Online Appendix for Home-Delivery-Subscription Services in Omnichannel Grocery Retailing

A. Model and Identification Assumptions of DiD

A.1. Stacked DiD

Overall, the stacked DiD modeling approach is appropriate for our empirical analysis because it involves a quasi-experimental design in which a group of consumers (treatment group) joined a subscription program at different points in time after its launch. Moreover, as part of this design, we have a control group comprising consumers who did not join the program during the period of analysis (i.e., the "never-treated" consumers). The validity of a stacked DiD modeling approach rests on three main assumptions (Baker et al. 2022):

- **Parallel Trends**. This implies that in the absence of treatment, the difference between the treatment and control groups remains constant over time. For stacked DiD, this must hold in a more generalized form across multiple time periods and groups. We conduct an event study to examine the coefficients on the lead indicators (before treatment) and lag indicators (after treatment). To satisfy this assumption, generally, the coefficients on lead indicators should not be significantly different from zero. This was the case in our study, as we show in Part G of this appendix. Statistically significant coefficients would suggest the existence of pre-treatment differences in trends between treated and control groups.

- **No Anticipation**. The stacked DiD modeling approach also assumes that consumers do not change their behavior in anticipation of treatment. This ensures that the observed dynamics are driven by the treatment effect itself rather than by behavioral changes in anticipation of treatment. Based on our interviews with the retailer, we verified there were no promotions incentivizing consumers to try the subscription program prior to its launch date. Moreover, no consumer had access to the program trials before the launch date. We also conducted a placebo test to verify that no spurious correlations were driving the results (see Section 6.3.2 of the paper and Part H of this appendix). Based on this evidence, it is unlikely that consumers modified their behavior in anticipation of subscription.

- **Stable Unit Treatment Value Assumption (SUTVA)**. According to this assumption, the potential outcomes for any consumer must not depend on other consumers' treatment status. Any "spillover" effects will violate this assumption and bias the results. In our particular case, if consumers can share their subscription with other consumers, this would constitute a violation of this assumption. We verified that this is not possible, as the subscription is tagged to a unique individual's payment account and home address. Given these restrictions, it is unlikely that a non-subscriber would use a subscriber's membership to place an online order for home delivery.

A.2. Two-Way Fixed Effects Regression

We provide a summary of the key issues related to two-way fixed effects (TWFE) regression models under a staggered DiD treatment scenario (Goodman-Bacon 2021, Callaway and Sant'Anna 2021). One issue is that TWFE models might not be an appropriate method to identify the Average Treatment Effect on the Treated (ATT) subjects if the treatment effects are heterogeneous and the timing of the treatment varies across individual subjects (consumers in our case).

Consider a standard DID specification like the one we have in Equation (1a) in the paper:

$$Y_{A,it} = \alpha_{A,i} + \beta_A \times TREAT_i \times AFTER_{it} + \theta_{A,t} + \varepsilon_{A,it}.$$

The parameter β_A attempts to estimate an average difference comparing the same unit across time, as well as comparing different units with and without treatment at the same point in time. The issue mainly resides in the second comparison. When comparing different units, there are three possible comparison groups that the TWFE model does not differentiate: (1) treated units vs. never treated units, (2) treated units vs. not yet treated units, and (3) later treated units vs. earlier treated units. The first two comparison groups do not pose an estimation problem. However, this is not necessarily the case for the third group because the estimation uses control observations that were already treated.

This issue is less apparent if the treatment effect is homogeneous (i.e., there is a one-time upward shift in the potential outcomes). However, if the treatment effect is heterogeneous and potentially changes over time, then incorporating the third comparison group would bias the estimates. This is mainly because the parallel trend

assumption might no longer hold for units receiving the treatment at an earlier time, even if it holds for all units before their respective treatment timing. Hence, by the time the later units receive the treatment, the parallel assumption condition may no longer hold between the earlier and the later treated units.

The literature also discusses the negative weights issue, which is related to the issue highlighted in the third comparison above. Essentially, when estimating ATT, we are comparing treated and untreated units:

$$ATT = \sum(\omega_1 Y_{treated,t} - \omega_0 Y_{untreated,t}).$$

In this case, all treated units receive a positive weight, whereas all untreated units have an effective negative weight due to the subtraction. However, if we compare earlier treated units with later treated units, then we have the following:

$$ATT = \sum(\omega_1 Y_{treated,t} - \omega_0 Y_{earlier_treated,t}).$$

Because the earlier treated units are used as the control units, they are entering the equation subtracting from the later treated units. However, since at time *t*, the control units are already treated, they are each receiving a "negative" weight, when the weight should in fact be positive. Therefore, the TWFE model would place negative weights on some of the treated units because the model uses them as controls.

We note, however, that our estimates do not suffer from the issue of "bad" comparisons (i.e., the third case) as we use only consumers who did not sign up to the subscription program (i.e., never treated units) as controls. By focusing on the first subscription period of each consumer, we also ensure that we observe their transactions over an identical 16-week period and that consumers are not treated multiple times.

The proposed estimator by Callaway and Sant'Anna (2021) resolves the above two issues by avoiding comparisons in the third case. In addition, this approach attempts to estimate the treatment effects by accounting for as much information as is reasonable. As such, it offers greater precision about the target parameter (estimand) and how to aggregate treatment effects across time/units. This approach is superior to the TWFE model, which assumes equal weights across time and units.

Consistent with Callaway and Sant'Anna's (2021) notations, under staggered versions of the parallel trends and no anticipation assumptions, one can identify $ATT(g,t)$, which is the group-time average treatment effect on the treated for cohort first treated in time *g* at time *t*, by comparing the expected change in outcome for cohort *g* between periods $g-1$ and *t* to that of a control group not-yet treated at period *t*. Formally, the model can be specified as

$$ATT(g,t) = \mathbb{E}\left[Y_{i,t} - Y_{i,g-1} \middle| G_i = g\right] - \mathbb{E}\left[Y_{i,t} - Y_{i,g-1} \middle| G_i = g'\right], \text{ for any } g' > t.$$

Note that since this equation holds for any comparison group $g' > t$, it also holds for the average over some set of comparisons $\mathcal{G}_{comp}$ such that $g' > t$ for all $g' \in \mathcal{G}_{comp}$, so that

$$ATT(g,t) = \mathbb{E}\left[Y_{i,t} - Y_{i,g-1} \middle| G_i = g\right] - \mathbb{E}\left[Y_{i,t} - Y_{i,g-1} \middle| G_i = \mathcal{G}_{comp}\right].$$

Callaway and Sant'Anna (2021) consider two options for $\mathcal{G}_{comp}$: using only never-treated units ($\mathcal{G}_{comp} = \{\infty\}$) or all not-yet-treated units ($\mathcal{G}_{comp} = \{g' : g' > t\}$). We use the former approach in the manuscript. We can then estimate $ATT(g,t)$ by replacing the expectations with their sample analogs, so that

$$\widehat{ATT}(g,t) = \frac{1}{N_g} \sum_{i:G_i=h} [Y_{i,t} - Y_{i,g-1}] - \frac{1}{N_{\mathcal{G}_{comp}}} \sum_{i:G_i \in \mathcal{G}_{comp}} [Y_{i,t} - Y_{i,g-1}].$$

This equation estimates the group-time average treatment effect, that is, the average treatment effect for group *g* at time *t*. We can obtain ATT aggregation across cohorts and time periods by:

$$\overline{ATT} = \frac{\sum(\omega_{g,t} * \widehat{ATT}(g,t))}{\sum(\omega_{g,t})},$$

where $\omega_{g,t}$ denotes a weight, which serves as a proxy for the amount of information used to estimate $\widehat{ATT}(g,t)$. Larger and more precise estimators are given more weights. We report the aggregated estimates across time and units in the tables below (Tables A1–A3). As observed, the estimates are mostly consistent with the those in the paper (see Tables 5–7). We note, however, that the current estimator by Callaway and Sant'Anna (2021) is not the best suited for count outcomes, such as those involving frequency of orders. As such, we should interpret the relevant coefficients with caution (i.e., in Column 2 in Table A2, and Columns 1 and 5 in Table A3), which have smaller magnitudes than those based on a Poisson regression model in the original

manuscript. Given this limitation and the fact that the majority of the coefficients estimated in the TWFE model are more conservative (i.e., smaller in absolute magnitude), we provide the results based on the Callaway and Sant'Anna (2021) estimator as a robustness check.

Table A1. Average Effects of Subscription Program (Product Revenue Aggregated Across and Per Channel)

Variable	Weekly Product Revenue (Home-Delivery) (1)	Weekly Product Revenue (CC) (2)	Weekly Product Revenue (Offline) (3)	Weekly Product Revenue (Aggregated) (4)
Treat × After	11.940*** (1.658)	–0.035 (0.095)	–5.324*** (0.949)	6.582*** (1.753)
Consumer Effects	Yes	Yes	Yes	Yes
Week Effects	Yes	Yes	Yes	Yes
# of observations	855,217	855,217	855,217	855,217

Notes. Cluster robust standard errors at the consumer level in parentheses. * p<0.05, ** p<0.01, *** p<0.001.

