More Than the Sum of Its Parts? Markups and the Role of Establishments

Markus Kondziella and Joshua Weiss*

November 24, 2025

Click here for the most recent version

Abstract

We study distortions to firm production along the intensive margin—how much to produce at each establishment—and the extensive margin—how many establishments to open. Using data on the universe of Swedish establishments in services industries, we show that 1) size-dependent *firm* markups are just a symptom of size-dependent *establishment* markups, i.e., firms with larger establishments set higher markups but firms with more establishments do not, and 2) each successive establishment at a firm tends to be smaller relative to its municipality-industry. In a model of competition between firms through establishments, we characterize the distortions implied by size-dependent establishment markups: firms inefficiently undervalue 1) production at larger establishments, 2) opening larger establishments, and 3) production at existing establishments relative to opening new establishments. Calibrating to our Swedish data, we find that firms' extensive margin decisions are responsible for only 9% of the losses implied by these distortions. Nonetheless, firms' extensive margin decisions are crucial for the design of optimal firm size-dependent policy and sharply limit its effectiveness. The reason is that compared to intensive margin distortions, extensive margin distortions imply a much lower optimal relative subsidy for large firm sales.

*Kondziella: University of St. Gallen and the Swiss Finance Institute, markus.kondziella@unisg.ch. Weiss: University of Bristol and IIES, Stockholm University, joshua.weiss@iies.su.se. For useful feedback, we thank Corina Boar, Timo Boppart, Maarten De Ridder, Niklas Engbom, Ricardo Lagos, Kieran Larkin, Virgiliu Midrigan, Kurt Mitman, John Morrow, Jane Olmstead-Rumsey, Paula Onuchic, Alessandra Peter, Riccardo Silvestrini, Kjetil Storesletten, Florian Trouvain, and Venky Venkateswaran. We also thank seminar and conference participants at various institutions.

1 Introduction

Firms in many services industries can expand along two margins: selling more at each establishment (the intensive margin) and opening new establishments (the extensive margin) to sell new goods or access new markets. Recent work shows that both margins are important for understanding the recent rise of large firms in services industries in the United States.[1] How do inefficient distortions differ for these two margins? How should we design policy to undo distortions if it can only target firm-level variables? Can such a policy be effective?

An important distortion to firms' production decisions is size-dependent markups.[2] Within industries, larger firms set higher markups, which implies misallocation: larger firms inefficiently underproduce and smaller firms overproduce. A relative subsidy for large firm sales can fully undo this misallocation. Although controversial, this at least suggests that a shift in production toward large firms with high markups—as occurred recently in the US—may reflect an improvement in allocative efficiency despite a rise in the average markup.[3]

Using data on the universe of Swedish firms and establishments in services industries, we argue that size-dependent *firm* markups are just a symptom of size-dependent *establishment* markups; that is, larger firms set higher markups only because they have larger establishments, which set higher markups. We show that this distinction matters because size-dependent *establishment* markups imply different distortions for firms' intensive and extensive margins of production: relative to a social planner, firms 1) undervalue producing at large establishments relative to small establishments, 2) undervalue opening large establishments relative to small establishments, and 3) undervalue producing at existing establishments relative to opening new establishments. By contrast, size-dependent *firm* markups distort both margins equally.

We compute the costs of these distortions in our Swedish data and find that 91% are due to the misallocation of production across the competitive equilibrium set of establishments. The remaining 9% comes from distortions to firms' extensive margin decisions (how many establishments to open). Thus, computing misallocation taking the set of establishments as given, as in Hsieh and Klenow (2009), captures almost all misallocation due to size-

[1] Hsieh and Rossi-Hansberg (2023) and Cao et al. (2022) show that large firms with large establishments opening new establishments accounts for the observed rise in industry concentration.

[2] Edmond et al. (2023) find that eliminating the misallocation associated with size-dependent markups can increase welfare by the equivalent of a more than 11% permanent increase in consumption. They also provide empirical evidence for, and a thorough discussion of, size-dependent firm markups.

[3] See Baqaee and Farhi (2020) for a general analysis. More specifically, Weiss (2020) shows that a shift toward intangible capital favors large firms, leading to a rise in industry concentration, markups, and welfare.

dependent establishment markups. However, even though firms' extensive margin decisions are nearly efficient, they are crucial for the design of *firm* size-dependent policy and sharply limit its effectiveness. If we take the set of establishments as given, then optimal firm size-dependent policy is effective: it eliminates 95% of misallocation. But this policy backfires badly when the set of establishments endogenously responds, leading to lower welfare than without policy. Moreover, even if we take this response into account, optimal firm size-dependent policy improves welfare only 40% as much as it does if we can hold fixed the set of establishments. The intuition is the following trade-off. Efficiently reallocating production across establishments requires a high relative subsidy on large firm production. But this inefficiently leads large firms to open too many establishments, which pushes smaller firms to shrink and close their establishments. This means we should be much less inclined to favor large firms than focusing on size-dependent *firm* markups would suggest.

The first contribution of the paper is empirical. We document three findings using firm- and establishment-level data on the universe of Swedish firms in services industries from 1997-2017, which is about three-fifths of the Swedish economy. First, multi-establishment firms are rare but earn almost half of all revenues. Their advantage over single-establishment firms is about 60% because they have many establishments and about 40% because their establishments are relatively large, where the two margins are positively correlated. Second, firms' *marginal* establishments are smaller than their *average* establishments. Specifically, a firm's second establishment tends to be smaller than its first, its third establishment tends to be smaller than its second, and so on, where an establishment's size is measured relative to other establishments of the same age, location, and industry. Third, relatively larger firms within industries set relatively higher markups because they have *larger establishments*, which are associated with higher markups. On the other hand, having *more establishments* is not associated with higher markups once we control for firms' average establishment size.

Next, we develop a model of competition between firms through establishments, which we analyze theoretically and then quantify using our data. Each firm begins with an initial unit measure of establishments (interpreted as their first establishment), which is characterized by its quality that serves as a demand shifter. Firms then choose how many additional establishments to open subject to a nonlinear cost function. Each successive establishment at a firm has a lower quality, which matches our second empirical finding that each successive establishment is smaller. We interpret this as capturing that a firm sees a set of establishment opportunities and conditional on opening N establishments, chooses the N where consumer tastes are best aligned with their products. Finally, firms choose production at each of their

establishments. Non-CES Kimball (1995) demand across establishments implies that larger establishments set higher markups, which matches our third empirical finding.

The second contribution of the paper is a general characterization of the inefficient distortions implied by size-dependent establishment markups in our model. Households inelastically supply labor, so the cost of distortions is the misallocation of labor across its various uses—opening establishments at each firm and producing at each establishment. If markups are constant across establishments, then the competitive equilibrium is efficient. If markups are increasing in establishment size—under *any* Kimball (1995) demand system where non-producing establishments are irrelevant—then on the margin, a social planner wants to 1) reallocate production from small establishments to large establishments, 2) reallocate establishments from firms with small marginal establishments to firms with large marginal establishments, and 3) close marginal establishments in order to increase production at any larger establishments and even some smaller establishments. Since firms' marginal establishments are their smallest establishments, 3) means it is beneficial to close establishments at relatively large firms in exchange for producing more at the establishments of much smaller firms. The intuition for these distortions is as follows. If larger establishments set higher markups, then firms undervalue each successive unit of an establishment's production by more relative to a social planner. Thus, 1) and 2) hold, respectively, because firms undervalue *marginal* and *average* production at large establishments more than at small establishments. Finally, 3) holds because firms undervalue an establishment's marginal production by more than they undervalue its average production.

We calibrate the model to match our Swedish data in order to quantify the losses from these distortions and the implications for firm size-dependent policy. We target the frequency of multi-establishment firms, their average number of establishments, and their share of total sales. Despite only targeting these averages for multi-establishment firms, the model does a good job of matching the tails of the distributions of firms sales and number of establishments, as well as the relationship between number of establishments and sales per establishment among multi-establishment firms. This relationship is important because inefficient distortions come from variation in establishment size.

To quantify misallocation, we efficiently reallocate labor in three steps. First, reallocating production labor across establishments within each firm raises consumption by only 0.06%. Second, reallocating production labor across firms yields additional gains of 1.28%. Finally, choosing the efficient set of establishments leads to further gains of only 0.13%. Thus, 91% of misallocation is of production labor across the competitive equilibrium set of establishments,

most of which is due to misallocation across firms rather than within firms. Our assumption of declining quality across a firm's successive establishments is crucial. That is, we find ten times as large gains from changing the set of establishments in an alternative model in which a firm's marginal establishment is not, on average, different from its other establishments. This explains why Afrouzi et al. (2023) find much higher misallocation from firms' extensive margin decisions in their similar model where the extensive margin is a firm's number of customers rather than its number of establishments. Specifically, they assume all a firm's customers are identical, whereas we match our empirical finding that each successive establishment at a firm is smaller. This introduces dramatically diminishing returns to changing the set of establishments, which reduces the gains from doing so.

Finally, we study firm size-dependent policy, which is a tax/subsidy scheme where a firm's payment depends on its total sales across all its establishments. If we can hold fixed the set of establishments from the competitive equilibrium, then the optimal firm size-dependent policy is effective: it eliminates 95% of production misallocation across these establishments. To do so, it offers a relative subsidy to large firms because they have large establishments that inefficiently underproduce. However, if we implement this policy and the set of establishments endogenously responds in equilibrium, then welfare *falls* relative to the competitive equilibrium without policy because large firms open way too many establishments. If we take into account this extensive margin response when designing policy, then optimal policy still gives a relative subsidy to large firms, but much less so. Moreover, the optimal policy eliminates only 35% of total misallocation, which is about 40% of the welfare improvement we get with a fixed set of establishments.

Summarizing our quantitative results, the third contribution of our paper is to show that firms' extensive margin decisions—how many establishments to open—are responsible for a small share of the welfare losses due to size-dependent establishment markups, but are critical for the design of firm size-dependent policy and significantly reduce its effectiveness.

Expanding through establishments. This paper relates to a recent literature studying firm expansion through opening new establishments. Hsieh and Rossi-Hansberg (2023) and Cao et al. (2022) show that the recent rise in industry concentration in the US services sector was due to the largest firms expanding by opening new establishments. Oberfield et al. (2024) and Becker et al. (2024) use spatial models to study heterogeneity across locations in the firms that open establishments and the markups they set. Instead, we abstract from location heterogeneity and instead focus on the misallocation of establishments and produc-

tion across firms, as well as the implications for firm size-dependent policy. To this end, we show that firms successive establishments tend to be smaller even controlling for location. We also contribute to this literature by demonstrating empirically what is often the case in models of local competition: size-dependent markups are set based on the size of a firm's establishments rather than the firm's number of establishments.

Size-dependent markups and misallocation. Our theoretical results build on previous work on the inefficient distortions associated with size-dependent markups in a competitive equilibrium. Dixit and Stiglitz (1977) study a model with single-establishment identical firms, non-CES demand (with CES as a special case), and a free entry condition. Zhelobodko et al. (2012) and Dhingra and Morrow (2019) include productivity heterogeneity and selection through a fixed cost. Behrens et al. (2020) add multiple sectors, where aggregation across sectors is CES for most of the analysis. These papers all argue that non-CES aggregation—which maps into size-dependent markups—distorts the intensive margin (production at each firm) and the extensive margin (number of firms). We differ in that we add a firm-specific entry choice (number of establishments) subject to a non-linear cost, where aggregation across all establishments both within and across firms is non-CES. If we had a linear cost for opening establishments, for example, then all firms' marginal establishments would be the same size, so there would be no misallocation of establishments across firms. We also simplify by focusing on the case where an establishment's demand elasticity is weakly falling in its output, whereas the referenced papers allow for a more general relationship.

Our results on the costs of these distortions and the policy implications relate to Edmond et al. (2023). They study the cost of size-dependent *firm* markups in a model of single-establishment firms. They show that firm size-dependent policy can eliminate these costs by giving a relative subsidy to larger firms who set higher markups. By contrast, we find that firm size-dependent policy is not effective for undoing the costs of size-dependent *establishment* markups because intensive and extensive margin distortions are not well aligned.

