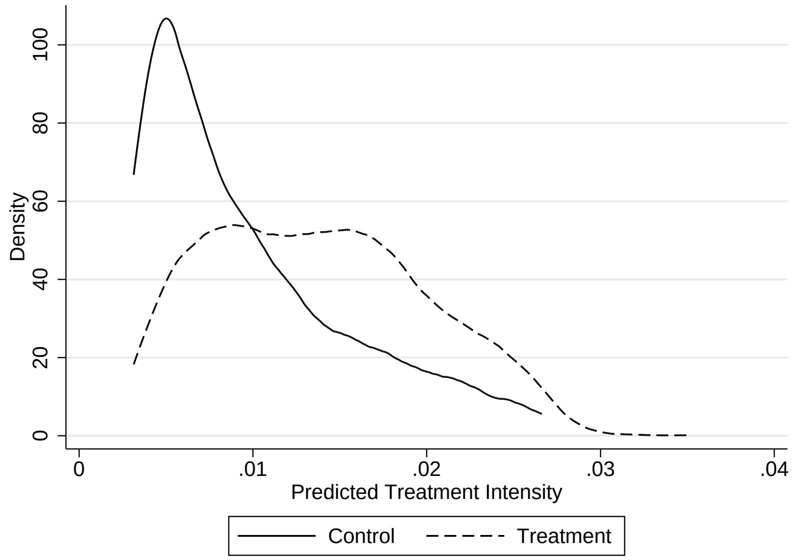
Online Appendix

Smaller Slices of a Growing Pie: The Effects of Entry in Platform Markets

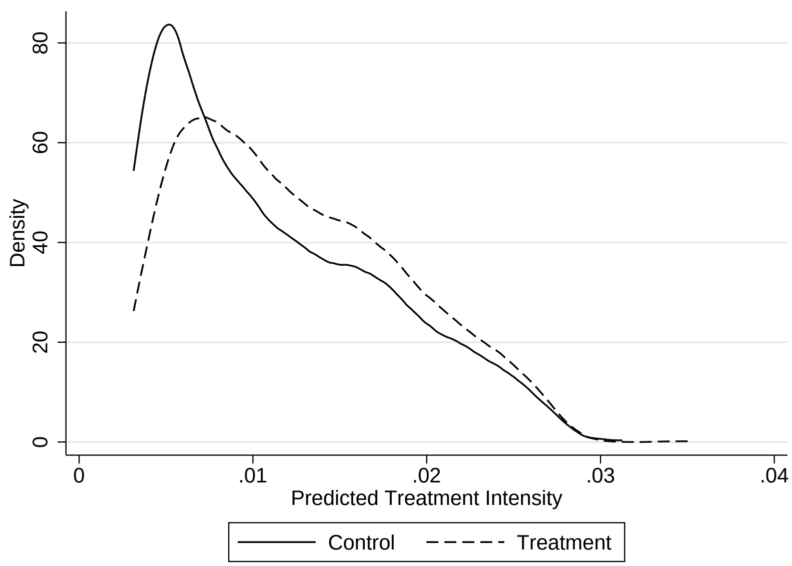
Oren Reshef

A Online Appendix Figures and Tables

Figure A1: Predicted Treatment Intensity by Treatment Assignment



(a) Original Distributions



(b) Distribution Using Within Bin Assignment

Note: The figure presents the distribution of propensity scores by treatment intensity. Propensity scores are estimated on the continuous change in share of restaurants on YTP. Treatment is an indicator for above median treatment intensity. Panel A presents the distributions of propensity scores in the original data by treatment assignment. Panel B presents the distributions of propensity scores when treatment status is assigned by propensity score bins.

Figure A2: Search Behavior

Note: This figure presents event-time estimates of the effect of treatment on various measures of search intensity on the platform. The dependent variables are the inverse hyperbolic sine transformation and should be interpreted as percent changes on a scale from 0 to 1. The unit of observation is city-week. The treatment indicator compares cities that experienced almost no change in the percentage of businesses available on the platform to cities that experienced meaningful changes. Vertical bars represent 95% confidence intervals, where standard errors are clustered at the city level.