Table A2. Average Effects of Subscription Program (Home-Delivery Channel)

Variable	Weekly Fulfillment Fee Revenue (Home-Delivery) (1)	Weekly Order Frequency (Home-Delivery) (2)	Weekly Average Order Size (Home-Delivery)		
			SKUs (3)	Items (4)	Euros (5)
Treat × After	–0.121* (0.050)	0.160*** (0.014)	3.382*** (0.484)	5.306*** (0.904)	10.685*** (1.579)
Consumer Effects	Yes	Yes	Yes	Yes	Yes
Week Effects	Yes	Yes	Yes	Yes	Yes
# of observations	855,217	855,217	855,217	855,217	855,217

Notes. Cluster robust standard errors at the consumer level in parentheses. * p<0.05, ** p<0.01, *** p<0.001.

Table A3. Average Effects of Subscription Program (CC and Offline Channels)

Variable	Weekly Order Frequency (CC) (1)	Weekly Average Order Size (CC)			Weekly Order Frequency (Offline) (5)	Weekly Average Order Size (Offline)		
		SKUs (2)	Items (3)	Euros (4)		SKUs (6)	Items (7)	Euros (8)
Treat × After	–0.003	–0.012	–0.067***	–0.046	–0.088***	–1.560***	–1.753***	–3.810***
	(0.002)	(0.030)	(0.053)	(0.096)	(0.017)	(0.282)	(0.344)	(0.814)
Consumer Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Week Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
# of observations	855,217	855,217	855,217	855,217	855,217	855,217	855,217	855,217

Notes. Cluster robust standard errors at the consumer level in parentheses. * p<0.05, ** p<0.01, *** p<0.001.

B. Model and Identification Assumptions of GRF

This part details the GRF estimator summarized in Section 5.4. In implementing the GRF approach, we follow the standard identifying assumption that once conditioned on the set of observed covariates, the status of treatment should be independent of the unobservables (Rosenbaum and Rubin 1983). Specifically, let the sample observed be represented by (X_i, Y_i, W_i), where X_i represents the set of covariates, and Y_i and W_i denote the outcome variable of interest and the treatment indicator, respectively. Then, based on Rubin's (1974) potential outcomes framework and under the assumption of unconfoundedness, that is, $\left\{Y_i^{(0)}, Y_i^{(1)}\right\} \perp W_i | X_i$, the conditional average treatment effect at a given x is $\beta(x) = \mathrm{E}[Y_i^{(1)} - Y_i^{(0)} | X_i = \mathrm{x}]$. The terms $Y_i^{(0)}$ and $Y_i^{(1)}$ are the potential outcomes experienced by consumer *i* without and with the treatment.

To estimate $\beta(x)$, we need to construct a suitable set of neighbors. A common approach is to first calculate a similarity weight $a_i(x)$ that measures the *i*-th observation's relevance to x. Then, we apply the local generalized method of moments to estimate the conditional average treatment effect $\beta(x)$ using the weighted set of neighbors. The standard method of moments estimates $\beta(x)$ by solving for

$$\hat{\beta}(x) = \left[\sum_{i=1}^{n} a_i(x)\left(\widetilde{W}_i - \overline{W_a}\right)\left(\widetilde{W}_i - \overline{W_a}\right)^T\right]^{-1} \sum_{i=1}^{n} a_i(x)\left(\widetilde{W}_i - \overline{W_a}\right)\left(\widetilde{Y}_i - \overline{Y_a}\right), \quad \text{(B1)}$$

where $\overline{W_a} = \sum a_i(x)\,\widetilde{W}_i$ and $\overline{Y_a} = \sum a_i(x)\,\widetilde{Y}_i$. The traditional approach to estimating $a_i(x)$ typically involves the use of a kernel function that assigns a weight based on some measure of distance. This approach, however, suffers from the curse of dimensionality when there are many covariates. Therefore, rather than using a kernel weighting, our forest-based approach calculates $a_i(x)$ by compiling the number of times the *i*-th observation is in the same causal-tree leaf as x. Mathematically, suppose we let $L_r(x)$ and R denote the observation-set flowing to the same leaf as x in the *r*-th tree and number of trees in a causal forest, respectively, we can then compute a_i based on Equation (B2).

$$a_i(x) = \frac{1}{R}\sum_{r=1}^{R} \frac{1[X_i \in L_r(x)]}{|L_r(x)|}. \quad \text{(B2)}$$

Wager and Athey (2018) specify the following assumptions for a valid analysis based on the GRF procedure:

- **Honesty and Regularity Assumptions**. The GRF methodology requires that the data-splitting approach be "honest" and that the trees satisfy "regularity" to ensure unbiasedness and consistent variance estimation. Honesty in this context means that the same subset of the data used to create each tree is not used to estimate treatment effects. Regularity concerns how splits are performed in each tree, ensuring that splits do not depend heavily on any individual observation. To ensure honesty, we implement cross-validation and out-of-bag (OOB) error estimation to validate the tree predictions, assess model performance and the generalizability of the treatment effects across different subsets of data. In this manner, we also prevent overfitting. To ensure regularity, we use bootstrap resampling techniques to validate the stability and reliability of the estimated treatment effects. This involves repeatedly sampling from the data, with replacement, to create multiple "bootstrap" datasets. For each dataset, we apply the GRF algorithm, examine the distribution and variability of the estimated effects, and then aggregate to summarize the effects.

- **Overlap Assumption.** The overlap assumption requires that every unit in the dataset have a positive probability of receiving each treatment, conditional on covariates. Thus, for any set of covariates, no estimated propensity score (i.e., the probability of receiving treatment given the covariates) should be 0 or 1. Ensuring overlap is key to causal inference validity using GRF, as it enables more reliable estimation of treatment effects across the observed range of covariate values. The overlap assumption can be checked by plotting a histogram of the propensity scores to ensure none are close to 0 or 1. This avoids areas where the propensity scores indicate a deterministic treatment assignment based on covariates and helps ensure that the model can adequately control for confounding and provide accurate estimates of treatment effects.

  To check the overlap assumption, we plot a propensity score histogram of total weekly revenue across all channels as an outcome (Figure B1). We want to observe none of the estimated propensity scores to be close to 0 or 1. In other words, we should not be able to deterministically decide the treatment status of a consumer based on the covariates. Based on the plot, we observe no values are close to one. However, there are some values that are close to zero. To address this limitation, we follow the approach by Li et al. (2018) to estimate the overlap-weighted average treatment effect. Li et al. (2018) show that this estimand is more appropriate when the treatment propensities $e(x)$ are close to 0 or 1 as it does not involve a division by estimated propensities.

$$\frac{\mathrm{E}[e(X)\big(1-e(X)\big)\big(Y(1)-Y(0)\big)]}{\mathrm{E}[e(X)(1-e(X)]},$$

  where $e(X) = \mathrm{P}[W_i = 1 \backslash X_i = x]$, and W_i is the treatment indicator and X_i is the set of covariates. We use this estimand for the heterogeneous analysis based on the revised GRF, which we discuss next.

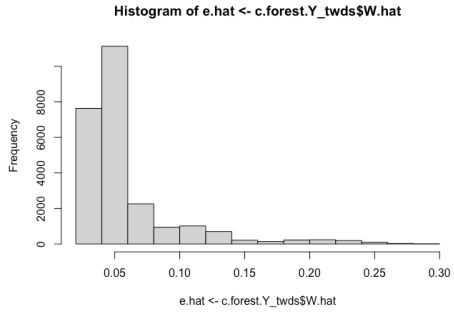


Figure B1. Overlap Assumption Plot

- **Exogeneity (or Unconfoundedness) Assumption**. For the conditional average treatment effect to be identified, the treatment assignment is required to be exogenous, that is, independent of unobservables, conditionally on covariates. Since we cannot ensure that the treatment assignment is exogenous, as consumers may self-select into the subscription program, we implement an instrumental-variables approach to identify the treatment effect of interest. Specifically, we incorporate an Instrumental Variable (IV) into the GRF framework to control for biases arising from unobservable factors that may affect both

the treatment and the outcome. This allows for the estimation of causal effects even in the presence of selection on unobservables.

We compute the percentage of subscribers residing in the same postal code as the focal consumer in a given week as the IV. Note that we exclude the focal consumer's subscription status in this calculation. We argue that the IV is valid; that is, it satisfies the relevance and exclusion restriction conditions. First, the IV is correlated with a consumer's subscription decision due to homophily or social influence (Ma et al. 2015, Bapna and Umyarov 2015). That is, consumers who are close together are more likely to behave similarly or exhibit similar purchasing behavior. In addition, the incremental *F*-Statistic value in the first-stage regression is 20.25 (larger than 10), indicating that the instrument is relevant (Staiger and Stock 1997). Please refer to the first-stage regression results in Table B1. The IV also satisfies the exclusion restriction condition because neighbors' subscriptions should not directly impact the focal consumer's purchase outcomes. Although the percentage of subscribers in a postal code could correlate with socioeconomic factors, we account for this correlation by including consumer fixed effects in our model. We present the results from this analysis below. The estimations (in Figures B2–B4) are consistent with those we obtained in the paper (see Figures 2–4).