In this context, perhaps the paper most closely related to ours is Afrouzi et al. (2023), who study a similar model but where the extensive margin is a firm's number of customers rather than number of establishments. A major difference in our results is that they find substantial misallocation due to the extensive margin and we find little. We show that this is due to a key difference between our models: they suppose all a firm's customers are identical, whereas we match our empirical finding that each successive establishment at a firm is smaller. Moreover, our focus is different in that we study firm size-dependent policy,

which they do not. We also show that declining size across a firm's successive establishments is important for our results in this context; otherwise, firm size-dependent policy is much more effective because intensive and extensive margin distortions are better aligned.

Finally, unlike Baqaee and Farhi (2020), we focus on size-dependent markups rather than all markup variation. We are interested in firms' decisions to open new establishments, so it is important to pick up the predictable component of that establishment's markup rather than ex-post variation. Moreover, we do not have establishment revenue data, so we cannot measure establishment markups directly. Instead, we infer the relationship between an establishment's size and its markup from the relationship between a firm's sales per establishment and the firm's markup.

The paper proceeds as follows. In Section 2, we describe our empirical analysis. In Section 3, we develop the model and prove theoretical results. In Section 4, we calibrate the model, which we use in Section 5 to study misallocation and firm size-dependent policy.

2 Empirical Analysis

2.1 Data

We use data on the universe of Swedish firms and establishments in services industries from 1997 to 2017. We merge information from two administrative datasets based on firms' tax forms. Both are from Statistics Sweden, the Swedish government agency responsible for producing official statistics. The first data set, Företagens Ekonomi, contains annual financial accounts of all Swedish firms. We use firms' annual sales and expenditures on intermediate inputs. The second data set, Registerbaserad arbetsmarknadsstatistik, contains the complete set of employer-employee linkages at a monthly frequency. For each worker, the data include their labor earnings and the IDs of the establishment and firm at which they are employed. We aggregate across workers within each establishment-year and merge with our firm-level data. This yields annual data on each firm's number of establishments and on each establishment's firm ID, wage bill, employment, municipality, and industry.

There are 291 municipalities in 2017, ranging from Stockholm with about 1 million workers in services industries to Bjurholm with 960 workers in services industries. Industries are at the 5-digit level, which is the finest level of disaggregation available. There are 423 ser-

vices industries, which have an average of 422 firms each.[4] To restrict attention to services industries, we only include firms and establishments in these industries, respectively, for our firm- and establishment-level analyses. We also only include firm-year observations with at least one employee and positive sales operating in the private economy. This leaves about 3.5 million firm-year observations, which cover about 60% of sales in the Swedish economy. Finally, we deflate nominal variables to 2017 SEK using the GDP deflator.

2.2 Intensive vs. Extensive Margins of Sales

We first demonstrate the importance of sales per establishment (the intensive margin) and number of establishments (the extensive margin) as determinants of services firms' sales. Table 1 compares single- and multi-establishment firms. The latter are the minority (2.7% of firms), but they earn about half the sales in services industries. The large average size of multi-establishment firms arises from two factors: their establishments sell almost five times as much as single-establishment firms' establishments, and they have about seven times as many establishments compared to each single-establishment firm.

Table 1: Single vs. multi-establishment firms

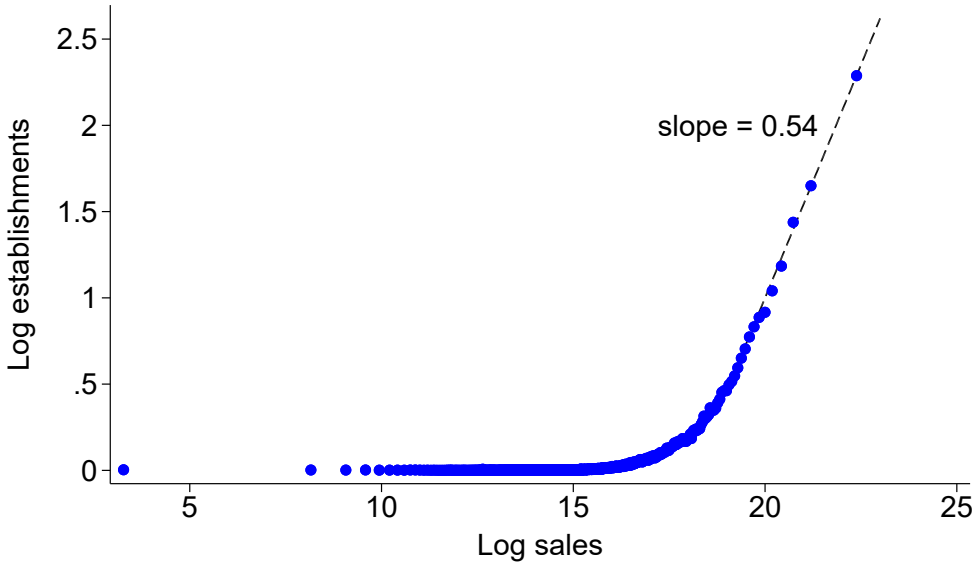
Multi-establishment firms		**Decomposing multi-establishment firm sales**	
Firm share	*Sales share*	*Relative sales per establishment*	*Relative number of establishments per firm*
2.7%	48.7%	4.7	7.2

The first two entries are the share of services firms that are multi-establishment, and the share of services sales earned by multi-establishment firms. The third entry is mean sales per establishment (deflated to 2017 SEK) across multi-establishment firms' establishments relative to across single-establishment firms' establishments. The fourth entry is mean establishments per firm across multi-establishment firms. We compute these in each year, and then take the average across years.

Figure 1 shows the role of the intensive and extensive margins across the firm size distribution. It plots the average log number of establishments at a firm as a function of its log sales. For small firms, the line is flat at 0 because nearly all small firms are single-

[4]Out of 821 industries, we keep those coded 45 and above, excluding "financial and insurance activities".

Figure 1: Drivers of firm sales



We group all firm-year observations into 1000 equally sized bins according to log sales (in 2017 SEK). Each dot plots average log number of establishments within a bin against average log sales.

establishment. For larger firms, the slope of the line is positive, increasing, and converges to about 0.54. This means larger firms tend to have more establishments and tend to sell more at each establishment. Among the largest firms, 54% of variation in sales comes from variation in their number of establishments, and the remaining 46% comes from variation in their sales per establishment.

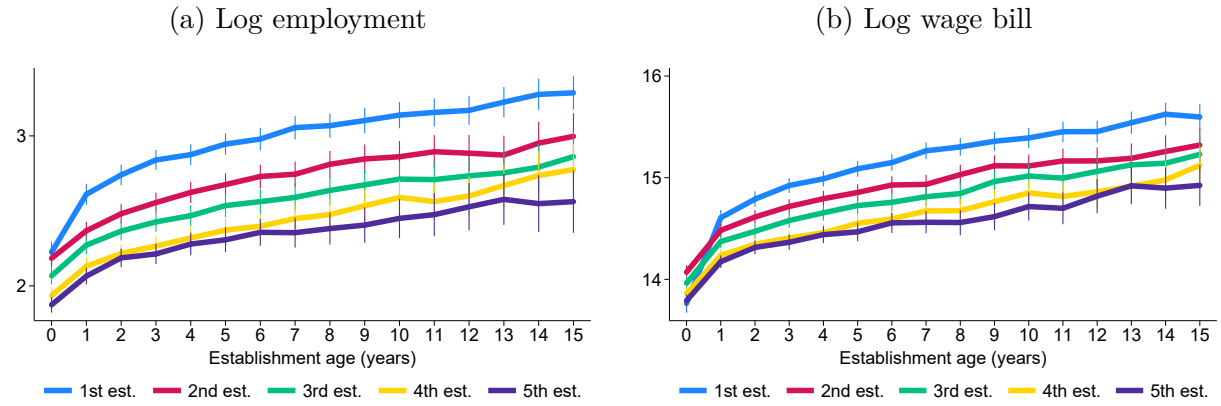
2.3 Declining Size at Successive Establishments

To study firms' extensive margin decisions—how many establishments to open—we want to know about their *marginal* establishments, not just their *average* establishments. To this end, we compare firms' successive establishments, i.e., we ask how the size of a firm's n^{th} establishment depends on n, which we call the establishment's "order". We use employment (number of employees) and the wage bill to measure establishment size because these are the measures for which we have establishment-level data.

We find that each successive establishment at a firm is smaller, conditional on establishment age. First, Figure 2 plots the average log wage bill and log employment across an establishment's life cycle conditional on survival for a firm's first five establishments. We

restrict attention to firms with at least five establishments, so the set of firms is the same for each line. For each successive establishment, the entire life cycle profile of the wage bill and employment is shifted down. For example, at an *establishment* age of ten years, the wage bill at a firm's third establishment tends to be about 10.5% smaller than the wage bill was at the firm's second establishment. In Appendix A, we show that the same stark pattern holds for firms with different numbers of establishments.

Figure 2: Size of successive establishments



Average log employment (number of employees) and wage bill (in 2017 SEK) for firms' first five establishments as functions of establishment age, conditional on survival. Averages are computed across firms with at least five establishments. The vertical bars are 95% confidence intervals.

Next, we conduct a more general parametric analysis, and confirm our findings, using the following regression:

$$\ln(\text{establishment size}_{i,j,n,t}) = \beta \ln(n) + \text{fixed effects}_{i,j,n,t} + \epsilon_{i,j,n,t}. \tag{1}$$

There is one observation for each establishment-year pair. On the left-hand side is the log size in year t of the n^{th} establishment opened by firm i in industry j, measured by the establishment's employment or wage bill. On the right-hand side, our main coefficient of interest is β on the log of the establishment's order, n. In our baseline specification, we include establishment age, firm, and year fixed effects. The residual is $\epsilon_{i,j,n,t}$. Before, we restricted attention to firms with at least five establishments to eliminate selection effects. In our regression, firm fixed effects make this unnecessary. As such, we include all establishments.

Columns 1 and 3 in Table 2 show the estimated coefficient β from our baseline regression,

Table 2: Size of successive establishments

	Log employment			*Log wage bill*		
β	-0.201	-0.197	-0.196	-0.198	-0.195	-0.194
	(0.021)	(0.020)	(0.020)	(0.020)	(0.019)	(0.020)
Fixed effects						
Establishment age	✓	✓	✓	✓	✓	✓
Firm	✓	✓	✓	✓	✓	✓
Year	✓	✓	✓	✓	✓	✓
Municipality-industry		✓	✓		✓	✓
Firm age			✓			✓
R^2	0.840	0.861	0.861	0.740	0.756	0.756
Observations	6,925,089	6,914,070	6,914,070	6,925,089	6,914,070	6,914,070

Estimates of the coefficient, β, on log establishment order from regression equation (1). Columns 1-3 use employment (number of employees) to measure establishment size and columns 4-6 use the wage bill (in 2017 SEK). Check marks indicate which fixed effects each regression includes. We cluster standard errors (in parentheses) at the firm level.

using employment and the wage bill, respectively, to measure establishment size. For both size measures, we estimate $\beta \approx -0.2$, so a 1% increase in establishment order is associated with a 0.2% decrease in establishment employment and the wage bill. As an example, this implies that a firm's third establishment is about 7.8% smaller than its second establishment, where the size of each is relative to what we expect given establishment age and year.[5] For both size measures, the estimated β is highly statistically significant.

Columns 2 and 5 in Table 2 add fixed effects for an establishment's municipality (location) crossed with its 5-digit industry. The results are nearly identical to the results from our baseline regression. This suggests that the decline in size of firms' successive establishments holds both in absolute terms as well as relative to the average size in each establishment's municipality-industry. In this sense, each successive establishment at a firm is smaller relative to its market. Finally, columns 3 and 6 add firm age fixed effects and again show the same results. As such, declining size across successive establishments is not due to firms opening later establishments when they are older.

[5]Going from the 2nd to the 3rd establishment, log establishment size falls by $0.2(\ln(3) - \ln(2)) \approx 0.081$.

2.4 Size-Dependent Markups

Previous work finds that *within industries* larger firms set higher markups.[6] We now investigate this relationship in services industries in Sweden. In particular, we estimate the relationship between a firm's markup and the two margins of its sales: its sales per establishment and its number of establishments.