Table A1: Market-Level Analysis

	(1)	(2)	(3)
	Panel A: Weekly Unique Users		
Treat*Post	0.364***	0.482***	10.271***
	(0.015)	(0.020)	(0.634)
	Panel B: Weekly Orders		
Treat*Post	0.367***	0.486***	10.343***
	(0.015)	(0.020)	(0.639)
	Panel C: Weekly Revenue		
Treat*Post	0.587***	0.762***	17.226***
	(0.032)	(0.042)	(1.202)
Observations	327993	226157	327993
# of Clusters	3964	2781	3964
Treatment Def.	Median	25<>75	Change

Note: This table reports regression coefficients from nine separate regressions, three per panel. The unit of observation is city-week, including both incumbent and newly added businesses. The dependent variables are the inverse hyperbolic sine transformation of the outcomes indicated in sub-headings and should be interpreted as percent changes on a scale from 0 to 1. Regressions include city and week-state fixed effects. Standard errors are in parentheses and are clustered at the city level.
* significant at 10%; ** significant at 5%; *** significant at 1%

Table A2: Placebo Trends

	(1)	(2)	(3)
	Panel A: Weekly Number of Orders		
Treat*Post	-0.008	-0.003	-0.003
	(0.007)	(0.009)	(0.013)
Treat*Post*Low		-0.010	0.007
		(0.012)	(0.015)
$\beta 1 + \beta 2$		-0.013	0.004
Pvalue		0.179	0.686
	Panel B: Weekly Revenue		
Treat*Post	-0.020	-0.017	-0.022
	(0.018)	(0.023)	(0.033)
Treat*Post*Low		-0.010	0.047
		(0.031)	(0.040)
Observations	1477208	1477208	740706
# of Clusters	2729	2729	2625
$\beta 1 + \beta 2$		-0.027	0.025
Pvalue		0.301	0.363
Treatment Def.	25<>75	25<>75	25<>75
Quality Def.		Median	25<>75

Note: This table reports regression coefficients from 9 separate regressions, 3 per panel. The unit of observation is business-week. The dependent variables are the per-business inverse hyperbolic sine transformation of weekly number of orders (Panel A) and weekly-revenue (Panel B), and should be interpreted as percent changes on a scale from 0 to 1. Post is counterfactually set to the middle of the pre-treatment period. The sum of the coefficients is presented below each panel along with the corresponding p-value. The interaction between post and quality level indicators is omitted for brevity. Regressions include business and week-state fixed effects. Standard errors are in parentheses and are clustered at the city level.
* significant at 10%; ** significant at 5%; *** significant at 1%

Table A3: Placebo Outcomes

	(1)	(2)	(3)		
	Panel A: Weekly Number of New Business				
Treat*Post	0.003	0.001	0.077		
	(0.003)	(0.004)	(0.084)		
Observations	321980	221935	321980		
# of Clusters	3788	2611	3788		
	Panel B: Weekly Number of Review				
Treat*Post	-0.001	-0.001	0.025	0.000	-0.000
	(0.001)	(0.001)	(0.034)	(0.001)	(0.002)
Treat*Post*Low				-0.002	-0.000
				(0.003)	(0.003)
Observations	31833645	17587953	31833645	17587953	10751909
# of Clusters	3964	2781	3964	2781	2047
$\beta 1 + \beta 2$				-0.002	-0.001
Pvalue				0.365	0.689
	Panel C: Average Weekly Rating				
Treat*Post	-0.000	-0.002	0.021	-0.001	-0.000
	(0.001)	(0.001)	(0.028)	(0.002)	(0.002)
Treat*Post*Low				-0.003	-0.004
				(0.002)	(0.003)
Observations	10236710	5685666	10236710	5685666	3444919
# of Clusters	3964	2781	3964	2781	2047
$\beta 1 + \beta 2$				-0.003	-0.005
Pvalue				0.049	0.009
Treatment Def.	Median	25<>75	Change	25<>75	25<>75
Quality Def.				Median	25<>75

Note: This table reports regression coefficients from 13 separate regressions, 3 in Panel A and 5 in Panels B and C. The unit of observation is business-city in Panel A, and business-week in Panels B and C. The sample includes only non-YTP affiliated businesses and users. The dependent variables are inverse hyperbolic sine transformations and should be interpreted as percent changes on a scale from 0 to 1. Outcomes are indicated in the sub-headers and described further in the text. The sum of the coefficients is presented below each panel along with the corresponding p-value. The interaction between post and quality level indicators is omitted for brevity. Regressions include business and week-state fixed effects. Standard errors are in parentheses and are clustered at the city level.
* significant at 10%; ** significant at 5%; *** significant at 1%

