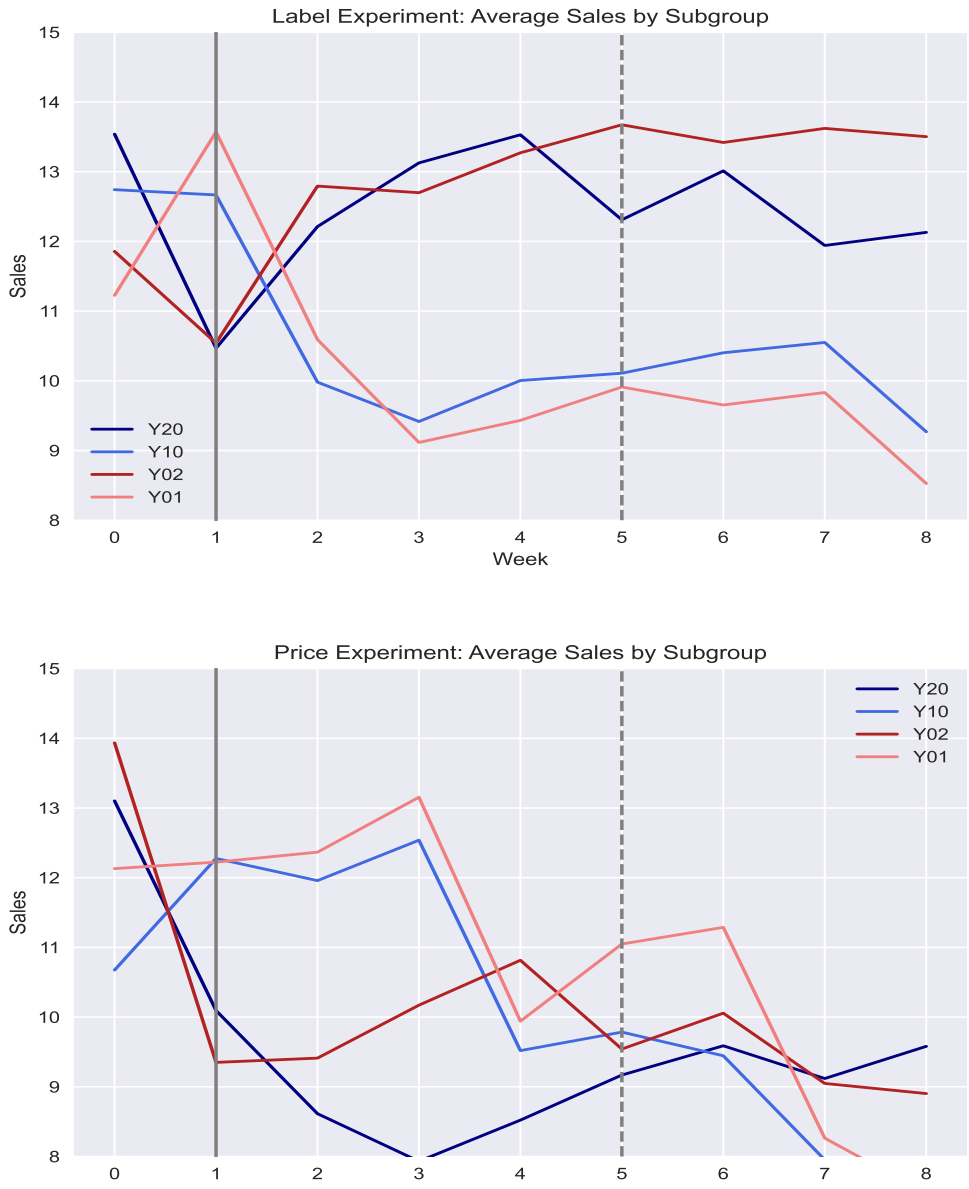


Figure 4: This plot shows the observed trend in sales for each of the four subgroups averaged over stores. The blue lines labeled Y20 and Y10 correspond to the set of treated and untreated coffees in treated stores, while the red lines labeled Y02 and Y01 are initially untreated but switch on during Phase 2 (dashed line). Note that the identification argument is conditional on store characteristics as well as the type of coffee so this plot should be interpreted by pairwise comparisons of the corresponding blue and red lines (e.g. (Y20, Y02), and (Y10, Y01)).

4 Estimation and Inference

For estimation and inference at the coffee-level, we first impose some simple restrictions on the error terms. Note that Assumption 1 implies that the coffee-level panel is i.i.d. across stores, but may be dependent within stores. As the goal of this paper is to analyze the empirical setting with weak assumptions, we do not impose any functional form restrictions and assume the true data generating process at the coffee-level is given by

$$Y_{jst} = m(X_{js}, D_{jst}) + \epsilon_{jst}$$

where $m(X_{js}, D_{jst}) := \mathbb{E}[Y_{jst}|X_{js}, D_{jst}]$, and so by construction $\mathbb{E}[\epsilon_{jst}|X_{js}, D_{jst}] = 0$. Recall that $X_{js} = (X_j^{(tc)}, X_s)$ is a high-dimensional vector consisting of a test coffee indicator $X_j^{(tc)}$ and store-level characteristics X_s, and $\mathbf{X}_s = (X_{1s}, X_{2s}, \ldots, X_{|J|s})$ denotes the collection of coffee-level covariates for a store.

Assumption 10. (Conditional Variance of Errors)

$$\begin{aligned}
\mathbb{E}[\epsilon_{jst}^2|\mathbf{X}_s, \mathbf{D}_{st}] &= \mathbb{E}[\epsilon_{jst}^2|X_{js}, D_{jst}] = \sigma^2(X_{js}) \\
\mathbb{E}[\epsilon_{jst}\epsilon_{j'st}|\mathbf{X}_s, \mathbf{D}_{st}] &= \mathbb{E}[\epsilon_{jst}\epsilon_{j'st}|X_{js}, X_{j's}, D_{jst}, D_{j'st}] = \sigma(X_{js}, X_{j's}) \;\; \forall\, j' \neq j
\end{aligned}$$

This assumption states that the conditional variance of the error for each unit depends only on its own covariates, and that the conditional covariance of the error between any two coffees depend only on the test coffee indicator and shared store-level covariates.

4.1 Semiparametric Estimation and Reduced Form

We first consider Phase 1 where the main parameters of interest, treatment effects and spillover effects, are point-identified.

For estimation and inference, the conventional way is to assume a linear parametric structure and estimate some variant on two way fixed effects models. However recent literature has pointed out that if there is substantial heterogeneity in treatment effects, the regression coefficients can be inconsistent, difficult to interpret, and even be the wrong sign entirely due to a "negative weighting" problem (Chaisemartin and D'Haultfoeuille 2022; Goodman-Bacon 2021; Sun and Abraham 2021).

In our empirical application, there is a strong possibility that treatment effects are heterogeneous across both units and time since consumers may receive diminishing marginal utility from repeatedly supporting social causes such as Fair Trade, and especially when prices were raised during the Price Experiment. Consumers with different levels of income,

education, age, and other demographics are also likely to have highly different valuations of ethical sourcing that may induce strong heterogeneity in treatment effects. Given the large variation in socioeconomic characteristics as noted in Section 2, the heterogeneity can be substantial. Furthermore, two way fixed effects models impose a particular parametric assumption on the conditional means which may be highly misleading if misspecified. In particular since the set of covariates we use to justify conditional parallel trends is high-dimensional, the risk of functional form misspecification is especially large.

Under conditional identification and the possibility of heterogeneous effects, Abadie (2005) proposes a semiparametric estimator for difference-in-differences that flexibly incorporates covariates through the propensity score. Related earlier work by Heckman, Ichimura, and Todd (1997) also considers matching estimators based on the propensity score. In light of the issues pointed out with standard fixed effects models under heterogeneity, the recent literature on difference-in-differences has seen a surge of interest in semiparametric and doubly robust methods including machine learning, see for instance Sant'Anna and Zhao (2020), Callaway and Sant'Anna (2021), and Chang (2020).

The following proposition then derives the relevant moment conditions for our parameters of interest.