We first discuss our estimation strategy. To proxy for a firm's relative markup, we use the ratio of its nominal sales to its expenditures on intermediate inputs. The idea is that if a firm chooses intermediate inputs flexibly and takes the intermediate input price as given, then their sales-to-intermediates ratio is equal to their markup divided by the elasticity of their output with respect to intermediates. We then assume that this output elasticity is the same for all firms in each industry-year. Thus, within an industry-year, the difference between two firms' log markups is equal to the difference between their log sales-to-intermediates ratios.[7]

Using our proxy for firms' relative markups, we run the following regression:

$$\ln(\underbrace{\text{sales}/\text{intermediates}_{i,j,t}}_{\text{relative markup}}) = \beta \ln(\text{size}_{i,j,t}) + \text{fixed effects}_{i,j,t} + \epsilon_{i,j,t}. \tag{2}$$

The unit of observation is a firm-year pair. On the left-hand side is our proxy for the relative markup of firm i in industry j and year t: its sales-to-intermediate ratio. On the right-hand side, the coefficient of interest, β, is on the log of some measure of firm size, which can be any subset of total sales, sales per establishment, and number of establishments. Based on our above discussion, we include industry-year fixed effects so that a firm's sales-to-intermediates ratio and size are relative to industry-year averages. Then, β is the elasticity of a firm's *relative* markup with respect to its *relative* size. Finally, $\epsilon_{i,j,t}$ is the residual.

Table 3 reports the estimation results. For our baseline regressions, we restrict attention to firm-year observations with more than one establishment so that when we use a firm's number of establishments to measure its size, the estimated coefficient β does not capture differences between single- and multi-establishment firms. In column 1, we measure a firm's size by its total sales. We find that within an industry-year, firms with higher sales tend to set higher markups. The point estimate, which is highly statistically significant, says that a firm with 1% higher sales has about a 0.06% higher markup. As an example, if firm A's

[6] For example, Edmond et al. (2023) studies size-dependent markups in US manufacturing data, Burstein et al. (2024) use French data, and Afrouzi et al. (2023) use Nielsen scanner data.

[7] For thorough discussions of this approach to estimating relative markups, see De Loecker and Warzynski (2012) and De Ridder et al. (2025).

Table 3: Size-dependent markups

	Baseline regressions				Robustness checks	
	Log markup				*Log markup*	
Log sales	0.062					
	(0.006)					
Log sales per estab.		0.087		0.087	0.186	0.047
		(0.009)		(0.009)	(0.002)	(0.008)
Log number of estabs.			0.021	0.008	-0.107	0.010
			(0.007)	(0.007)	(0.004)	(0.007)
Only multi-establishment	✓	✓	✓	✓		✓
Industry-year fixed effects	✓	✓	✓	✓	✓	✓
IV for sales per estab.						✓
R^2	0.542	0.545	0.532	0.545	0.376	0.547
Observations	93,164	93,164	93,164	93,164	3,454,250	86,272

Each column reports the results from estimating regression equation (2) using different measures of firm size. In each case, we report the estimated coefficients for the included firm size measures. Check marks indicate whether we only include multi-establishment firm-year observations or use all observations, whether we include industry-year fixed effects, and whether we instrument for a firm's sales per establishment using its previous year's value (see Appendix A for details). We cluster standard errors (in parentheses) at the firm level.

markup is 1.2 and industry competitor firm B has twice the sales of firm A, then this implies that firm B's markup is about 1.25. By comparison, using US manufacturing data, Edmond et al. (2023) estimate a coefficient of 0.031, so half as steep a relationship between a firm's relative sales and relative markup.[8]

Our main result on size-dependent markups is in columns 2-4 in Table 3. Together, they demonstrate that this positive relationship between a firm's size and its markup is driven by firms' sales per establishment, not their number of establishments. Column 2 uses a firm's sales per establishment as the size measure in regression equation (2), column 3 uses a firm's number of establishments, and column 4 uses both.[9] When we consider each size measure separately, a firm's markup has a positive statistically significant relationship with the firm's

[8] They report this result in Table C4 of the online appendix.

[9] For column 4, $\beta \ln(\text{size}_{i,j,t})$ on the right-hand side of regression equation (2) becomes $\beta_1 \ln(\text{sales per establishment}_{i,j,t}) + \beta_2 \ln(\text{number of establishments}_{i,j,t})$.

sales per establishment (column 2) and with its number of establishments (column 3). But a 1% increase in sales via sales per establishment implies about four times as big an increase in the markup as a 1% increase in sales via number of establishments. Moreover, when we include both size measures, the coefficient on sales per establishment is unaffected and still highly statistically significant, whereas the coefficient on number of establishments falls by more than half and is no longer statistically significant at the 90% level. Finally, the R^2 is higher when we use sales per establishment to measure size than when we use total sales or number of establishments, and including number of establishments along with sales per establishment does not increase the R^2.

Columns 5 and 6 in Table 3 report the results of two robustness checks that confirm our main finding. In both cases, we use both sales per establishment and number of establishments to measure firm size, so they are comparable to column 4. Column 5 uses all firm-year observations rather than just those with more than one establishment. This increases the coefficient on sales per establishment, which suggests that the relationship between a firm's sales per establishment and its markup is steeper among single-establishment firms than among multi-establishment firms. As a consequence, the regression model predicts too high markups for multi-establishment firms, so the coefficient on number of establishments becomes negative to compensate. Next, column 6 instruments for a firm's sales per establishment using its previous year's value (the first stage results are in Appendix A). This accounts for the possibility that noise in sales mechanically raises the markup and sales per establishment, leading to a positive estimated relationship. The coefficient on sales per establishment falls but remains positive and highly statistically significant. This suggests that there is some noise in sales driving the high coefficient on sales per establishment in our baseline regression, but this noise does not fully account for the positive relationship between a firm's sales per establishment and its markup.

3 Model and Qualitative Results

We now develop the model, discuss its solution, and characterize inefficient distortions in the competitive equilibrium.

3.1 Model

There is a single period. A representative household consumes the numeraire final good, inelastically supplies labor (the only input) in a perfectly competitive market, and owns all firms. There is a unit mass of firms. Each firm is characterized by a baseline quality, which is the quality of its initial unit measure of establishments. A firm then chooses how many further establishments to open subject to a cost function in labor, where each successive establishment has lower quality. Through each of its establishments, a firm uses labor to produce a differentiated variety, where the establishment's quality serves as a demand shifter. Perfectly competitive final good producers aggregate all the establishment varieties into the final good, which they sell to the household.

Representative household. The representative household maximizes final good consumption C subject to its budget constraint:

$$C \leq W\bar{L} + \Pi,$$

where W is the wage, $\bar{L}$ is the inelastic labor supply, Π are profits from firms, and the final good price is normalized to 1. We take final good consumption C as our measure of welfare.

Final good producers and demand. Perfectly competitive final good producers aggregate varieties from firms' establishments into final good output Y. They sell the final good to the representative household, so their output must equal final good consumption:

$$C = Y.$$

Specifically, final good producers purchase varieties from a double continuum of establishments: each firm $i \in [0, 1]$ sells a different variety at each of its establishments $n \in [0, N(i)]$. If final good producers purchase $y(i, n)$ from firm i's establishment n, then final good output Y is given implicitly by the Kimball (1995) aggregator:

$$\int_0^1 \int_0^{N(i)} \varphi(i,n)\Upsilon(q(i,n))dndi = 1 \qquad q(i,n) \equiv \frac{y(i,n)}{Y}, \tag{3}$$

where $\varphi(i, n)$ is the quality of firm i's establishment n, $q(i, n)$ is the relative real output of firm i's establishment n, and $\Upsilon(\cdot)$ is twice continuously differentiable, strictly increasing,

and strictly concave. That is, given variety purchases $y(\cdot,\cdot)$, final good output Y is such that (3) holds. The aggregation technology is constant returns-to-scale because multiplying Y and each $y(i,n)$ by the same positive constant leaves (3) unchanged. A special case of the aggregator is CES (constant elasticity of substitution), where $\Upsilon(q) = q^{\frac{\bar{\sigma}-1}{\bar{\sigma}}}$ for some $\bar{\sigma} > 1$.

Final good producers choose demand $y(\cdot,\cdot)$ and final good output Y to maximize profits

$$Y - \int_0^1 \int_0^{N(i)} p(i,n)y(i,n)dndi$$

subject to the aggregation technology (3), where $p(i,n)$ is the price at firm i's establishment n. The resulting relative demand for each variety, $q(i,n) \equiv y(i,n)/Y$, is given by[10]

$$p(i,n) = \varphi(i,n)\Upsilon'(q(i,n))D \qquad D \equiv \left(\int_0^1 \int_0^{N(i)} \varphi(i,n)q(i,n)\Upsilon'(q(i,n))dndi\right)^{-1}, \tag{4}$$

where $D > 0$ is a demand index. Given an establishment's relative output $q(i,n)$, its price $p(i,n)$ moves one-for-one with its quality $\varphi(i,n)$. Given quality, an establishment's price is falling in its relative output because $\Upsilon(\cdot)$ is strictly concave. Finally, an establishment's price does not depend on output at its firm's other establishments because it only depends on other establishments through *economy-wide* aggregates, D and Y.

Firms, establishments, and production. Each firm $i \in [0,1]$ chooses a measure $N(i) \geq 1$ establishments to open subject to labor cost

$$\frac{\kappa}{\theta+1}\left(N(i)^{\theta+1} - 1\right),$$

where $\kappa > 0$ scales the cost and $\theta \in \mathbb{R}$ determines the curvature. The first unit measure of establishments are free, i.e., the cost of setting $N(i) = 1$ is zero. This captures in our continuous setting that a firm has an initial establishment and decides how many more to open. Firm i's marginal cost of opening further establishments is $\kappa N(i)^\theta$. This can be increasing or decreasing in firm i's measure of establishments $N(i)$ because θ can be above

[10]The first order condition for $y(i,n)$ is $p(i,n) \geq (\lambda/Y)\varphi(i,n)\Upsilon'(y(i,n)/Y)$, where λ is the Lagrange multiplier on the aggregation constraint and where the inequality holds with equality if $y(i,n) > 0$. For $y(i,n) = 0$, the inequality can be strict if $\Upsilon'(0) < \infty$ (unlike with CES) because then there is a finite price threshold above which demand is 0. Without loss of generality, we suppose the price is weakly below this threshold. Multiply both sides of the first order condition by $y(i,n)$, integrate across establishments, and plug in that final good producer profits are zero (implied by perfect competition) to get $Y = \lambda/D$. Plug back into the first order condition to get (4).

or below 0. We will impose a lower bound on θ so that firm problems are well-behaved.

At each establishment $n \in [0, N(i)]$, firm i produces a different variety. We use the terms establishment and variety interchangeably. At all establishments, productivity is the same: output of the firm i establishment n variety is the labor used for producing it, $l(i,n)$. Hence the marginal cost of production is the wage W.

Establishments (varieties) differ in quality. Firm i's establishment n has quality

$$\varphi(i,n) = \bar{\varphi}(i)\max\{n,1\}^{-\rho}. \tag{5}$$

The first unit measure of establishments $n \in [0,1]$ have quality $\varphi(i,n) = \bar{\varphi}(i)$, which we call firm i's baseline quality, or sometimes just firm i's quality. Then quality declines at each successive establishment with elasticity $\rho > 0$. So firm i's highest quality establishments are its first unit measure, $n \in [0,1]$, and its lowest quality establishment is its last, $n = N(i)$.[11] We interpret this quality decline across successive establishments as capturing that a firm sees all potential establishment opportunities and picks the best ones. Thus a firm's marginal or last establishment is lower quality than its other establishments.

Firm i chooses its measure $N(i)$ of establishments as well as the price $p(i,n)$, labor $l(i,n)$, and output $y(i,n)$ for each variety $n \in [0, N(i)]$ to maximize profits:

$$\int_0^{N(i)} [p(i,n)y(i,n) - Wl(i,n)]dn - \frac{W\kappa}{\theta+1}\left(N(i)^{\theta+1} - 1\right)$$

subject to production, $y(i,n) = l(i,n)$, and demand (4) for all establishments $n \in [0, N(i)]$. The integral in firm i's profits is the production profits from its measure $N(i)$ establishments. The remaining term is the cost of opening establishments beyond the initial unit measure.