Table B1. First-Stage Estimates

Variable	(1) Estimate
Percentage of Subscribers	6.151*** (0.106)
Consumer Effects	Yes
Week Effects	Yes
# of observations	855,217

Notes. Cluster robust standard errors at the consumer level in parentheses.
* $p<0.05$, ** $p<0.01$, *** $p<0.001$.

Figure B2. Distribution of Subscription Program Effect (Product Revenue Aggregated Across/Per Channel)

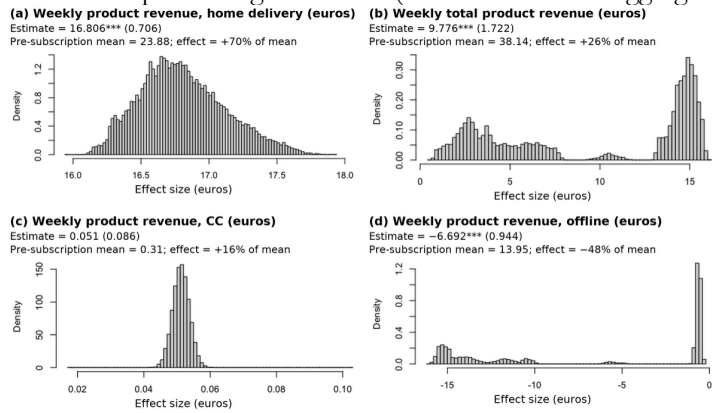


Figure B3. Distribution of Subscription Program Effect
(Fulfillment Revenue and Order Frequency and Size in Home-Delivery Channel)

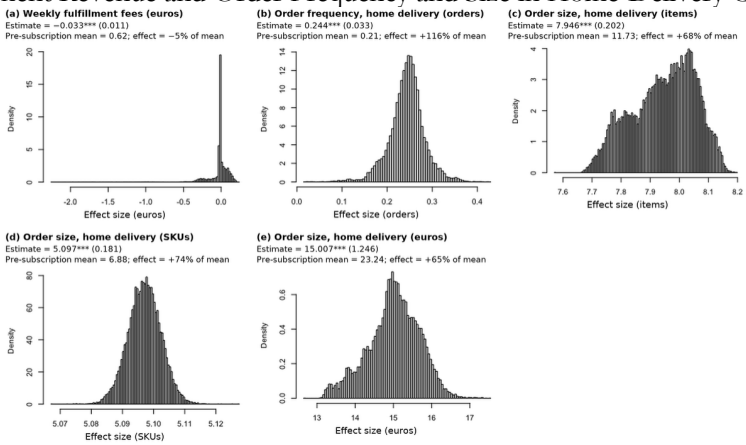
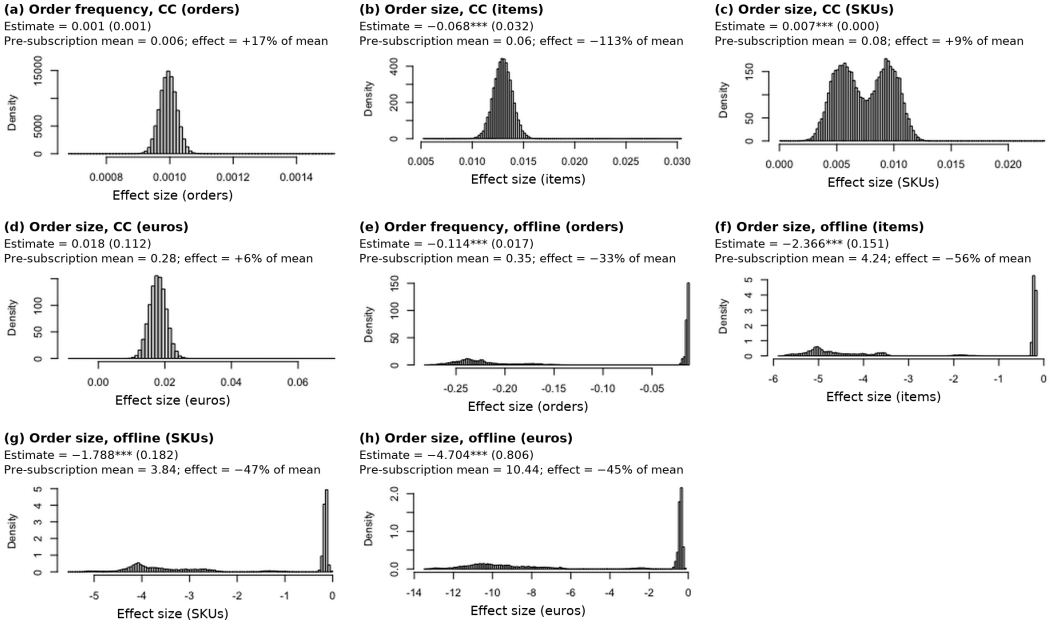


Figure B4. Distribution of Subscription Program Effect
(Order Frequency and Size in CC and Offline Channels)



C. Output Variables' Descriptive Statistics

Table C1. Descriptive Statistics

Output Variables	Subscriber		Nonsubscriber
	Pre-subscription	Post-subscription	
	Average (SD) [Min, Max]	Average (SD) [Min, Max]	Average (SD) [Min, Max]
Weekly Product Revenue in Home-Delivery ($R_{H,it}$)	23.88 (58.94) [0.00, 935.75]	36.40 (64.58) [0.00, 958.08]	14.68 (51.83) [0.00, 996.80]
Weekly Product Revenue in CC ($R_{C,it}$)	0.31 (5.18) [0.00, 176.75]	0.19 (3.82) [0.00, 220.71]	0.20 (5.21) [0.00, 203.92]
Weekly Product Revenue Offline ($R_{O,it}$)	13.95 (36.78) [0.00, 651.54]	9.50 (30.00) [0.00, 815.03]	12.78 (36.87) [0.00, 996.98]
Weekly Total Product Revenue ($R_{A,it}$)	38.14 (67.45) [0.00, 935.75]	46.09 (70.62) [0.00, 967.35]	27.66 (63.19) [0.00, 996.98]
Weekly Fulfillment Fees ($F_{A,it}$)	0.62 (1.84) [0.00, 16.80]	0.42 (0.54) [0.00, 1.125]	0.45 (1.59) [0.00, 46.00]
Weekly Order Frequency in Home-Delivery ($V_{H,it}$)	0.21 (0.43) [0.00, 5.00]	0.42 (0.58) [0.00, 7.00]	0.12 (0.35) [0.00, 12.00]
Weekly Order Frequency in CC ($V_{C,it}$)	0.006 (0.108) [0.00, 7.00]	0.005 (0.072) [0.00, 3.00]	0.003 (0.060) [0.00, 4.00]
Weekly Order Frequency Offline ($V_{O,it}$)	0.35 (0.72) [0.00, 7.00]	0.29 (0.67) [0.00, 7.00]	0.32 (0.73) [0.00, 7.00]
Weekly Average Order Size (SKUs) per Home-Delivery Order ($BQ_{H,it}$)	6.88 (15.70) [0.00, 113.00]	10.92 (17.29) [0.00, 127.00]	3.92 (12.80) [0.00, 215.00]
Weekly Average Order Size (SKUs) per CC Order ($BQ_{C,it}$)	0.08 (1.52) [0.00, 59.00]	0.06 (1.14) [0.00, 67.00]	0.06 (1.51) [0.00, 115.00]
Weekly Average Order Size (SKUs) per Offline Order ($BQ_{O,it}$)	3.84 (9.70) [0.00, 96.00]	2.52 (7.59) [0.00, 169.00]	3.51 (9.72) [0.00, 126.00]
Weekly Average Order Size (Items) per Home-Delivery Order ($BK_{H,it}$)	11.73 (29.96) [0.00, 389.00]	18.00 (32.76) [0.00, 482.00]	7.86 (28.59) [0.00, 1246.00]
Weekly Average Order Size (Items) per CC Order ($BK_{C,it}$)	0.06 (1.14) [0.00, 41.00]	0.10 (1.92) [0.00, 92.00]	0.10 (2.50) [0.00, 482.00]
Weekly Average Order Size (Items) per Offline Order ($BK_{O,it}$)	4.24 (11.42) [0.00, 120.00]	3.26 (10.04) [0.00, 197.00]	4.29 (12.60) [0.00, 468.00]
Weekly Average Order Size (Euros) per Home-Delivery Order ($B_{H,it}$)	23.24 (57.42) [0.00, 935.75]	34.12 (60.39) [0.00, 887.90]	14.30 (49.88) [0.00, 996.80]
Weekly Average Order Value (Euros) per CC Order ($B_{C,it}$)	0.28 (4.68) [0.00, 176.75]	0.18 (3.53) [0.00, 173.43]	0.20 (5.10) [0.00, 203.92]
Weekly Average Order Size (Euros) Value per Offline Order ($B_{O,it}$)	10.44 (27.41) [0.00, 347.13]	7.02 (22.68) [0.00, 461.46]	9.37 (27.87) [0.00, 996.98]

