Online Appendix

for

Advertising as Information for Ranking E-Commerce Search Listings

Joonhyuk Yang Navdeep S. Sahni Harikesh S. Nair Xi Xiong

A Heterogeneous Treatment Effects

A.1 Effects by user activeness

More active users, by virtue of using the platform more, are more likely to get exposed to our treatment relative to other users. Hence, we expect the estimate for the difference between treatment A and B to be measured more precisely for active users.

In implementing the idea, however, we avoid to group users based on their activeness during the experiment period, since the in-experiment usage behavior can be endogenously determined by the treatment conditions. Instead, we use users' search engine usage behavior in the 30-day pre-experiment period with an underlying assumption that users' activeness between pre- and in-experiment periods are positively correlated even in the absence of our treatments.

Specifically, we define *inactive* users as those who made no searches in the 30-day period prior to the experiment. On the contrary, *active* users are those who searched at least once during the same period. Among the 7,687,390 users assigned to either treatment A or B groups, more than half submitted at least one search in the pre-experiment period, constituting the active user group. On average, active users, by our definition, made about 2.35 times more searches than inactive users during the experiment period (p-value $< .001$). Also, active users had about 2.19 (1.89) times more searches that had at least one (advertised) new product than inactive users (p-values $< .001$), which suggests that the treatment intensity was greater for active users.[1]

To check how the differences in our outcome measures between treatment A and B vary by user activeness, we estimate the following regression:

$$Y_i = \beta_0 + \beta_1 \cdot \text{Active}_i + (\beta_2 \cdot \text{Inactive}_i + \beta_3 \cdot \text{Active}_i) \cdot \text{Treatment B}_i + \varepsilon_i, \tag{A.1}$$

where Inactive_i and Active_i are binary variables that indicate user activeness. β_2 and β_3 are the parameters of interest, each of which separately reports the estimated average treatment effect for the two groups of users.

Table A.1 reports the estimation results. We find that the average treatment effects are positive and statistically significant ($p < .05$) for active users. The magnitudes of estimates and the percent difference from baseline are also greater than the numbers we report in Table 3. For instance, conversion is 0.30% greater and GMV is 0.73% greater in treatment B than in treatment A. On the other hand, the estimates are negative but statistically insignificant for inactive users (p-values = 0.705, 0.770, and 0.866, respectively).

[1] For confidentiality purposes, we do not disclose the actual numbers.

Table A.1: Treatment effects by user activeness

	DV: Conversion‡		DV: log(1+#Orders)‡		DV: log(1+GMV)‡	
	Estimate	Robust SE	Estimate	Robust SE	Estimate	Robust SE
β_3: Treatment B						
× Active users	0.00550**	0.00238	0.00201**	0.00077	0.00626**	0.00261
Percent difference†	0.30%		0.47%		0.73%	
95% CI	[.05,.56%]		[.12%,.83%]		[.13%,1.3%]	
β_2: Treatment B						
× Inactive users	−0.00119	0.00342	−0.00041	0.00094	−0.00186	0.00351
Percent difference†	−0.12%		−0.18%		−0.30%	
95% CI	[-.79%,.55%]		[-.97%,.62%]		[-1.4%,.81%]	
β_1: Active user	0.82165**	0.00295	0.26388**	0.00086	0.97238**	0.00309
β_0: Intercept	1.00000**	0.00242	0.26388**	0.00067	0.81397**	0.00248

Notes: Table reports the estimation results of the multiple regressions in Equation 1 for each of three outcome variables. Inactive (active) users indicate those who submitted no (at least one) search query term in the 30-day period prior to the experiment. N=7,687,390.

† The percent differences for #Orders and GMV are computed after inversing the log transformation. The 95% confidence intervals are computed based on one million random draws for each parameter estimate.

‡ For confidentiality purposes, we mask the actual values of estimates and standard errors. For conversion, masking is done after estimation to take advantage of the properties of binomial approximation of a binary variable. We divide the estimates and standard errors by the estimated intercept. For #Orders and GMV, masking is done before estimation by multiplying each variable by an undisclosed constant. The two constants may or may not be the same.

* p-value < 0.1; ** p-value < 0.05

A.2 Effects by user treatment score based on query terms

Our treatment intensity varies across user search queries because queries vary in their tendency to show advertised new products in treatment B, relative to A. Hence, we expect a larger effect size among users who tend to submit query terms with higher treatment intensity.