Proposition 5. Under Assumptions 2-5, average treatment effects on the treated and spillover effects are identified by the following moment conditions:

$$\theta_t^{treat} = \mathbb{E}\left[\frac{(Y_t - Y_0)\left(\mathbf{1}_{\{D=2\}}p^{(0)}(X) - \mathbf{1}_{\{D=0\}}p^{(2)}(X)\right)}{p^{(0)}(X)\mathbb{P}(D=2)}\right]$$

$$\theta_t^{spill} = \mathbb{E}\left[\frac{(Y_t - Y_0)\left(\mathbf{1}_{\{D=1\}}p^{(0)}(X) - \mathbf{1}_{\{D=0\}}p^{(1)}(X)\right)}{p^{(0)}(X)\mathbb{P}(D=1)}\right]$$

where $p^{(d)}(X) = \mathbb{P}(D = d|X)$. The proof is in Appendix A.3.

Rearranging the moment condition for treatment effects, we can see that

$$\theta_t^{treat} = \mathbb{E}\left[\frac{Y_t - Y_0}{\mathbb{P}(D=2)}\left(\mathbf{1}_{\{D=2\}} - \mathbf{1}_{\{D=0\}}\frac{p^{(2)}(X)}{p^{(0)}(X)}\right)\right]$$

where we may interpret $w(X) := \frac{p^{(2)}(X)}{p^{(0)}(X)}$ as a weighting function that ensures covariate balance between the treated and untreated groups. For any covariates such that $w(X)$ is small so that they are unlikely to be treated, the distribution of $Y_t - Y_0$ is weighted down in the untreated group. For any covariates such that $w(X)$ is large so that they are likely to be treated, the distribution is weighted up in the untreated group. It is precisely this idea that

allows propensity score-based weighting methods to adjust for selection bias (Abadie 2005; Heckman, Ichimura, and Todd 1997), though potentially at the cost of efficiency (Hirano, Imbens, and Ridder 2003).

By the sample analogue principle, this then motivates the following semiparametric estimator

$$\hat{\theta}_t^{treat} = \frac{1}{n} \sum_{j=1;s=1}^{|J|;S} \frac{Y_{jst} - Y_{js0}}{\hat{g}^{(2)}} \Big(\mathbf{1}\{D_{jst} = 2\} - \mathbf{1}\{D_{jst} = 0\} \frac{\hat{p}^{(2)}(X_{js})}{\hat{p}^{(0)}(X_{js})} \Big)$$

where $n = S|J|$, and $\hat{g}^{(d)} = \frac{1}{n} \sum_{j;s} \mathbf{1}\{D_{jst} = d\}, \hat{p}^{(d)}(X)$ are consistent estimators for $\mathbb{P}(D = d), \mathbb{P}(D = d|X)$ respectively. Due to the high-dimensional covariates however, estimation of the propensity score becomes challenging with classical nonparametric methods. This would typically require either imposing some parametric restrictions on the propensity score, or using Double Machine Learning[12] as proposed by Chernozhukov et al. (2018).

In our application however, this problem simplifies considerably since propensity scores are actually known if we include the type of coffee in the covariates. Recall that $X_{js} = (X_j^{(tc)}, X_s)$ where $X_j^{(tc)} = \mathbf{1}\{j \in J^{(tc)}\}$ is a binary indicator for if it is a test coffee. Since we know the treatment assignment mechanism, then by construction with our choice of covariates the propensity scores are known and given by

$$p^{(2)}(X_{js}) = \mathbb{P}(D_{jst} = 2|X_{js}) = \begin{cases} 1/2 & \text{if } X_j^{(tc)} = 1 \\ 0 & \text{if } X_j^{(tc)} = 0 \end{cases}$$

$$p^{(1)}(X_{js}) = \mathbb{P}(D_{jst} = 1|X_{js}) = \begin{cases} 0 & \text{if } X_j^{(tc)} = 1 \\ 1/2 & \text{if } X_j^{(tc)} = 0 \end{cases}$$

$$p^{(0)}(X_{js}) = \mathbb{P}(D_{jst} = 0|X_{js}) = \begin{cases} 1/2 & \text{if } X_j^{(tc)} = 1 \\ 1/2 & \text{if } X_j^{(tc)} = 0 \end{cases}$$

Remark. Hirano, Imbens, and Ridder (2003) shows that for average treatment effects, even if true propensity scores were known then estimating it anyways could lead to efficiency gains. However for average treatment effects on the treated, Hahn (1998) shows that the asymptotic semiparametric efficiency bound is lower if true propensity scores were known[13]. Removing $X_j^{(tc)}$ from the set of covariates could still justify identification if we

[12] Chernozhukov et al. (2018) shows that simply plugging in a machine learning estimator for infinite-dimensional nuisance parameters complicates inference and can lead to a large bias in estimates, as machine learning estimators generally converge slower than the parametric $\sqrt{n}$-rate due to overfitting or regularization. The general idea behind Double Machine Learning is to Neyman-orthogonalize the moment condition such that it is locally insensitive to noisy estimation of the nuisance parameters, which allows the orthogonalized estimator to attain the $\sqrt{n}$-rate when using cross-fitting with i.i.d. data.

[13] For selection on observables settings.

were to incorporate additional coffee-level characteristics such as average historical sales that controls for differences between the coffee types. This could potentially allow for the usage of more coffees in the control group and improve estimates in finite samples. However in doing so we would only obtain partial knowledge of the propensity scores which would require high-dimensional estimation. Due to the additional complications from cluster sampling, we do not further consider the problem of efficiency in this paper.

With the known propensity scores, note that the weighting function simplifies into $w(X) = \frac{p^{(2)}(X)}{p^{(0)}(X)} = X^{(tc)} \in \{0,1\}$ depending on whether it is a test coffee or not. Furthermore since $D = 2$ implies $X^{(tc)} = 1$ a.s., the semiparametric estimator for treatment effects then simplifies into

$$\hat{\theta}_t^{treat} = \frac{1}{n} \sum_{j=1;s=1}^{|J|S} \left(\frac{Y_{jst} - Y_{js0}}{\hat{g}^{(2)}} \mathbf{1}\{D_{jst} = 2\} - \frac{Y_{jst} - Y_{js0}}{\hat{g}^{(2)}} \mathbf{1}\{D_{jst} = 0\} \right) X_j^{(tc)}$$

where the term $\hat{g}^{(2)} = \frac{1}{n} \sum_{j;s} \mathbf{1}\{D_{jst} = 2\}$ is just a scaling constant to balance against the fact that we are dividing by the full sample size n.

This implies that even though we have not made any functional form assumptions on the data generating process, allowed for arbitrary heterogeneity, and relied on a high-dimensional set of conditioning variables for identification, we can still estimate θ_t^{treat} by a simple difference-in-differences between test coffees in treated stores, and test coffees in untreated stores. This is of course not true in general[14], but holds given that the distributions of covariates we conditioned on were balanced for these two specific subgroups. We may then estimate the simple linear model

$$Y_{jst} = \alpha + \tau_t + \delta D_{jst} + \theta_t D_{jst} \tau_t + \epsilon_{jst}$$

on the subset of the data containing test coffees, where τ_t is a binary time indicator. For inference, we use cluster-robust standard errors at the store-level. A potential concern here however is the sample size, as clustered standard errors can be heavily biased in small samples (MacKinnon, Nielsen, and Webb 2023). Recall that there are only two test coffees in each of the 26 stores, of which half are untreated. Recent work by Egami and Yamauchi (2023) suggests the usage of additional pre-treatment periods to improve the precision of point estimates and efficiency. In our setting, even though the sample size is small during the experiment we observe a long history of sales in these stores dating back two years. We

[14] For example, if we were to add further coffee-level characteristics which were not used in the matching process.

incorporate the additional historical data[15] and found substantial gains in efficiency.

The case for spillover effects follows analogously, with

$$\hat{\theta}_t^{spill} = \frac{1}{n} \sum_{j=1;s=1}^{|J|S} \left(\frac{Y_{jst} - Y_{js0}}{\hat{g}^{(1)}} \mathbf{1}\{D_{jst} = 1\} - \frac{Y_{jst} - Y_{js0}}{\hat{g}^{(1)}} \mathbf{1}\{D_{jst} = 0\} \right) (1 - X_j^{(tc)})$$

which becomes a simple difference-in-differences between the set of untreated coffees in treated stores and untreated coffees in untreated stores.