Establishment size and markups. To solve a firm's profit maximization problem, we first characterize its optimal production and pricing decisions at each establishment. Decisions are independent across a firm's establishments because a firm is small relative to the economy and establishments are only related through economy-wide aggregates. Hence, if firm i's establishment n has strictly positive sales, then its markup of price over marginal cost,

[11] We can allow for each successive establishment to be *higher* quality ($\rho < 0$) but when we calibrate the model, this would contradict our empirical finding that each successive establishment at a firm tends to be smaller. Moreover, it does not admit such a clear interpretation.

$\mu(i,n) \equiv p(i,n)/W$, must satisfy the usual expression:

$$\mu(i,n) = \frac{\sigma(q(i,n))}{\sigma(q(i,n)) - 1} \qquad \sigma(q(i,n)) \equiv \frac{-\Upsilon'(q(i,n))}{q(i,n)\Upsilon''(q(i,n))} = \frac{-p(i,n)}{y(i,n)} \frac{\partial y(i,n)}{\partial p(i,n)}, \quad (6)$$

where $\sigma(q(i,n))$ is the demand elasticity (of quantity with respect to price) at firm i's establishment n, which depends on the establishment's relative output $q(i,n) \equiv y(i,n)/Y$. To finish solving the optimization problem at each establishment, we make the following assumption that the demand elasticity is sufficiently high and is weakly decreasing in an establishment's size.

Assumption 1. *For all q, $\sigma(q) \equiv \frac{-\Upsilon'(q)}{q\Upsilon''(q)}$ is weakly decreasing. Moreover, there exists a $\bar{q} > 0$ such that $\sigma(q) > 1$ for all $q \leq \bar{q}$.*

Assumption 1 implies that given an establishment's quality $\varphi(i,n) = \bar{\varphi}(i)n^{-\rho}$, there is a unique relative output and price pair that satisfies expression (6) and demand curve (4).[12] That is, $\bar{\varphi}(i)$ and $n^{-\rho}$ do not matter separately. Furthermore, an establishment's sales, price, markup, and profits are weakly increasing in its quality. The restriction that larger establishments face a weakly lower demand elasticity—and so set weakly higher markups—matches our empirical results on size-dependent markups.

Measure of establishments. We next characterize each firm's optimal measure of establishments. Let $\pi(\varphi)$ be the profits of an establishment with quality φ. Then, firm i's benefit of opening its marginal establishment (increasing its measure $N(i)$ of establishments) is the profits it will earn from that establishment, $\pi(\bar{\varphi}(i)N(i)^{-\rho})$. The cost of opening the marginal establishment is $W\kappa N(i)^{\theta}$, which is the wage times the marginal establishment opening cost in labor. To ensure the benefit and cost have at most one crossing point, we make the following assumption that the establishment opening cost function is not too concave.

Assumption 2. *$\theta > -\rho$, where θ and ρ are the elasticities of a firm's marginal establishment opening cost and marginal establishment quality with respect to its measure of establishments.*

[12] The second part of Assumption 1 is necessary because expression (6) implies that all establishments with strictly positive relative output q set q such that they face a demand elasticity $\sigma(q)$ strictly above 1. As such, if the second statement is violated, then the aggregator constraint (3) cannot be satisfied. We suppose $\bar{q}$ is sufficiently high so that there is an equilibrium.

Assumption 2 implies that the higher a firm's measure of establishments, the lower the profits at its marginal establishment relative to the cost of opening that marginal establishment.[13] Moreover, this profit-to-cost ratio goes to zero as a firm's measure of establishments goes to infinity. As such, each firm has a unique and finite optimal measure of establishments. In particular, firm i does not open any establishments beyond its initial unit measure if its baseline quality is sufficiently low; specifically, if $\bar{\varphi}(i) \leq \varphi^*$ where $\pi(\varphi^*) = W\kappa$. In this case, we say that firm i is a "single-establishment" firm. On the other hand, if $\bar{\varphi}(i) > \varphi^*$, then firm i is a "multi-establishment" firm and chooses the unique $N(i) > 1$ that satisfies

$$\pi(\bar{\varphi}(i)N(i)^{-\rho}) = W\kappa N(i)^{\theta}. \tag{7}$$

This optimal $N(i)$ is uniquely given by, and increasing in, firm i's baseline quality $\bar{\varphi}(i)$.

Firm size and markups. To compute firm-level outcomes, we aggregate across a firm's establishments. Firm i's sales are $S(i) = \int_0^{N(i)} p(i,n)y(i,n)dn$ and its production labor is $L(i) = \int_0^{N(i)} l(i,n)dn$. We define a firm's markup as its sales over production costs in order to align with our definition of an establishment's markup, which is the same ratio. Hence, a firm's markup is the cost-weighted average of its establishment markups:

$$\mu(i) \equiv \frac{S(i)}{WL(i)} = \int_0^{N(i)} \frac{l(i,n)}{L(i)}\mu(i,n)dn.$$

Aggregation. The economy aggregates so that final good output is

$$Y = ZL_p,$$

where $L_p = \int_0^1 L(i)di$ is labor used for production (rather than for opening establishments) and Z is aggregate productivity:[14]

$$Z = \left(\int_0^1 \int_0^{N(i)} q(i,n)dndi\right)^{-1}. \tag{8}$$

Hence, aggregate productivity depends on each firm's measure of establishments and its rel-

[13] Note that the elasticity of establishment profits $\pi(\varphi)$ with respect to quality φ is greater than 1.

[14] Write $Z^{-1} = L_p/Y = \int_0^1 \int_0^{N(i)} (l(i,n)/Y)dndi$. Using $y(i,n) = l(i,n)$ and $q(i,n) = y(i,n)/Y$ yields (8).

ative output at each establishment. The distribution of establishment qualities then affects aggregate productivity through relative outputs. For example, if all establishment qualities double and establishment purchases $y(\cdot,\cdot)$ remain constant, then final good output Y must rise until the Kimball aggregation constraint (3) is again satisfied. As a result, relative output $q(i,n)$ is lower at all establishments, so aggregate productivity Z is higher.

Equilibrium. In equilibrium, the representative household maximizes consumption C subject to its budget constraint, taking as given wage income $W\bar{L}$ and firm profits Π. Each firm i chooses their measure of establishments $N(i)$ and the price $p(i,n)$, labor $l(i,n)$, and output $y(i,n)$ at each establishment to solve their profit maximization problem, taking as given final good output Y, the wage W, and the demand index D. Final good producers choose demand $y(i,n)$ for each establishment, and final good output Y to solve their profit maximization problem, taking as given the price $p(i,n)$ at each establishment. Perfect competition implies that final good producer profits are zero.

Household consumption must equal final good output: $C = Y$. Final good output Y and demand for each variety $y(\cdot,\cdot)$ must satisfy aggregation (3). Demand for each variety must equal output, which must equal the labor used to produce it: $y(i,n) = l(i,n)$. The labor used by firms for production and opening establishments must equal the household's inelastic labor supply: $\int_0^1 \left(L(i) + \frac{\kappa}{\theta+1}(N(i)^{\theta+1} - 1)\right) di = \bar{L}$.

We can find an equilibrium as follows. Given final good output Y and the demand index relative to the wage D/W, use the demand curve (4) and the optimal markup expression (6) to get establishment relative output and profits relative to the wage as functions of establishment quality. The latter implies each firm's measure of establishments $N(i)$ using (7). Then check whether the Kimball aggregator constraint holds and the labor market clears. Hence, there are two aggregate variables (Y and D/W) that must satisfy two equations (Kimball aggregator constraint and labor market clearing). We can then compute the remaining equilibrium variables.

3.2 Inefficient Distortions in the Competitive Equilibrium

We now provide a general characterization of the inefficient distortions to firms' decisions in a competitive equilibrium. We later calibrate the model to compute the costs of these distortions and to assess the usefulness of firm size-dependent policy for reducing these costs. The competitive equilibrium can be inefficient due to misallocation of the fixed labor supply

across its various uses—opening establishments at each firm and producing at each establishment. On the margin, we can raise welfare (household consumption) by shifting labor from uses with a low marginal social value to uses with a high marginal social value. Hence, to state our result, we define $V_{prod}(i,n)$ to be the marginal social value of labor used to produce at firm i's establishment n and define $V_{estab}(i)$ to be the marginal social value of labor used to open and produce at firm i's marginal establishment. Formally, suppose a social planner has an infinitesimal quantity Δ of labor beyond the inelastic labor supply $\bar{L}$. Hold fixed the allocation of the $\bar{L}$ labor across its various uses from the competitive equilibrium. If the planner uses the additional Δ labor to increase production at firm i's establishment n, then household consumption increases by $V_{prod}(i,n)\Delta$. If the planner uses the additional labor to open establishments at firm i and produce the competitive equilibrium output, $y(i,N(i))$, at those establishments, then household consumption increases by $V_{estab}(i)\Delta$.[15] We can now state our result.

Theorem 1. *If the demand elasticity $\sigma(q)$ is constant in relative output q, then the competitive equilibrium is efficient.*

If the demand elasticity $\sigma(q)$ is always strictly decreasing in relative output q, then the competitive equilibrium is inefficient and the following are true:

1. **Misallocation of production.** *$V_{prod}(i,n) > V_{prod}(j,m)$ if firm i's establishment n is strictly larger than firm j's establishment m, i.e., $q(i,n) > q(j,m)$;*

2. **Misallocation of establishments.** *$V_{estab}(i) > V_{estab}(j)$ if firm i's marginal establishment is strictly larger than firm j's marginal establishment, i.e., $q(i,N(i)) > q(j,N(j))$, where firm i and firm j are both single-establishment or both multi-establishment;*

3. **Production vs. establishment opening.** *$V_{prod}(i,n) > V_{estab}(j)$ if firm i's establishment n is weakly larger than firm j's marginal establishment, i.e., $q(i,n) \geq q(j,N(j))$.*

Proof. See Appendix C.1. ■

The theorem first states that if the demand elasticity is constant (CES demand), then we cannot improve welfare relative to the competitive equilibrium by reallocating labor. Second,

[15] This means the planner opens $\frac{\Delta}{\kappa N(i)^{\theta} + y(i,N(i))}$ establishments because each establishment uses $\kappa N(i)^{\theta}$ in establishment-opening labor and $y(i,N(i))$ in production labor.

if the demand elasticity is falling in an establishment's relative output—as in our calibrated model—then on the margin, we can improve household consumption by 1) reallocating production from small establishments to large establishments, 2) reallocating establishments from firms with small marginal establishments to firms with large marginal establishments, and 3) closing marginal establishments in exchange for producing more at other weakly larger establishments. To understand the strength of 3), recall that a firm's marginal establishment is its smallest establishment. As such, we can improve household consumption by closing establishments at a firm in exchange for producing more *at any of the firm's other establishments* or at many smaller firms' establishments. Indeed, 3) is not even the strongest possible statement: by continuity, $V_{prod}(i,n) > V_{estab}(j)$ for some establishments that are strictly smaller than firm j's marginal establishment, i.e., with $q(i,n) < q(j,N(j))$. We later illustrate Theorem 1 in our calibrated model in Figure 5.

Intuitively, the allocation of labor in the competitive equilibrium is determined by equalizing the marginal private value (the effect on a firm's profits) of the various uses of labor. So uses with a high marginal social value are those that are more undervalued by firms. So the marginal social value of using labor to produce at a firm's establishment is high if the firm more deeply undervalues the establishment's *marginal* output. On the other hand, the marginal social value of using labor to open and produce at a firm's marginal establishment is high if the firm more deeply undervalues the establishment's *total* output.

Now, firms undervalue output at an establishment—relative to the social value—because they face a finite demand elasticity: to sell an additional unit at an establishment, a firm must cut the price on all the establishment's other units. The lower an establishment's demand elasticity, the more the firm must cut its price, and so the more it undervalues additional output. It follows that if the demand elasticity is constant in an establishment's relative output, then firms undervalue marginal output and total output at all establishments by the same amount. As such, the competitive equilibrium is efficient.

