INSTITUTE OF RETAIL ECONOMICS

WHY DO FIRMS COMPETE ON PRICE COMPARISON WEBSITES? THE IMPACT ON PRODUCTIVITY, PROFITS, AND WAGES

CHARLIE LINDGREN, YUJIAO LI &
NIKLAS RUDHOLM

Why do firms compete on price comparison websites? The impact on productivity, profits, and wages

Charlie Lindgren, Yujiao Li and Niklas Rudholm

This paper investigates how firm entry into a price comparison website marketplace affects firm productivity, profits, and wages. We want to answer the key research question: Why do firms compete on price comparison websites? A substantial literature indicates that competition in such marketplaces is fierce, leading to lower prices for products sold. We suggest that participation in these marketplaces also leads to increased productivity, i.e. output increases when holding constant the level of inputs used. This leads to increased profits, motivating firms to enter price comparison websites despite fierce competition. Our results indicate that for the full sample of firms, PriceSpy participation increases output by almost 12% when holding the level of inputs constant. Also, investigation of who gains from the increased productivity shows that, for entering firms, operating profits increase by 9% and gross wages by 14% when studying the full sample of firms. That labor gains more from PriceSpy participation is even clearer when studying the impact on wholesale and retail firms separately. For those firms, wages increased by 16–17% after entry, while no statistically significant impact was found regarding operating profits.

Keywords: Online retailing; e-commerce; price comparison websites; productivity, value added.

JEL codes: D22, D24, D33, L81.

1. Introduction

> The explosive growth of the Internet promises a new age of perfectly competitive markets. With perfect information about prices and products at their fingertips, consumers can quickly and easily find the best deals. In this brave new world, retailers' profit margins will be competed away, as they are all forced to price at cost.
>
> *The Economist*, November 20, 1990

At the beginning of the Internet era, the introduction of online retailing was expected to create almost perfectly competitive markets, with no excess profits[1] for retailers competing in online marketplaces. While these predictions have not been realized, there is a literature indicating that online competition in general (Brynjolfsson and Smith, 2000; Clay et al., 2001), and competition on price comparison websites in particular (Brown and Goolsbee, 2002; Haynes and Thompson, 2008; Tang et al., 2010; Lindgren et al., 2020), indeed lowers prices.

Despite reports of increased competition and lower prices, firms have chosen to enter price comparison website marketplaces at an increasing pace. The increase in the use of the price comparison website PriceSpy in Sweden from 2013 to 2016 is remarkable. Lindgren et al. (2020) reported detailed statistics on this development for an example product category, i.e. games for the PlayStation 4 console. The data show that in 2013 there were about 20 retailers marketing some 20 games on the PriceSpy

[1] We use the term "excess profits" to represent all economic profits, i.e. all profits above a normal return on investment given in a competitive market. This separates the concept of economic profits from the operating profits found in annual reports and studied in the empirical part of the paper.

website, while by 2016 this had increased to almost 60 retailers marketing approximately 600 products.

Why do firms choose to compete in a marketplace with fierce competition that reduces prices? The purpose of this paper is to investigate how entry into the PriceSpy marketplace affects productivity, operating profits, and gross wages to answer the main research question: Why do firms compete on price comparison websites?

In this paper, we suggest that the willingness to compete on price comparison websites is due to the influence entry has on the supply function of the firms. Entry into price comparison website marketplaces creates a shift in the firm's supply function, leading to a reduced unit cost for the products sold. This in turn leads to lower prices, larger quantities sold, and increased excess profits, even when firms hold the levels of inputs, labor and capital, constant. The increased productivity creates an increase in excess profits to be shared between shareholders and labor depending on their respective bargaining power, and this is what motivates firms to enter the price comparison website marketplace, despite the fierce competition.

Empirically investigating the impact of PriceSpy market participation on productivity, profits, and wages is not easy. Firms are free to select whether and when to enter or exit the PriceSpy marketplace, creating the problem of self-selection into treatment. In this paper, we use a two-step procedure to address this problem to the extent possible. In the first step, we control for differences in observables between entering firms and potential control-group firms, with a special focus on output development in the pre-entry period. This procedure reduces heterogeneity in pre-entry output between the two groups and makes the pre-entry trends in our main outcome variable output parallel for the entering and selected control-group firms. Then, in a second step, we use a within-firm difference-in-difference translog production function estimator to investigate how entry into the PriceSpy marketplace affects output while holding inputs constant.

Our results indicate that firms entering the PriceSpy marketplace from 2005 to 2015 experienced an increase in output, while holding inputs constant, of 11.63%. For retail firms the increase was 17.35% and for wholesale firms it was 12.75%, indicating that non-retail or wholesale firms entering the PriceSpy website did not gain as much as did retail and wholesale firms. The results for firms from industries other than retail or wholesale indicate that output increased by an average of 6.18% when entering. The group of other firms is very heterogeneous, however, including firms from all types of industries, making it difficult to say precisely why this is the case. One possible explanation is that the retail and wholesale firms that entered had more experience in online retailing in general, and thus a better understanding of how to use the PriceSpy market to increase sales.

Turning to the results regarding who gains more from PriceSpy participation, capital or labor, the results indicate that gross wages increase by 12.75–17.35% when firms enter PriceSpy, depending on the industry, while operating profits increase by 9.42% when analyzing the full sample of firms. Since both operating profits and gross wages increase when firms enter the PriceSpy marketplace, this indicates that there is an increase in excess profits due to entry. However, for the retail and wholesale firms in our sample, we did not find any statistically significant impact of PriceSpy market participation on operating profits, all of the increase being from firms in industries other than retail or wholesale, for which we found an increase in operating profits of 13.88%. This suggests that most of the gains from PriceSpy entry go to labor in the retail and wholesale industries, while the gains are shared more equally between capital and labor when firms from other industries enter.

The rest of the paper is organized as follows. In Section 2, we discuss the theoretical background to the research questions studied here. Section 3 presents the empirical analysis, beginning in Section 3.1 with control group selection and a description of the estimation methods. Section 3.2 presents the data collection and preparation methods, regarding both the PriceSpy entry dates and the annual report data, together with some descriptive statistics. Then, in Section 3.3, we present the results of the empirical analysis. Finally, Section 4 summarizes and discusses our results.

2. Theoretical background

From previous literature we know that firms entering price comparison website marketplaces lower their prices as they enter (Brown and Goolsbee, 2002; Haynes and Thompson, 2008; Tang et al., 2010; Lindgren et al., 2020). However, the fact that firms voluntarily choose to enter suggests that entry also leads to increased profits, otherwise firms would have no incentive to enter these markets. As such, our model of what happens when firms enter these marketplaces needs to incorporate both these features, a combination of lower prices and higher profits.

Assume linear demand (D) and marginal revenue (MR) curves, and that the total cost curve (TC) can be represented by the function $TC = a + bQ - cQ^2 + dQ^3$, where $b > c > d$. The average total cost can then be written $ATC = \frac{a}{Q} + b - cQ + dQ^2$, while the marginal cost is given by $MC = b - 2cQ + 3dQ^2$, and when represented in a graph, the marginal and average total cost curves have the general shape depicted in Fig. 1. For low volumes of output, the marginal cost falls to a certain minimum, after which it increases with output. Firms are assumed to compete in prices, creating a Bertrand oligopoly market with differentiated offers to consumers. Thus, the firms' marginal revenue (MR) is equated to the marginal cost (MC) to find the profit-maximizing price

(P). Even in the long-run equilibrium, this price will exceed the marginal cost due to product offers being heterogeneous and the oligopolistic nature of the market. This situation is depicted in Fig. 1a below, with excess profit for the firm shown by the marked area in the graph.

A firm entering the Swedish PriceSpy marketplace must already have its own website and a warehouse set up to convey the online sales to the carriers delivering the product to consumers, since PriceSpy does not provide such services.[2] We also assume that the firm does not alter the level of inputs, capital, and labor when entering. Since there is now access to a new and larger marketplace, firm demand is assumed to increase due to entry (D_1 in Fig. 1a shifts to D_2 in Fig. 1b). However, note that if the only impact of PriceSpy marketplace entry were to increase demand, without any effect on firm supply, this would make the firm increase its prices upon entry, contrary to previous findings. This suggests that entry also affects the marginal cost curve of the firm, creating a shift in the firm supply function (S_1 in Fig. 1a shifts to S_2 in Fig. 1b).

By comparing Fig. 1a and 1b, we see that entry into the PriceSpy marketplace will lead to a reduction in price, an increase in quantity sold, and an increase in excess profits (i.e. the marked areas given by $[P - ATC] \times Q$ in the graphs). Note also that this happens even though the use of labor and capital is assumed to remain constant in the analysis. This leads to the first research question: Does entry into the PriceSpy marketplace increase output when the level of inputs (i.e. capital and labor) are held constant? This question will be studied using a within-firm difference-in-difference translog production function approach investigating how output changes when firms enter the PriceSpy marketplace while holding the levels of capital and labor constant.

[2] Swedish firms typically use outside carriers such as PostNord, Schenker, or DHL for delivery services.

If we can show that output increases while the levels of labor and capital are held constant, this will increase excess profits for the entering firms if the output increase is enough to offset the price decrease due to entry. The increase in excess profits will then be divided between labor and capital depending on the relative bargaining power of capital owners and labor, leading to the following research questions: First, is there an increase in excess profits, measured as the sum of compensation to capital owners and labor, when firms enter the PriceSpy marketplace (as depicted in Fig. 1c)? If so, the reduction in price must be due to a shift in the supply function rather than due to increased competition facing the entering firms.[3] Second, who gains more if there is an increase in excess profits caused by PriceSpy market participation, capital owners or labor? These questions will be studied using a difference-in-difference approach investigating how compensation to capital owners (measured as operating profits) and compensation to labor (measured as gross wages) change when firms enter the PriceSpy marketplace.

[3] If entry into the PriceSpy marketplace mainly leads to increased competition, this would also cause reduced prices for the entering firms. However, the increase in competition would then decrease, rather than increase, the excess profits of the entrants.