Table C2. Weekly Summary Statistics for Home-Delivery Output Variables During the Pre-Subscription Launch Period

Variable	Subscriber Avg	Nonsubscriber Avg	Difference	*p*-value
Weekly Product Revenue Via Home Delivery [euros, pretax]				
Week 1. Jan 2015	23.630	9.325	14.305	0.000
Week 2. Jan 2015	24.584	14.298	10.286	0.000
Week 3. Jan 2015	22.891	23.734	–0.842	0.747
Week 4. Jan 2015	25.401	22.230	3.171	0.127
Week 5. Jan 2015	24.194	17.957	6.237	0.000
Week 1-5 Average	24.230	17.508	6.722	0.000
Weekly Order Frequency Via Home Delivery				
Week 1. Jan 2015	0.172	0.096	0.076	0.000
Week 2. Jan 2015	0.211	0.121	0.090	0.000
Week 3. Jan 2015	0.198	0.181	0.017	0.298
Week 4. Jan 2015	0.223	0.195	0.028	0.064
Week 5. Jan 2015	0.215	0.152	0.063	0.000
Week 1-5 Average	0.208	0.149	0.059	0.000
Weekly Home-Delivery Order Size [SKUs]				
Week 1. Jan 2015	6.949	2.928	4.020	0.000
Week 2. Jan 2015	7.063	3.667	3.396	0.000
Week 3. Jan 2015	7.000	6.248	0.753	0.235
Week 4. Jan 2015	7.647	6.099	1.548	0.004
Week 5. Jan 2015	7.093	4.898	2.196	0.000
Week 1-5 Average	7.186	4.768	2.418	0.000
Weekly Home-Delivery Order Size [Items]				
Week 1. Jan 2015	15.132	4.953	–10.179	0.000
Week 2. Jan 2015	9.956	7.233	–2.723	0.025
Week 3. Jan 2015	11.028	11.840	0.812	0.533
Week 4. Jan 2015	12.470	10.984	–1.486	0.165
Week 5. Jan 2015	13.541	8.887	–4.655	0.000
Week 1-5 Average	12.487	8.779	–3.708	0.000
Weekly Home-Delivery Order Size [euros, pretax]				
Week 1. Jan 2015	23.537	9.279	–14.258	0.000
Week 2. Jan 2015	24.115	13.957	–10.158	0.000
Week 3. Jan 2015	22.623	23.174	0.551	0.827
Week 4. Jan 2015	25.149	21.709	–3.440	0.086
Week 5. Jan 2015	23.972	17.479	–6.493	0.000
Week 1-5 Average	23.971	17.120	–6.851	0.000

Notes. Averages computed based on all weeks, including weeks with no purchase. Table 4 in the paper reports the five-week averages.

D. Balancing Subscriber and Nonsubscriber Groups

To demonstrate that the generated forest balances the subscriber (treatment) and nonsubscriber (control) groups well, we construct a series of weighted and unweighted regressions and compare their estimates. Specifically, we use one covariate at a time (e.g., population density) as the outcome variable and the treatment indicator (i.e., whether a consumer is a subscriber) as the independent variable in these regressions. We specify the model as follows:

$$X = \varphi_1 Treatment + \varphi_0, \quad \text{(D1)}$$

where X equals each covariate. A statistically significant φ_1 coefficient implies a significant difference between the subscriber and nonsubscriber groups. Thus, φ_1 should be statistically nonsignificant for a covariate to be balanced across the two groups. Table D1 shows the results from the covariate-balance analysis. Except for the coefficients for two variables (Urban Area in Metroplex 1 (M = 1) and Distance to Nearest Store), all of the φ_1 estimates in the unweighted column are statistically significant at the 0.05 level (i.e., t-statistic = 1.96), suggesting that subscribers have observed characteristics that are different than those of nonsubscribers. The weighted sample constructed by the GRF procedure alleviates these differences. According to the t-statistic values in the weighted column in Table D1, none of the φ_1 estimates in this column are statistically significant (p= 0.05). Thus, based on these results, our approach adequately balances the data.

Table D1. Covariate Balance Test

		Unweighted		Weighted	
		Estimate	***t*-statistic**	**Estimate**	***t*-statistic**
Postal-Code Level Variables					
Population Density		–20924.00 (4499.00)	–4.65	–2531.00 (2120.00)	–1.19
Male Population Amount		–9330.80 (2047.30)	–4.56	–1096.1 (864.90)	–1.13
Buying Power Index		–8.21 (1.30)	–6.28	–1.56 (1.01)	–1.54
Degree of Population Ageing		–4.89 (1.11)	–4.40	–0.72 (0.53)	–1.37
University Degree		–0.01(0.00)	–6.66	–0.00 (0.00)	–1.35
Household Size		0.03 (0.01)	5.28	0.00 (0.00)	1.58
Rural / Urban Area	Rural Area ($M = 0$)	0.04 (0.01)	3.77	0.01 (0.00)	1.01
	Urban Area in Metroplex 1 ($M = 1$)	0.01 (0.01)	1.09	0.00 (0.00)	0.33
	Urban Area in Metroplex 2 ($M = 2$)	–0.05 (0.01)	–3.73	–0.01 (0.01)	1.54
Consumer Level Variables					
Distance to Nearest Store		0.03 (0.05)	0.63	0.02 (0.02)	0.79
Home Delivery Order Size in euros (month before subscription launch)		47.64 (3.65)	13.05	–3.31 (2.08)	–1.59
Home Delivery Spending (month before subscription launch)		173.81 (1.77)	98.14	–16.34 (9.42)	–1.73
Home Delivery Depth (month before subscription launch)		0.48 (0.03)	18.01	–0.02 (0.01)	–1.66
Customer Type (pre-subscription launch)					
Customer Type 1		–0.09 (0.01)	–7.01	–0.04 (0.04)	–1.11
Customer Type 2		0.01 (0.00)	4.95	0.01 (0.00)	1.38
Customer Type 3		0.06 (0.01)	4.72	0.03 (0.02)	1.80
Customer Type 4		0.02 (0.00)	5.45	0.00 (0.00)	1.13

Notes. Std. errors are in parentheses.

E. Regression Results Based on Log-Linear Specification

To obtain percentage estimates for the treatment effects, we follow the specification for the linear models in Equation (1) but take the logarithm of the dependent variables. Specifically,

$$\ln(Y_{A,it}+1) = \alpha_{A,i} + \beta_A \times TREAT_i \times AFTER_{it} + \theta_{A,t} + \varepsilon_{A,it}, \quad \text{(E1a)}$$
$$\ln(Y_{H,it}+1) = \alpha_{H,i} + \beta_H \times TREAT_i \times AFTER_{it} + \theta_{H,t} + \varepsilon_{H,it}, \quad \text{(E1b)}$$
$$\ln(Y_{C,it}+1) = \alpha_{C,i} + \beta_C \times TREAT_i \times AFTER_{it} + \theta_{C,t} + \varepsilon_{C,it}, \quad \text{(E1c)}$$
$$\ln(Y_{O,it}+1) = \alpha_{O,i} + \beta_O \times TREAT_i \times AFTER_{it} + \theta_{O,t} + \varepsilon_{O,it}. \quad \text{(E1d)}$$

We add 1 to the dependent variables in the log-linear model to handle zero values during log transformation. All other terms are defined similarly as Equation (1). We retain the original specification for Poisson models. Based on the new specification, we can interpret the coefficients of interest for linear models as percentages (e.g., $\beta_A \times 100\%$). To interpret the coefficients of interest as a percentage in the Poisson models, we apply the following transformation, for example, $[\exp(\beta_H) - 1] \times 100\%$. As an illustration, based on Column (2) of Table E2, the coefficient is 0.758, so that the average percentage increase in weekly order frequency for the home-delivery channel equals 113.4% ($\exp(0.758) - 1] \times 100\%$). We provide all the results based on Equation (E1) in Tables E1 through E3. All estimates' signs and statistical significance are consistent with those reported in Tables 5 through 7 of the paper. We summarize the percentage values estimated for all effects in Table E4.