With a demand elasticity that is falling in an establishment's relative output (as in our calibrated model), firms more deeply undervalue each successive unit of output at an establishment. One consequence is that the larger an establishment, the more firms undervalue its marginal output and its total output. Thus, we can improve welfare by reallocating production from small establishments to large establishments and by reallocating establishments from firms with small marginal establishments to firms with large marginal establishments. Another consequence is that firms undervalue an establishment's marginal output more than they undervalue its total output. Thus, we can improve welfare by closing marginal establish-

ments in exchange for producing more at larger (or even somewhat smaller) establishments.

4 Quantifying the Model

We now calibrate the model to match our data on services industries in the Swedish economy from 1997-2017. We then test the model using untargeted data moments and describe key features of the equilibrium. In the following sections, we use the calibrated model to study misallocation and policy.

4.1 Calibration

Kimball Aggregator Function. For the aggregator function $\Upsilon(\cdot)$, we use the Klenow and Willis (2016) specification of the *derivative*:

$$\Upsilon'(x) = \frac{\bar{\sigma}-1}{\bar{\sigma}} \exp\left(\frac{1-x^{\frac{\epsilon}{\bar{\sigma}}}}{\epsilon}\right),$$

where $\bar{\sigma}$ and ϵ are parameters. It follows that the demand elasticity (of quantity with respect to price) and markup at firm i's establishment n are:

$$\sigma(q(i,n)) = \bar{\sigma} q(i,n)^{\frac{-\epsilon}{\bar{\sigma}}} \qquad\qquad \mu(i,n) = \frac{\bar{\sigma} q(i,n)^{\frac{-\epsilon}{\bar{\sigma}}}}{\bar{\sigma} q(i,n)^{\frac{-\epsilon}{\bar{\sigma}}} - 1}.$$

Hence, $\bar{\sigma}$ is the demand elasticity at relative output 1 and $|\epsilon/\bar{\sigma}|$ is the super elasticity—the elasticity of the demand elasticity with respect to quantity. If $\epsilon = 0$, then we have CES demand, so the demand elasticity and markup are constant in an establishment's relative output. If $\epsilon > 0$, then the demand elasticity is falling in an establishment's relative output, so larger establishments set higher markups. If $\epsilon < 0$, then larger establishments face a higher demand elasticity and set lower markups. Thus, $\epsilon \geq 0$ satisfies Assumption 1 and matches our empirical finding that firms with higher sales per establishment set higher markups.

Our aggregator function differs from Klenow and Willis (2016) in that we pin down $\Upsilon(\cdot)$ by setting $\Upsilon(0) = 0$, whereas they set $\Upsilon(1) = 1$. This distinction is important because if $\epsilon > 0$, then there is a strictly positive quality cutoff below which an establishment has zero sales in equilibrium. If $\Upsilon(0) \neq 0$, which is generically the case with $\Upsilon(1) = 1$, then these non-producing establishments affect economic outcomes. For example, if $\Upsilon(0) > 0$, then opening non-producing establishments raises aggregate productivity. This is particularly relevant

when we turn to policy and can incentivize firms to open many low quality establishments. To avoid this oddity, we set $\Upsilon(0) = 0$. The resulting aggregator function is

$$\Upsilon(q) = \frac{\bar{\sigma} - 1}{\epsilon} \int_0^{q^{\epsilon/\bar{\sigma}}} \tilde{q}^{\frac{\bar{\sigma}}{\epsilon}-1} \exp\left(\frac{1-\tilde{q}}{\epsilon}\right) d\tilde{q}.$$

Parameters. We set the inelastic labor supply to $\bar{L} = 0.987$ so that final good output is $Y = 1$. This leaves six parameters to calibrate: 1) κ, the establishment opening cost shifter; 2) θ, the elasticity of the establishment opening cost; 3) ω, the Pareto tail parameter for the baseline quality distribution; 4) $\bar{\sigma}$, the demand elasticity at relative output of 1; 5) $\epsilon/\bar{\sigma}$, the demand super elasticity; and 6) ρ, the rate at which establishment quality falls across a firm's successive establishments.

Table 4: Calibration Targets and Calibrated Parameter Values

Moment	**Data**	**Model**
fraction of firms with multiple establishments	0.027	0.027
average number of establishments across all firms	1.17	1.17
sales share of multi-establishment firms	0.49	0.49
log(markup) on log(sales per establishment)	0.087	0.087
wage bill decline across successive establishments	0.20	0.20
sales-weighted average markup	1.15	1.15

Parameter	**Value**
κ (establishment opening cost shifter)	0.192
θ (establishment opening cost elasticity)	0.076
ω (baseline quality tail parameter)	16.43
$\bar{\sigma}$ (demand elasticity at relative output of 1)	8.356
$\epsilon/\bar{\sigma}$ (demand super elasticity)	0.507
ρ (establishment quality elasticity)	0.059

To calibrate these parameters, we exactly match five moments in our Swedish data on services industries, plus an estimate of the sales-weighted average markup in Sweden from Sandström (2020). The five moments are 1) the share of firms that are multi-establishment; 2) the average number of establishments across all firms; 3) the share of sales earned by

multi-establishment firms; 4) the estimated elasticity of the markup with respect to sales per establishment among multi-establishment firms, controlling for industry-year fixed effects (column 2 of Table 3); and 5) the rate at which the log wage bill or log employment falls across a firm's successive establishments (Table 2).[16] For our last data moment, Sandström (2020) estimates the economy-wide sales-weighted average markup in Sweden from 1997-2017 (the same time period as our data) to be about 1.15.[17] Table 4 lists the data moments and calibrated parameter values.

All six moments jointly determine all six parameters. Nonetheless, there is an intuitive mapping between particular moments and parameters. The demand parameters $\bar{\sigma}$ and ϵ determine the relationship between the markup and relative output at each establishment, so they map into the elasticity of the markup with respect to sales per establishment and the average markup. The rate at which establishment quality declines across a firm's successive establishments, ρ, is mostly pinned down by the rate at which the wage bill declines with each successive establishment. The establishment opening cost shifter κ and the tail parameter ω of the firm baseline quality distribution map into the fraction of firms with multiple establishments and the sales share of multi-establishment firms. Specifically, a higher κ means firms open fewer establishments, so there are fewer multi-establishment firms and each multi-establishment firm has lower sales. A higher ω pushes these moments in the same direction, but particularly affects the highest quality firms, so it has a stronger effect on the sales share of multi-establishment firms relative to the share of firms with multiple establishments. Finally, the elasticity θ of the marginal establishment opening cost maps into the average size of multi-establishment firms' establishments, which is determined by the average number of establishments and the sales share of multi-establishment firms. Specifically, a higher θ means multi-establishment firms open fewer establishments, which raises the average size of their establishments.

4.2 Untargeted Moments and Model Fit

We now show the model does a good job of matching the relationship between sales per establishment and number of establishments among multi-establishment firms, as well as

[16] We run regression (1) on our model data, controlling for firm fixed effects. To create establishment-level data, we generate an establishment for each unit measure of establishments a firm controls; that is, if firm i has a measure of establishments $N(i) = 4.7$, then we take their first unit measure of establishments ($n \in [0, 1]$) to be one establishment, their second unit measure ($n \in [1, 2]$) to be another, up until ($n \in [3, 4]$).

[17] Sandström (2020) does not report the *cost-weighted* average markup, so we use the sales-weighted average. In our calibrated model, both weightings yield an average markup of 1.15.

the tails of the distributions of firm sales and number of establishments. This is the case even though we only target the share of firms that are multi-establishment, the sales share of multi-establishment firms, and the average number of establishments. That said, the model does not perfectly match the data. We argue that splitting firm baseline quality variation into a component known before firms choose their measure of establishments and a component known after would likely improve the model fit along various dimensions.

First, we compare the relationship between a firm's sales per establishment and its number of establishments in the model and in the data. This is important because it determines the extent to which large firms have large establishments vs. many establishments. As we saw in Theorem 1, variation in establishment size underlies the distortions in our model. To measure this relationship in our model and in the data, we regress a firm's log sales per establishment on its log number of establishments.[18] In the model, we get a coefficient of 0.131 and in the data, we get 0.143, which is highly statistically significant.[19]

Table 5: Concentration in Largest x% of Firms

	Top 5%	**Top 1%**	**Top 0.1%**
Sales share			
Model	52%	45%	35%
Data	77%	58%	33%
Establishments share			
Model	18.8%	14.6%	10.5%
Data (sales-based)	16.6%	10.1%	4.1%
Data (establishment-based)	18.7%	13.3%	7.2%

The share of total sales and establishments at the largest x% of firms in the model and in the data. We compute the data share in each year and average across years. The first two data rows use sales to categorize the largest x% of firms; the last data row uses number of establishments. These two methods of categorization are equivalent in the model.

Second, Table 5 compares the concentration of firm sales and number of establishments in our model and in the data. For sales, we compute the share of total sales at the largest x% of

[18] In the data, a unit of observation is a firm-year pair, and we control for industry-year fixed effects. In both the model and data, we restrict attention to observations with more than one establishment. We regress log sales per establishment on log number of establishments rather than the other way around because we expect that in the data, sales per establishment is noisier.

[19] Standard error is 0.014 (clustered at the firm level), R^2 is 0.374, and number of observations is 93,433.

firms for various values of x. In the data, we compute these shares in each year and average across years. The model shares undershoot the data shares for the largest 5% and 1% of firms because it misses the presence of large single-establishment firms. Specifically, firms in the model perfectly know their quality before choosing their measure of establishments, so all single-establishment firms are smaller than all multi-establishment firms. Since we target the sales share of multi-establishment firms in the data, we therefore underestimate the sales share of the largest single-establishment firms.

For number of establishments, we compute the share of total establishments at the largest x% of firms for various values of x. In the data, we compute these shares ranking firms based on their sales and based on their number of establishments (the two rankings are equivalent in the model). The model shares overshoot the data shares, more when ranking firms based on sales and more the higher up the distribution we go. These again reflect that in the model, firms perfectly know their quality before choosing their measure of establishments, so there is a tight relationship between a firm's sales and its number of establishments. If some variation in firm quality occurs after firms choose their measure of establishments, then the highest quality and largest firms would have fewer establishments.

4.3 Initial Equilibrium

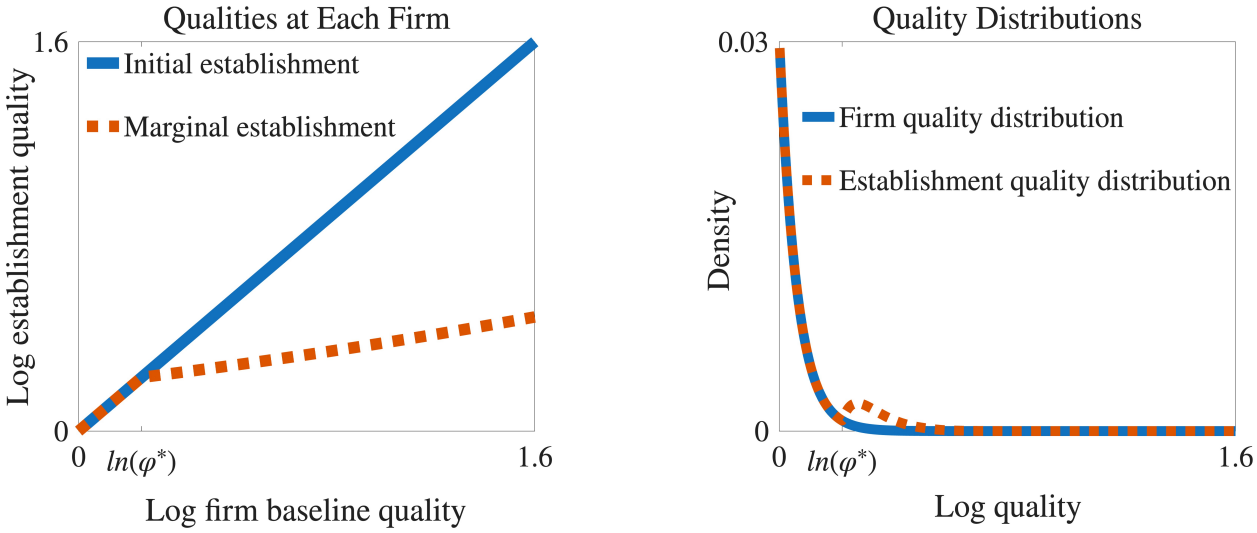
We now illustrate features of the competitive equilibrium of our calibrated model. First, 5.7% of labor is used for establishment opening costs, leaving 94.3% for production. Firms with more establishments face a higher marginal cost of opening establishments because the elasticity of the marginal establishment opening cost is strictly positive ($\theta > 0$). For example, a firm with more sales than 99.9% (99.99%) of firms has 18 (134) establishments and faces a 25% (45%) higher marginal establishment opening cost than single-establishment firms.