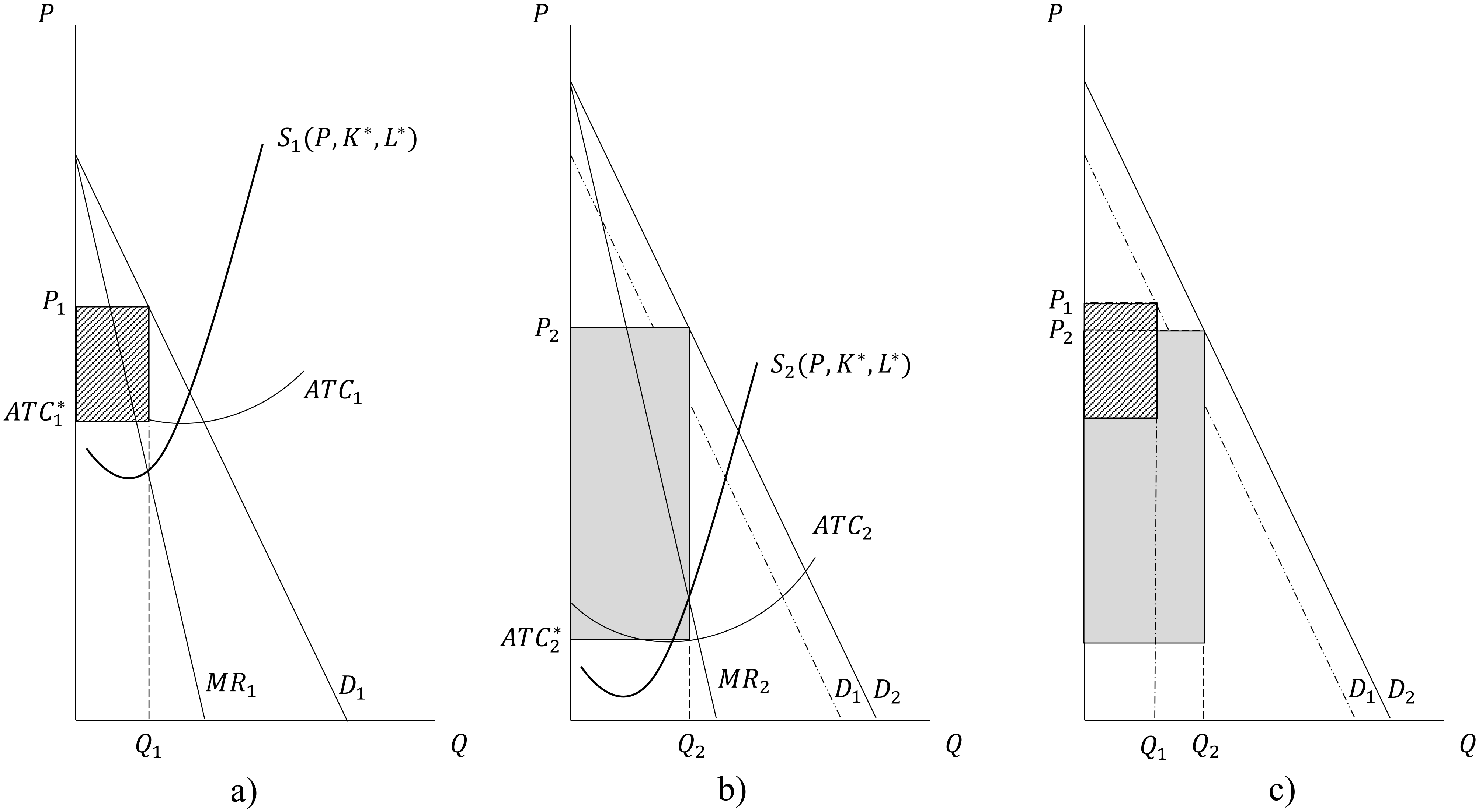

Fig. 1. Effects of PriceSpy participation on demand, marginal and average costs, price, and excess profits.

3. Empirical analysis

To accurately estimate the effects of PriceSpy participation on productivity, operating profits, and gross wages, we use a two-step procedure. First, we use coarsened exact matching (CEM) as a pre-processing method to prune observations from the data so that the remaining observations are better balanced between the treatment and control groups, with a special focus on the pre-treatment trends in the main outcome variable, i.e. consumer price index (CPI)-adjusted sales, our measure of output. Then, after verifying that the treatment and control groups have parallel trends in the outcome variable in the pre-treatment period, we use a within-firm difference-in-difference translog production function estimator to estimate the effect of PriceSpy participation on productivity, operating profits, and gross wages.

3.1 Control group selection and estimation methods

3.1.1 Finding a suitable control group using CEM

To estimate the impact of PriceSpy participation, a control group of firms similar with respect to the pre-entry characteristics of the firms entering the PriceSpy website needs to be identified. Since we use difference-in-difference analysis in the second step of the analysis, a special focus will be on investigating whether the identification assumption of parallel trends in the outcome variable in the absence of treatment is fulfilled.

A firm is considered treated after having entered the PriceSpy website, while firms that have never been on PriceSpy are defined as not treated and thus included in the donor pool of potential control-group firms. Our goal is to find control-group firms that give an accurate measure of the counterfactual outcome for firms entering PriceSpy, meaning that treated and control-group firms should preferably differ only in terms of treatment assignment, and would in the absence of treatment have had identical development of the outcome variable of interest.

To identify such firms in the donor pool of potential controls, we use CEM (Blackwell et al., 2009; Iacus et al., 2011, 2012). In propensity score matching, improving the balance in one covariate might lead to increased imbalance in other covariates, while in CEM improved balance in one covariate does not affect the imbalance of other covariates. This is the case since the maximum level of imbalance between treated and control-group firms is set for each covariate by the researcher in CEM (Iacus et al., 2011, 2012). Furthermore, CEM has been shown to reduce model dependence, implying that empirical findings will be more robust to the choices of estimation model and model specification (Ho et al., 2007; Iacus et al., 2011).

As our main focus is to reduce imbalance in the outcome variable in the pre-treatment period, we match on the levels of CPI-adjusted sales two, three, four, and five years before PriceSpy entry for the treated firms.[4] The CEM is set to generate 1:1 matching, so that the numbers of treated and matched firms are equal in the matched dataset used in the difference-in-difference estimations. The continuous variable CPI-adjusted sales is coarsened into 10 equally sized bins, making the maximum allowed difference in CPI-adjusted sales in each bin approximately 10%. In addition, we group firms into retailers, wholesalers, and firms from other industries, and force the matching process to accept only firms from the same type of industry. The same goes for the year of entry of the treated firms: the matching is forced to find control firms in the same industry and that, in the same year as the year of entry of the treated firms, have CPI-adjusted sales that differ by at most 10% from those of the entry firms in the second, third, fourth, and fifth years before entry.

Table 1 presents descriptive statistics for the CPI-adjusted sales expressed in logarithms, $\ln Q_{i,t}$, as well as for the covariates used in the production function

[4] We use a two-year lag to reduce the possibility that any pre-entry adjustments by treated firms might affect the results.

difference-in-difference model used to estimate the impact of PriceSpy participation on output for all firms in the sample.[5] The statistics are presented separately for treated and control-group firms, and contain data from both before and after the matching procedure. The data clearly indicate that the matching has improved the balance in the outcome variable, and in most of the covariates as well.[6]

The identifying assumption in the difference-in-difference regression model presented in equation (4) below is that firms in the entry and control groups would have had parallel trends in the outcome variable in the absence of treatment. The development of output in the absence of PriceSpy entry for the entering firms is of course impossible to observe empirically, but we can observe the pre-entry trends in the outcome variable, $\ln Q_{i,t}$, for both the entry- and control-group firms. Fig. 2 presents the raw trends of $\ln Q_{i,t}$, while Fig. 3 presents the type of underlying trends suggested by Pope and Pope (2015) with which to evaluate the parallel trend assumption.[7] To produce the Pope and Pope (2015) trends, the regression presented in equation (4) below is run without the treatment-effect variable, and the residuals from this regression are presented in Fig. 3. As such, these residuals are supposed to represent the underlying trend in the outcome variable after having controlled for the impact of the other dependent variables in the regression.

As can be seen in Figs. 2 and 3, the trends seem to be parallel in the period leading up to entry, and there is also a slight indication of a treatment effect even in these descriptive statistics. Also note the negative trend in CPI-adjusted sales in Fig. 2 in the

[5] These statistics are also presented industry by industry in Appendix A.

[6] Regarding the covariates for which the matching is less successful in reducing imbalance, it should be noted that since these are included in the estimation of equation (4), reduction in balance is not as important as it is regarding the outcome variable.

[7] In Appendix B, these trends are separately presented for treatment- and control-group firms in the retail, wholesale, and other industries.

years leading up to entry, indicating that the firms entering PriceSpy, at least on average, might be doing so to address a downward trend in output.

Table 1. Means and standard deviations of dependent and independent variables used in estimating equation (4), before and after CEM, all industries.

Variable	**All industries**			
	Before CEM		**After CEM**	
	Treated	**Control**	**Treated**	**Control**
$\ln Q_{i,t}$	8.77 (2.17)	7.57 (2.07)	8.70 (2.14)	8.65 (2.16)
$\ln K_{i,t-1}$	5.17 (2.43)	5.49 (2.52)	5.11 (2.43)	5.68 (1.27)
$\ln L_{i,t-1}$	1.67 (1.33)	1.03 (1.04)	1.64 (1.30)	1.49 (1.27)
$\ln K_{i,t-1}^2$	32.68 (30.49)	36.55 (31.04)	32.05 (30.13)	38.55 (31.98)
$\ln L_{i,t-1}^2$	4.58 (8.10)	2.16 (4.18)	4.39 (7.71)	3.82 (6.65)
$\ln L_{i,t-1} \ln K_{i,t-1}$	12.33 (15.61)	8.11 (9.79)	12.03 (15.17)	11.94 (13.55)

Note: The differences for the treated firms before and after CEM are due to the loss of 30 firms from the full sample of firms that could not be matched using the chosen criteria.