Table E1. Average Effects of Subscription Program (Product Revenue Aggregated Across and Per Channel)

Variable	**ln Weekly Product Revenue (Home-Delivery) (1)**	**ln Weekly Product Revenue (CC) (2)**	**ln Weekly Product Revenue (Offline) (3)**	**ln Weekly Product Revenue (Aggregated) (4)**
Treat × After	0.656*** (0.025)	–0.008* (0.004)	–0.278*** (0.019)	0.300*** (0.026)
Consumer Effects	Yes	Yes	Yes	Yes
Week Effects	Yes	Yes	Yes	Yes
# of observations	868,692	868,692	868,692	868,692

Notes. Cluster robust standard errors at the consumer level in parentheses. * p<0.05, ** p<0.01, *** p<0.001.

Table E2. Average Effects of Subscription Program (Fulfillment Revenue and Home-Delivery Channel)

Variable	Weekly Fulfillment Fee Revenue (Home-Delivery) (1)	Weekly Order Frequency (Home-Delivery) (2)	Weekly Average Order Size (Home-Delivery)		
			SKUs (3)	Items (4)	Euros (5)
Treat × After	–0.067*** (0.009)	0.758*** (0.027) †	0.487*** (0.018)	0.555*** (0.021)	0.640 *** (0.024)
Consumer Effects	Yes	Yes	Yes	Yes	Yes
Week Effects	Yes	Yes	Yes	Yes	Yes
# of observations	868,692	868,590 ‡	868,692	868,692	868,692

Notes. Cluster robust standard errors at the consumer level in parentheses. * p<0.05, ** p<0.01, *** p<0.001.
† The corresponding percentage change for weekly order frequency in the home delivery channel is: 113.40% = [exp(0.758) – 1] × 100%.
‡ All observations with zero outcomes are automatically dropped when estimating the Poisson model.

Table E3. Average Effects of Subscription Program (CC and Offline Channels)

Variable	Weekly Order Frequency (CC) (1)	Weekly Average Order Size (CC)			Weekly Order Frequency (Offline) (5)	Weekly Average Order Size (Offline)		
		SKUs (2)	Items (3)	Euros (4)		SKUs (6)	Items (7)	Euros (8)
Treat × After	–0.583* †	–0.005*	-0.005*	–0.008*	–0.352*** †	–0.201***	-0.210***	–0.256***
	(0.285)	(0.002)	(0.003)	(0.004)	(0.025)	(0.013)	(0.014)	(0.018)
Consumer Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Week Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
# of observations	34,892 ‡	868,692	868,692	868,692	431,008 ‡	868,692	868,692	868,692

Notes. Cluster robust standard errors at the consumer level in parentheses. * p<0.05, ** p<0.01, *** p<0.001.
† The corresponding percentage change for weekly order frequency in the CC and offline channel, respectively, is: –44.18% = [exp(–0.583) – 1] × 100% and –29.67% = [exp(–0.352) – 1] × 100%.
‡ All observations with zero outcomes are automatically dropped when estimating the Poisson model.

Table E4. Average Effects of Subscription Program (in Percentage Terms)

Variable	Weekly Fulfillment Revenue	Weekly Product Revenue				Weekly Order Frequency			Average Order Size SKUs (Items) [Euros]		
		Home	CC	Offline	Aggregate	Home	CC	Offline	Home	CC	Offline
Effect	–6.70%	65.60%	–0.80%	–27.80%	30.00%	113.40%	–44.18%	–29.67%	48.70% (55.50%) [64.00%]	–0.50% (-0.51%) [–0.80%]	–20.10% (-21.00%) [–25.60%]

F. Accounting for Weekly Budgets and Discounts

We specify the regression in Equation (F1) to account for the possibility of consumers making spending decisions based on their weekly budgets and the discounts available at the focal retailer, which may lead them to shift spending from competing firms to the retailer after subscribing.

$$R_{A,it} = \alpha_{A,i} + \beta_A \times TREAT_i \times AFTER_{it} + \beta_1 \times \ln(BUDGET_i) + \beta_2 \times \ln(DISCOUNT_{it}) + \theta_{A,t} + \varepsilon_{A,it}, \quad \text{(F1)}$$

where, $R_{A,it}$ represents the weekly total amount of product revenue by consumer i in week t across the home-delivery, CC, and offline channel, $BUDGET_{it}$ is the baseline budget of consumer i , and $DISCOUNT_{it}$ is the amount of discounts (in dollars) received by consumer i in week t. The parameters β_1 and β_2 are the semi-elasticities of budget and discount amounts, respectively. Furthermore, the indicator $TREAT_i$ equals 1 if consumer i is in the treatment group; 0 otherwise. The indicator $AFTER_{it}$ equals 1 if week t corresponds to the posttreatment period (i.e., after customer i subscribed to the delivery subscription program); 0 otherwise. The term α_i denotes consumer fixed effects to account for consumer heterogeneity, and θ_t accounts for weekly time fixed effects due to seasonality or temporal shocks. The inclusion of the two-way fixed effects allows us to control for both time-invariant unobserved consumer characteristics and common time-trends and week-to-week variations. The term ε_{it} represents independent and identically distributed Gaussian random shocks. For robust inferences, we estimate clustered robust standard errors at the consumer level. Table F1 presents the estimates for the predictors' coefficients in Equation (F1).

Table F.1. Average Effects of Subscription Program (Product Revenue Aggregated Across and Per Channel)

Variable	Weekly Product Revenue (Aggregated across Channels) (1)
TREAT ×*AFTER*	5.890*** (0.511)
ln(*BUDGET*)	0.002* (0.001)
ln(*DISCOUNT*)	0.001* (0.000)
Consumer Effects	Yes
Week Effects	Yes
# of observations	868,692

Notes. Clustered robust standard errors at the consumer level in parentheses. * p<0.05, ** p<0.01, *** p<0.001.

G. Parallel Trend Plots

Figure G1. Parallel trend plots (Overall and Home-Delivery)

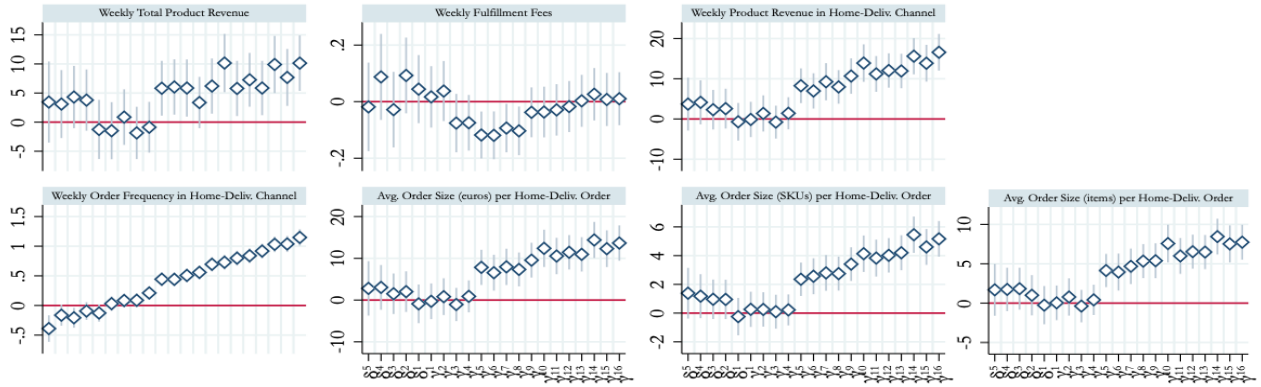
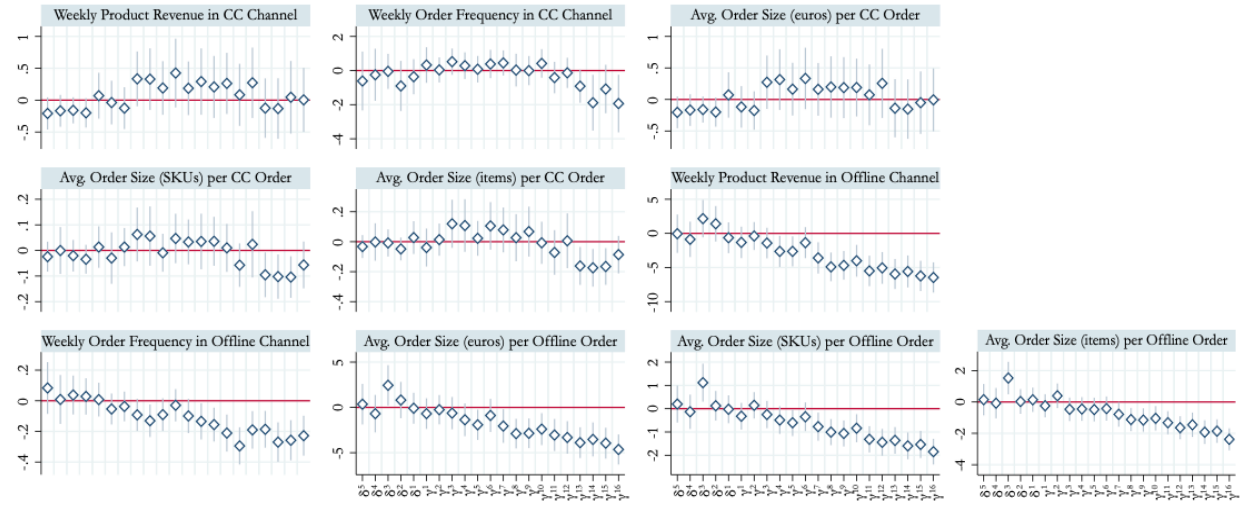


Figure G2. Parallel trend plots (CC and Offline)



Placebo Tests

Table H1. Average Placebo Effects of Subscription Program (Product Revenue Aggregated Across and Per Channel)

Variable	Weekly Product Revenue (Home-Delivery)	Weekly Product Revenue (CC)	Weekly Product Revenue (Offline)	Weekly Product Revenue (Aggregated)
TREAT ×*AFTER*	2.687 (2.055)	–0.042 (0.024)	–2.790 (1.398)	–0.145 (2.269)
Consumer Effects	Yes	Yes	Yes	Yes
Week Effects	Yes	Yes	Yes	Yes
Matching	No	No	No	No
# of observations	52,238	52,238	52,238	52,238

Notes. Cluster robust standard errors at the consumer level in parentheses. * p<0.05, ** p<0.01, *** p<0.001.