To implement this idea, we start by splitting users into two groups—one for estimating each query term's treatment intensity, and another for estimating the treatment effect. We randomly select 10% of users for the first task. Using data on this subset of users, we compute for each query term, the difference in the average ad-propensity score of all new products in the search results between treatment B and A conditions. To be more specific, consider a query term q. Suppose a user in our 10% sample and in treatment A condition submitted q and retrieved a search results page. From the search results page, the user sees a certain number of new products, wherein the sum of these new products' ad-propensity scores is s_q^A. This value can vary across time, so we take the average across all occasions in which q is submitted under treatment A during our experiment period. We do the same for the users under treatment B and obtain s_q^B. Then, we take the difference, $\Delta s_q = s_q^B - s_q^A$, which we define as the *query treatment score* for q. A higher value of Δs_q indicates that users who search for q are more likely to see advertised new products under treatment B than under treatment A.[2]

[2]Because we only utilize a small portion of our sample to compute Δs_q, it may not cover all the query terms

We then compute the *user treatment score* for each user in the remaining 90% of the data. The user treatment score of user i, Treatment score$_i$, is defined as the average of s_q's for all query terms submitted by the user during the *pre*-experiment period and only the *first* query term submitted during the in-experiment period.[3] For instance, Treatment score$_i > 0$ indicates that, prior to the experiment, user i submitted queries with higher treatment intensity relative to the average user. On average, users with Treatment score$_i > 0$ made 1.31 times more searches and had about 1.51 (1.35) times more searches that had at least one (advertised) new product than users with Treatment score$_i <= 0$ during the experiment period ($p < .001$).

To see how the treatment effect varies by Treatment score$_i$, we estimate the following regression:

$$Y_i = \beta_0 + \beta_1 \cdot 1[\text{Treatment score}_i > 0] + \{\beta_2 \cdot 1[\text{Treatment score}_i \leq 0] + \beta_3 \cdot 1[\text{Treatment score}_i > 0]\} \cdot \text{Treatment B}_i + \varepsilon_i. \quad \text{(A.2)}$$

Table A.2: Treatment effects by treatment score

	DV: Conversion‡		DV: log(1+#Orders)‡		DV: log(1+GMV)‡	
	Estimate	Robust SE	Estimate	Robust SE	Estimate	Robust SE
β_3: Treatment B						
× 1[Treatment score>0]	0.00476*	0.00261	0.00170*	0.00098	0.00530	0.00327
Percent difference†	0.27%		0.38%		0.60%	
95% CI	[-.02%,.57%]		[-.05%,.80%]		[-.13%,1.3%]	
β_2: Treatment B						
× 1[Treatment score≤0]	0.00041	0.00251	0.00070	0.00086	0.00220	0.00311
Percent difference†	0.04%		0.25%		0.32%	
95% CI	[-.45%,.53%]		[-.35%,.86%]		[-.55%,1.2%]	
β_1: 1[Treatment score>0]	0.73422**	0.00256	0.27933**	0.00092	0.91141**	0.00319
β_0: Intercept	1.00000**	0.00177	0.32467**	0.00061	1.21073**	0.00220

Notes: Table reports the estimation results of the multiple regressions in equation (A.2) for each of three outcome variables using randomly selected 90% of users. Using the data on the remaining 10%, we compute for each search query term the difference in the ad-propensity score of new products in the search results between treatment B and A conditions. A user's treatment score is defined as the average of the differences for all query terms submitted by the user during the pre-experiment period and the first query term submitted during the experiment period. N=6,918,315.

† The percent differences for #Orders and GMV are computed after inversing the log transformation. The 95% confidence intervals are computed based on one million random draws for each parameter estimate.

‡ For confidentiality purposes, we mask the actual values of estimates and standard errors. For conversion, masking is done after estimation to take advantage of the properties of binomial approximation of a binary variable. We divide the estimates and standard errors by the estimated intercept. For #Orders and GMV, masking is done before estimation by multiplying each variable by an undisclosed constant. The two constants may or may not be the same.

* p-value < 0.1; ** p-value < 0.05

and/or differ to the values from the entire sample. In our case, the 10% sample covers 85.2% of searches in all pre-experimental searches and the first searches during the experiment. The correlation in the values of Δs_q between the 10% sample and the full data is 0.904.