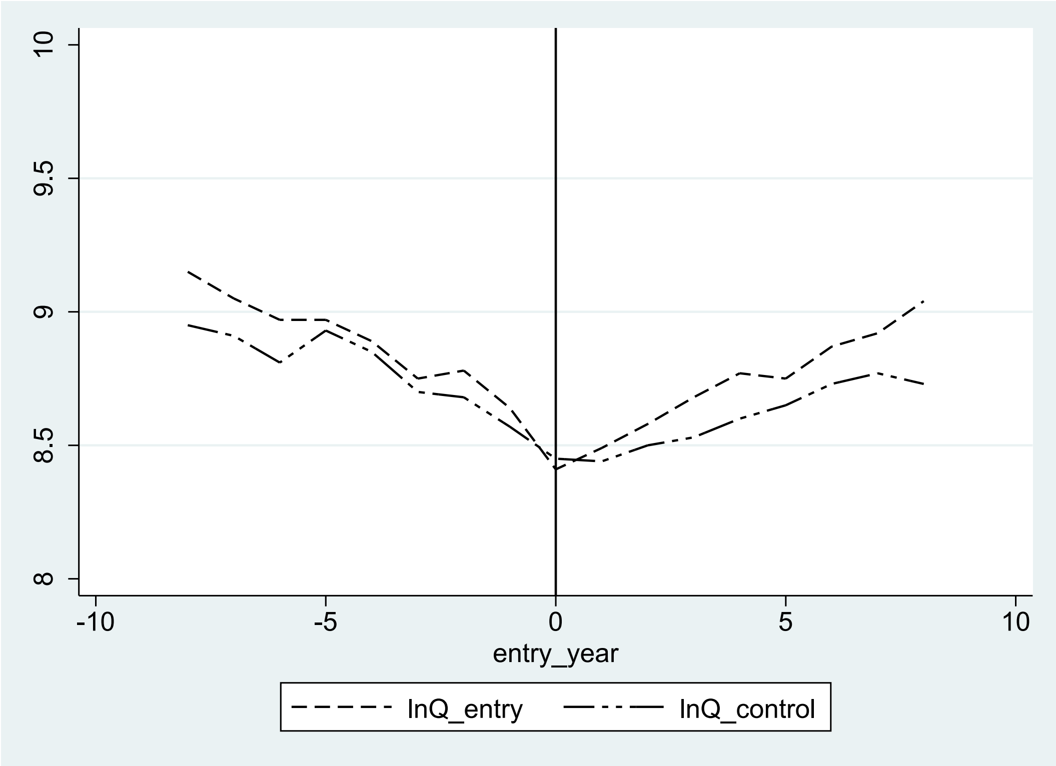

Fig. 2. Pre- and post-entry trends in ln $Q_{i,t}$.

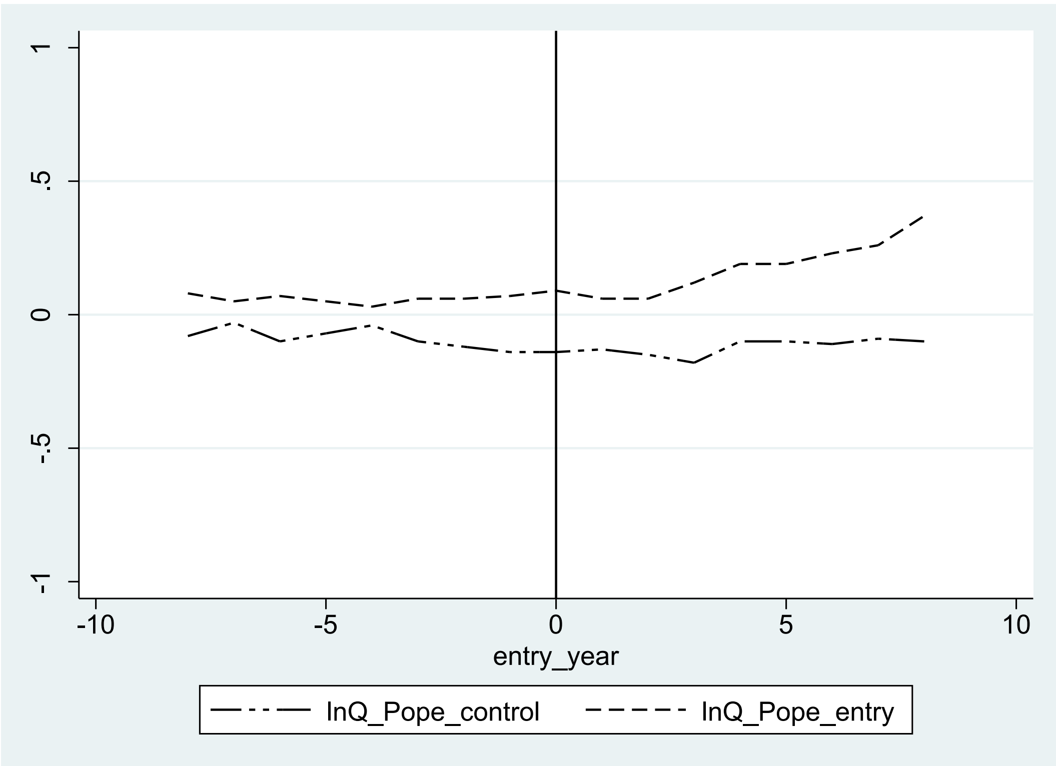

Fig. 3. Pre- and post-entry trends in ln $Q_{i,t}$ (Pope and Pope, 2015).

3.1.2 A translog difference-in-difference model

Following Han et al. (2018), our empirical model uses both cross-sectional and temporal variation in the data to estimate the impact of PriceSpy participation on output while holding the levels of inputs constant. Firms are assumed to use a technology that can be represented by the transcendental logarithmic (translog) production function developed by Christensen et al. (1971). This functional form is a second-order Taylor series approximation of an arbitrary production function, and can be written as follows:

$$\ln Q_{i,t} = \beta_1 \ln L_{i,t-1} + \beta_2 \ln K_{i,t-1} + \beta_3 \ln L_{i,t-1}^2 + \beta_4 \ln K_{i,t-1}^2 + \beta_5 \ln L_{i,t-1} \ln K_{i,t-1} + R_{i,t}, \quad (1)$$

where $Q_{i,t}$ is a measure of output and $L_{i,t-1}$ and $K_{i,t-1}$ are measures of the labor and capital inputs, respectively, both lagged one period to alleviate a potential endogeneity problem. Finally, $R_{i,t}$ is the remainder term of the Taylor series approximation, which in most empirical work is assumed to contain a constant and a random error term, making Eq. (1) a traditional OLS regression model to be estimated. However, as we are interested in measuring how entry by firms into the PriceSpy marketplace affects output when holding the level of inputs constant, i.e. whether entry on average causes a positive and statistically significant shift in the production function of the affected firms, our remainder term needs to take this into account. As such, we suggest the following remainder:

$$R_{i,t} = \beta_0 + \beta_6 TR_{i,t} + \omega_{i,t}, \quad (2)$$

where β_0 is a constant and $TR_{i,t}$ is an indicator variable equal to one for firms that have entered the PriceSpy marketplace in periods after entry, and zero otherwise. Our key variable of interest is $TR_{i,t}$, as this will provide an estimate of the treatment effect, i.e.

how the output of firms entering the PriceSpy marketplace compares with their own output before entry, and with the output of control-group firms throughout the study period, holding the levels of inputs (i.e. labor and capital) constant.[8] A positive parameter estimate for β_6 will indicate an increase in productivity in the sense that output has increased for given levels of inputs. Finally, $\omega_{i,t}$ represents other factors affecting output, $Q_{i,t}$.

The identification of β_6 could be confounded if there is a correlation between $TR_{i,t}$ and $\omega_{i,t}$, and since firms are free to select both whether and when to enter or exit the PriceSpy marketplace, this correlation is unlikely to be zero. Using the variation of the timing of PriceSpy entry across firms, $\omega_{i,t}$ will be specified as a function of firm- and time-specific fixed effects, γ_i and γ_t, and a residual, $\varepsilon_{i,t}$. Following Arcidiacono et al. (2020) in their study of the impact of Walmart entry on sales of incumbent retailers in the USA, our identification assumption will be that entry into the PriceSpy marketplace is uncorrelated with the error term, $\varepsilon_{i,t}$, conditional on the firm and time fixed effects. The remainder term can now be written:

$$R_{i,t} = \beta_0 + \beta_6 TR_{it} + \gamma_i + \gamma_t + \varepsilon_{i,t}, \tag{3}$$

Combining equations (1) and (3), we get:

$$\ln Q_{i,t} = \beta_0 + \beta_1 \ln L_{i,t-1} + \beta_2 \ln K_{i,t-1} + \beta_3 \ln L_{i,t-1}^2 + \beta_4 \ln K_{i,t-1}^2$$
$$+ \beta_5 \ln L_{i,t-1} \ln K_{i,t-1} + \beta_6 TR_{it} + \gamma_i + \gamma_t + \varepsilon_{i,t}, \tag{4}$$

which is a generalized difference-in-difference model. This type of difference-in-difference model is one of the tools most frequently used in applied economics research

[8] To obtain the change in output due to entry into the PriceSpy marketplace in percentage terms, the formula $100 \times [exp(\beta_6) - 1]$ is used (Wooldridge, 2010).

to evaluate the effects of public interventions and other treatments of interest on relevant outcome variables (Abadie, 2005).

If entry into the PriceSpy marketplace causes a shift in the supply function, as in our theoretical model, we expect $\beta_6 > 0$, i.e. that firms entering PriceSpy become more productive in that they increase their output for given levels of labor and capital in the period after entry. All variables used in estimating equation (4) will be described in Section 3.2, where we also present descriptive statistics.

3.2 Data collection and preparation

3.2.1 Identifying PriceSpy entry dates

Data collection regarding PriceSpy entry dates was conducted using the Wayback Machine, and the procedure is described in detail in Appendix C. The data collection process is briefly summarized below.

We use data covering the 2005–2015 period, and the data collection and analysis followed six steps: (1) sampling, (2) organizing and defining the boundaries of the web crawl, (3) crawling, (4) website variable operationalization, (5) integration with annual report data, and (6) analysis of the combined dataset.

Sampling involves collecting data on entry/exit into the PriceSpy website for as many firms as possible. Two approaches were used, i.e. "carbon dating" of webpages (SalahEldeen and Nelson, 2013) and retrieval of posted firm lists, both from the PriceSpy website, using the Wayback Machine. We also collected the historical number of firms stated by PriceSpy to validate our data.