4.2 Estimation of Partially Identified Regions

We now consider estimation of the parameters of interest in Phase 2 after the treatment crossover. Recall that the partially identified regions are given by

$$\Theta_t^{treat} = \left[L_t^{treat}, U_t^{treat} \right]$$
$$\Theta_t^{spill} = \left[L_t^{spill}, U_t^{spill} \right]$$

To simplify the notation, let

$$h_t^{\vee}(X) = \max\{\theta_t^{(2,0)}(X), \theta_t^{(1,0)}(X)\}$$
$$h_t^{\wedge}(X) = \min\{\theta_t^{(2,0)}(X), \theta_t^{(1,0)}(X)\}$$

be the upper and lower bounds on the counterfactual trends from Assumption 9. The bounds on the partially identified regions can then be written as

$$L_t^{treat} = \mathbb{E}[Y_t - Y_0 | D^{(1)} = 0, D^{(2)} = 2] - \int h_t^{\vee}(X) \; dP_{X|\{D^{(1)}=0, D^{(2)}=2\}}$$
$$U_t^{treat} = \mathbb{E}[Y_t - Y_0 | D^{(1)} = 0, D^{(2)} = 2] - \int h_t^{\wedge}(X) \; dP_{X|\{D^{(1)}=0, D^{(2)}=2\}}$$
$$L_t^{spill} = \mathbb{E}[Y_t - Y_0 | D^{(1)} = 0, D^{(2)} = 1] - \int h_t^{\vee}(X) \; dP_{X|\{D^{(1)}=0, D^{(2)}=1\}}$$
$$U_t^{spill} = \mathbb{E}[Y_t - Y_0 | D^{(1)} = 0, D^{(2)} = 1] - \int h_t^{\wedge}(X) \; dP_{X|\{D^{(1)}=0, D^{(2)}=1\}}$$

Since X is high-dimensional, for estimation of the nuisance functions $h^{\vee}, h^{\wedge}$ we consider using a plug-in machine learning estimator[16] by nonparametrically fitting the conditional means, taking the minimum or maximum, then integrating over the empirical conditional

[15] Identification would require parallel trends to also hold in the historical data, but presumably this was also used in the initial matching of stores so our assumptions on parallel trends should also extend to the historical data.

[16] This would likely substantially complicate inference on the partially identified set, though we only consider estimation of these bounds in this paper.

distributions of the data.

More precisely, we let $\hat{m}^{(d,0)}(X)$ be a machine learning estimator for $\theta_t^{(d,0)}(X) = \mathbb{E}[Y_t - Y_{t-1}|X, D^{(1)} = d, D^{(2)} = 0]$, and form the plug-in estimators

$$\hat{h}_t^{\vee}(X) = \max\{\hat{m}^{(2,0)}(X), \hat{m}^{(1,0)}(X)\}$$
$$\hat{h}_t^{\wedge}(X) = \min\{\hat{m}^{(2,0)}(X), \hat{m}^{(1,0)}(X)\}$$

The estimator for the bounds are then given by

$$\hat{L}_t^{treat} = \frac{1}{n_{02}} \sum_{j=1;s=1}^{|J|;S} \Big((Y_{jst} - Y_{js0}) - \hat{h}_t^{\vee}(X_{js})\Big)\mathbf{1}\{D_{jst}^{(1)} = 0, D_{jst}^{(2)} = 2\}$$

$$\hat{U}_t^{treat} = \frac{1}{n_{02}} \sum_{j=1;s=1}^{|J|;S} \Big((Y_{jst} - Y_{js0}) - \hat{h}_t^{\wedge}(X_{js})\Big)\mathbf{1}\{D_{jst}^{(1)} = 0, D_{jst}^{(2)} = 2\}$$

$$\hat{L}_t^{spill} = \frac{1}{n_{01}} \sum_{j=1;s=1}^{|J|;S} \Big((Y_{jst} - Y_{jst}) - \hat{h}_t^{\vee}(X_{js})\Big)\mathbf{1}\{D_{jst}^{(1)} = 0, D_{jst}^{(2)} = 1\}$$

$$\hat{U}_t^{spill} = \frac{1}{n_{01}} \sum_{j=1;s=1}^{|J|;S} \Big((Y_{jst} - Y_{jst}) - \hat{h}_t^{\wedge}(X_{js})\Big)\mathbf{1}\{D_{jst}^{(1)} = 0, D_{jst}^{(2)} = 1\}$$

where $n_{0d} = \sum_{j;s} \mathbf{1}\{D_{jst}^{(1)} = 0, D_{jst}^{(2)} = d\}$.

5 Empirical Results

5.1 Label Experiment

In the Label Experiment, recall that test coffees in treated stores were assigned a Fair Trade label as shown in Figure 1. Alternative coffees in treated stores were assigned a generic label to avoid any general effects from labeling unrelated to Fair Trade. This experiment was conducted from January-March 2009. After four weeks, the treated coffees had their labels immediately removed while another set of coffees in other stores were assigned labels.

Table 4 presents the full results across both phases of the experiment, and Figure 5 shows an event-study visualization of the two phases. For Phase 1 (Weeks 1-4), the table shows point estimates of dynamic treatment and spillover effects. We also show estimates for aggregated effects across time, which may be interpreted as the average effect of the treatment throughout all of Phase 1.

For Phase 2 (Weeks 5-8), the table shows estimates of the partially identified region. For estimation of the infinite-dimensional nuisance functions $h_t^{\wedge}, h_t^{\vee}$ in the bounds, we used a

random forest to nonparametrically fit the conditional means. Since the stores were matched on a large set of socioeconomic and store-level characteristics, it is likely that there are similar treatment effect dynamics between the treated coffees in Phase 1 and the treated coffees in Phase 2. For example if variables such as income or education play a role in consumers' preferences of Fair Trade, then the matching likely induces some degree of homogeneity in treatment effects across the two groups even though they are treated at different times. Interestingly, the estimated partially identified region seems to capture similar trends and dynamics in the treatment effects that the first group of coffees experienced during Phase 1.

A surprising result is that both treatment effects and spillover effects are near zero at the start of the experiment, but both are positive during Weeks 3 and 4. This suggests that on average, consumers tended to purchase both more Fair Trade and non-Fair Trade coffees due to the labeling. It could be possible that simply having large labels on the bulk coffee bins induced new purchases from consumers who otherwise would not have purchased any coffee from the stores.

However the sample size of treated coffees for each week is quite small and so the point estimates may be imprecise, as can be seen from the large standard errors. Even though we incorporated a large set of historical data and found substantial gains in efficiency, the average effects are insignificant at conventional levels after controlling for correlation within stores. This result is not particularly surprising, since when taken together with Figure 4, we can see that the average sales between the treated test coffees and untreated test coffees are similar during Phase 1 despite one group being treated. However as the average weekly sales of coffees is around 12 pounds, the cumulative treatment effect from Fair Trade labeling in Phase 1 does result in an average increase of 1.15 pounds. This corresponds to approximately a 9.5% increase[17], which is similar to the findings of 10% from Hainmueller, Hiscox, and Sequeira (2015) under different assumptions including homogeneous effects.

Average effects also mask potentially insightful heterogeneity. In addition to time-varying dynamics, there is a strong possibility of heterogeneity across units given the large differences in socioeconomic characteristics noted earlier such as income, public welfare programs, and age. Figure 6 shows the distribution of conditional treatment and spillover effects, which we non-parametrically estimated using machine learning based on the difference-in-differences decomposition in Proposition 1 and 2. A particularly interesting result is that while the average effects appear stable, the distribution of spillovers has increasingly large left tails

[17]This is of course a very crude approximation. While it is common to use a log transformation for these types of interpretations, parallel trends holding in levels generally does not imply it holds in logs unless we are willing to make much stronger assumptions on the full distribution of potential outcomes (Roth and Sant'Anna 2023).

as time passes. This may indicate a growing shift in the distribution of tastes due to Fair Trade labeling, while average effects remain mostly stable due to the matched design.

	Treatment Effects			Spillover Effects		
Week	$\hat{\theta}_t^{treat}$	$se(\hat{\theta}_t^{treat})$	95% CI	$\hat{\theta}_t^{spill}$	$se(\hat{\theta}_t^{spill})$	95% CI
$t = 1$	0.206	2.117	[-3.943, 4.355]	-0.165	1.560	[-3.221, 2.892]
$t = 2$	-0.302	1.438	[-3.121, 2.516]	-0.027	1.015	[-2.017, 1.962]
$t = 3$	0.706	1.436	[-2.108, 3.520]	0.984	1.199	[-1.367, 3.334]
$t = 4$	0.538	1.530	[-2.461, 3.536]	1.231	1.225	[-1.170, 3.631]
$t = 5$	[0.052, 1.620]	-	-	[0.113, 1.720]	-	-
$t = 6$	[-0.486, 1.619]	-	-	[-0.352, 1.821]	-	-
$t = 7$	[0.297, 2.356]	-	-	[0.401, 2.478]	-	-
$t = 8$	[0.285, 2.510]	-	-	[-0.744, 1.514]	-	-

Num. Obs.: 17674, Clusters: 26

Note: This table presents estimates of treatment and spillover effects for the Label Experiment. Weeks 1-4 correspond to Phase 1, where we have point-identification of the effects of interest. Weeks 5-8 correspond to Phase 2, where the reported estimates are the partially identified regions. For the point estimates, standard errors were clustered at the store-level. For the partially identified regions, nuisance functions were estimated nonparametrically with random forests. The number of observations include a large set of historical data.