Next, Figure 3 displays the set of open establishments. The left panel shows the interval of establishment qualities at each firm. A firm operates establishments with quality between the two lines, where its initial unit measure of establishments ($n \in [0, 1]$) are its highest quality establishments, and its marginal establishment ($n = N(i)$) is its lowest quality establishment. Firms with baseline quality above the cutoff φ^* are multi-establishment ($N(i) > 1$), so there is a gap between the two lines. Higher quality firms have higher quality marginal establishments because they have more establishments, so they face higher marginal establishment opening costs (since $\theta > 0$), which means they require higher profits at their marginal establishments. The right panel depicts the resulting distribution of establishment qualities. The density

of firm baseline quality, which is exogenous, is also the density of establishment quality across firms' initial unit measures of establishments. The density of establishment quality is endogenous due to firms' establishment opening decisions. The gap between the two densities shows the set of additional establishments firms open beyond their initial unit measures.

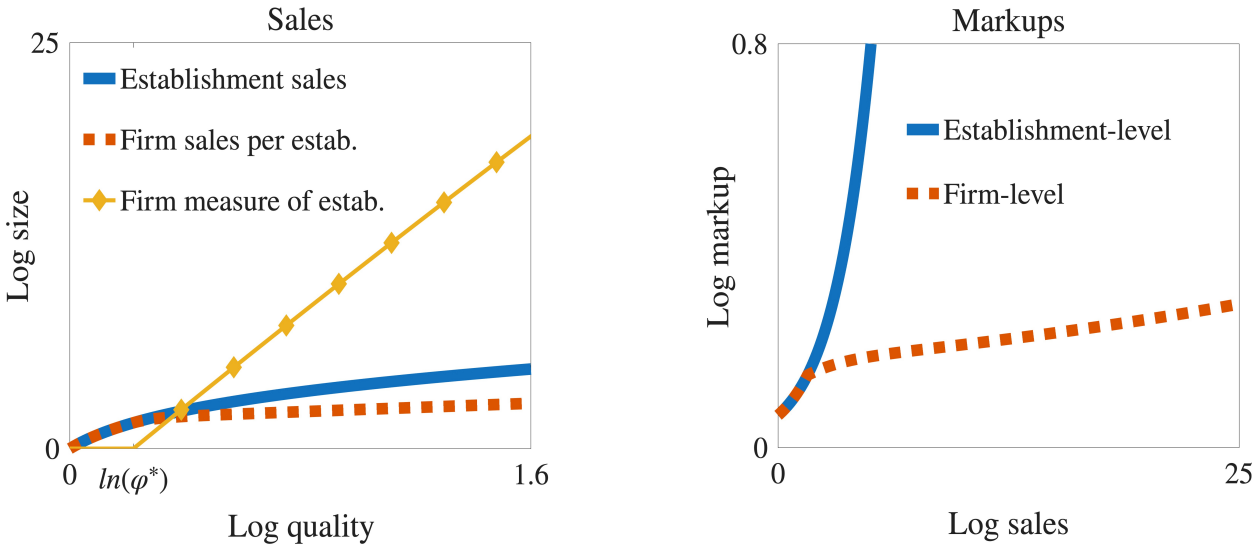
Finally, Figure 4 illustrates sales and markups across firms and their establishments. The left panel shows that higher quality establishments sell more. Then, higher quality firms sell more per establishment because they have higher quality establishments. They also have more establishments, which brings down the average quality of their establishments. As a result, sales per establishment becomes relative flat among multi-establishment firms (with baseline quality above φ^*). This means that variation in sales among larger firms is increasingly driven by variation in measure of establishments rather than in sales per establishment, as in the data (Section 2.2). The right panel shows that larger establishments and larger firms set higher markups. The relationship between firm sales and firm markups becomes much flatter than the relationship between establishment sales and establishment markups because large firms' sales are mostly driven by their measure of establishments, which do not directly affect their markups.

Figure 3: The Set of Establishments



Left panel: log quality at a firm's initial unit measure of establishments and at its marginal establishment, as functions of its log baseline quality. Right panel: the density of the distribution of firm baseline quality and of establishment quality; the former integrates to the measure of firms, which is 1, and the latter integrates to the measure of establishments, which is greater than 1. In both panels, φ^* is the firm baseline quality cutoff above which firms are multi-establishment.

Figure 4: Sales and Markups



Left panel: log establishment sales as a function of log establishment quality and log firm sales per establishment and measure of establishments as functions of log firm baseline quality; φ^* is the firm baseline quality cutoff above which firms are multi-establishment. Right panel: log establishment markup as a function of log establishment sales and log firm markup as a function of log firm sales. In both panels, sales are normalized to 1 for establishments with the minimum quality of 1.

5 Quantitative Results

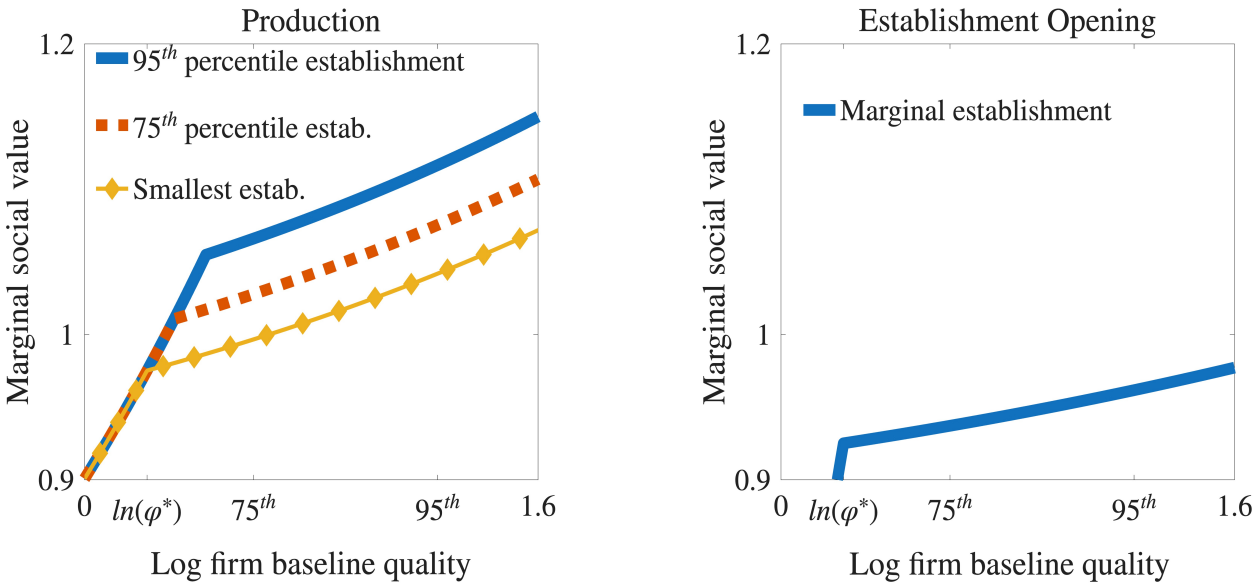
5.1 Misallocation

We now solve a planner problem to compute the costs of inefficient distortions in the competitive equilibrium of our calibrated model. The planner chooses each firm's measure of establishments and how much to produce at each establishment in order to maximize consumption (equivalently, final good output) subject to the Kimball aggregation technology (3), firms' establishment opening and production technologies, and the inelastic labor supply $\bar{L}$. Therefore, the planner can improve on the competitive equilibrium by reallocating the fixed labor supply across its various uses.

To build intuition, Figure 5 illustrates the planner's marginal incentives in the competitive equilibrium, characterized qualitatively in Theorem 1. First, the left panel depicts $V_{prod}(i,n)$—the marginal social value of production at firm i's establishment n—at each firm's 95^{th} percentile largest establishment, 75^{th} percentile largest establishment, and smallest establishment. As expected from Theorem 1, these lines are increasing and the first line

sits above the second, which sits above the third. This means the planner wants to reallocate production from small firms' establishments to large firms' establishments and wants to reallocate production within firms from smaller establishments to larger establishments. The dispersion in marginal social values is bigger across firms (moving along the x-axis) than within firms (moving along the y-axis), so the marginal gains from reallocating production are bigger across firms than within firms.

Figure 5: Marginal Social Values in the Competitive Equilibrium



Left panel: $V_{prod}(i,n)$ at firm i's 95^{th} percentile ($n = 0.95N(i)$), 75^{th} percentile ($n = 0.75N(i)$), and smallest ($n = N(i)$) establishments as functions of firm i's log baseline quality. Right panel: $V_{estab}(i)$ as a function of firm i's log baseline quality. $V_{prod}(i,n)$ and $V_{estab}(i)$ are normalized so that the production weighted average of $V_{prod}(i,n)$ across all establishments is 1. In both panels, φ^* is the baseline quality cutoff above which firms are multi-establishment; m^{th} on the x-axis denote the log baseline quality below which are firms that earn m% of all sales in the competitive equilibrium.

Next, the right panel of Figure 5 depicts $V_{estab}(i)$—the marginal social value of opening and producing the competitive equilibrium quantity at firm i's marginal establishment—at each firm. As expected from Theorem 1, this line is increasing, so the planner wants to reallocate establishments from small multi-establishment firms (with small marginal establishments) to large multi-establishment firms (with large marginal establishments). Moreover, this line sits well below the lines in the left panel because marginal establishments are small and the marginal value of opening an establishment is lower than the marginal value of producing at any larger establishment. Indeed, we can see it is considerably lower. Therefore,

the planner wants to close establishments at firms in exchange for producing at even much smaller firms' establishments.

First best allocation. We solve the planner's problem, which yields the first best allocation of the measure of establishments at each firm and production at each establishment. Going from the competitive equilibrium to the first best increases aggregate consumption by 1.47%. To understand the quantitative significance of this number, it is useful to compare to Edmond et al. (2023), who study a similar model with size-dependent markups but with only single-establishment firms. In their calibration with quality heterogeneity and the same aggregate markup,[20] they find similar static losses from misallocation, but these effects are then amplified more than three-fold because output is used as an intermediate input and to build capital over time.[21] Thus, we understate the welfare cost of distortions because we exclude these amplifying macroeconomic forces.