Organizing and defining the boundaries of the web crawl involved finding out what part of the legacy PriceSpy content was of interest and within the scope of the data collection. This proved challenging, as the site has seen several changes over such a long period, so a trial-and-error approach to finding out the structure and changes over time was necessary.

Crawling the site was performed with R (R Team Core, 2017) and Ruby code. Website variable operationalization was then conducted by structuring the data using HTML nodes and regular expressions, with care taken to not omit firms and identifying firms such as sole proprietorships and foreign firms.

Finally, after data quality was assessed and found satisfactory, we combined our collected data with the annual report data described in Section 3.2.2. for analysis.

3.2.2 Annual report data and descriptive statistics

Griffith and Harmgardt (2005) and Reynolds et al. (2005) discussed how to measure output in retailing. When studying productivity in retailing, increased productivity is typically measured as the increase in sales or value added per worker (Reynolds et al., 2005), sometimes also accounting for other inputs such as capital.

First, it should be noted that using value added as the measure of output is not an option in our setting. This is the case because value added consists of approximately two-thirds wages, creating severe endogeneity problems if estimating equation (4) using labor as one of the independent variables.[9] Second, to make it possible to compare the different goods sold by different types of retailers, controlling for price is crucial (Griffith and Harmgardt, 2005).

As such, the sales of the firms included in this study must be discounted using a relevant price index; output, $Q_{i,t}$, is thus measured for each firm (index i) and year (index t), and is defined as sales of firm i in year t discounted by the Swedish consumer price index (CPI). The log transformation of output, $Q_{i,t}$, also has the benefit of making

[9] We have, however, estimated a traditional difference-in-difference model using value added as the outcome variable. For all firms, the results indicate an increase in productivity of 8%, while for retail firms it was 13% and for wholesale firms 10%. For firms in other industries, the result was not statistically significant at conventional levels.

the parameter estimate related to the effect of PriceSpy marketplace entry on firm output interpretable in percentage terms after some calculations (see footnote 8).

Following Håkansson et al. (2019), labor ($L_{i,t-1}$) is measured as the number of employees of firm i at time $t-1$, while capital ($K_{i,t-1}$) is measured as the value of the capital stock, i.e. the value of the land, buildings, and machinery of firm i at time $t-1$. Since the variables are log transformed, the parameter estimates from the estimation of equation (4) can be interpreted as elasticities. The annual report data from Bisnode cover the 2005–2015 period, and the means and standard deviations for all variables included in the estimation of equation (4) are presented in Table 2.

Table 2

Dependent and independent variables; means and standard deviations, variable descriptions, and data source, after CEM.

Variable	All industries	Retail	Wholesale	Other industries	Variable description	Data source
$\ln Q_{i,t}$	8.68 (2.15)	8.99 (1.78)	9.06 (2.30)	8.27 (2.24)	Output, measured as sales of firm i in year t discounted by CPI.	Bisnode/Statistics Sweden/own calculations
$\ln K_{i,t-1}$	5.37 (2.48)	5.19 (2.34)	5.47 (2.44)	5.47 (2.60)	Sum of the value of the land, buildings, and machinery of firm i at time $t-1$.	Bisnode/ own calculations
$\ln L_{i,t-1}$	1.57 (1.29)	1.75 (1.20)	1.61 (1.30)	1.42 (1.32)	Number of employees of firm i at time $t-1$.	Bisnode/ own calculations
$\ln K^2_{i,t-1}$	35.04 (31.16)	32.38 (28.23)	35.85 (29.94)	36.69 (33.66)	$\ln K_{i,t-1}$ squared.	Bisnode/ own calculations
$\ln L^2_{i,t-1}$	4.12 (7.24)	4.49 (7.54)	4.33 (6.88)	3.76 (7.17)	$\ln L_{i,t-1}$ squared.	Bisnode/ own calculations
$\ln L_{i,t-1} \ln K_{i,t-1}$	11.99 (14.45)	12.10 (14.26)	12.70 (13.99)	11.56 (14.80)	$\ln L_{i,t-1}$ multiplied by $\ln K_{i,t-1}$	Bisnode/ own calculations

3.3 Estimation results

The results of estimating equation (4) are presented in Table 3. The main variable of interest is $TR_{i,t}$, which provides an estimate of how the output of firms entering the PriceSpy marketplace compares with their own output before entry, and with the output of control-group firms throughout the study period, holding the levels of labor and capital constant. The effect in percentage terms of PriceSpy market participation on output while holding inputs constant is presented in the row marked *Effect in %* in Table 3.

Table 3

Estimation results; dependent variable $\ln Q_{i,t}$, translog difference-in-difference model.

	All industries		**Retail**		**Wholesale**		**Other industries**	
	coef.	**s.e.**	**coef.**	**s.e.**	**coef.**	**s.e.**	**coef.**	**s.e.**
$\ln K_{i,t-1}$	0.02	0.03	0.01	0.02	0.08***	0.03	0.01	0.05
$\ln L_{i,t-1}$	0.81***	0.08	0.79***	0.09	0.87***	0.12	0.77***	0.13
$\ln K^2_{i,t-1}$	0.01***	0.002	0.01**	0.004	0.004	0.004	0.01***	0.004
$\ln L^2_{i,t-1}$	0.01	0.01	0.0002	0.02	0.001	0.02	0.02	0.02
$\ln L_{i,t-1} \ln K_{i,t-1}$	–0.03***	0.01	–0.02	0.02	–0.03	0.02	–0.02	0.02
$TR_{i,t}$	0.11***	0.02	0.16***	0.02	0.12***	0.04	0.06**	0.03
Effect in %	11.63		17.35		12.75		6.18	
n	26882		9365		5792		11725	
Firm F.E.	Yes		Yes		Yes		Yes	
Year F.E.	Yes		Yes		Yes		Yes	
R^2	0.20		0.21		0.19		0.20	

*** significant at the 1% level, and ** significant at the 5% level. *Effect in %* is calculated using the formula $100 \times [exp(\beta_6) - 1]$.

The results indicate that entry into the PriceSpy marketplace increased output by, on average, 11.63% when analyzing the impact on all entering firms, irrespective of industry. For retail firms the increase was 17.35%, for wholesale firms it was 12.75%, and for firms in other industries it was 6.18%.[10]

Finally, we also analyze who gains the most from the increases in productivity due to PriceSpy market participation, shareholders or employees, using operating profits and gross wages as our dependent variables in a traditional difference-in-difference model.[11] Operating profits is a measure of firm profit that includes all operating incomes and expenses except interest expenses and income tax expenses, while gross wages refers to the total gross pre-tax compensation paid by employers to employees for work done, both measured during an accounting period, i.e. one year. The results of these estimations are presented in Table 4.

Table 4
Estimation results, dependent variables gross wages and operating profits, difference-in-difference model.

	All industries		**Retail**		**Wholesale**		**Other industries**	
	coef.	**s.e.**	**coef.**	**s.e.**	**coef.**	**s.e.**	**coef.**	**s.e.**
Gross wages$_{i,t}$	0.14***	0.02	0.16***	0.02	0.15***	0.04	0.12***	0.03
Effect in %	15.03		17.35		16.18		12.75	
Operating profit$_{i,t}$	0.09***	0.03	0.09	0.06	0.04	0.06	0.13**	0.05
Effect in %	9.42		9.42		4.08		13.88	

*** significant at the 1% level, and ** significant at the 5% level. *Effect in %* calculated as in Table 3.

[10] In Appendix D, we also present results of estimating a Cobb–Douglas production function specification. These results are similar to those presented in Table 3, indicating that our results are robust regarding the choice of production function, translog or Cobb–Douglas.
[11] A production function model is not an option in this setting since using capital and labor in estimating profits or wages would create severe endogeneity problems.

The results indicate that gross wages increased by, on average, 12.75–17.35% depending on the industry, while operating profits increased by 9.42% when analyzing the full sample of firms irrespective of industry. However, for the retail and wholesale firms in our sample, we do not find any statistically significant impact of PriceSpy market participation on operating profits, while the increase in operating profits was 13.88% for firms in industries other than retail and wholesale.

4. Summary and discussion.

The purpose of this paper has been to investigate how entry into the Swedish PriceSpy marketplace affects productivity, profits, and wages, to answer our main research question: Why do firms compete on price comparison websites?

Our interest in this question comes from previous research into how the increased use of PriceSpy and other price comparison website marketplaces has affected pricing. Lindgren et al. (2020) showed that for all 15 product categories under study, competition on PriceSpy caused a reduction in price, and for all categories except one, the result was statistically significant at the 1% level. This finding is unsurprising, since reductions in price due to competition in online markets or on price comparison websites has previously been reported by, among others, Brynjolfsson and Smith (2000), Clay et al. (2001), Brown and Goolsbee (2002), Haynes and Thompson (2008), and Tang et al. (2010).

What is more surprising is that despite the increased competition and reduced prices, firms are entering such marketplaces at an accelerating rate. Lindgren et al. (2020), for example, reported that the number of firms marketing games for the PlayStation 4 console increased from 20 firms marketing 20 games on the PriceSpy website in 2013 to 60 firms marketing 600 products in 2016. It is difficult to imagine such a development if the only effects on the participating firms were increased competition and lower prices.

In this paper, we suggest that entry into the PriceSpy marketplace also reduces marginal cost, shifting the supply function of the firm. Such a shift explains why prices fall after entering the PriceSpy website, something that would not be the case if PriceSpy participation only affected demand. As such, this explains the pattern observed in previous research that PriceSpy participation leads to reductions in price, but also to the testable prediction that firms entering the PriceSpy marketplace will increase their output even when holding the levels of inputs constant.

Our results indicate that for all firms entering the PriceSpy marketplace, there was an increase in output, while holding inputs constant, of 11.63%, while for retail firms the increase was 17.35%, for wholesale firms 12.75%, and for firms from other industries 6.18%, clearly suggesting that PriceSpy participation increases productivity. Also, as the numbers show, we found that retail and wholesale firms that entered the PriceSpy website increased their output more for a given level of inputs than did other firms. One possible explanation is that the retail and wholesale firms that entered the PriceSpy marketplace had more experience in online retailing in general, which could have given them a better understanding of how to use the PriceSpy market to increase sales. Investigating the precise reasons for this result would be an interesting avenue for future research, and answers could perhaps be found using qualitative research methods such as interviews with store managers.