[3] By doing so, we try to minimize the concern that users' choice of search query terms during in-experiment period can be affected by treatment conditions.

The parameters of interest are β_2 and β_3, which represent the separate treatment effects for low and high treatment intensity groups.

Estimates reported in Table A.2 show that the estimates are positive across all outcome variables regardless of the user treatment score. However, the treatment effect is greater, and more precise for user groups with strictly positive treatment score. For instance, conversion in treatment B is .27% greater than in treatment A for users with higher treatment score while the difference is about .04% for users with lower treatment score. We detect similar patterns for #Orders and GMV.

B Additional Tables and Figures

Table B.1: Randomization checks: comparing pre-experimental user behavior

Variable[†]	Difference between A and B: p-value[‡]
1[visit]	0.263
#Searches	0.995
#Searches|visit	0.727
#ProdViewedPerSearch	0.646
#ClicksPerSearch	0.521
#OrdersPerSearch	0.294
#TotalNewProdViewed	0.986
#TotalClicksNewProd	0.762
#TotalOrdersNewProd	0.718

Notes: Table compares the pre-experimental usage of users in the three treatment groups during the 30-day period prior to the first date of the experiment. N(treatment A)=3,841,923; N(treatment B)=3,845,467. We do not report the raw means and standard deviations for confidentiality purposes.

[†] 1[visit] is a binary variable that indicates whether a user submitted a search query (1 if submitted or 0 otherwise); #Searches is the number of searches unconditional on 1[visit]; #Search|visit is the number of searches conditional on 1[visit]; #ProdViewedPerSearch is the average number of products viewed per search; #ClicksPerSearch is the average number of clicks per search; #OrdersPerSearch is the average number of orders per search; #TotalNewProdViewed is the total number of new products viewed; #TotalClicksNewProd is the total number of clicks for new products; and #TotalOrdersNewProd is the total number of orders for new products.

[‡] The p-values are from a two-sided t-test for the equality of means with unequal variances assumption.

Table B.2: Results from trimmed mean comparisons: Treatment A versus B

	Trimming				
	$\gamma = 0$	$\gamma = 0.05$	$\gamma = 0.10$	$\gamma = 0.15$	$\gamma = 0.2$
	(A) #Orders				
Treatment A					
Mean	1.0000	0.5804	0.4188	0.3002	0.2272
SD	2.5979	1.0910	0.7946	0.5941	0.5327
Treatment B					
Mean	1.0058	0.5830	0.4213	0.3017	0.2290
SD	2.7474	1.0936	0.7979	0.5952	0.5344
Percent difference	0.58%	0.44%	0.61%	0.51%	0.78%
p-value[†]	0.002	0.020	0.010	0.037	0.001
	(B) Gross Merchandise Value (GMV)				
Treatment A					
Mean	1.0000	0.3312	0.1979	0.1167	0.0583
SD	6.2325	0.7368	0.4327	0.2709	0.1568
Treatment B					
Mean	1.0027	0.3327	0.1994	0.1178	0.0591
SD	6.7274	0.7384	0.4351	0.2729	0.1553
Percent difference	0.27%	0.46%	0.75%	0.99%	1.41%
p-value	0.548	0.084	0.019	0.022	0.032
N	7,687,390	6,918,652	6,149,914	5,381,174	4,612,436

Notes: In each column, the mean and the standard deviation (SD) are computed using 5,000 bootstrap samples of observations in treatment group. For confidentiality purposes, numbers are reported relative to the mean value of each variable at $\gamma = 0$, which is set to one.

[†] The p-values are from the percentile bootstrap method for testing the equality of means.

Table B.3: DID results on seller responses (all data)

	NewSKU	log(NewSKU)	AvgRating	Share45	Share45New
β: Treated	0.236	0.272*	−0.001	−0.001	0.001
	(0.249)	(0.132)	(0.002)	(0.003)	(0.001)

Notes: Table reports the estimation results of Equation 4. NewSKU is the number of new products added to the platform; AvgRating is the average star rating (1,...,5) of all listed products; Share45 is the share of star ratings 4 and 5; Share45New is the share of star ratings 4 and 5 for new products. Standard errors are clustered at the product category level. $N = 460$. Fixed effects are not reported for brevity.
* p-value < 0.1; ** p-value < 0.05