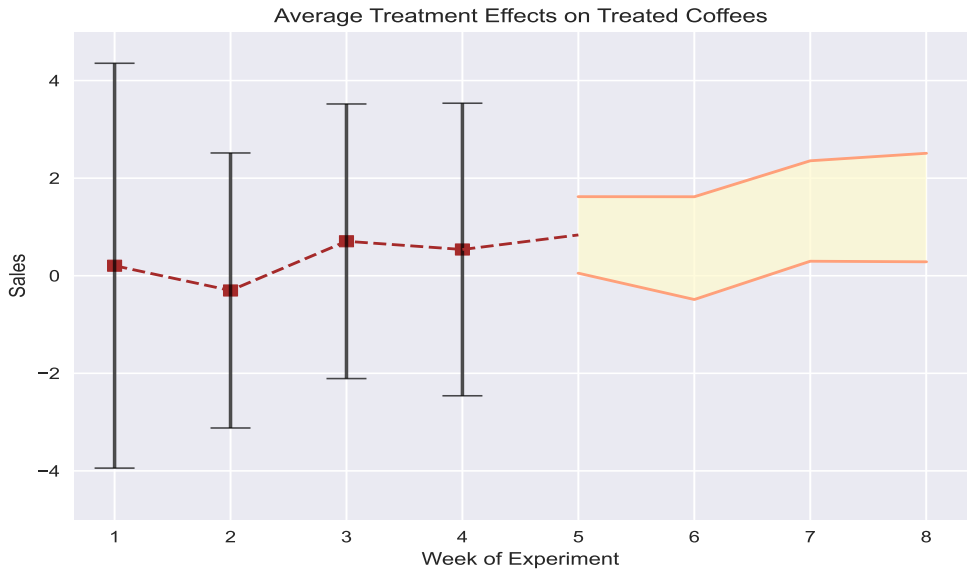
Table 4: Label Experiment Results

5.2 Price Experiment

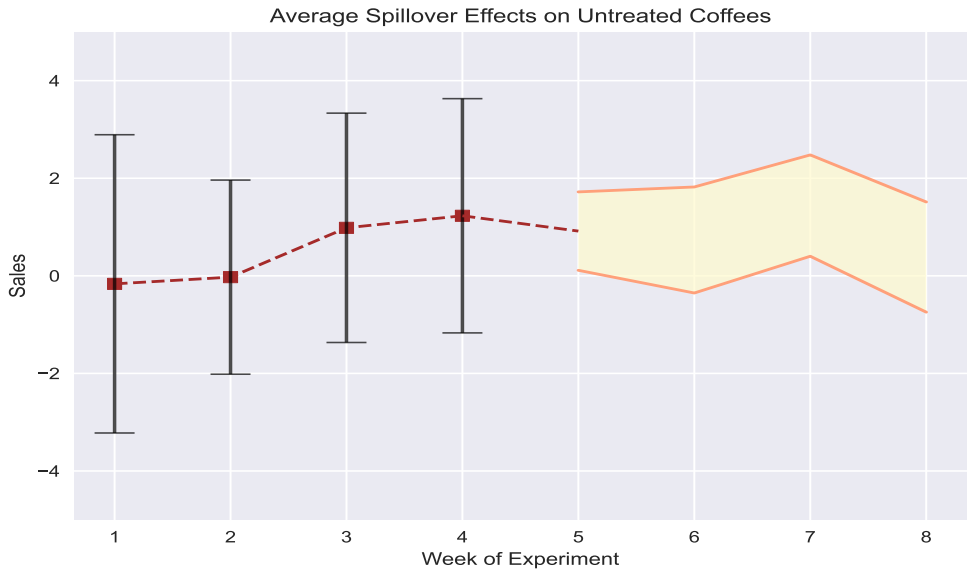
In the Price Experiment, test coffees were assigned a Fair Trade label as before but also had their prices raised by $1 per pound. This corresponds to an 8.3% increase in prices for French Roast and 9.1% increase for Coffee Blend. However an important implication of this is that when compared to other coffees in the stores, French Roast now becomes the most expensive coffee and Coffee Blend no longer one of the cheapest. This experiment was conducted several months after the Label Experiment ended.

Table 5 and Figure 7 shows the results. In contrast to the Label Experiment, here we see a noticeably large decrease in sales after prices were raised. An interesting result in the dynamics of the treatment effects is that the downward trend plateaus after two weeks, where the distribution of tastes seem to stabilize afterwards. When taken together with the previous results, this may suggest that average consumer preferences for Fair Trade gradually shift rather than abruptly change.

Another interesting result is that treatment effects and spillover effects are highly asymmetrical. If consumers would simply substitute towards a cheaper option, then we should see a growing trend in spillover effects. However the spillover effects are mostly near-zero, which may indicate that consumers had opted to not purchase any coffee from the store entirely due to the price raise. The partially identified region also captures similar dynam-

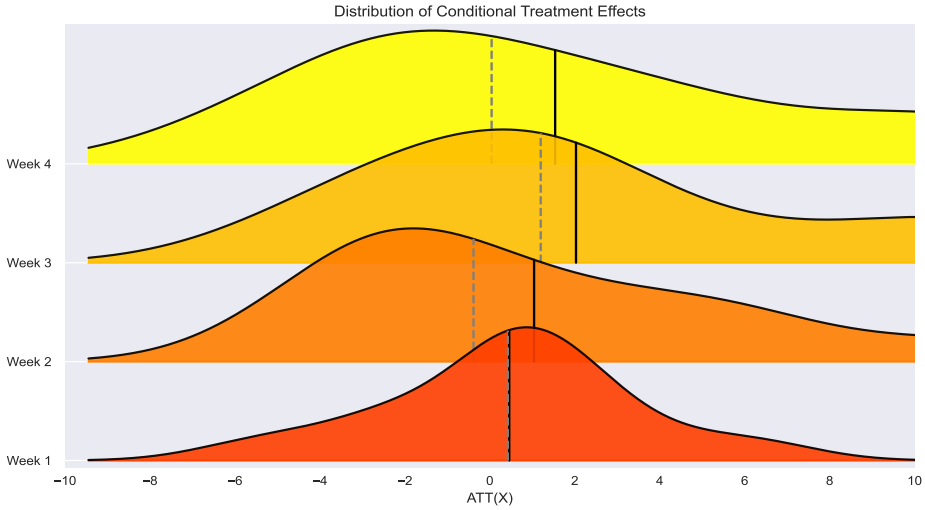


(a) Label Experiment: Treatment Effects

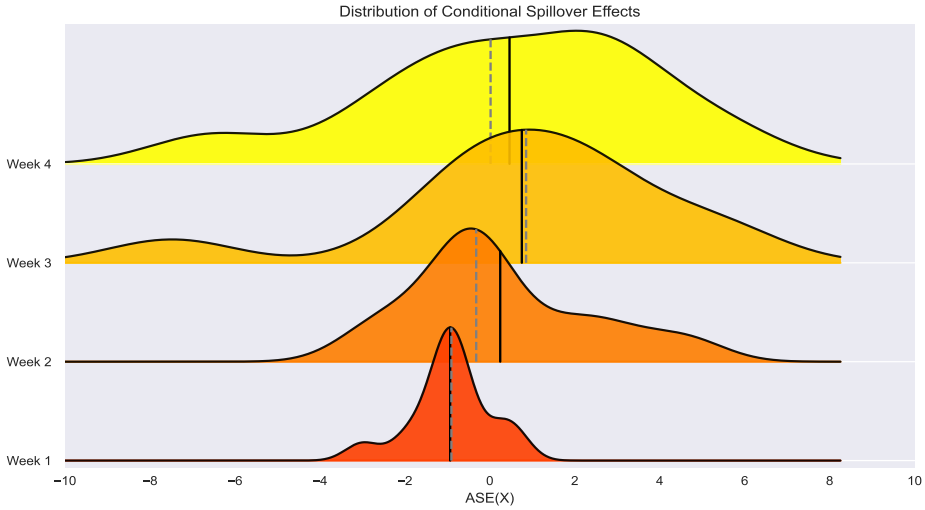


(b) Label Experiment: Spillover Effects

Figure 5: Estimates of dynamic treatment and spillover effects over time for the Label Experiment, along with their corresponding 95% confidence intervals. The shaded yellow region indicates the estimated partially identified region for coffees treated during Phase 2.