We also investigated whether PriceSpy participation increases excess profits, and if so, who gains more from this increase, capital or labor. When analyzing the full sample of firms, we found that operating profits increased by 9.42% and gross wages by 15.03%. Since there is a statistically significant increase in both operating profits and gross wages, we conclude that PriceSpy entry creates an increase in excess profits to be divided among labor and capital owners.

However, the results also indicate that there is no increase in operating profits in retail or wholesale firms, suggesting that most of the gains from PriceSpy entry go to labor in these industries. This could be due to a change in the composition of the labor force if, for example, firms employ more computer specialists after entry. Alternatively, labor unions have a strong position in Sweden, with 67% of the labor force belonging to a union in 2018 (Kjellberg, 2019), and wages are covered by collective bargaining agreements. The fact that labor reaps most of the gains made by retail and wholesale firms might also be because labor unions have superior bargaining power in these industries. Discriminating between these possible explanations is outside the scope of the present paper, but would surely be an interesting avenue for future research.

Finally, we know from the trends of CPI-adjusted sales presented in Fig. 2 that firms entering the PriceSpy marketplace on average have negative trends in the period leading up to entry. This could be an indication that PriceSpy entry is seen by firm managers as a necessary step to avoid having to exit the market altogether. An empirical investigation into the motivations of firm managers entering the PriceSpy marketplace would thus be another interesting avenue for future research.

Acknowledgments

Research funding from the Swedish Retail and Wholesale Council, grant number 2018:773, is gratefully acknowledged. The authors would also like to thank Kenneth Carling, Lena Nerhagen, Siril Yella, and the participants in the Microdata Analysis Seminar (June 12, 2020) for their valuable comments and suggestions. We also thank Anton Gidehag for research assistance regarding the CEM modelling.

References

Abadie, A. (2005). Semiparametric difference-in-difference estimators. *The Review of Economic Studies, 72*(1), 1–19.

Arcidiacono, P., Ellickson, P. B., Mela, C. F., & Singleton, J. D. (2020). The competitive effects of entry: Evidence from supercenter expansion. *American Economic Journal*, in press.

Arora, S. K., Li, Y., Youtie, J., & Shapira, P. (2016). Using the Wayback Machine to mine websites in the social sciences: A methodological resource. *Journal of the Association for Information Science and Technology, 67*(8), 1904–1915.

Blackwell, M., Iacus, S. M., King, G., & Porro, G. (2009). CEM: Coarsened exact matching in Stata. *The Stata Journal, 9*(4), 524–546.

Brown, J. R., & Goolsbee, A. (2002). Does the Internet make markets more competitive? Evidence from the life insurance industry. *Journal of Political Economy, 110*(3), 481–507.

Brynjolfsson, E., & Smith, M. D. (2000). Frictionless commerce? A comparison of Internet and conventional retailers. *Management Science, 46*(4), 563–585.

Christensen, L. R., Jorgenson, D. W., & Lau, L. J. (1971). Conjugate duality and the transcendental logarithmic production function. *Econometrica, 39*, 255–256.

Clay, K., Krishnan, R., & Wolff, E. (2001). Prices and price dispersion on the web: Evidence from the book industry. *Journal of Industrial Economics, 49*(4), 521–539.

Griffith, R., & Harmgardt, H. (2005). Retail productivity. *International Review of Retail, Distribution, and Consumer Research, 15*(3), 281–290.

Han, M., Mihaescu, O., Li, Y., & Rudholm, N. (2018). Comparison and one-stop shopping after big-box retail entry: A spatial difference-in-difference analysis, *Journal of Retailing and Consumer Services, 40,* 175–187.

Haynes, M., & Thompson, S. (2008). Price, price dispersion and number of sellers at a low entry cost shopbot. *International Journal of Industrial Organization, 26*, 459–472.

Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis, 15*(3), 199–236.

Håkansson, J., Li, Y., Mihaescu, O., & Rudholm, N. (2019). Big-box retail entry in urban and rural areas: Are there productivity spillovers to incumbent retailers? *International Review of Retail, Distribution, and Consumer Research, 29*(1), 23–45.

Iacus, S. M., King, G., & Porro, G. (2011). Multivariate matching methods that are monotonic imbalance bounding. *Journal of the American Statistical Association, 106*(493), 345–361.

Iacus, S. M., King, G., & Porro, G. (2012). Causal inference without balance checking: Coarsened exact matching. *Political Analysis, 20*(1), 1–24.

Kjellberg, A. (2019). *Kollektivavtalens täckningsgrad samt organisationsgraden hos arbetsgivarförbund och fackförbund* (in Swedish). Studies in Social Policy, Industrial Relations, Working Life and Mobility. Research Report 2019:1, Department of Sociology, Lund University.

Lindgren, C., Daunfeldt, S.-O., & Rudholm, N. (2020). Pricing in retail markets with low search costs: Evidence from a price comparison website. Mimeo, Dalarna University.

Reynolds, J., Howard, E., Dragun, D., Rosewell, B., & Ormerod, P. (2005). Assessing the productivity of the UK retail sector. *International Review of Retail, Distribution, and Consumer Research, 15*(3), 237–280.

R Team Core, O. (2017). R: A language and environment for statistical computing.

SalahEldeen, H. M., & Nelson, M. L. (2013). Carbon dating the web: Estimating the age of web resources. Proceedings of the 22nd International Conference on World Wide Web, 1075–1082.

Tang, Z., Smith, M. D., & Montgomery, A. (2010). The impact of shopbot use on price and price dispersion: Evidence from online book retailing. *International Journal of Industrial Organization*, 28, 579–590.

Wooldridge, J. M. (2010). *Econometric Analysis of Cross Section and Panel Data*. Cambridge, MA: MIT Press.

Appendix A. Descriptive statistics before and after CEM, industry by industry.

Table 1

Means and standard deviations, dependent and independent variables, before and after CEM.

Variable	Retail				Wholesale				Other industries			
	Before CEM		**After CEM**		**Before CEM**		**After CEM**		**Before CEM**		**After CEM**	
	Treated	**Control**	**Treated**	**Control**	**Treated**	**Control**	**Treated**	**Control**	**Treated**	**Control**	**Treated**	**Control**
$\ln Q_{i,t}$	9.12	8.33	9.07	8.89	9.11	8.39	9.08	9.05	8.44	7.58	8.29	8.25
	(1.79)	(1.83)	(1.77)	(1.79)	(2.27)	(2.30)	(2.26)	(2.34)	(2.15)	(2.21)	(2.23)	(2.26)
$\ln K_{i,t-1}$	5.13	4.95	5.10	5.30	5.46	5.17	5.42	5.52	4.93	5.46	4.98	6.06
	(2.35)	(2.13)	(2.36)	(2.31)	(2.52)	(2.24)	(2.56)	(2.30)	(2.27)	(2.51)	(2.41)	(2.69)
$\ln L_{i,t-1}$	1.83	1.41	1.81	1.67	1.76	1.29	1.74	1.48	1.54	0.96	1.48	1.36
	(1.27)	(0.92)	(1.26)	(1.13)	(1.36)	(1.09)	(1.35)	(1.25)	(1.26)	(1.10)	(1.29)	(1.35)
$\ln K_{i,t-1}^2$	31.80	29.01	31.54	33.37	36.19	31.76	35.99	35.69	29.49	36.11	30.58	44.04
	(28.88)	(22.42)	(28.92)	(27.37)	(32.06)	(24.85)	(32.50)	(26.76)	(25.49)	(30.54)	(29.71)	(36.54)
$\ln L_{i,t-1}^2$	4.99	2.84	4.85	4.09	4.94	2.86	4.86	3.75	3.95	2.12	3.85	3.66
	(8.34)	(3.77)	(8.14)	(6.79)	(7.91)	(4.44)	(7.85)	(5.62)	(6.27)	(4.39)	(7.30)	(7.01)
$\ln L_{i,t-1} \ln K_{i,t-1}$	12.54	9.01	12.38	11.78	13.66	9.49	13.63	11.65	10.99	8.92	11.01	12.23
	(15.33)	(8.53)	(15.19)	(13.06)	(15.70)	(9.77)	(15.76)	(11.61)	(12.37)	(10.96)	(14.80)	(14.79)

Note: The differences for the treated firms before and after CEM are due to the loss of 30 firms (in the full sample) that could not be matched when using the chosen criteria.

Appendix B. Trends for retail, wholesale, and other industries.

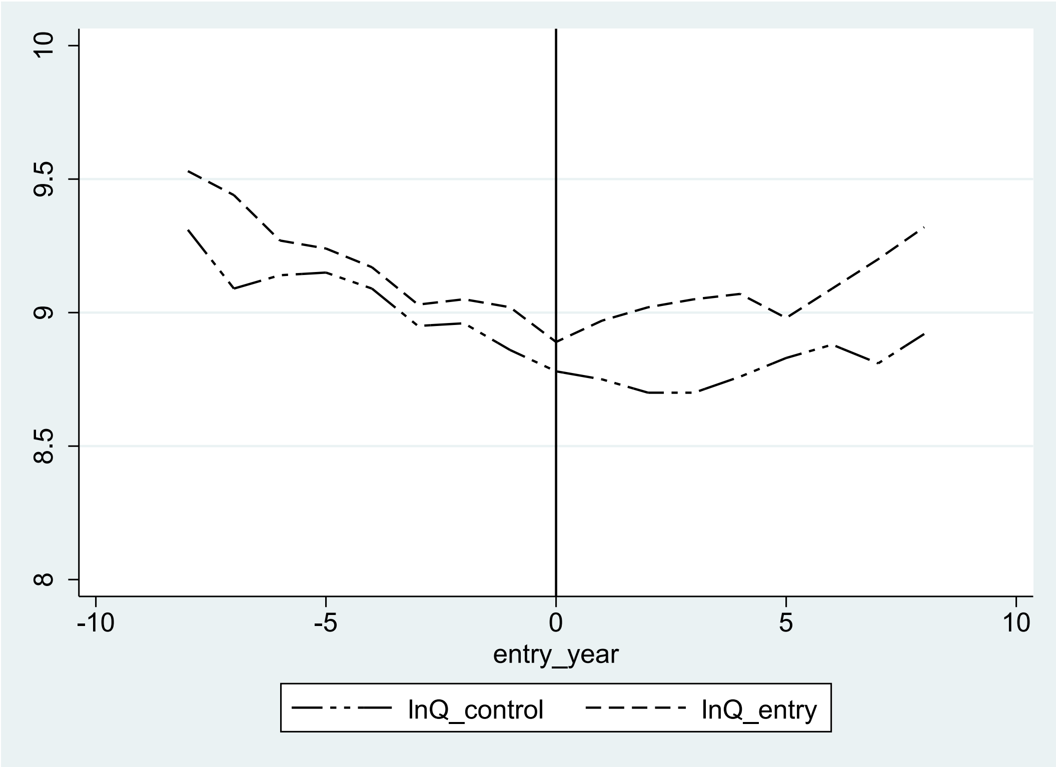

Fig. B1. Pre- and post-entry trends in $\ln Q_{i,t}$, retail.