Table B.4: DID results on sales (all data)

	Total	New	Non-new	log(Total)	log(New)	log(Non-new)
	(A) #Orders					
β: Treated	0.114	1.045	0.099	0.054	0.391**	0.040
	(0.100)	(0.666)	(0.103)	(0.048)	(0.166)	(0.049)
	(B) Gross Merchandising Value (GMV)					
β: Treated	0.053	1.860*	−0.020	0.025	0.570***	−0.006
	(0.096)	(1.000)	(0.107)	(0.074)	(0.166)	(0.078)

Notes: Table reports the estimation results of Equation 4 for #Order or GMV. Standard errors are clustered at the product category level. $N = 460$. Fixed effects are not reported for brevity.
* p-value < 0.1; ** p-value < 0.05

Table B.5: DID results on seller responses (cohort 1)

	NewSKU	log(NewSKU)	AvgRating	Share45	Share45New
β: Treated	0.244	0.137	0.003*	0.004*	0.0004
	(0.202)	(0.199)	(0.002)	(0.002)	(0.002)

Notes: Table reports the estimation results of Equation 4. NewSKU is the number of new products added to the platform; AvgRating is the average star rating (1,...,5) of all listed products; Share45 is the share of star ratings 4 and 5; Share45New is the share of star ratings 4 and 5 for new products. Standard errors are clustered at the product category level. $N = 276$. Fixed effects are not reported for brevity.
* p-value < 0.1; ** p-value < 0.05

Table B.6: DID results on sales (cohort 1)

	Total	New	Non-new	log(Total)	log(New)	log(Non-new)
	(A) #Orders					
β: Treated	0.136**	1.932	0.106**	0.011	0.427	−0.006
	(0.052)	(1.507)	(0.048)	(0.063)	(0.375)	(0.066)
	(B) Gross Merchandising Value (GMV)					
β: Treated	0.106	2.851**	−0.005	−0.034	0.694**	−0.079
	(0.091)	(1.223)	(0.119)	(0.077)	(0.306)	(0.084)

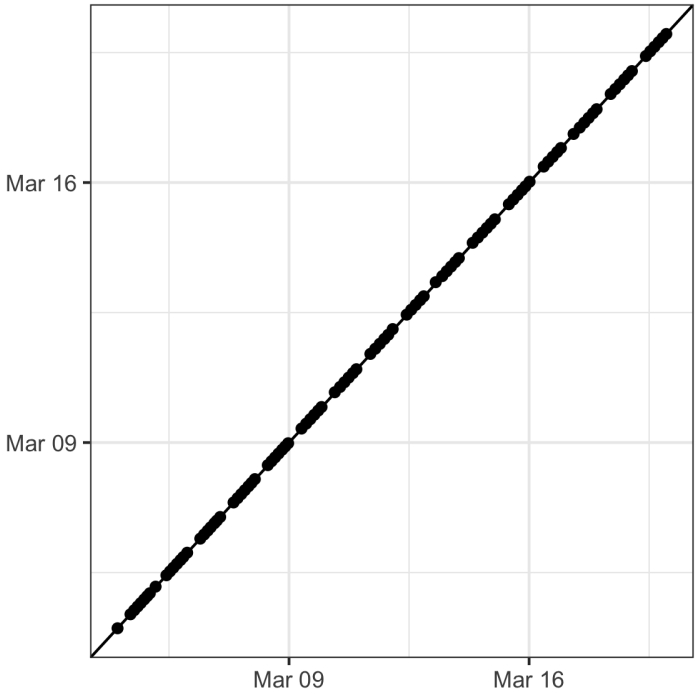
Notes: Table reports the estimation results of Equation 4 for #Order or GMV. Standard errors are clustered at the product category level. $N = 276$. Fixed effects are not reported for brevity.
* p-value < 0.1; ** p-value < 0.05

Table B.7: DID results on sponsored clicks

	log(Sponsored clicks)	log(All clicks)
β: Treated	0.110	0.064
	(0.082)	(0.108)