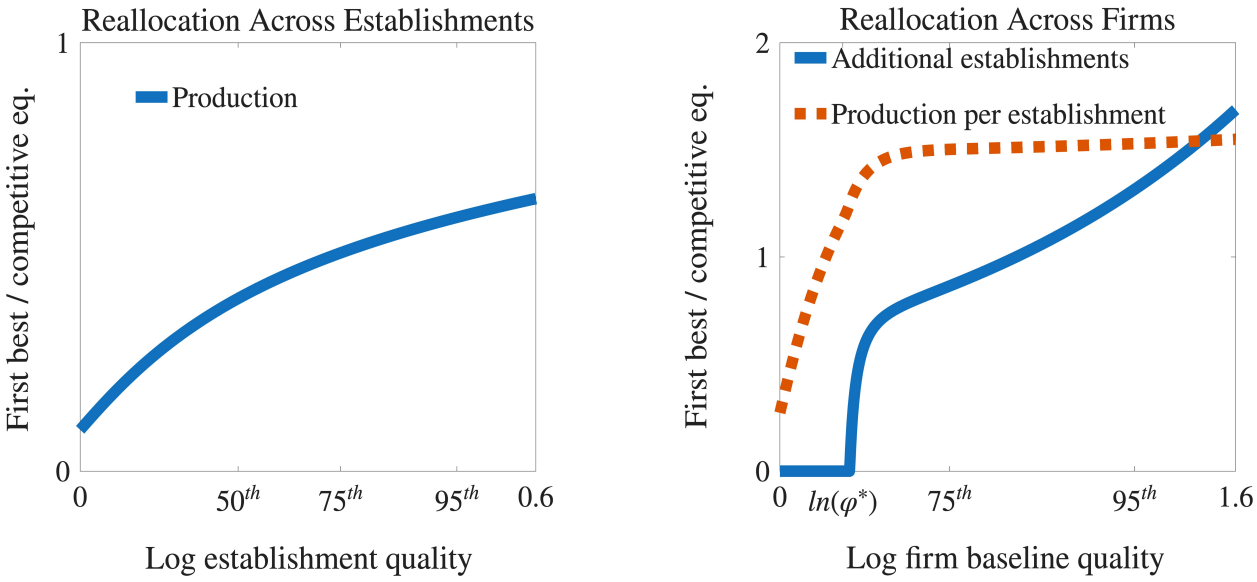
The 1.47% increase in consumption comes from a 1.45% increase in aggregate productivity and a shift of 0.02% of the labor supply from opening establishments to production. Figure 6 illustrates how the planner reallocates labor to generate these gains, which is as expected given marginal incentives in the competitive equilibrium (see Figure 5). The left panel of Figure 6 depicts each establishment's production (equivalently, labor) in the first best relative to the competitive equilibrium. The line is increasing, which indicates a reallocation of production from small (low quality) establishments to large (high quality) establishments. The line sits below 1 because the planner opens enough high quality establishments that production at each establishment falls.

The right panel of Figure 6 depicts each firm's additional establishments (beyond its initial unit measure) and production per establishment in the first best relative to the competitive equilibrium. For additional establishments, the line is 0 for single-establishment and small multi-establishment firms, then is increasing and ultimately passes 1. Therefore, the planner does not open establishments at single-establishment firms, closes establishments at small multi-establishment firms—pushing some into single-establishment status—and opens establishments at large multi-establishment firms. In total, the planner closes 11.2% of ad-

[20] Their version with quality rather than productivity heterogeneity is in online appendix E.1.

[21] Specifically, Edmond et al. (2023) implement a size-dependent tax/subsidy that eliminates markup dispersion, but leaves the aggregate markup unchanged. Initially, aggregate productivity rises by 1.75%. Taking into account the transition path to a new steady state, the representative household's welfare rises by the equivalent of a 5.58% permanent increase in consumption. There are further welfare gains of 6.44% (permanent consumption equivalent) from eliminating the aggregate markup, which inefficiently lowers the labor supply. We keep labor supply fixed, so this effect is also absent from our model.

Figure 6: Reallocation in the First Best



Left panel: establishment production (labor) in the first best relative to the competitive equilibrium, as a function of log establishment quality; m^{th} denotes the log quality below which are establishments that earn m% of all sales in the competitive equilibrium. Right panel: a firm's measure of additional establishments (beyond its initial unit measure) and production per establishment in the first best relative to the competitive equilibrium, as functions of its log baseline quality; φ^* is the baseline quality cutoff above which firms are multi-establishment in the competitive equilibrium; m^{th} denotes the log baseline quality below which are firms that earn m% of all sales in the competitive equilibrium.

ditional establishments, which is 1.6% of all establishments. The labor used for opening establishments falls by only 0.35% (a tiny 0.02% of labor supply) because convex establishment opening costs make it expensive to open establishments at large firms with many establishments. Finally, for production per establishment, the line is increasing because the planner reallocates production from small firms with small establishments to large firms with large establishments. The line flattens out among multi-establishment firms because larger firms get more new establishments, so they see a bigger drop in average establishment quality, which compensates for their initially higher average establishment quality.

Sources of misallocation. We decompose the 1.47% increase in consumption from implementing the first best into three sources of misallocation. Table 6 summarizes the results, including in the alternative model that we discuss below. First, we compute the misallocation of production within firms. Suppose the planner can only reallocate production labor across

establishments within each firm, but cannot change each firm's measure of establishments or total production labor. This improves consumption by 0.06%. Second, we compute the misallocation of production across firms. Suppose the planner can reallocate production labor across all establishments, but cannot change each firm's measure of establishments. This improves consumption by 1.34%, which implies that the gain from reallocating production across firms is 1.28% of competitive equilibrium consumption (1.34% - 0.06%). Finally, we compute extensive margin misallocation, i.e., the gains from changing the set of establishments. This is the remaining increase in consumption that results from going to the first best, which is 0.13% of competitive equilibrium consumption (1.47% - 1.34%).

Table 6: Sources of Misallocation

	Production within firms	**Production across firms**	**Extensive margin**	**Total misallocation**
Original model	0.06%	1.28%	0.13%	1.47%
Alternative model	0.06%	1.28%	1.33%	2.67%

Intuitively, it is not surprising that the misallocation of production across firms is much bigger than the misallocation of production within firms. As discussed earlier, we saw in the left panel of Figure 5 that there is much more dispersion in the marginal social value of production across firms than within firms, i.e., there are much higher marginal gains from reallocating production across firms than within firms. However, it is surprising that extensive margin misallocation is so low. Comparing the left and right panels of Figure 5, there appear to be big marginal gains from closing establishments at firms in exchange for producing more at even far smaller firms' establishments. Moreover, Afrouzi et al. (2023) study a similar model but where a firm's extensive margin is its number of customers rather than its number of establishments. They find a much higher level of misallocation and argue that this is due mostly to the extensive margin, i.e., there are large gains from efficiently choosing each firm's number of customers.

Why is extensive margin misallocation so low? We now investigate the effects of declining quality across a firm's successive establishments and in particular, demonstrate its importance for our finding of low extensive margin misallocation. This is the key difference between our model and that of Afrouzi et al. (2023); they assume all a firm's customers are identical. We deviate from this assumption to match our empirical finding that each

successive establishment at a firm is smaller.

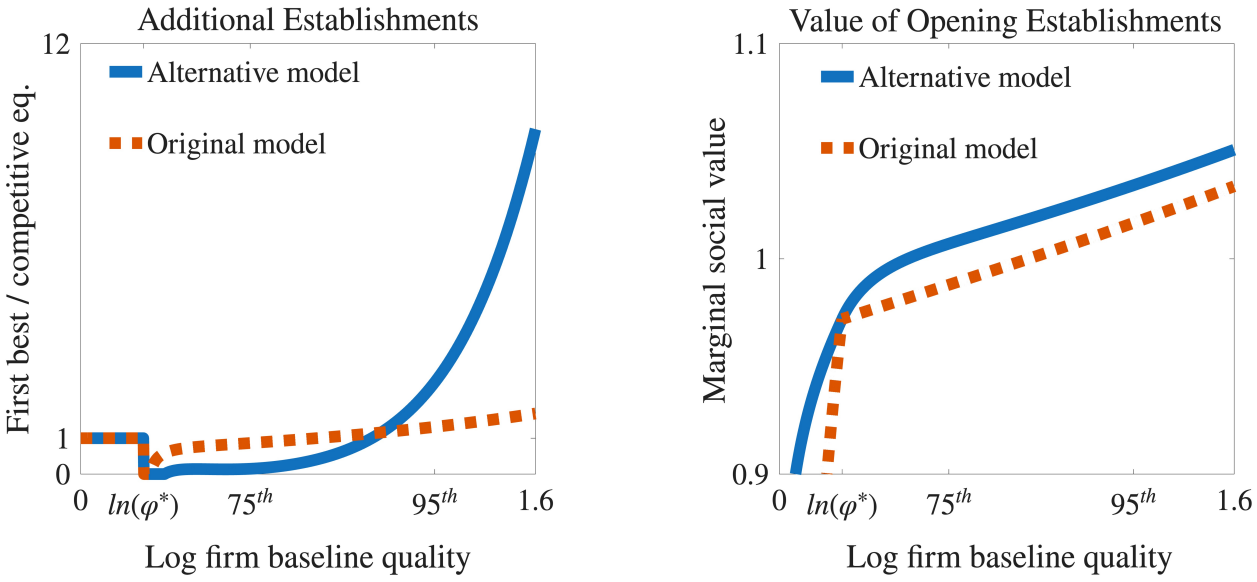
To evaluate the role of declining quality across a firm's successive establishments, we develop an alternative model in which quality at each of a firm's establishments is drawn from the same firm-specific distribution. The competitive equilibrium of the alternative model exactly replicates the joint distribution of establishment sales and markups from the competitive equilibrium of our original model, as well as the distribution of establishments across firms. As such, we interpret the alternative model as what we would think if we saw the data generated by our original model, but did not know the order of each firms' establishments, i.e., did not know that a firm's marginal establishment was any different from its others. We describe the alternative model in more detail in Appendix B.

We solve the planner's problem in the alternative model. Implementing the first best increases consumption by 2.67% relative to the competitive equilibrium. By construction, the gains from reallocating production across the competitive equilibrium set of establishments are the same as in the original model at 1.34%. Thus, there is a *ten-fold increase in extensive margin misallocation*—the gains from changing the set of establishments—relative to the original model: they rise from 0.13% of competitive equilibrium consumption (9% of total misallocation) to 1.33% (50% of total misallocation).

The left panel of Figure 7 compares the planner's optimal changes to the set of establishments in the alternative and original models. Relative to the original model planner, the alternative model planner closes more establishments at low quality multi-establishment firms and opens many more establishments at high quality firms. In total, the planner closes 25.9% of additional establishments, which is 3.7% of all establishments (compared to 11.2% and 1.6% in the original model). Due to convex establishment opening costs, the labor used for opening establishments *increases* by 78.2%, which is 4.5% of labor supply (compared to a *decrease* of 0.35% in the original model, which is 0.02% of labor supply). Thus, absent declining quality across firms' successive establishments, there are substantial gains from reducing production labor, closing establishments at low quality multi-establishment firms, and using the freed up labor to open establishments at high quality firms.

The stark difference in extensive margin misallocation between the original and alternative models is due to two effects of declining quality across firms' successive establishments. First, declining quality heightens the diminishing returns to changing the set of establishments. As the planner increases (decreases) a firm's measure of establishments, the firm's marginal establishment quality falls (rises), which dissipates the marginal social gains from further increases (decreases). Second, declining quality means each firm's marginal estab-

Figure 7: Alternative vs. Original Model



Left panel: a firm's measure of additional establishments (beyond its initial unit measure) in the first best of the alternative and original models relative to the competitive equilibrium, as functions of log firm baseline quality. Right panel: $V_{estab}(i)$ as a function of firm i's log baseline quality, computed using the competitive equilibrium set of establishments but with production efficiently allocated across establishments; $V_{estab}(i)$ is relative to $V_{prod}(i,n)$, which is equalized across all establishments. In both panels, φ^* is the baseline quality cutoff above which firms are multi-establishment in the competitive equilibrium of either model; m^{th} denotes the log baseline quality below which are firms that earn m% of all sales in the competitive equilibrium of either model.

lishment is its smallest establishment. This implies a low marginal social value of opening establishments at each firm in the competitive equilibrium, as we saw in Theorem 1 and Figure 5. Importantly, this does not necessarily imply low extensive margin misallocation because there can be gains from closing establishments. To isolate this second effect, the right panel of Figure 7 plots the marginal social value of opening each firm's competitive equilibrium marginal establishment in the alternative and original models. A value above (below) 1 indicates marginal social gains from opening (closing) establishments. The marginal social value of opening establishments is higher for all firms in the alternative model than in the original model, as expected, but is on average closer to 1.[22] Thus, we conclude that extensive

[22]Weighting each multi-establishment firm equally (or by its measure of establishments), the average absolute difference from 1 is $4.6 \cdot 10^{-4}$ (0.014) in the alternative model and $6.8 \cdot 10^{-4}$ (0.016) in the original model. We exclude single-establishment firms because the planner does not want to change their measure of establishments in either model.

margin misallocation is so much higher in the alternative model because of the first effect: the planner faces less diminishing returns to changing the set of establishments. We can see this clearly among high quality firms, where the planner opens many more establishments in the alternative model than in the original model even though the marginal social value of opening high quality firms' marginal establishments is only slightly higher.