Table H2. Average Placebo Effects of Subscription Program (Home-Delivery Channel)

Variable	Weekly Fulfillment Fee Revenue (Home-Delivery) (1)	Weekly Order Frequency (Home-Delivery) (2)	Weekly Average Order Size (Home-Delivery) SKUs (3)	Items (4)	Euros (5)
TREAT ×*AFTER*	–0.045 (0.063)	0.104 (0.124)	1.470 (0.757)	2.466 (1.390)	2.877 (2.053)
Consumer Effects	Yes	Yes	Yes	Yes	Yes
Week Effects	Yes	Yes	Yes	Yes	Yes
Matching	No	No	No	No	No
# of observations	52,238	12,342 †	52,238	52,238	52,238

Notes. Cluster robust standard errors at the consumer level in parentheses. * p<0.05, ** p<0.01, *** p<0.001.
† All observations with zero outcomes are automatically dropped when estimating the Poisson model.

Table H3. Average Placebo Effects of Subscription Program (CC and Offline Channels)

Variable	Weekly Order Frequency (CC)	Weekly Average Order Size (CC)			Weekly Order Frequency (Offline)	Weekly Average Order Size (Offline)		
		SKUs	Items	Euros		SKUs	Items	Euros
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
TREAT ×*AFTER*	–0.609 (1.181)	–0.003	-0.014	–0.042	–0.070 (0.071)	–0.815	-0.858	–2.451
		(0.005)	(0.009)	(0.024)		(0.617)	(0.014)	(1.860)
Consumer Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Week Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Matching	No	No	No	No	No	No	No	No
# of observations	198 †	52,238	52,238	52,238	14,240 †	52,238	52,238	52,238

Notes. Cluster robust standard errors at the consumer level in parentheses. * p<0.05, ** p<0.01, *** p<0.001.
† All observations with zero outcomes are automatically dropped when estimating the Poisson model.

I. Test for Unconfoundedness

I.1. Test for Confoundedness in Control Group Assignment

As an additional robustness check, we use late subscribers, rather than nonsubscribers as control subjects. We chose subscribers who joined the program between May 1 and December 31, 2015, right after the last treatment subject's subscription date (April 30), and repeat the analyses in Section 6.1 of the main paper. Tables I1–I3 present the results from this analysis. These are consistent with those in Tables 5–7 in the paper.

Table I1. Average Effects of Subscription Program (Product Revenue Aggregated Across and Per Channel)

Variable	(1) Weekly Product Revenue (Home-Delivery)	(2) Weekly Product Revenue (CC)	(3) Weekly Product Revenue (Offline)	(4) Weekly Product Revenue (Aggregated)
TREAT ×*AFTER*	8.210*** (0.806)	–0.345*** (0.104)	–5.171*** (0.519)	2.694** (0.857)
Consumer Effects	Yes	Yes	Yes	Yes
Week Effects	Yes	Yes	Yes	Yes
# of observations	140,442	140,442	140,442	140,442

Notes. Cluster robust standard errors at the consumer level in parentheses. * p<0.05, ** p<0.01, *** p<0.001.

Table I2. Average Effects of Subscription Program (Home-Delivery Channel)

Variable	Weekly Fulfillment Fee Revenue (Home-Delivery)	Weekly Order Frequency (Home-Delivery)	Weekly Average Order Size (Home-Delivery)		
			SKUs	Items	Euros
	(1)	(2)	(3)	(4)	(5)
TREAT ×*AFTER*	–0.357*** (0.033)	0.506*** (0.028)	2.827*** (0.204)	5.201** (0.366)	7.253*** (0.741)
Consumer Effects	Yes	Yes	Yes	Yes	Yes
Week Effects	Yes	Yes	Yes	Yes	Yes
# of observations	140,442	140,442	140,442	868,692	140,442

Notes. Cluster robust standard errors at the consumer level in parentheses. * p<0.05, ** p<0.01, *** p<0.001.

Table I3. Average Effects of Subscription Program (CC and Offline Channels)

Variable	Weekly Order Frequency (CC)	Weekly Average Order Size (CC)			Weekly Order Frequency (Offline)	Weekly Average Order Size (Offline)		
		SKUs	Items	Euros		SKUs	Items	Euros
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
TREAT ×*AFTER*	–0.817** (0.278)	–0.140***	–0.005	–0.335***	–0.316*** (0.029)	–1.451***	–2.146***	–3.656***
		(0.031)	(0.007)	(0.094)		(0.127)	(0.154)	(0.366)
Consumer Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Week Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
# of observations	11,093 †	140,442	868,692	140,442	81,727 †	140,442	868,692	140,442

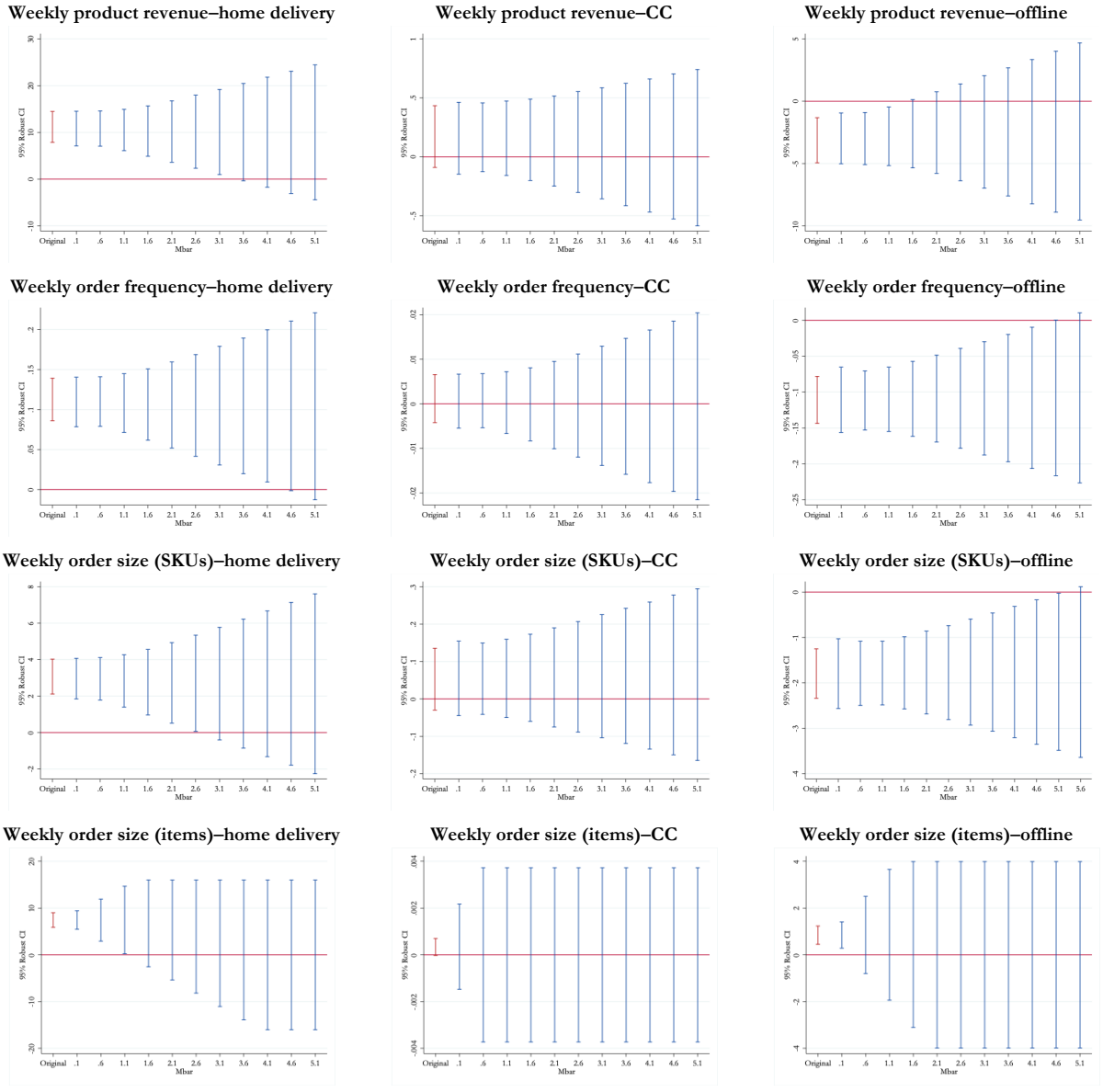
Notes. Cluster robust standard errors at the consumer level in parentheses. * p<0.05, ** p<0.01, *** p<0.001.
† All observations with zero outcomes are automatically dropped when estimating the Poisson model.