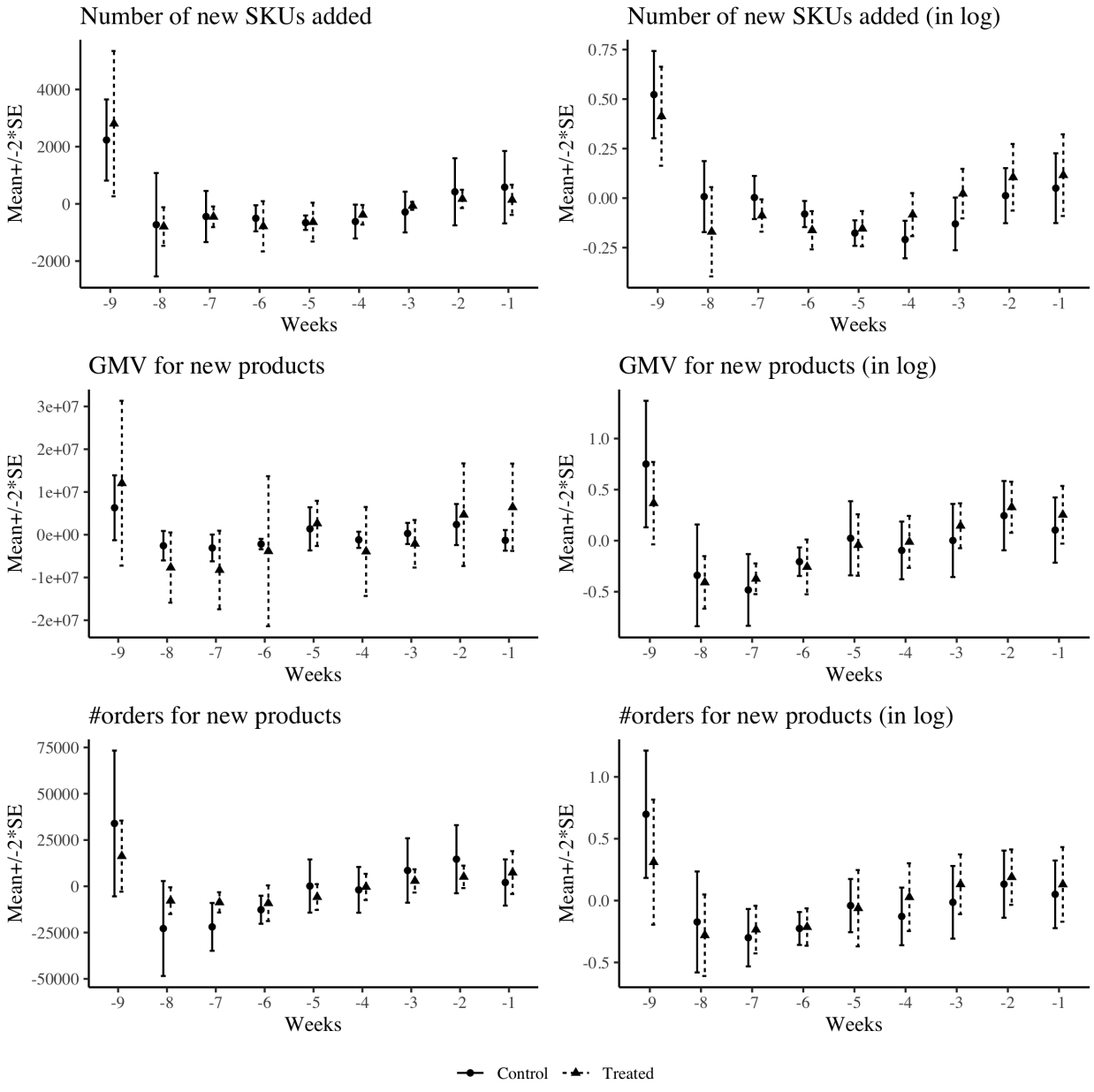
Notes: Table reports the estimation results of Equation 4 for #Sponsored clicks or All clicks. 'Sponsored clicks' is the number of search clicks for sponsored listings and 'All clicks' is the number of search clicks for all listings. $N = 391$. Fixed effects are not reported for brevity.
* p-value < 0.1; ** p-value < 0.05