(a) Label Experiment: Conditional Treatment Effects



(b) Label Experiment: Conditional Spillover Effects

Figure 6: These plots show $\hat{\theta}_t^{treat}(X), \hat{\theta}_t^{spill}(X)$ over the empirical conditional distributions of $X|D = 2$ and $X|D = 1$ respectively for the Label Experiment. The black line indicates the mean, and the dashed grey line indicates the median.

ics as those treated in Phase one, with both bounds on treatment effects in the negative region. However there is a slight upward trend in the partially identified treatment effects for Phase 2, with a corresponding downward trend in spillover effects. After the treatment switch, the treated coffees had their prices lowered back to regular pricing and the labels removed. A possible explanation for the trend in the bounds is due to carryover effects, where consumers who opted to exit the market entirely may re-enter after seeing the price hike was reverted. As the distribution of tastes seem to have stabilized after two weeks, it is also possible that when the treatment was removed some consumers view it as a discount rather than a return to regular pricing.

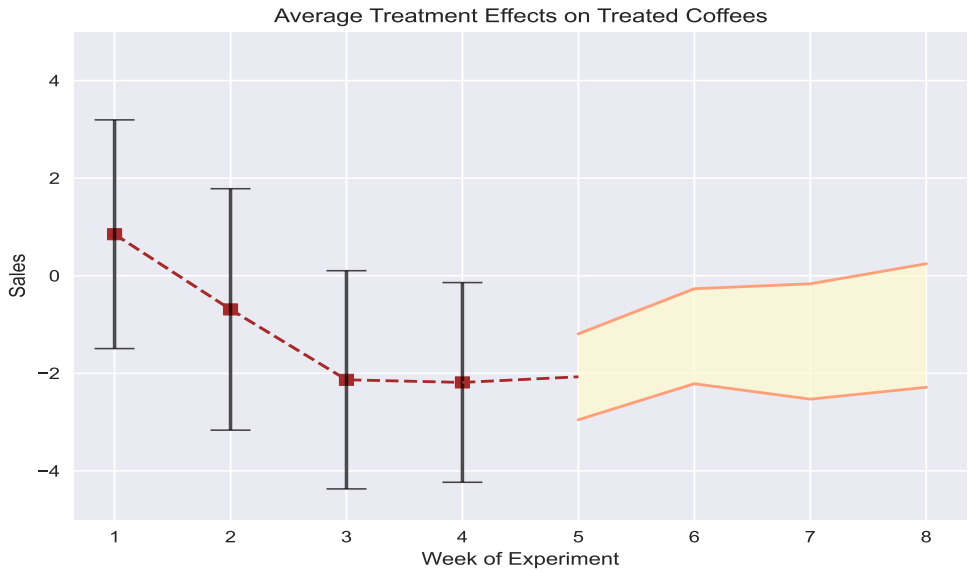
Figure 8 shows the distribution of conditional treatment and spillover effects. In contrast to the Label Experiment, there is a noticeably heavy left tail on conditional treatment effects and a corresponding heavy right tail on spillover effects. This may indicate that some consumers in particular are especially sensitive to the price hike, and substitute immediately to other options. This may be especially prevalent in communities with low average income who are more sensitive to prices.

	Treatment Effects			Spillover Effects		
Week	$\hat{\theta}_t^{treat}$	$se(\hat{\theta}_t^{treat})$	95% CI	$\hat{\theta}_t^{spill}$	$se(\hat{\theta}_t^{spill})$	95% CI
$t = 1$	0.850	1.196	[-1.494, 3.194]	9.478	1.261	[-1.992, 2.950]
$t = 2$	-0.691	1.262	[-3.166, 1.783]	0.017	1.257	[-2.446, 2.481]
$t = 3$	-2.133	1.141	[-4.369, 0.103]	-0.191	1.433	[-3.00, 2.617]
$t = 4$	-2.187	1.044	[-4.234, -0.140]	0.003	0.506	[-0.988, 0.994]
$t = 5$	[-2.952, -1.191]	-	-	[-0.207, 1.531]	-	-
$t = 6$	[-2.214, -0.267]	-	-	[0.115, 2.053]	-	-
$t = 7$	[-2.530, -0.166]	-	-	[-2.396, -0.114]	-	-
$t = 8$	[-2.286, 0.246]	-	-	[-2.584, -0.146]	-	-

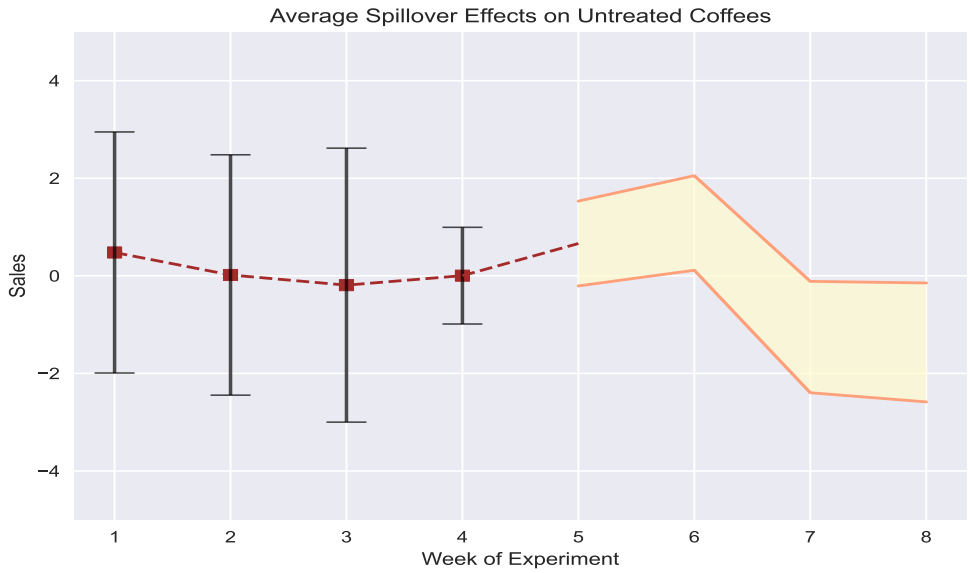
Num. Obs.: 5908, Clusters: 26

Note: This table presents estimates of treatment and spillover effects for the Price Experiment. Weeks 1-4 correspond to Phase 1, where we have point-identification of the effects of interest. Weeks 5-8 correspond to Phase 2, where the reported estimates are the partially identified regions. For the point estimates, standard errors were clustered at the store-level. For the partially identified regions, nuisance functions were estimated nonparametrically with random forests. The number of observations includes a large set of historical data.