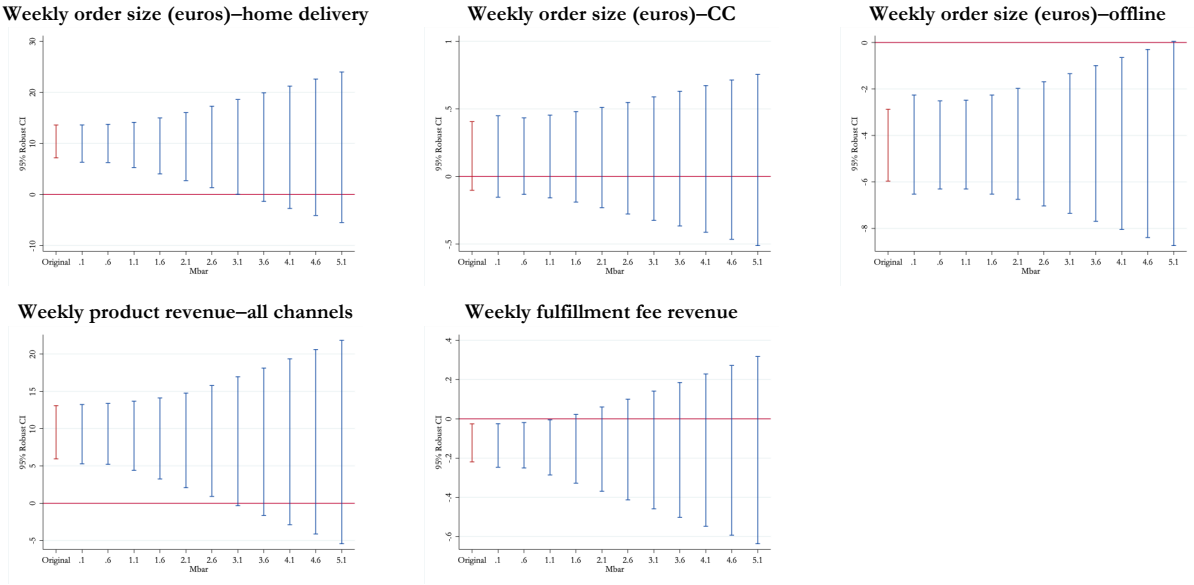
I.2. Time-Varying Unobservables

The presence of time-varying unobservables threatens the parallel trend assumption that is necessary for identification of the causal effect in DiD models. Therefore, following Rambachan and Roth (2023), we identify the largest potential violation of parallel trends between two consecutive periods post-subscription, relative to that observed pre-subscription. Mathematically, this can be formalized by imposing a constraint limiting the post-subscription violation of parallel trends to not exceed the maximum pre-subscription violation by a factor of $\bar{M}$. As an illustration, having $\bar{M} = 2$ would mean that the post-subscription violation of parallel trends cannot be two times larger than the maximum violation in the pre-subscription period.

Accordingly, we generate robust confidence sets of the size of the effect on each output variable for different values of $\bar{M}$. This is denoted by the vertical bars across the values of $\bar{M}$ on the horizontal axis in the plots in Figure I1. In the case of weekly product revenue via the home delivery and offline channels, for instance, we observe that the "breakdown values" for a null effect (i.e., the effect size includes a zero) are $\bar{M} = 3.5$ and 1.5, respectively. We note that these breakdown values, as well as those in the other plots in Figure I1, are significantly higher than those of the other studies which have tested for similar sensitivity and reported robustness for $\bar{M}$ ranging between 0.001 and 0.15 (Ang 2021, Miller et al. 2021, Alpert et al. 2022, Bleemer 2022). Therefore, our results are statistically significant across a broad range of possible deviations in parallel trends, suggesting minimal impact from confounding factors.

Figure I1. Bounds on Relative Magnitudes





J. Difference-in-Differences (DiD) and Propensity Score Matching (PSM)

We use the PSM technique to identify subjects in the control group (nonsubscribers) matching individual subjects in the treatment group (subscribers) based on similarities in observed characteristics. To that end, we use the set of matching covariates in Table 3 of the paper. We provide the results from this analysis below.

Table J1. Average Effects of Subscription Program (Product Revenue Aggregated Across and Per Channel)

Variable	Weekly Product Revenue (Home-Delivery)	Weekly Product Revenue (CC)	Weekly Product Revenue (Offline)	Weekly Product Revenue (Aggregated)
TREAT ×*AFTER*	14.767*** (0.786)	0.254 (0.241)	–5.412*** (0.529)	9.609*** (0.874)
Consumer Effects	Yes	Yes	Yes	Yes
Week Effects	Yes	Yes	Yes	Yes
Matching	Yes	Yes	Yes	Yes
# of observations	102,332	102,332	102,332	102,332

Notes. Cluster robust standard errors at the consumer level in parentheses. * p<0.05, ** p<0.01, *** p<0.001.

Table J2. Average Effects of Subscription Program (Home-Delivery Channel)

Variable	Weekly Fulfillment Fee Revenue (Home-Delivery) (1)	Weekly Order Frequency (Home-Delivery) (2)	Weekly Average Order Size (Home-Delivery)		
			SKUs (3)	Items (4)	Euros (5)
TREAT ×*AFTER*	–0.089** (0.032)	0.846*** (0.036)	4.703*** (0.219)	8.600** (0.352)	13.878*** (0.746)
Consumer Effects	Yes	Yes	Yes	Yes	Yes
Week Effects	Yes	Yes	Yes	Yes	Yes
Matching	Yes	Yes	Yes	Yes	Yes
# of observations	102,332	91,975 †	102,332	102,332	102,332

Notes. Cluster robust standard errors at the consumer level in parentheses. * p<0.05, ** p<0.01, *** p<0.001.
† All observations with zero outcomes are automatically dropped when estimating the Poisson model.

Table J3. Average Effects of Subscription Program (CC and Offline Channels)

Variable	Weekly Order Frequency (CC) (1)	Weekly Average Order Size (CC)			Weekly Order Frequency (Offline) (5)	Weekly Average Order Size (Offline)		
		SKUs (2)	Items (3)	Euros (4)		SKUs (6)	Items (7)	Euros (8)
TREAT ×*AFTER*	–0.157 (0.307)	–0.007 (0.029)	0.005 (0.010)	0.252 (0.237)	–0.323*** (0.030)	–1.487*** (0.136)	2.434*** (0.163)	–3.845*** (0.393)
Consumer Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Week Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Matching	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
# of observations	7,267 †	102,332	102,332	102,332	31,297 †	102,332	102,332	102,332

Notes. Cluster robust standard errors at the consumer level in parentheses. * p<0.05, ** p<0.01, *** p<0.001.
† All observations with zero outcomes are automatically dropped when estimating the Poisson model.

K. Control for Selection on Unobservables Using an IV in the DiD analysis

We adopt the choice of IV in Guo and Liu (2023). We evaluate whether the chosen IV is valid and satisfies the relevance and exclusion restriction condition. First, the IV is correlated with a consumer's subscription decision due to homophily or social influence (Ma et al. 2015, Bapna and Umyarov 2015). That is, consumers who are close together are more likely to behave similarly or exhibit similar purchasing behavior. Furthermore, the incremental *F*-Statistic in the first-stage regression is 20.25 (larger than 10), indicating the instrument is relevant (Staiger and Stock 1997). Please see the first-stage regression results in Table K1. The IV also satisfies the exclusion restriction condition because neighbors' subscription enrollments should not directly impact the focal consumer's purchase outcomes. Although the percentage of subscribers in a postal code could correlate with socioeconomic factors, the model accounts for this correlation by including consumer fixed effects. Furthermore, even if we assume that specific seasons like Christmas drive both the subscription enrollments and consumer purchase outcomes, our inclusion of weekly time effects accounts for this correlation. We present the results from this analysis below. The coefficient estimates for *TREAT* ×*AFTER* (in Tables K2–K4) are largely consistent with those in the paper (see Tables 5–7).