Table A4: Sensitivity to Geographical Market Definition

	By 5-Digits Zip Code				By County			
	(1) Orders	(2) Orders	(3) Revenue	(4) Revenue	(5) Orders	(6) Orders	(7) Revenue	(8) Revenue
Treat*Post	0.049***	0.066***	0.136***	0.180***	0.048*	0.049**	0.094*	0.087
	(0.008)	(0.011)	(0.019)	(0.028)	(0.019)	(0.018)	(0.045)	(0.047)
Treat*Post*Low	-0.076***	-0.108***	-0.158***	-0.234***	-0.100***	-0.111***	-0.184**	-0.202***
	(0.011)	(0.014)	(0.026)	(0.034)	(0.023)	(0.022)	(0.057)	(0.058)
Observations	2268248	1283178	2268248	1283178	1613929	1273658	1613929	1273658
# of Clusters	6666	4197	6666	4197	1260	921	1260	921
$\beta 1 + \beta 2$	-0.027	-0.042	-0.022	-0.054	-0.052	-0.062	-0.090	-0.114
Pvalue	0.000	0.000	0.232	0.024	0.000	0.000	0.015	0.004
Treatment Def.	Median	25<>75	Median	25<>75	Median	25<>75	Median	25<>75
Quality Def.	25<>75	25<>75	25<>75	25<>75	25<>75	25<>75	25<>75	25<>75

Note: This table reports regression coefficients from 8 separate regressions. The unit of observation is business-week. Geographic market definitions are indicated in sub-headings and described further in the text. The dependent variables are the per-business inverse hyperbolic sine transformations of weekly number of orders and weekly-revenue, and should be interpreted as percent changes on a scale from 0 to 1. The sum of the coefficients is presented below each panel along with the corresponding p-value. The interaction between post and quality level indicators is omitted for brevity. Regressions include business and week-state fixed effects. Standard errors are in parentheses and are clustered at the city level.
* significant at 10%; ** significant at 5%; *** significant at 1%

Table A5: Sensitivity to Outliers

	Dropping Top & Bottom 5% of Cities				Randomization Inference			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Orders	Orders	Revenue	Revenue	Orders	Orders	Revenue	Revenue
Treat*Post	0.032**	0.033*	0.082**	0.090*	0.050***	0.065***	0.119***	0.158***
	(0.011)	(0.016)	(0.028)	(0.040)	[0.000]	[0.000]	[0.000]	[0.000]
Treat*Post*Low	-0.058***	-0.060**	-0.131***	-0.134*	-0.098***	-0.119***	-0.201***	-0.250***
	(0.015)	(0.021)	(0.039)	(0.056)	[0.000]	[0.000]	[0.000]	[0.000]
Observations	832690	423002	832690	423002	2173124	1321540	2173124	1321540
# of Clusters	2483	1422	2483	1422	3862	2714	3862	2714
$\beta 1 + \beta 2$	-0.026	-0.027	-0.049	-0.044				
Pvalue	0.016	0.082	0.089	0.302				
Treatment Def.	Median	25<>75	Median	25<>75	Median	25<>75	Median	25<>75
Quality Def.	25<>75	25<>75	25<>75	25<>75	25<>75	25<>75	25<>75	25<>75

Note: This table reports regression coefficients from 8 separate regressions. The unit of observation is business-week. Columns (1)-(4) exclude outliers cities. The dependent variables are the per-business inverse hyperbolic sine transformations of weekly number of orders and weekly-revenue, and should be interpreted as percent changes on a scale from 0 to 1. The sum of the coefficients is presented below each panel along with the corresponding p-value. The interaction between post and quality level indicators is omitted for brevity. Regressions include business and week-state fixed effects. In Columns (1)-(4), standard errors are in parentheses and are clustered at the city level. In Columns (4)-(8), randomization inference p-values based on 2000 draws are reported in square brackets.
* significant at 10%; ** significant at 5%; *** significant at 1%