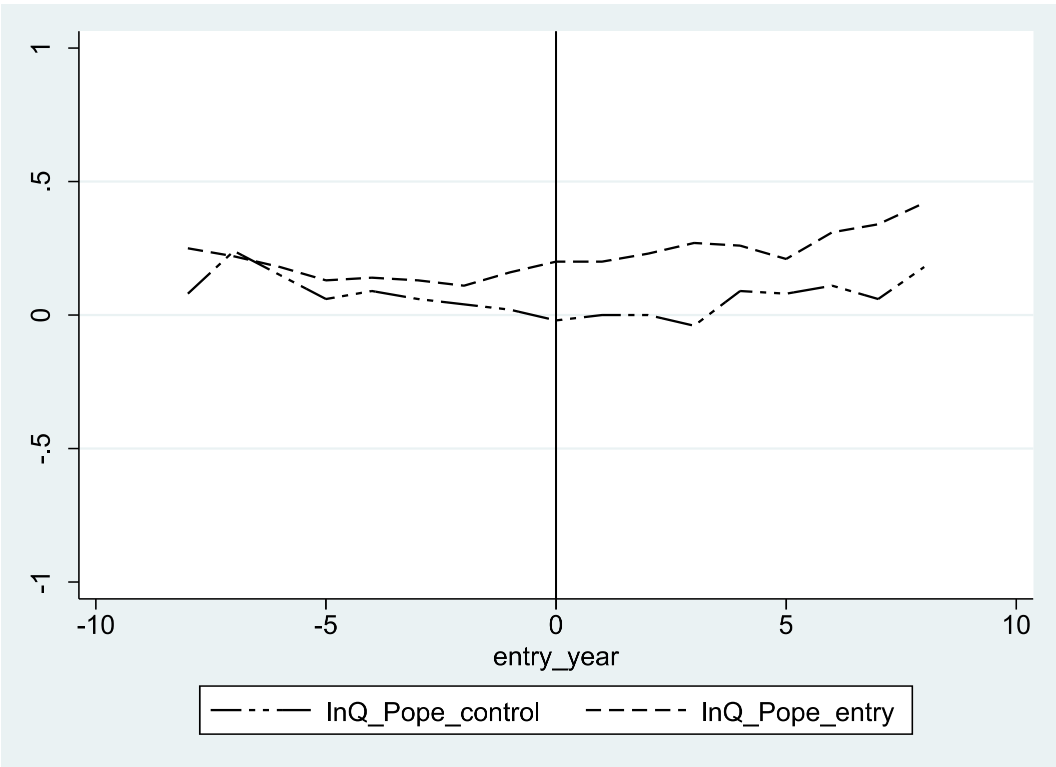

Fig. B2. Pre- and post-entry trends in $\ln Q_{i,t}$ (Pope and Pope, 2015), retail.

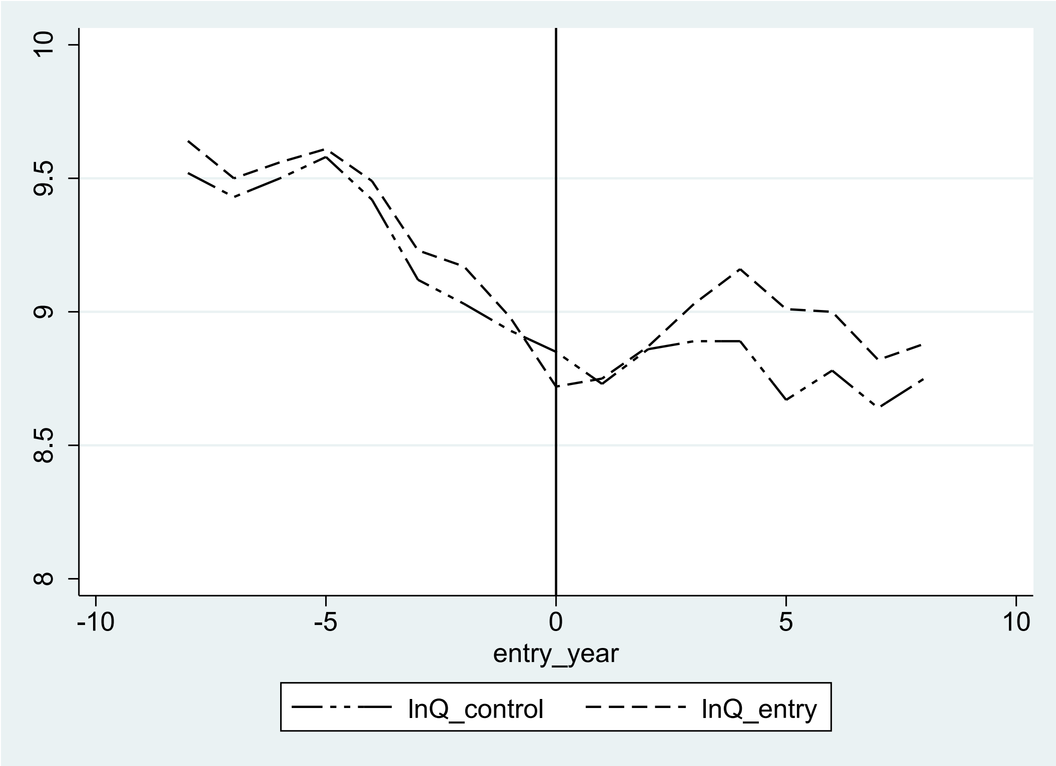

Fig. B3. Pre- and post-entry trends in $\ln Q_{i,t}$, wholesale.

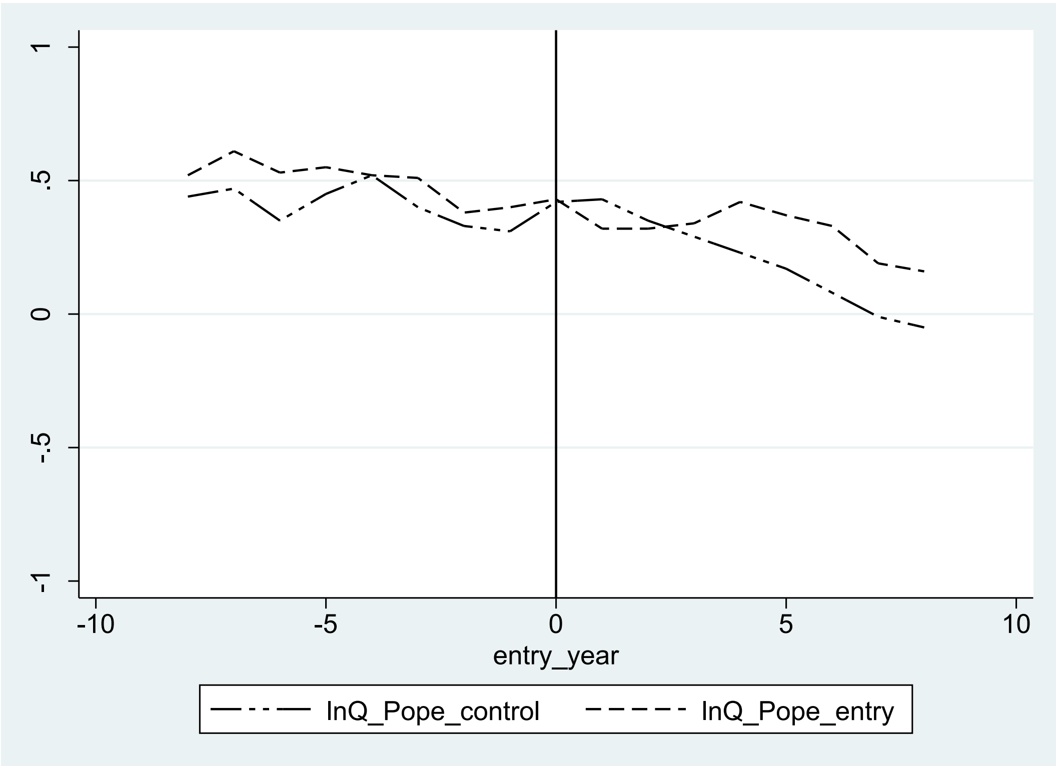

Fig. B4. Pre- and post-entry trends in $\ln Q_{i,t}$ (Pope and Pope, 2015), wholesale.