Table 5: Price Experiment Results

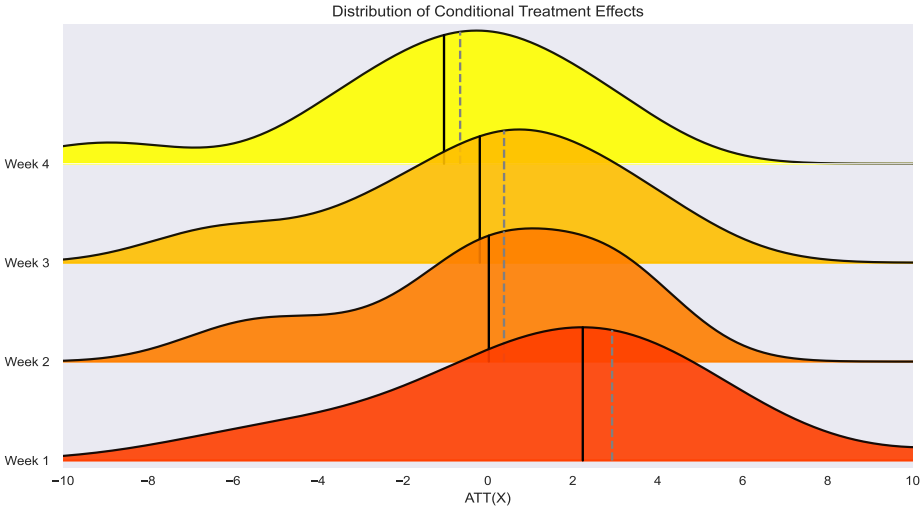


(a) Price Experiment: Treatment Effects

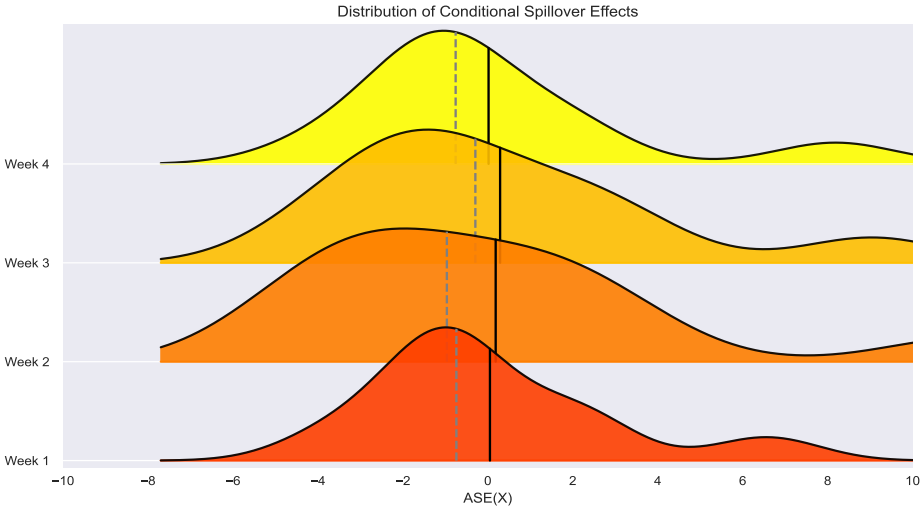


(b) Price Experiment: Spillover Effects

Figure 7: Estimates of dynamic treatment and spillover effects over time for the Price Experiment, along with their corresponding 95% confidence intervals. The shaded yellow region indicates the estimated partially identified region for coffees treated during Phase 2.



(a) Price Experiment: Conditional Treatment Effects



(b) Price Experiment: Conditional Spillover Effects

Figure 8: These plots show $\hat{\theta}_t^{treat}(X), \hat{\theta}_t^{spill}(X)$ over the empirical conditional distributions of $X|D = 2$ and $X|D = 1$ respectively for the Price Experiment. The black line indicates the mean, and the dashed grey line indicates the median.

6 Discussion

In this paper, we complement the analysis of the field experiment conducted by Hainmueller, Hiscox, and Sequeira (2015) by focusing on heterogeneity in valuations of ethical sourcing and price sensitivities. Overall, we found that consumers do prefer ethically sourced products when they are offered without a price premium. However in contrast to survey responses, the average consumer is sensitive to prices and unwilling to pay a price premium of $8 - 9\%$ for Fair Trade coffees. By analyzing the distribution of estimated conditional treatment effects, we find strong heterogeneity in both preferences for Fair Trade and aversion towards price premiums. These results are similar to those found by Hainmueller, Hiscox, and Sequeira (2015) which also indicate strong heterogeneity in the valuation of ethical sourcing.

A potential limitation of our study are in the relatively small sample sizes, where each week consists of only 170 observations across the stores. As we further restrict this to only comparing specific types of coffees across treated and untreated stores, it is difficult to conclusively determine whether the time-varying dynamics are due to changing valuations of ethical sourcing (such as decreasing marginal utilities from repeatedly supporting social initiatives) or simply naturally occurring consumption cycles in bulk purchasing habits. However comparing both treatment and spillover effects mitigates this to an extent.

Another area that could be developed further is on efficiency. While incorporating the historical data made a large difference, it may be worth exploring other cluster-robust methods for inference, especially since the number of clusters is relatively small. Alternatively, incorporating additional coffee-level characteristics and estimating propensity scores could allow for the usage of more coffees in the control group rather than limiting it to only comparisons between specific coffee types. This could benefit inference, though weaken the identification argument in Phase 1 which is currently strongly justified by the matched design. As for Phase 2, the partially identified bounds could potentially be sharpened further. For example the current bounds require both the treated group and those subject to spillovers to not have any carryover effects in order to collapse to a point. In the case where only one group has no carryover effect, then we already have point identification but only one side of the bounds would collapse. This scenario is possible, but somewhat unlikely since both groups belong to the same set of stores and are likely highly related.

Overall, this paper analyzed the empirical setting under a different set of assumptions and found similar results with existing work, which strengthens the validity of these findings. By focusing on heterogeneity, we provide new insights complementary to existing work on the time-varying dynamics and variation in ethical valuation induced by socioeconomic differences.

References

Abadie, Alberto (Jan. 2005). "Semiparametric difference-in-differences estimators". en. In: *The Review of Economic Studies* 72.1, pp. 1–19. ISSN: 1467-937X, 0034-6527. DOI: 10.1111/0034-6527.00321. URL: https://academic.oup.com/restud/article-lookup/doi/10.1111/0034-6527.00321 (visited on 03/06/2023).

Andreoni, James (1990). "Impure Altruism and Donations to Public Goods: A Theory of Warm-Glow Giving". In: *The Economic Journal* 100.401, pp. 464–477. ISSN: 00130133, 14680297. URL: http://www.jstor.org/stable/2234133 (visited on 03/06/2023).

Arnould, Eric J., Alejandro Plastina, and Dwayne Ball (Sept. 2009). "Does Fair Trade Deliver on Its Core Value Proposition? Effects on Income, Educational Attainment, and Health in Three Countries". en. In: *Journal of Public Policy & Marketing* 28.2, pp. 186–201. ISSN: 0743-9156, 1547-7207. DOI: 10.1509/jppm.28.2.186. (Visited on 09/21/2023).

Bacon, Christopher M. et al. (2008). "Are Sustainable Coffee Certifications Enough to Secure Farmer Livelihoods? The Millenium Development Goals and Nicaragua's Fair Trade Cooperatives". In: *Globalizations* 5.2, pp. 259–274. DOI: 10.1080/14747730802057688.

Callaway, Brantly and Pedro H.C. Sant'Anna (Dec. 2021). "Difference-in-Differences with multiple time periods". en. In: *Journal of Econometrics* 225.2, pp. 200–230. ISSN: 03044076. DOI: 10.1016/j.jeconom.2020.12.001. URL: https://linkinghub.elsevier.com/retrieve/pii/S0304407620303948 (visited on 03/06/2023).

Chaisemartin, Clément de and Xavier D'Haultfoeuille (Jan. 2022). *Two-Way Fixed Effects and Differences-in-Differences with Heterogeneous Treatment Effects: A Survey*. Working Paper 29691. National Bureau of Economic Research. DOI: 10.3386/w29691. URL: http://www.nber.org/papers/w29691.

Chang, Neng-Chieh (May 2020). "Double/debiased machine learning for difference-in-differences models". en. In: *The Econometrics Journal* 23.2, pp. 177–191. ISSN: 1368-4221, 1368-423X. DOI: 10.1093/ectj/utaa001. URL: https://academic.oup.com/ectj/article/23/2/177/5722119 (visited on 03/06/2023).

Chernozhukov, Victor et al. (Jan. 2018). "Double/debiased machine learning for treatment and structural parameters". In: *The Econometrics Journal* 21.1, pp. C1–C68. ISSN: 1368-4221. DOI: 10.1111/ectj.12097. URL: https://doi.org/10.1111/ectj.12097.

Egami, Naoki and Soichiro Yamauchi (Apr. 2023). "Using multiple pretreatment periods to improve difference-in-differences and staggered adoption designs". en. In: *Political Anal-*

ysis 31.2, pp. 195–212. ISSN: 1047-1987, 1476-4989. DOI: 10.1017/pan.2022.8. URL: https://www.cambridge.org/core/product/identifier/S1047198722000080/type/journal_article (visited on 04/26/2024).

Elliott, Kimberly Ann and Richard B. Freeman (2003). *Can Labor Standards Improve Under Globalization?* Washington, DC: Institute for International Economics. ISBN: 9780881323320.

Goodman-Bacon, Andrew (2021). "Difference-in-differences with variation in treatment timing". In: *Journal of Econometrics* 225.2. Themed Issue: Treatment Effect 1, pp. 254–277. ISSN: 0304-4076. DOI: https://doi.org/10.1016/j.jeconom.2021.03.014. URL: https://www.sciencedirect.com/science/article/pii/S0304407621001445.