Table A6: Sensitivity to Initial Differences

	Propensity Score Weighting				Blocked Treatment Assignment			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Orders	Orders	Revenue	Revenue	Orders	Orders	Revenue	Revenue
Treat*Post	0.044**	0.066***	0.105**	0.152***	0.034***	0.066***	0.092***	0.173***
	(0.014)	(0.016)	(0.034)	(0.039)	(0.009)	(0.014)	(0.021)	(0.032)
Treat*Post*Low	-0.067***	-0.089***	-0.154***	-0.186***	-0.054***	-0.103***	-0.135***	-0.253***
	(0.018)	(0.020)	(0.046)	(0.052)	(0.014)	(0.021)	(0.030)	(0.044)
Observations	814388	501376	814388	501376	2173127	1016860	2173127	1016860
# of Clusters	3412	2467	3412	2467	3863	2770	3863	2770
$\beta 1 + \beta 2$	-0.023	-0.023	-0.049	-0.035	-0.020	-0.037	-0.043	-0.079
Pvalue	0.045	0.070	0.111	0.333	0.026	0.003	0.037	0.005
Treatment Def.	Median	25<>75	Median	25<>75	Median	25<>75	Median	25<>75
Quality Def.	25<>75	25<>75	25<>75	25<>75	25<>75	25<>75	25<>75	25<>75

Note: This table reports regression coefficients from 8 separate regressions. The unit of observation is business-week. In Columns (1)-(4), observations are weighted by the inverse probability score, and in Columns (5)-(8), treatment status is assigned by propensity score bins. See text for additional details. The dependent variables are the per-business inverse hyperbolic sine transformations of weekly number of orders and weekly-revenue, and should be interpreted as percent changes on a scale from 0 to 1. The sum of the coefficients is presented below each panel along with the corresponding p-value. The interaction between post and quality level indicators is omitted for brevity. Regressions include business and week-state fixed effects. Standard errors are in parentheses and are clustered at the city level.
* significant at 10%; ** significant at 5%; *** significant at 1%

Table A7: Sensitivity to City-Level Time Trends

	(1) Orders	(2) Orders	(3) Revenue	(4) Revenue
Treat*Post	0.060***	0.067***	0.139***	0.165***
	(0.010)	(0.012)	(0.025)	(0.033)
Treat*Post*Low	-0.102***	-0.126***	-0.217***	-0.274***
	(0.014)	(0.016)	(0.035)	(0.043)
Observations	2173244	1321619	2173244	1321619
# of Clusters	3875	2725	3875	2725
$\beta 1 + \beta 2$	-0.042	-0.059	-0.078	-0.109
Pvalue	0.000	0.000	0.001	0.000
Treatment Def.	Median	25<>75	Median	25<>75
Quality Def.	25<>75	25<>75	25<>75	25<>75

Note: This table reports regression coefficients from 4 separate regressions. The unit of observation is business-week. The dependent variables are the per-business inverse hyperbolic sine transformation of weekly number of orders and weekly-revenue, and should be interpreted as percent changes on a scale from 0 to 1. The sum of the coefficients is presented below each panel along with the corresponding p-value. The interaction between post and quality level indicators is omitted for brevity. Regressions include business and week-state fixed effects. In addition, the specification allows for city-level time trends in establishment outcome. Standard errors are in parentheses and are clustered at the city level. * significant at 10%; ** significant at 5%; *** significant at 1%

Table A8: Effect of Entry on Incumbents' Subsequent Ratings - Robustness Checks

	Placebo Test				Type of Raters			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Treat*Post	0.003	-0.001	0.008	-0.007	0.001	0.001	0.003	0.002
	(0.008)	(0.010)	(0.009)	(0.015)	(0.003)	(0.003)	(0.003)	(0.004)
Treat*Post*Low			-0.059**	-0.027			-0.008	-0.004
			(0.026)	(0.020)			(0.008)	(0.010)
Observations	834964	488933	393194	233178	1464204	859456	695502	413378
# of Clusters	3356	2243	3110	2095	3827	2655	3680	2569
$\beta 1 + \beta 2$			-0.051	-0.034			-0.005	-0.002
Pvalue			0.050	0.109			0.539	0.810
Treatment Def.	Median	25<>75	Median	25<>75	Median	25<>75	Median	25<>75
Quality Def.			25<>75	25<>75			25<>75	25<>75

Note: This table reports regression coefficients from 8 separate regressions. In Columns (1)-(4), the unit of observation is business-week; in Columns (5)-(8) the unit of observation is user who rated a YTP restaurant. The dependent variables are the inverse hyperbolic sine transformations of the average rating received in a given week (Columns (1)-(4)) and the average rating given by user (Column (5)-(8)), and should be interpreted as percent changes on a scale from 0 to 1. The sum of the coefficients is presented below each panel along with the corresponding p-value. The interaction between post and quality level indicators is omitted for brevity. Regressions include business and week-state fixed effects. Standard errors are in parentheses and are clustered at the city level.
* significant at 10%; ** significant at 5%; *** significant at 1%