Figure B.1: Randomization check: comparing user arrival date and time



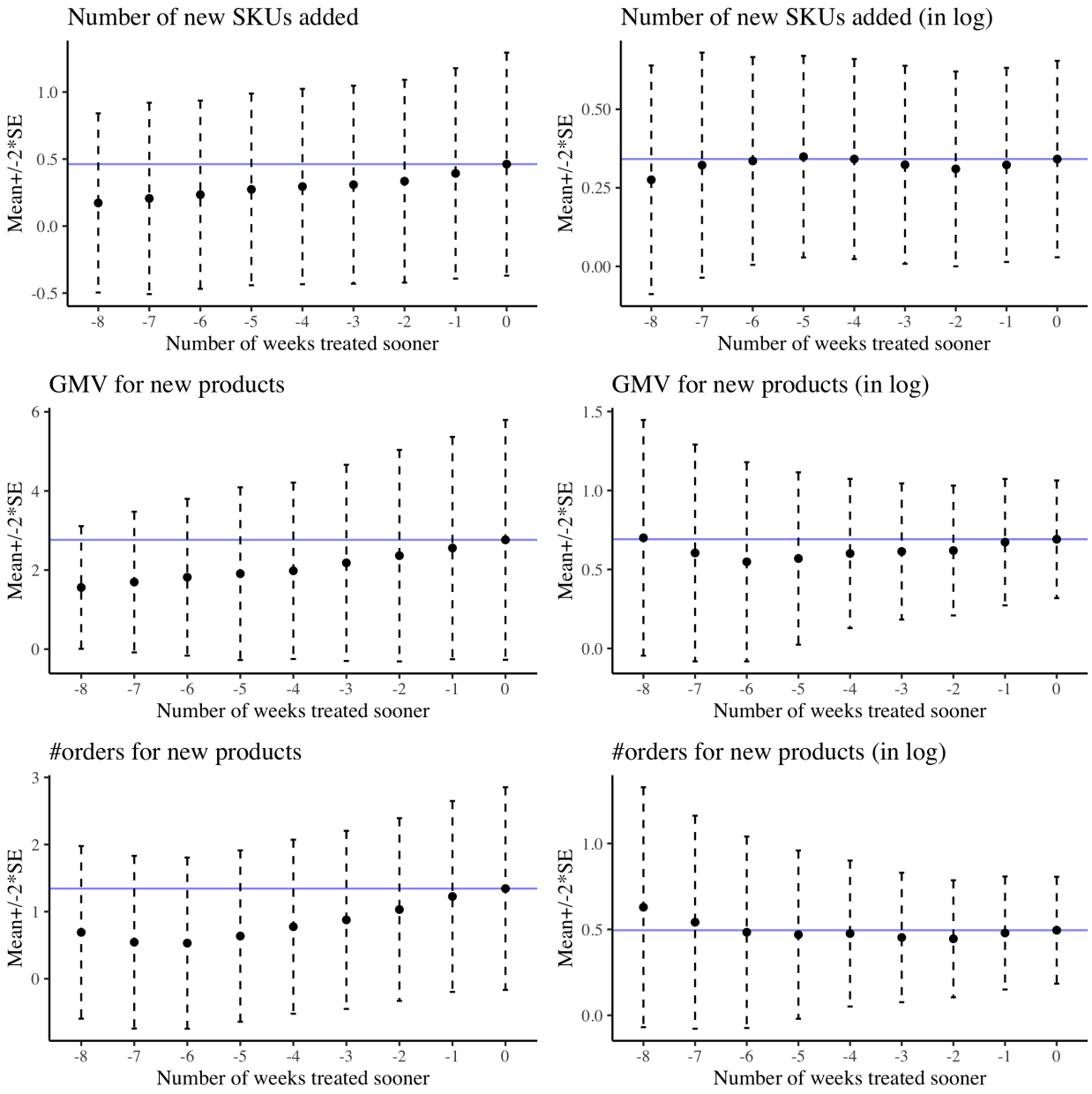
Notes: Figure compares the arrival date and time of users in the the treatment groups by reporting the n-th quantile-quantile values where $n = (0, 1, ..., 100)$. A 45-degree line is overlaid in the plot.

Figure B.2: Pre-trends between treated and non-treated product categories



Notes: Figure compares the pre-trends in each of three outcome variables between treated and non-treated product categories. Each plot reports the means and the confidence intervals of residuals from a regression in which we regress each of the outcome variable on product category fixed effects.

Figure B.3: Placebo test



Notes: Figures report the parameter estimate of β in Equation 4 under various scenarios of alternative adoption timing. The estimate is expected to decrease in the number of weeks pushed sooner.