Hahn, Jinyong (Mar. 1998). "On the role of the propensity score in efficient semiparametric estimation of average treatment effects". In: *Econometrica* 66.2, p. 315. ISSN: 00129682. DOI: 10.2307/2998560. URL: https://www.jstor.org/stable/2998560?origin=crossref (visited on 05/01/2024).

Hainmueller, Jens (2017). *Additional replication data for: Consumer Demand for Fair Trade: Evidence from a Multi-Store Field Experiment*. Version V1. DOI: 10.7910/DVN/GSBOEU. URL: https://doi.org/10.7910/DVN/GSBOEU.

Hainmueller, Jens, Michael J. Hiscox, and Sandra Sequeira (May 2015). "Consumer Demand for Fair Trade: Evidence from a Multistore Field Experiment". en. In: *Review of Economics and Statistics* 97.2, pp. 242–256. ISSN: 0034-6535, 1530-9142. DOI: 10.1162/REST_a_00467. URL: https://direct.mit.edu/rest/article/97/2/242-256/58230 (visited on 03/06/2023).

Heckman, J. J., H. Ichimura, and P. E. Todd (Oct. 1997). "Matching as an econometric evaluation estimator: evidence from evaluating a job training programme". en. In: *The Review of Economic Studies* 64.4, pp. 605–654. ISSN: 0034-6527, 1467-937X. DOI: 10.2307/2971733. URL: https://academic.oup.com/restud/article-lookup/doi/10.2307/2971733 (visited on 05/01/2024).

Hertel, Shareen, Lyle Scruggs, and C. Patrick Heidkamp (Sept. 2009). "Human Rights and Public Opinion: From Attitudes to Action". en. In: *Political Science Quarterly* 124.3, pp. 443–459. ISSN: 0032-3195, 1538-165X. DOI: 10.1002/j.1538-165X.2009.tb00655.x. URL: https://academic.oup.com/psq/article/124/3/443/6964008 (visited on 09/21/2023).

Hirano, Keisuke, Guido W. Imbens, and Geert Ridder (July 2003). "Efficient estimation of average treatment effects using the estimated propensity score". en. In: *Econometrica*

71.4, pp. 1161–1189. ISSN: 0012-9682, 1468-0262. DOI: 10.1111/1468-0262.00442. URL: http://doi.wiley.com/10.1111/1468-0262.00442 (visited on 04/29/2024).

Jaffee, Daniel (Feb. 2008). "'Better, But Not Great': The Social and Environmental Benefits and Limitations of Fair Trade for Indigenous Coffee Producers in Oaxaca, Mexico". In: *Sociology Faculty Publications and Presentations* 131.

MacKinnon, James G., Morten Ørregaard Nielsen, and Matthew D. Webb (Feb. 2023). "Cluster-robust inference: A guide to empirical practice". en. In: *Journal of Econometrics* 232.2, pp. 272–299. ISSN: 03044076. DOI: 10.1016/j.jeconom.2022.04.001. URL: https://linkinghub.elsevier.com/retrieve/pii/S0304407622000781 (visited on 05/02/2024).

Méndez, V. et al. (Sept. 2010). "Effects of Fair Trade and Organic Certifications on Small-scale Coffee Farmer Households in Central America and Mexico". In: *Renewable Agriculture and Food Systems* 25, pp. 236–251. DOI: 10.1017/S1742170510000268.

Reinstein, David A (June 2011). "Does One Charitable Contribution Come at the Expense of Another?" In: *The B.E. Journal of Economic Analysis & Policy* 11.1. ISSN: 1935-1682. DOI: 10.2202/1935-1682.2487. URL: https://www.degruyter.com/document/doi/10.2202/1935-1682.2487/html (visited on 09/22/2023).

Roth, Jonathan and Pedro H. C. Sant'Anna (2023). "When is parallel trends sensitive to functional form?" en. In: *Econometrica* 91.2, pp. 737–747. ISSN: 0012-9682. DOI: 10.3982/ECTA19402. URL: https://www.econometricsociety.org/doi/10.3982/ECTA19402 (visited on 05/02/2024).

Sant'Anna, Pedro H.C. and Jun Zhao (2020). "Doubly robust difference-in-differences estimators". In: *Journal of Econometrics* 219.1, pp. 101–122. ISSN: 0304-4076. DOI: https://doi.org/10.1016/j.jeconom.2020.06.003. URL: https://www.sciencedirect.com/science/article/pii/S0304407620301901.

Sun, Liyang and Sarah Abraham (2021). "Estimating dynamic treatment effects in event studies with heterogeneous treatment effects". In: *Journal of Econometrics* 225.2. Themed Issue: Treatment Effect 1, pp. 175–199. ISSN: 0304-4076. DOI: https://doi.org/10.1016/j.jeconom.2020.09.006. URL: https://www.sciencedirect.com/science/article/pii/S030440762030378X.

Appendix: Proofs

A.1

Proof of Proposition 1.

Let $\theta_t^{treat}(X) := \mathbb{E}[Y_t(2) - Y_t(0)|X, D = 2]$.

First note that under Assumption 4, by a telescoping sum

$$\begin{aligned}
\mathbb{E}[Y_t(0) - Y_0(0)|X, D = 2] &= \mathbb{E}\left[\sum_{k=1}^{t} Y_k(0) - Y_{k-1}(0)|X, D = 2\right] \\
&= \sum_{k=1}^{t} \mathbb{E}\left[Y_k(0) - Y_{k-1}(0)|X, D = 2\right] \\
&= \sum_{k=1}^{t} \mathbb{E}\left[Y_k(0) - Y_{k-1}(0)|X, D = 0\right] \\
&= \mathbb{E}[Y_t(0) - Y_0(0)|X, D = 0]
\end{aligned}$$

so that conditional parallel trends also holds between time t and time 0. Adding and subtracting $\mathbb{E}[Y_t(0)|X, D = 2]$ along with no anticipation effects from Assumption 2, we then have

$$\begin{aligned}
\mathbb{E}[Y_t(0)|X, D = 2] &= \mathbb{E}[Y_0(0)|X, D = 2] + \mathbb{E}[Y_t(0) - Y_0(0)|X, D = 2] \\
&= \mathbb{E}[Y_0|X, D = 2] + \mathbb{E}[Y_t(0) - Y_0(0)|X, D = 0] \\
&= \mathbb{E}[Y_0|X, D = 2] + \mathbb{E}[Y_t - Y_0|X, D = 0]
\end{aligned}$$

This implies that

$$\begin{aligned}
\theta_t^{treat}(X) &= \mathbb{E}\left[Y_t(2) - Y_t(0)\middle|X, D = 2\right] \\
&= \mathbb{E}[Y_t - Y_0|X, D = 2] - \mathbb{E}[Y_t - Y_0|X, D = 0]
\end{aligned}$$

By the support condition in Assumption 3, it follows by the Law of Iterated Expectations that

$$\theta_t^{treat} = \int \theta_t^{treat}(X)\, dP_{X|D=2}$$

is non-parametrically point identified.

A.2

Proof of Proposition 3.

Let $\theta_t^{treat}(X) := \mathbb{E}[Y_t(0,2) - Y_t(0,0)|X, D^{(1)} = 0, D^{(2)} = 2]$.

By Assumption 6 along with Assumption 8, we then have

$$\begin{aligned}
&\mathbb{E}[Y_t(0,0)|X, D^{(1)} = 0, D^{(2)} = 2] \\
&\quad = \mathbb{E}[Y_0(0,0)|X, D^{(1)} = 0, D^{(2)} = 2] + \mathbb{E}[Y_t(0,0) - Y_0(0,0)|X, D^{(1)} = 0, D^{(2)} = 2] \\
&\quad = \mathbb{E}[Y_0|X, D^{(1)} = 0, D^{(2)} = 2] + \sum_{k=1}^{t} \mathbb{E}[Y_t(0,0) - Y_{t-1}(0,0)|X, D^{(1)} = 0, D^{(2)} = 2] \\
&\quad = \mathbb{E}[Y_0|X, D^{(1)} = 0, D^{(2)} = 2] + \sum_{k=1}^{t} \delta_k(X)
\end{aligned}$$

This implies that

$$\theta_t^{treat}(X) = \mathbb{E}[Y_t - Y_0|X, D^{(1)} = 0, D^{(2)} = 2] - \sum_{k=1}^{t} \delta_k(X)$$

Using the bounds on $\delta_t(X)$ from Assumption 9 then gives the upper bound on conditional treatment effects

$$\theta_t^{treat}(X) \leq \mathbb{E}[Y_t - Y_0|X, D^{(1)} = 0, D^{(2)} = 2] - \min\{\theta_t^{(2,0)}(X), \theta_t^{(1,0)}(X)\}$$

By the Law of Iterated Expectations, monotonicity of integrals, and with support conditions in Assumption 7, this implies that

$$\theta_t^{treat} \leq \mathbb{E}[Y_t - Y_0|D^{(1)} = 0, D^{(2)} = 2] - \int \min\{\theta_t^{(2,0)}(X), \theta_t^{(1,0)}(X)\}\ dP_{X|\{D^{(1)}=0, D^{(2)}=2\}}$$

Similarly for the lower bound, we have

$$\theta_t^{treat} \geq \mathbb{E}[Y_t - Y_0|D^{(1)} = 0, D^{(2)} = 2] - \int \max\{\theta_t^{(2,0)}(X), \theta_t^{(1,0)}(X)\}\ dP_{X|\{D^{(1)}=0, D^{(2)}=2\}}$$

A.3

Proof of Proposition 6.