Table A9: Heterogeneity by Differentiation Between Incumbents and Entrants

	Num. of Orders (Prc.)		Revenue (Prc.)	
	(1)	(2)	(3)	(4)
Share of Type*Post	0.010*	0.029**	0.002	0.064*
	(0.006)	(0.015)	(0.014)	(0.037)
Observations	3223311	3223311	3223311	3223311
# of Clusters	1917	1917	1917	1917
Similarity Definition	Positive Change	Continuous	Positive Change	Continuous

Note: This table reports regression coefficients from four separate regressions. The unit of observation is business-week. The sample includes only cities that received above-median changes in the share of businesses on YTP. The dependent variables are the inverse hyperbolic sine transformations of weekly number of orders and weekly revenue, and should be interpreted as percent changes on a scale from 0 to 1. Coefficients represent the interaction between the measure of similarity and a dummy for post implementation and treatment status. In Columns (1) and (2), the measure is an indicator for whether any business in the same category were added, whereas in Columns (3) and (4), the measure is the share of businesses of the same food category as the incumbent out of the total number of added business. Regressions include business and week-state fixed effects. Standard errors are in parentheses and are clustered at the city level.
* significant at 10%; ** significant at 5%; *** significant at 1%

B Online Data Appendix

Sample selection and main results While Yelp keeps data on when restaurants join and exit YTP, in cases where transactions were made prior to a business 'entering' the platform or after the business 'exited' the platform, I always code entry as the earliest date of the two and exit as the later date. For this reason, I leave a margin of 8 weeks at the beginning and end of my sample to separate between businesses with zero sales from businesses that have left the platform. Sales and revenue are coded as missing for the week before entry or after exit.[1] The final data used in the analysis consist of 88 weeks from March 2017 to December 2018. I limit the analysis to cities in which there are ten or more businesses on the standard Yelp platform, since in very small places treatment intensities are extremely large mechanically. I excluded businesses that are marked by Yelp as bogus, spammy, or that are removed from users' search results. 310 businesses in 291 cities are dropped from the analysis, which amounts to less than 30,000 observations in my data. The final sample consists of 3,956 cities. For the main part for the analysis, I use only the incumbent businesses, which joined YTP prior to the partnership with Grubhub; there are a total of 56,493 incumbent businesses and over 4 million business-week observations.

[1] The main results are robust to two alternative methods to address attrition: First, excluding any establishment that ultimately leaves the platform from the analysis all together, and second, coding leaving businesses' weekly number of orders and weekly revenue as zero after existing the platform.

The Yelp system does not store historic businesses star-ratings. I calculate businesses rating on the eve of integration excluding reviews that are marked by Yelp as untrustworthy. Restaurant categories are based on Yelp's classification. Generally, Yelp collects little demographic information on its users. Users are encouraged to enter their gender and date of birth, but I found these fields to be mostly missing in my sample. Thus, I do not use individual-level characteristics in my analysis. For each user, however, I do observe the full history of transactions on YTP, which is used to differentiate between new and repeating consumers.

Search Orders data and search data are handled by different parts of the organization and, more importantly, are stored in different data clusters. Consequently, joining the two datasets is not a trivial task. To identify the search sessions which lead to an order, I develop an algorithm that matches each order with the most recent search session conducted by the user prior to finalizing the order. The algorithm has several issues. First, it will not be able to match an order to a search session if the user was not signed in during the search process. Second, when a user performs multiple searches on the same day, the algorithm only picks up the last session. This might be an issue if users use multiple search sessions to choose a restaurant from which to order. Though these issues create additional noise and reduce statistical power, they are unlikely to bias the results in any particular direction.

Placebo tests and alternative explanations To test the validity of the research design, I consider three outcomes: (1) the weekly flow of new restaurants onto Yelp; (2)the weekly flow of new ratings per business; and (3) the average rating given. I include only businesses that are classified as either 'food' or 'restaurants' and exclude businesses that are marked as bogus or spammy. Importantly, businesses participating in YTP are also excluded from the analysis. I also exclude reviews that are marked as untrustworthy, were removed by Yelp, or that are given by consumers that use YTP.

To test alternative explanations for the increase in ratings, I test whether selection of more lenient reviewers is driving the results. To test for reviewer leniency, I construct the leave-out-average of all reviews given by a user.