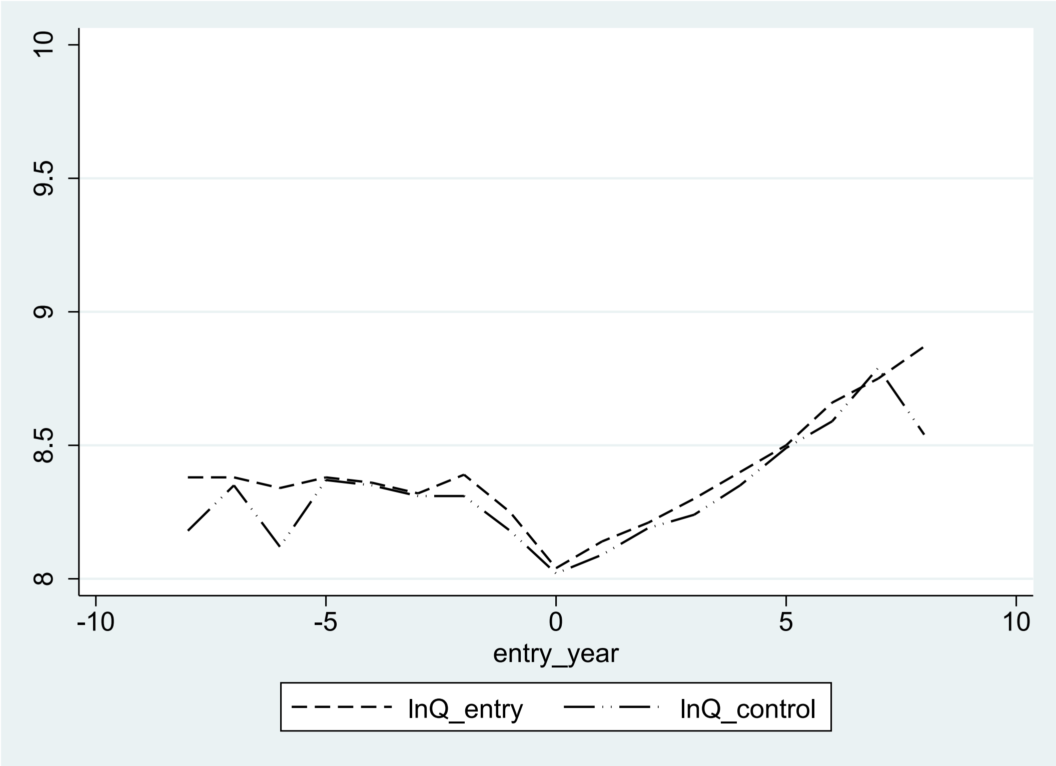

Fig. B5. Pre- and post-entry trends in $\ln Q_{i,t}$, other industries.

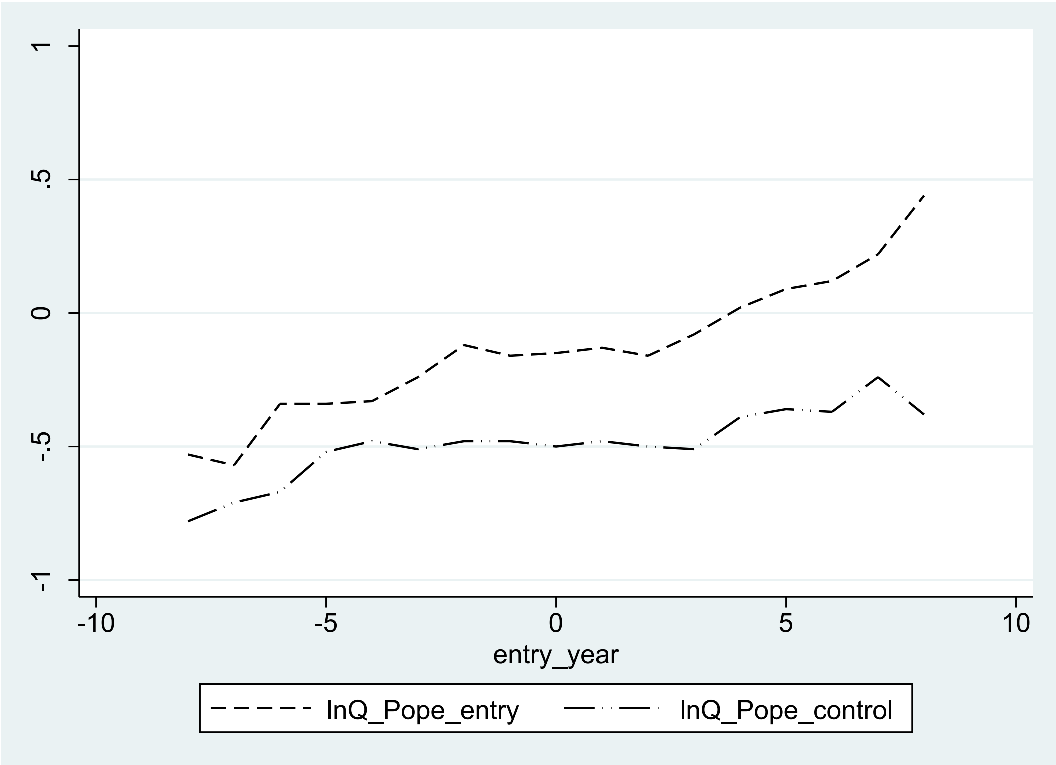

Fig. B6. Pre- and post-entry trends in $\ln Q_{i,t}$ (Pope and Pope, 2015), other industries.

Appendix C. Collection of PriceSpy entry dates.

Data on the dates of entry on PriceSpy for this paper were collected using the Wayback Machine, a large-scale data source used for analyzing web content launched in 2001. The final collected data cover the time span between 2002-12-16 and 2020-02-08, encompassing almost the entire lifetime of PriceSpy as a price comparison website, which began sometime in early 2000. However, since the annual report data cover only the 2005–2015 period, these are the years used in the final, combined dataset.

As we wanted to analyze legacy content that is not currently readily available, namely, panel data on firm behavior in terms of entry on the price comparison website PriceSpy in Sweden, to complement annual report data from Bisnode, there was a need to scrape the web using the Wayback Machine and to structure the subsequently collected data. There are generally six key steps in using the Wayback Machine for social science research: (1) sampling, (2) organizing and defining the boundaries of the web crawl, (3) crawling, (4) website variable operationalization, (5) integration with other data sources, and (6) analysis.[12] This appendix describes steps (1)–(5) in our context, while step (6) is left to the main article.

(1) Sampling. Sampling in this study involved collecting data on as many firms as possible that had entered and exited the price comparison site PriceSpy. We identified two main approaches to discovering whether a firm had entered or exited the website:

- Estimating the creation date of each firm's webpage on PriceSpy by "carbon dating" the date of creation (SalahEldeen and Nelson, 2013)
- Using information on active firms as listed on the PriceSpy website itself

[12] For a comprehensive outline of these steps as well as a literature review on the use of the Wayback Machine, see Arora et al. (2016).

The latter option is preferred to the former, since if we could secure all firm lists for all dates, we could simply compare each day t with $t + 1$ to get a perfect panel on when firms enter or leave the website. The option of carbon dating the websites would entail estimating the dates of entry or exit, and we cannot be certain that these dates align exactly with the true dates of entry/exit. Carbon dating has been shown to give 75.90% coverage and 32.78% correct dates when considering a "gold standard" test dataset (SalahEldeen and Nelson, 2013), and while this may be satisfactory in some research areas, we instead mainly relied on the firm lists posted on PriceSpy as our main source of firm entry data. Nevertheless, there is another important distinction between these approaches, in that we cannot retrieve the firm organization numbers from the posted firm lists; these numbers are necessary to efficiently connect the collected data to the Bisnode annual report data, and this information is only accessible through the firms' individual webpages on PriceSpy. Therefore, we had reason to combine the two approaches to reduce the sampling error as much as possible as well as to retrieve the firm organization numbers. It is fortunate that PriceSpy provides website-specific numerical firm ID variables that let us easily combine the resulting data from the two approaches.

Another concern regarding the sampling is to verify the validity of our data, for example, to determine whether or not we succeeded in capturing the number of firms on the website at a given time t. Again, fortunately, PriceSpy has diligently over the years provided an official count of the firms present on its site, which is also retrievable using the Wayback Machine. This gives us the means to compare our collected entry/exit data to an accurate measure of the number of firms present on the website over time.[13] The upper bound on our web scraping procedure is thus the number of

[13] This firm count was not posted on PriceSpy in the early years, such as 2002 and 2003. On the other

firms stated by PriceSpy, and with this number in hand, we can assess whether the data quality is sufficient for our purposes in terms of acceptable firm entry coverage.

(2) Organizing and defining the boundaries of the web crawl. The aforementioned firm list of webpages and individual firm webpages is within the scope of what was scraped from the site; for example, the unique uniform resource locators (URLs) for these two types of webpages were only necessary when web scraping, while other links on the website as a whole, such as product pages, reviews, and the home page, were outside the scope of this research. The web scraping code could be used to scrape the entire site over time, but we did not deem this additional complexity necessary for answering our research question.

A challenge when web scraping a legacy website, especially as far back in time as in our study, is that the structure and layout of the website, including the URLs, vary over time. The URLs of the individual firms' webpages were consistent throughout, with URLs supporting a "get" parameter towards the end as in "https://www.prisjakt.nu/butiksinfo.php?f=ID," where the ID is substituted for a website-specific numerical variable in order to uniquely identify a certain firm. The firm list webpages, on the other hand, have seen changes over the years, in terms of both URLs and content structure, which posed a challenge. After studying the change of the site over time, we were able to identify four types of URLs that PriceSpy transitioned between over the years; these are presented in Table C1, with the addition of a one-time web scrape of the PriceSpy firm list and individual firm webpages as of 2020-02-08 (e.g. not using the Wayback Machine). The URLs existing during 2003–2005 and 2004–2005 clearly overlap, likely due to some transition in the site

hand, the number of firms was low at that time, so we could simply count the number of firms present on the website for those years.

development during these years. We collected data from both these sources and removed any duplicates. After 2005, the webpages were more standardized, but when reaching the year 2012, PriceSpy chose to split the firm lists into sections based on the numerical and alphabetical initials of the firm names. This caused issues, since the Wayback Machine tends to collect data more frequently on webpages such as the main page of a website, while it captures web pages less frequently when considering hyperlinks in terms of crawling *depth* on the website—in our case, the sections of firm lists after 2012 as opposed to the whole firm lists presented on a single webpage before then.

Table C1
URLs and contents of firm lists on PriceSpy from 2003 to 2020.

Year and URL	**Content**
2003–2005 http://prisjakt.nu/foretag.php	Full list of firms on PriceSpy displayed on same webpage
2004–2005 http://prisjakt.nu/index.php?lista=butikslista	Full list of firms on PriceSpy displayed on same webpage
2005–2020 https://www.prisjakt.nu/butiksinfo.php	Full list of firms on PriceSpy displayed on same webpage until 2012, when a maximum of only 200 firms was shown in web page section “show all”
2012–2020 https://www.prisjakt.nu/butiksinfo.php?&begins_with=X	Store lists split into sections when substituting “get” variable X into the URL as: • “num”: names of firms beginning with 0–9 • “A” to “Z”: names of firms beginning with corresponding letters A–Z • "%C3%85", "%C3%84", and "%C3%96": initial letter of firm in Swedish alphabet Å, Ä, Ö

(3) Crawling. With boundary conditions set, we proceeded to web scrape and store HTML documents retrieved from the Wayback Machine. In this effort, we used code

written in R (R Team Core, 2017) and Ruby; the latter programming approach was found to be more successful, but the results of both were merged in the final dataset, omitting any duplicates.[14] When web scraping the individual firm webpages, we initially tried to use the subset of firm ID variables contained in the firm list webpages, but found it easier to loop over all integer values until we found no more firms for which to download HTML data. We found that the highest ID variable used by PriceSpy was 34797.[15] The final dataset consists of 6.10 GB and 5.04 GB of HTML documents retrieved with R and Ruby, respectively.