Let $\tilde{D}(\omega) = \mathbf{1}_{\{D=2\}}(\omega)$ be a binary indicator. First note that by no anticipation in Assumption 2,

$$\begin{aligned}\mathbb{E}\left[\frac{(Y_t - Y_0)\tilde{D}}{\mathbb{P}(\tilde{D}=1)}\right] &= \frac{1}{\mathbb{P}(\tilde{D}=1)}\Big(\mathbb{E}[(Y_t - Y_0)\tilde{D}|\tilde{D}=1]\mathbb{P}(\tilde{D}=1) + \mathbb{E}[(Y_t - Y_0)\tilde{D}|\tilde{D}=0]\mathbb{P}(\tilde{D}=0)\Big) \\ &= \mathbb{E}[Y_t - Y_0|\tilde{D}=1] = \mathbb{E}[Y_t - Y_0|D=2] \\ &= \mathbb{E}[Y_t(2) - Y_0(0)|D=2]\end{aligned}$$

From Assumptions 3-4,

$$\begin{aligned}\mathbb{E}[Y_t(0) - Y_0(0)|D=2] &= \mathbb{E}[(Y_t(0) - Y_0(0))\tilde{D}|\tilde{D}=1] \\ &= \mathbb{E}\left[\frac{(Y_t(0) - Y_0(0))\tilde{D}}{\mathbb{P}(\tilde{D}=1)}\right] \\ &= \mathbb{E}\left[\frac{\mathbb{E}[Y_t(0) - Y_0(0)|X, \tilde{D}=1]\mathbb{P}(\tilde{D}=1|X)}{\mathbb{P}(\tilde{D}=1)}\right] \\ &= \mathbb{E}\left[\frac{\mathbb{E}[Y_t(0) - Y_0(0)|X, D=0]\mathbb{P}(\tilde{D}=1|X)}{\mathbb{P}(\tilde{D}=1)}\right] \\ &= \mathbb{E}\left[\frac{\mathbb{E}[Y_t - Y_0|X, D=0]\mathbb{P}(\tilde{D}=1|X)}{\mathbb{P}(\tilde{D}=1)}\right] \qquad (*)\end{aligned}$$

We then have

$$\begin{aligned}\mathbb{E}[(Y_t - Y_0)(1-D)|X] &= \mathbb{E}[Y_t - Y_0|X, D=0]\mathbb{P}(D=0|X) - \mathbb{E}[Y_t - Y_0|X, D=2]\mathbb{P}(D=2|X) \\ &= \mathbb{E}[Y_t - Y_0|X, D=0]\mathbb{P}(D=0|X) - \mathbb{E}[Y_t - Y_0|X, \tilde{D}=1]\mathbb{P}(\tilde{D}=1|X) \\ &= \mathbb{E}[Y_t - Y_0|X, D=0]\mathbb{P}(D=0|X) - \mathbb{E}[(Y_t - Y_0)\tilde{D}|X]\end{aligned}$$

This implies that

$$\mathbb{E}[Y_t - Y_0|X, D=0] = \frac{\mathbb{E}\left[(Y_t - Y_0)(1 - D + \tilde{D})\Big|X\right]}{\mathbb{P}(D=0|X)}$$

Combining with $(*)$ then gives

$$\mathbb{E}[Y_t(0) - Y_0(0)|D=2] = \mathbb{E}\left[\frac{(Y_t - Y_0)(1 - D + \tilde{D})\mathbb{P}(\tilde{D}=1|X)}{\mathbb{P}(D=0|X)\mathbb{P}(\tilde{D}=1)}\right]$$

Putting it all together, and it follows that

$$\begin{aligned}
\theta_t^{treat} &= \mathbb{E}[Y_t(2) - Y_t(0)|D=2] \\
&= \mathbb{E}[Y_t(2) - Y_0(0)|D=2] - \mathbb{E}[Y_t(0) - Y_0(0)|D=2] \\
&= \mathbb{E}\left[\frac{(Y_t - Y_0)\Big(\tilde{D}(\mathbb{P}(D=0|X) - \mathbb{P}(\tilde{D}=1|X)) - (1-D)\mathbb{P}(\tilde{D}=1|X)\Big)}{\mathbb{P}(D=0|X)\mathbb{P}(\tilde{D}=1)}\right] \\
&= \mathbb{E}\left[\frac{(Y_t - Y_0)\Big(\mathbf{1}_{\{D=2\}}(\mathbb{P}(D=0|X) - \mathbb{P}(D=2|X)) - (1-D)\mathbb{P}(D=2|X)\Big)}{\mathbb{P}(D=0|X)\mathbb{P}(D=2)}\right] \\
&= \mathbb{E}\left[\frac{(Y_t - Y_0)\Big(\mathbf{1}_{\{D=2\}}\mathbb{P}(D=0|X) - \mathbf{1}_{\{D\neq 2\}}(1-D)\mathbb{P}(D=2|X)\Big)}{\mathbb{P}(D=0|X)\mathbb{P}(D=2)}\right] \\
&= \mathbb{E}\left[\frac{(Y_t - Y_0)\Big(\mathbf{1}_{\{D=2\}}\mathbb{P}(D=0|X) - \mathbf{1}_{\{D=0\}}\mathbb{P}(D=2|X)\Big)}{\mathbb{P}(D=0|X)\mathbb{P}(D=2)}\right]
\end{aligned}$$

The proof for θ_t^{spill} is similar. By defining $\tilde{D} = \mathbf{1}_{\{D=1\}}$, and starting from $(*)$,

$$\mathbb{E}[Y_t - Y_0|D=0, X] = \frac{\mathbb{E}\Big[(Y_t - Y_0)(2 - D - \tilde{D})\Big|X\Big]}{2\mathbb{P}(D=0|X)}$$

This gives

$$\mathbb{E}[Y_t(0) - Y_0(0)|D=1] = \mathbb{E}\left[\frac{(Y_t - Y_0)(2 - D - \tilde{D})\mathbb{P}(\tilde{D}=1|X)}{2\mathbb{P}(D=0|X)\mathbb{P}(\tilde{D}=1)}\right]$$

It follows that

$$\begin{aligned}
\theta_t^{spill} &= \mathbb{E}[Y_t(1) - Y_t(0)|D=1] \\
&= \mathbb{E}[Y_t(1) - Y_0(0)|D=1] - \mathbb{E}[Y_t(0) - Y_0(0)|D=1] \\
&= \mathbb{E}\left[\frac{(Y_t - Y_0)\Big(\tilde{D}(2\mathbb{P}(D=0|X) + \mathbb{P}(\tilde{D}=1|X)) - (2-D)\mathbb{P}(\tilde{D}=1|X)\Big)}{2\mathbb{P}(D=0|X)\mathbb{P}(\tilde{D}=1)}\right] \\
&= \mathbb{E}\left[\frac{(Y_t - Y_0)\Big(\mathbf{1}_{\{D=1\}}2\mathbb{P}(D=0|X) - \mathbf{1}_{\{D\neq 1\}}(2-D)\mathbb{P}(D=1|X)\Big)}{2\mathbb{P}(D=0|X)\mathbb{P}(D=1)}\right] \\
&= \mathbb{E}\left[\frac{(Y_t - Y_0)\Big(\mathbf{1}_{\{D=1\}}\mathbb{P}(D=0|X) - \mathbf{1}_{\{D=0\}}\mathbb{P}(D=1|X)\Big)}{\mathbb{P}(D=0|X)\mathbb{P}(D=1)}\right]
\end{aligned}$$