Search result rankings are constructed using the average weekly rank across all search results in which the business appears. Businesses that are marked as bogus, spammy, or that are removed from users' search are excluded.

C Identification and Robustness Checks Online Appendix

C.1 Validity of the Research Design

This section presents additional tests to support the parallel trend assumption, which is the key identifying assumption of the difference-in-differences research design.

Pre-trends The first suggestive evidence of parallel trends absent of treatment is to examine whether the main outcomes trend similarly in the prior to the Grubhub partnership. Graphic evidence is presented in Figures 3 and 4. Table A2 presents a more formal placebo test in which I counterfactually set the integration date to the middle of the pre-treatment period. I do not find any significant effects of the placebo on the average effect (Column

(1)) or when examining the effect on high- or low-quality firms separately (Columns (2) and (3)). This results suggests that the main results are not driven by initial differences in trends between treated and control cities.

Placebo on non-YTP outcomes A second potential concern is that the break in trends is driven by other unobserved changes at the city level that are unrelated to the Grubhub partnership. If that is the case, then we can expect to find significant differences in other city-level outcomes, not directly related to food ordering. I conduct several placebo tests to examine whether the partnership is correlated with outcomes of *non-YTP* businesses, such as the number of new restaurants on Yelp, restaurants average weekly ratings, and the number of new weekly reviews per business. Table A3 presents the results. I do not find any significant changes in non-YTP outcomes. This null results suggest that the main findings are not driven by unobserved changes in the city, the restaurant industry, or Yelp usage.[2]

C.2 Robustness of the Main Results

Market definition-geographic area Table A4 presents the estimation results using alternative geographical definitions for the relevant market. Note that the number of observations decreases since not all observations include zip-code and county data. Columns (1)-(4) and (5)-(8) present the results when using the 5-digit code and county as the relevant markets, respectively. The first two columns in each group show the effect on number of orders and the last two columns in each group show the effect on weekly revenue. Qualitatively, the results are similar to the main estimation results: entry leads to more sales and higher revenue for high-quality restaurants, and vice versa for low-quality businesses. The point-estimates of the effects vary across specifications and market definition. Nevertheless, estimates are centered around the main results, do not change signs, and are generally statistically significant. For instance, Columns (2) and (6) estimate the effect of entry on weekly orders using the sharp definitions for treatment and rating. They find a treatment effect of 4.9% to 6.6% for high-types and -6.2% to -4.2% for low-types. In comparison, using the city as the relevant market, I find an effect of 6.5% for high-types and -5.4% for low-types (Column (2) in Table 3).

Initial differences between treatment and control As mentioned in the main text, though there are similar trends in treated and control markets prior to integration, there are substantial differences between markets. Specifically, larger cities are, on average, more likely to be treated.

To test whether the main results are driven by a few large cities, I perform two separate robustness checks. The results are presented in Appendix Table A5. First, I exclude the cities that are in the top and bottom 5% in terms of the number of businesses on YTP prior to integration. This exercise turns out to be quite restrictive: To begin with, many small towns had only a handful of businesses prior to integration, so there is substantial mass at 5%. Additionally, the largest cities, with the most incumbent businesses, naturally contribute the most observations to the analysis. Thus, excluding outliers reduces both the

[2] In contrast, in section 5.3 I find that treatment does effect the average and high-quality weekly ratings of YTP restaurants in treated cities compared to untreated cities.

number of clusters and the number of observations substantially. As Columns (1)-(4) show, the qualitative results are similar to the main specification and are statistically significant. The estimated size of the effects diminishes in comparison to the main analysis, but remains with the 95% confidence interval.

Second, I estimate p-values using randomization inference instead of a traditional sampling-based approach. Randomization inference performs better in settings with a concentration of leverage, the degree to which individual observations of the right-hand side variables take extreme values and are influential (Young, 2016). Due to the large number of observations and the time it takes to run these specifications I perform only 1000 iterations for each specification. All of the p-values on the coefficients of interest are zero, i.e., the estimated effects were larger than all of the 1000 randomized treatment effects. Taking these results together, I conclude that it is unlikely that the estimates are driven solely by outliers.