Table K1. First-Stage Estimates

Variable	(1) Estimate
Percentage of Subscribers	6.151*** (0.106)
Consumer Effects	Yes
Week Effects	Yes
# of observations	855,217

Notes. Cluster robust standard errors at the consumer level in parentheses.
* $p<0.05$, ** $p<0.01$, *** $p<0.001$.

Table K2. Average Effects of Subscription Program (Product Revenue Aggregated Across and Per Channel)

Variable	(1) Weekly Product Revenue (Home-Delivery)	(2) Weekly Product Revenue (CC)	(3) Weekly Product Revenue (Offline)	(4) Weekly Product Revenue (Aggregated)
TREAT ×*AFTER*	11.813*** (0.637)	–0.148 (0.078)	–5.478*** (0.437)	6.186*** (0.686)
gr	–7.662 (5.940)	–6.209 (6.359)	4.177 (3.942)	–9.694 (8.996)
Consumer Effects	Yes	Yes	Yes	Yes
Week Effects	Yes	Yes	Yes	Yes
# of observations	855,217	855,217	855,217	855,217

Notes. Cluster robust standard errors at the consumer level in parentheses. * $p<0.05$, ** $p<0.01$, *** $p<0.001$.

Table K3. Average Effects of Subscription Program (Home-Delivery Channel)

Variable	Weekly Fulfillment Fee Revenue (Home-Delivery) (1)	Weekly Order Frequency (Home-Delivery) (2)	Weekly Average Order Size (Home-Delivery) SKUs (3)	Items (4)	Euros (5)
TREAT ×*AFTER*	–0.212*** (0.028)	0.762*** (0.027)	3.783*** (0.181)	5.625*** (0.280)	10.434*** (0.599)
gr	0.410** (0.127)	2.094** (0.654)	–3.936* (1.534)	–0.600 (2.709)	–10.962* (4.283)
Consumer Effects	Yes	Yes	Yes	Yes	Yes
Week Effects	Yes	Yes	Yes	Yes	Yes
# of observations	855,217	855,217	855,217	855,217	855,217

Notes. Cluster robust standard errors at the consumer level in parentheses. * $p<0.05$, ** $p<0.01$, *** $p<0.001$.

Table K4. Average Effects of Subscription Program (CC and Offline Channels)

Variable	Weekly Order Frequency (CC) (1)	Weekly Average Order Size (CC) SKUs (2)	Items (3)	Euros (4)	Weekly Order Frequency (Offline) (5)	Weekly Average Order Size (Offline) SKUs (6)	Items (7)	Euros (8)
TREAT ×*AFTER*	–0.508 (0.304)	–0.049* (0.023)	0.004 (0.007)	–0.132* (0.066)	–0.329*** (0.024)	–1.534*** (0.116)	–2.273*** (0.139)	–3.893*** (0.319)
gr	–2.368 (1.458)	–3.804 (2434)	0.591* (0.259)	–6.608 (6.337)	0.655 (0.638)	0.904 (1.260)	0.052 (2.641)	3.395 (3.631)
Consumer Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Week Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
# of observations	855,217	855,217	855,217	855,217	855,217	855,217	855,217	855,217

Notes. Cluster robust standard errors at the consumer level in parentheses. * $p<0.05$, ** $p<0.01$, *** $p<0.001$.

L. Hierarchical Clustering Procedure and Results

To identify the subscriber segments in our data, we use a hierarchical clustering approach, as opposed to a *k*-means clustering procedure, which requires no a priori specification of the number of clusters. This approach is based on two features: the subscriber-level estimation of the average effects on weekly gross profit (sales – COGS – delivery cost – picking cost) and the estimation of the average effects on weekly total cost (COGS + delivery cost + picking cost). The hierarchical clustering procedure comprises three steps. First, we standardize the two features to have a mean of 0 and a standard deviation of 1. This guarantees that the clustering algorithm will not depend on an arbitrary variable unit. Second, to decide on the measure of dissimilarity between two clusters of observations (i.e., linkage), we compute an agglomerative coefficient through four commonly used linkage clustering methods: average, single, complete, and Ward's Minimum Variance (Han et al. 2011). The agglomerative coefficient specifies the strength of the clustering structure (values closer to 1 suggest a strong clustering structure). As shown in Table L1, the Ward's Minimum Variance method identifies the strongest clustering structure among the methods assessed. Thus, we employ Ward's Minimum Variance as our method.

Table L1. Agglomerative Coefficients

Average	Single	Complete	Ward
0.995	0.951	0.996	0.999

Third, we generate a dendrogram (Figure L1) to identify the number of clusters. To obtain this number, we cut the dendrogram with a horizontal line at a height where the line can traverse the maximum distance up and down without intersecting the merging point. In our case, this occurs when the cut is between heights 0 and 20. As shown in Figure L2, the cut between these heights yields 3 segments: Segment 1 (N = 639), Segment 2 (N = 424), and Segment 3 (N = 399). We checked the validity of using these segments by performing an "elbow" method. As shown in Figure L3, the plot from this method shows that the elbow occurs at 3 segments.

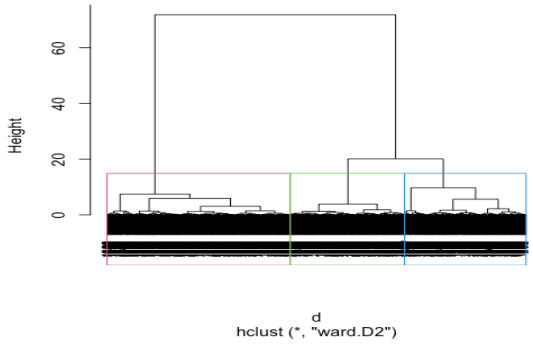


Figure L1. Cluster Dendrogram



Figure L2. Clustering Plot

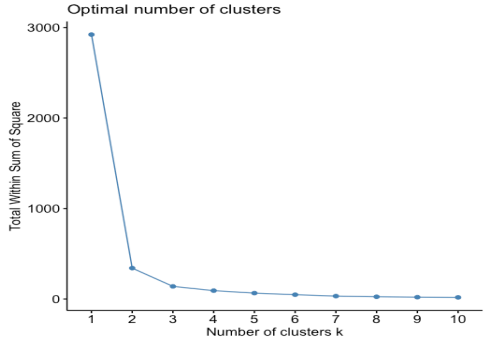


Figure L3 "Elbow" Plot

References

Alpert, A., Evans, W.N., Lieber, E.M., Powell, D. 2022. Origins of the opioid crisis and its enduring impacts. *The Quarterly Journal of Economics,* **137**(2): 1139–1179.

Ang, D. 2021. The effects of police violence on inner-city students. *The Quarterly Journal of Economics,* **136**(1): 115–168.

Baker, A.C., Larcker, D.F., Wang, C.C. 2022. How much should we trust staggered difference-in-differences estimates? *Journal of Financial Economics,* **144**(2): 370–395.

Bapna, R., Umyarov, A. 2015. Do your online friends make you pay? A randomized field experiment on peer influence in online social networks. *Management Science,* **61**(8): 1902–1920.

Bleemer, Z. 2022. Affirmative action, mismatch, and economic mobility after California's proposition 209. *The Quarterly Journal of Economics,* **137**(1): 115–160.

Callaway, B., Sant'Anna, P.H. 2021. Difference-in-differences with multiple time periods. *Journal of Econometrics*, **225**(2), 200–230.

Goodman-Bacon, A. 2021. Difference-in-differences with variation in treatment timing. *Journal of Econometrics,* **225**(2): 254–277.

Guo, F. and Liu, Y. 2023. The effectiveness of membership-based free shipping: An empirical investigation of consumers' purchase behaviors and revenue contribution. *Journal of Marketing,* 87(6): 869-888

Han, J, J. Pei, M. Kamber. 2011. Data Mining: Concepts and Techniques (Elsevier).

Li, F., Morgan, K.L. and Zaslavsky, A.M. 2018. Balancing covariates via propensity score weighting. *Journal of the American Statistical Association,* **113**(521): 390–400.

Ma, L., Krishnan, R. and Montgomery, A.L. 2015. Latent homophily or social influence? An empirical analysis of purchase within a social network. *Management Science,* **61**(2): 454–473.

Miller, S., Johnson, N., Wherry, L.R. 2021. Medicaid and mortality: new evidence from linked survey and administrative data. *The Quarterly Journal of Economics,* **136**(3): 1783–1829.

Rambachan, A., Roth, J. 2023. A more credible approach to parallel trends. *Review of Economic Studies*, 90(5): 2555–2591.

Rosenbaum, P.R., Rubin, D.B. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1): 41–55.

Rubin, D.B. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5): 688–701.

Staiger, D., Stock, J.H. 1997. Instrumental variables regression with weak instruments. *Econometrica*, **65**(3): 557–586.

Wager, S., Athey, S. 2018. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association,* **113**(523): 1228–1242.