(4) Website variable operationalization. The HTML documents were then converted into structured data using a combination of HTML nodes and regular expressions. Care was taken to make sure that firms were not mistakenly omitted due to changes in HTML code over time. This meant some trial and error in choosing nodes and regular expressions in order to narrow the gap between the reported numbers of firms and the ones collected. This especially held true for the organization numbers, as a subset of firms does not have this number reported due to being sole proprietorships or run from a foreign country. To mitigate this issue, we also collected an additional variable on the firm webpage that indicates whether a firm is run from a foreign country and, if so, what country it is. There are instances of firms that transition between different organization numbers; our panel data capture this while still maintaining the firm-specific ID variable supplied by PriceSpy. Finally, we also collected firm names as reported by PriceSpy, but these are subordinate to the organization numbers needed to connect the data to the Bisnode financial reports.

[14] We were not successful in running Ruby code on the 2012–2020 section URLs.
[15] Note that this does not mean that this is the maximum number of unique firms in our data, as the firm ID variable is not consistently spread over all integers between 1 and 34797 according to some domain-specific way of assigning these values at the PriceSpy website.

(5) Integration with other data sources. Our final data on firms consist of 7144 unique firm IDs; the entries of these firms are shown along with the official PriceSpy firm count variable in Fig. C1. We can see that the web scraping procedure was very successful during the years 2002–2010 and 2012–2016 (as it happens, all in the month of March save for 2002), while between 2010 and 2012 as well as after 2016 there are gaps. These gaps are due to the results available on the Wayback Machine and cannot be circumvented, since once a webpage has not been collected by the Wayback Machine, it is no longer retrievable.

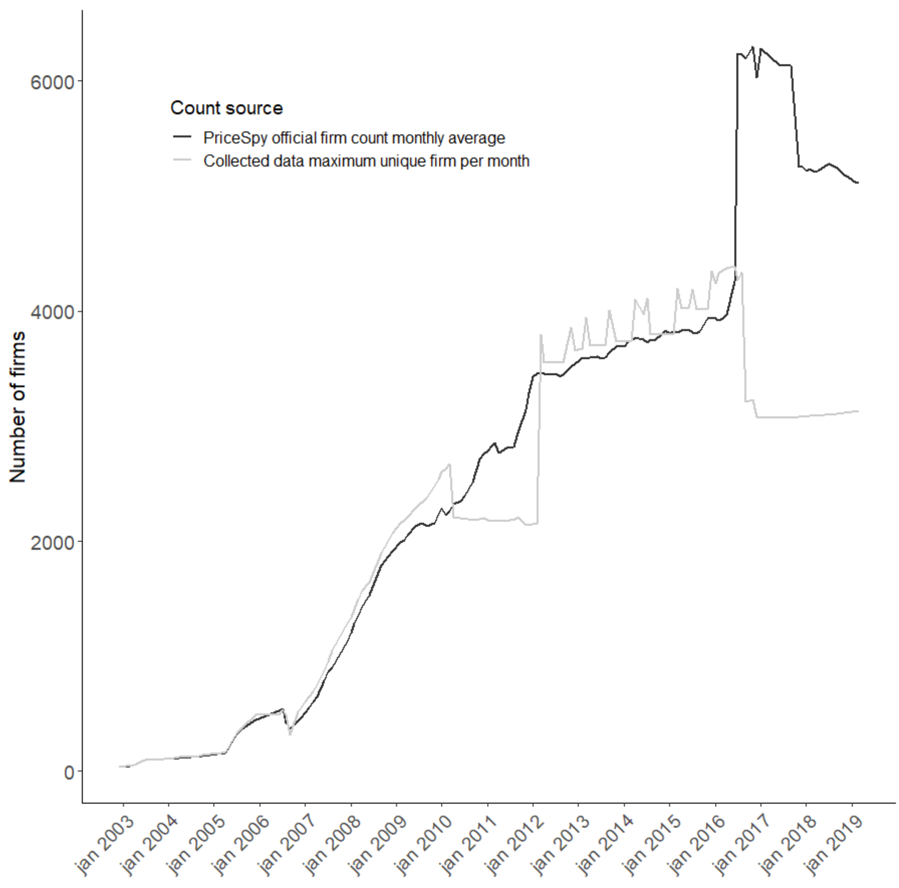

Fig. C1. PriceSpy official firm count versus collected firm count data per month.

It should be noted that the one-time scrape on the final date 2020-02-08 without using the Wayback Machine resulted in a subset of firms not found longitudinally in our data, likely entering some time after 2016 when the Wayback Machine did not have any entry data. The data after 2016-07-31 are therefore omitted due to insufficient data quality and for the obvious reason that this period did not give us any useful information on firm entry.

The structured data were then merged with the financial report data from Bisnode for the years 2005–2015.

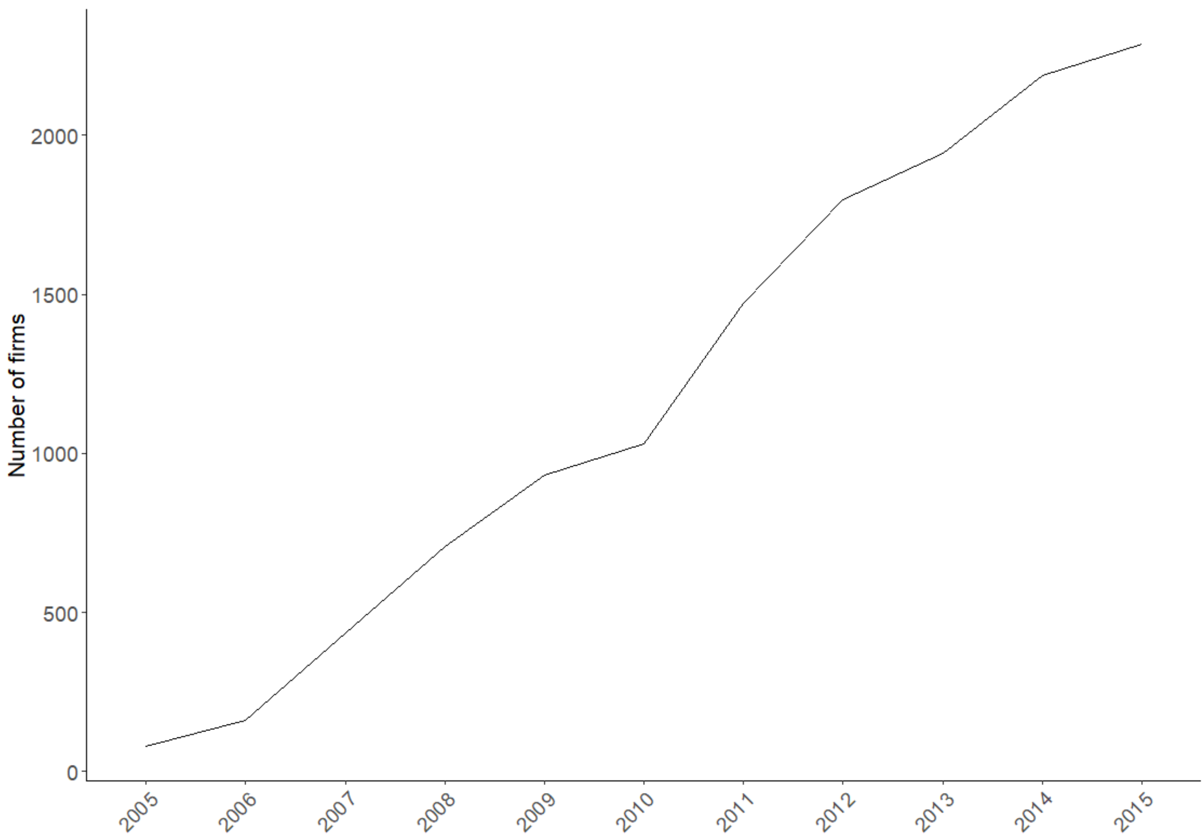

Fig. C2. Firms included in Bisnode financial report data, annual frequency.

As the Bisnode data were annual in frequency, we used 31 December of each year to determine whether a firm had entered that year and assigned a binary indicator variable for whether a firm did or did not exist on the PriceSpy website. Fig. C2 gives the plot of yearly firm entry, in which we can see that the underreporting during 2010–

2011 seen in Fig. C1 is not as significant when considering our final sample, which would indicate that the firms not included during this period were more likely to be sole proprietorships or foreign firms without organization numbers.

Appendix D. Estimation results using a Cobb-Douglas model.

Table D1

Estimation results, dependent variable $\ln Q_{i,t}$, Cobb–Douglas model.

	All industries		**Retail**		**Wholesale**		**Other industries**	
	coef.	**s.e.**	**coef.**	**s.e.**	**coef.**	**s.e.**	**coef.**	**s.e.**
$\ln K_{i,t-1}$	0.07***	0.007	0.05***	0.01	0.06***	0.01	0.10***	0.01
$\ln L_{i,t-1}$	0.70***	0.03	0.70***	0.05	0.70***	0.06	0.70***	0.05
$TR_{i,t}$	0.11***	0.02	0.16***	0.02	0.12***	0.04	0.06**	0.03
Effect in %	11.63		17.35		12.75		6.18	
n	26882		9365		5792		11725	
Firm F.E.	Yes		Yes		Yes		Yes	
Year F.E.	Yes		Yes		Yes		Yes	
R^2	0.19		0.21		0.18		0.19	

*** significant at the 1% level, and ** significant at the 5% level. *Effect in %* is calculated using the formula $100 \times [exp(\beta_6) - 1]$.