In addition, in Appendix Table A6, I attempt to correct for differences across the entire distribution: firstly, I use inverse propensity score weighting to correct for the bias (Hirano et al., 2003). I estimate the propensity score using a third-order polynomial logit model with the total number of businesses, the total number of businesses on YTP pre-integration, and the share of businesses on YTP pre-integration as predictors. Though there is generally common support on the full interval, I trim propensity above and below the 90th and 10th percentiles to correct for differences in mass. The results are presented in Columns (1)-(4). Though the point estimates are slightly lower, the estimated effect of market expansion on high-quality firms remains positive and statistically significant. The differences between the effects of high- compared to low-quality firms are both economically and statistically significant. The total effects on low-quality firms are both smaller in magnitude and noisier than in the main specification.

Secondly, since treatment and control look very different on observables, I conduct an additional analysis, which takes advantage of the fact that treatment intensity is continuous. First, I run a linear probability model of treatment intensity (change in share of businesses on YTP) on a third-order polynomial with the total number of businesses, total number of businesses on YTP pre-integration, and the share of businesses on YTP pre-integration as predictors. Figure A1a presents the distribution of *predicted* treatment intensity by treatment status, where treatment is defined by the median (actual) treatment intensity. City characteristics clearly predict treatment intensity, though there is substantial overlap between the two groups. Secondly, I divide markets into 20 bins based on predicted treatment intensity, with an equal number of markets within each bin. I then assign markets into treatment and control based on their relative treatment intensity in their respective bin. Intuitively, bins with higher predicted intensity will tend to have higher thresholds to be included as treatment. Accordingly, some markets with very high predicted intensity might be coded as control even though their true intensity is relatively high, and vice versa for low intensity markets. Figure A1b presents the distribution of predicted treatment intensity by treatment status, where treatment is defined by the median *within bin* treatment intensity. It is clear the the distribution of treated and control markets is much more similar than in the original sample.

The estimation results, using the new treatment definition, are presented in Appendix Table A6 Columns (5)-(8). The point estimates on all coefficients are similar in magnitude to the main specification and are statistically significant. Taken together, these results suggest that initial differences in market characteristics do not drive the main effects.

C.3 Robustness to Heterogeneous Treatment Effects

Recent econometric literature suggests that the two-way fixed effects (TWFE) my be biased when there are heterogeneous treatment effects across cohorts or time. The main intuition is that estimated treatment effect may be "contaminated" by treatments of other groups or at other times. Formally, this implied that the average treatment effect (ATE) is not a convex combination of the estimated average treatment effects on treated (ATT) for each group-time combination, i.e., while the ATE is a weighted average of all group-period ATTs, some weights might be negative (De Chaisemartin and d'Haultfoeuille, 2020, Callaway and Sant'Anna, 2021, Goodman-Bacon, 2021, Sun and Abraham, 2021).

The vast majority of work focuses on TWFE with staggered adoption—when units are exposed to treatments at different times (Athey and Imbens, 2022, Borusyak et al., 2021, Callaway and Sant'Anna, 2021, Goodman-Bacon, 2021, Sun and Abraham, 2021). In these settings, using previously treated units as control for newly treated units may lead to biased estimates. Many of these papers also mention in passing the robustness of the traditional two-by-two DiD estimator, in which all units receive the (binary) treatment simultaneously, to heterogeneous treatment effects. This intuitive argument is formalized in De Chaisemartin and D'Haultfoeuille (2022) which find that the ATE is not biased when (1) treatment status changes at most once, (2) treatment is binary, and (3) the is no variation in treatment timing. Importantly, all three conditions hold in the main specification(s) studied in this paper: the shock of the YTP-Grubhub partnership occurs only once and at given point in time and affects all markets simultaneously.[3] Thus, the research design even if treatment effects are heterogeneous, the results will remain unbiased.

In addition, to test the need for alternative estimators formally, I conduct the test suggested in De Chaisemartin and d'Haultfoeuille (2020), De Chaisemartin and D'Haultfoeuille (2022). Since the main issue with heterogeneous treatment effects is that some of the weights assigned to the ATTs are negative, this test estimates the prevalence of negative weights in computing the ATE. Because the main specification estimates the impact on high- and low-rated businesses separately, I conduct the test twice for high-rated and low-rated businesses separately (sharp definition). In both cases, I find that the sum of negative weights comprises a tiny percentage of the positive weights. For example, for high-rated businesses the sum of positive weights is 1.0002351 while the sum of negative weights is -0.0002351, implying a ratio of over 4,000 (and similarly over 1,500 for low-rated businesses) in favor of the positive weights.

The test offers another measure to test the likelihood that treatment effect heterogeneity is driving the results, by computing the standard deviations in treatment effects required to produce the estimated coefficient if the true average of ATTs is equal to zero, or is the opposite sign. The authors also offer a threshold to gain a sense of whether the standard deviation is small or large, by comparing testing whether the estimated coefficient is larger

[3] While the treatment can be thought of as continuous, in practice, all main specifications use a binary definition of treatment assignment.

than sqrt(3)x or 2sqrt(3)x, where x is the derived measure. For high-rated businesses these numbers are 0.57 and 0.36, over 13 and 8 times larger than the estimated coefficient (and similarly over 7 and 6 times for low-rated businesses). Naturally, the probability of have ATTs of opposite signs are even smaller. Thus, I conclude that the main results are robust to heterogeneous treatment effects across groups or times.

Finally, in an unshown analysis, I follow Wooldridge (2021), which argues that one can explicitly model heterogeneity in treamtent effects and the estimate remain unbiased. I allow treatment effects to change over time by dividing the post period into three 11-weeks bins. Since all units are treated at the same time, I group units by geography rather than treatment cohorts, and allow the effect to vary by eastern, western, midwestern, and southern states. The ATE of all coefficients and their interactions is practically identical to the main specification.

C.4 Investment in Ratings - Alternative Explanations

In section 5.3, I find that entry increased subsequent investments in ratings. This section addresses the alternative explanations to explain this result.

Rating inflation To address the concern that results are driven by rating inflation (Filippas et al., 2022), I first test a placebo specification in which integration is counterfactually coded at the middle of the pre-treatment period. The results are presented in Columns (1)-(4) of Table A8. Columns (1) and (2) find null average effects which hare are both statistically and economically insignificant. Columns (3) and (4) decompose the average effects by quality type; there are null effects for high-type firms, and marginally significant *negative* effects on low-type. Taken together these results suggest that differential trends in rating inflation are not driving the increases in ratings following integration. In addition, this concern is somewhat mitigated by the test conducted to support the identifying assumption; I test for changes in ratings trends for *non-YTP* businesses, and find null (and slightly negative) effects of entry into YTP on subsequent ratings of non-YTP businesses (Table A3).

User selection The second concern is that selection into specific services is correlated with rating behavior (Fradkin et al., 2018). For instance, if users who use delivery services also tend to rate more leniently, then increases in online ordering might mechanically drive up the ratings of restaurants. To alleviate this concern, I test for differential changes in raters' leniency. The results are presented in Panel B of Table A8. For each review rating, I calculate the average rating given by the user through their activity on Yelp. I do not find any evidence that raters' leniency changes following integration: the average effects and effects by quality-type on leniency are statistically and economically insignificant under all specifications.

References

Athey, Susan and Guido W. Imbens, "Design-based analysis in Difference-In-Differences settings with staggered adoption," *Journal of Econometrics*, 2022, *226* (1), 62–79.

Borusyak, Kirill, Xavier Jaravel, and Jann Spiess, "Revisiting event study designs: Robust and efficient estimation," *Working paper*, 2021.

Callaway, Brantly and Pedro HC Sant'Anna, "Difference-in-differences with multiple time periods," *Journal of Econometrics*, 2021, *225* (2), 200–230.

Chaisemartin, Clément De and Xavier d'Haultfoeuille, "Two-way fixed effects estimators with heterogeneous treatment effects," *American Economic Review*, 2020, *110* (9), 2964–96.

_ **and Xavier D'Haultfoeuille**, "Two-way fixed effects and differences-in-differences with heterogeneous treatment effects: A survey," Technical Report 2022.

Filippas, Apostolos, John Horton, and Joseph Golden, "Reputation inflation," *Marketing Science*, 2022, *forthcoming.*

Fradkin, Andrey, Elena Grewal, and David Holtz, "The determinants of online review informativeness: Evidence from field experiments on Airbnb," 2018.

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

Hirano, Keisuke, Guido W Imbens, and Geert Ridder, "Efficient estimation of average treatment effects using the estimated propensity score," *Econometrica*, 2003, *71* (4), 1161–1189.

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

Wooldridge, Jeff, "Two-way fixed effects, the two-way mundlak regression, and difference-in-differences estimators," *Working paper*, 2021.

Young, Alwyn, "Channeling fisher: Randomization tests and the statistical insignificance of seemingly significant experimental results," *London School of Economics, Working Paper*, 2016.