5.2 Firm Size-Dependent Policy

We have seen that in our calibrated model, 1) nearly all misallocation in the competitive equilibrium is the misallocation of production across the competitive equilibrium set of establishments; 2) nearly all this misallocation is across firms rather than within firms; and 3) this is primarily because there are rapidly diminishing returns to changing the set of establishments. Together, these suggest that firm size-dependent policy can be effective at undoing the losses from misallocation, as it is in models of single-establishment firms with size-dependent markups such as in Edmond et al. (2023). However, we now show that is not the case. Instead, firms' extensive margin decisions are important for the design of optimal firm size-dependent policy and sharply limit its effectiveness.

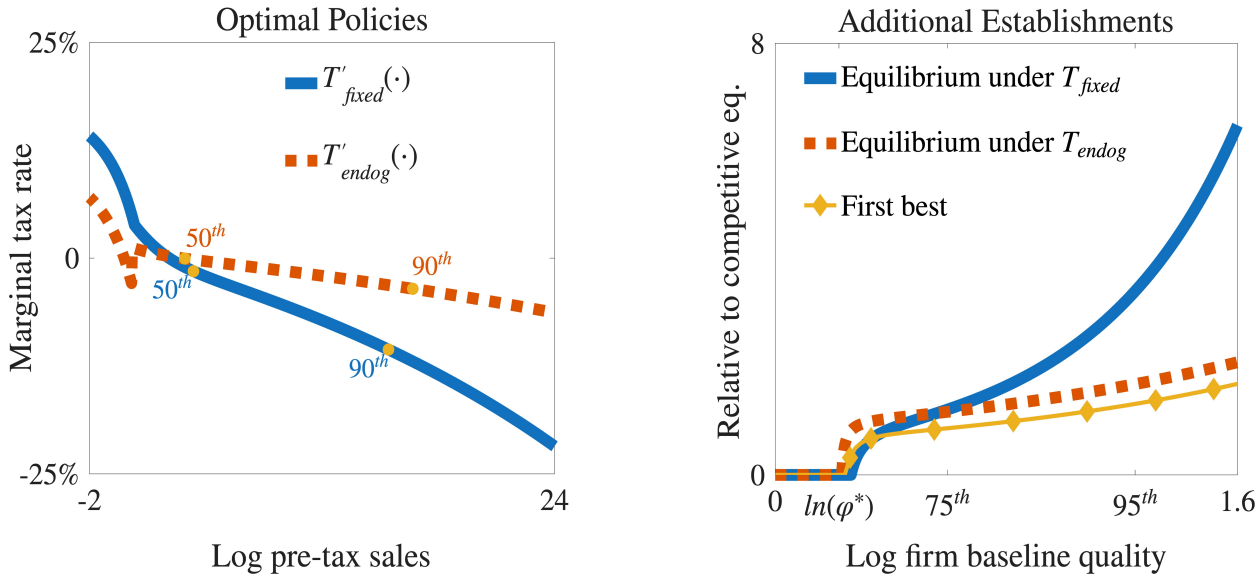
Formally, in our calibrated model, we study a firm size-dependent tax/subsidy $T(\cdot)$. If a firm has total nominal sales S across its establishments, then it pays $T(S)$, which is a tax if positive and a subsidy if negative. Hence, when choosing production at an establishment or how many establishments to open, its marginal revenue is multiplied by $1 - T'(S)$. Any revenue (positive or negative) from the policy is given lump sum to the representative household. Without loss of generality, we restrict attention to revenue-neutral policies with $T(0) = 0$.[23]

First, we choose the policy $T(\cdot)$ to maximize equilibrium consumption while ignoring firms' extensive margin decisions. That is, hold fixed the set of establishments from the competitive equilibrium without policy. Then, the firm size-dependent policy $T(\cdot)$ is common knowledge and firms choose production at each establishment to maximize profits in a competitive equilibrium. The optimal policy, $T_{fixed}(\cdot)$, improves aggregate consumption by 1.26%. Recall that efficiently allocating production across the competitive equilibrium set of establishments increased consumption by 1.34%. Thus, the optimal policy undoes 94%

[23]Revenue-neutrality means that in equilibrium, $\int_0^1 T(S(i))di = 0$. This is without loss of generality because if we multiply $1 - T'(\cdot)$ by $x > 0$, then the equilibrium wage also multiplies by x, so we decrease total revenue from the policy but do not affect firm incentives. $T(0)$ is also irrelevant because it simply shifts the household's total income between firm profits and lump sum transfers.

of this misallocation. The left panel of Figure 8 illustrates the optimal marginal tax rate $T'_{fixed}(\cdot)$ as a function of firm sales. Larger firms face a lower marginal tax rate, so policy shifts production from small firms to large firms. Since larger firms have larger establishments in the competitive equilibrium, this reallocates production from small establishments that inefficiently overproduce in the competitive equilibrium to large establishments that inefficiently underproduce.

Figure 8: Optimal Firm Size-Dependent Policies



Left panel: a firm's marginal tax rates under optimal firm size-dependent policies as functions of its log pre-tax sales; for each policy, m^{th} denotes the log pre-tax sales below which are firms that earn m% of all sales in the competitive equilibrium with policy (holding fixed the set of establishments from without policy for T_{fixed}). Right panel: each firm's additional establishments (beyond its initial unit measure) in the competitive equilibrium with each policy and in the first relative to the competitive equilibrium without policy, as functions of the firm's log baseline quality; m^{th} denotes the log baseline quality below which are firms that earn m% of all sales in the competitive equilibrium without policy.

Now, is the policy $T_{fixed}(\cdot)$ still effective if the set of establishments responds endogenously in equilibrium? The answer is no. We need to take the extensive margin into account when designing firm size-dependent policy. Formally, suppose the policy $T_{fixed}(\cdot)$ is common knowledge and then firms choose their measure of establishments and production at each establishment to maximize profits in a competitive equilibrium. Then, the policy backfires: consumption *falls* by 1.57% relative to no policy. To illustrate why, the right panel of Figure 8 depicts the equilibrium set of establishments under policy $T_{fixed}(\cdot)$ and in the first best

relative to the competitive equilibrium without policy. Large firms respond to their relative subsidy under $T_{fixed}(\cdot)$ by opening many more establishments than is optimal in the first best. Thus, aggregate productivity rises by 6.03% relative to no policy, but at too high a cost: 6.76% of the labor supply shifts from production to opening establishments.

Intuitively, optimal policy under a fixed set of establishments backfires when the set of establishments is endogenously determined in equilibrium because there is substantial disagreement between how it is socially optimal to incentivize firms on the intensive margin (production at existing establishments) and on the extensive margin (opening/closing establishments). We saw this disagreement earlier in Figure 5, which showed that in the competitive equilibrium without policy, the marginal social value of opening establishments at a firm is well below the marginal social value of production even at much smaller firms. Hence, the optimal policy under a fixed set of establishments, $T_{fixed}(\cdot)$, gives too big a relative subsidy to large firms than is optimal under an endogenous set of establishments.

Finally, is firm size-dependent policy effective if we take into account that the set of establishments is endogenously determined in equilibrium? To answer this, suppose the policy is common knowledge and then firms choose their measure of establishments and production at each establishment to maximize profits in a competitive equilibrium. In this case, the optimal policy, $T_{endog}(\cdot)$, increases aggregate consumption by 0.51% relative to no policy. Thus, it undoes only 34% of the total losses from misallocation (0.51% out of 1.47%). Moreover, the gains from this policy are only 40% of the gains from $T_{fixed}(\cdot)$ when the set of establishments are held fixed. Hence, firms' extensive margin decisions sharply limit the effectiveness of firm size-dependent policy.

The limited effectiveness of firm size-dependent policy when the set of establishments is endogenously determined in equilibrium is again because there is disagreement between how it is socially optimal to incentivize firms on the intensive margin and on the extensive margin. The left panel of Figure 8 plots the optimal marginal tax rate $T'_{endog}(\cdot)$ and the right panel shows the resulting equilibrium set of establishments relative to the competitive equilibrium without policy. $T_{endog}(\cdot)$ still gives a relative subsidy to large firms, but in order to balance the disagreement between the intensive and extensive margins, the relative subsidy is much smaller than under $T_{fixed}(\cdot)$. Indeed, $T'_{endog}(\cdot)$ is increasing for a brief interval to prevent some single-establishment firms from becoming multi-establishment. As a result of their lower relative subsidy, large firms open fewer establishments than under $T_{fixed}(\cdot)$, but still more than in the first best.

Importance of declining quality across successive establishments. To conclude our analysis of firm size-dependent policy, we show that declining quality across a firm's successive establishments plays an important role in our results thus far, as with our finding of low extensive margin misallocation. To do so, we again use our alternative model from Section 5.1 that shuts down declining quality. By construction, if the set of establishments is held fixed from the competitive equilibrium without policy, then the optimal firm size-dependent policy is the same in the alternative model as in the original model and increases consumption by 1.26%. But now, if we take into account that the set of establishments is endogenously determined in equilibrium, then the optimal policy increases consumption by 1.45% (compared to 0.51% in the original model). Therefore, in the alternative model, firm size-dependent policy is more effective with an endogenous set of establishments than with a fixed set of establishments. This is because firms' marginal establishment are larger in the alternative model than in the original model, so there is less disagreement between how it is socially optimal to incentivize firms on the intensive margin and on the extensive margin. Finally, for completeness, if we implement the optimal policy assuming a fixed set of establishments and then the set of establishments endogenously responds, consumption falls by 0.82%. Thus, it is still crucial to take into account the extensive margin when designing firm size-dependent policy in the alternative model.

References

Afrouzi, H., Drenik, A. and Kim, R. (2023), 'Concentration, market power, and misallocation: The role of endogenous customer acquisition', *Working paper* .

Baqaee, D. R. and Farhi, E. (2020), 'Productivity and misallocation in general equilibrium', *Quarterly Journal of Economics* **135**(1), 105–163.

Becker, J., Edmond, C., Midrigan, V. and Yi Xu, D. (2024), 'Local concentration, national concentration, and the spatial distribution of markups', *Working paper* .

Behrens, K., Mion, G., Murata, Y. and Suedekum, J. (2020), 'Quantifying the gap between equilibrium and optimum under monopolistic competition', *Quarterly Journal of Economics* **135**(4), 2299–2360.

Burstein, A., Carvalho, V. M. and Grassi, B. (2024), 'Bottom-up markup fluctuations', *Working paper* .

Cao, D., Hyatt, H. R., Mukoyama, T. and Sager, E. (2022), 'Firm growth through new establishments', *Working paper* .

De Loecker, J. and Warzynski, F. (2012), 'Markups and firm-level export status', *American Economic Review* **102**(6), 2437–2471.

De Ridder, M., Grassi, B. and Morzenti, G. (2025), 'The hitchhiker's guide to markup estimation: Assessing estimates from financial data', *Working paper* .

Dhingra, S. and Morrow, J. (2019), 'Monopolistic competition and optimum product diversity under firm heterogeneity', *Journal of Political Economy* **127**(1), 196–232.

Dixit, A. K. and Stiglitz, J. E. (1977), 'Monopolistic competition and optimum product diversity', *American Economic Review* **67**(3), 297–308.

Edmond, C., Midrigan, V. and Xu, D. Y. (2023), 'How costly are markups?', *Journal of Political Economy* **131**(7), 1619–1675.

Hsieh, C.-T. and Klenow, P. J. (2009), 'Misallocation and manufacturing tfp in china and india', *Quarterly Journal of Economics* **124**(4), 1403–1448.

Hsieh, C.-T. and Rossi-Hansberg, E. (2023), 'The industrial revolution in services', *Journal of Political Economy Macroeconomics* **1**(1), 3–42.

Kimball, M. S. (1995), 'The quantitative analytics of the basic neomonetarist model', *Journal of Money, Credit, and Banking* **27**(4), 1241–1277.

Klenow, P. J. and Willis, J. L. (2016), 'Real rigidities and nominal price changes', *Economica* **83**(331), 443–472.

Oberfield, E., Rossi-Hansberg, E., Sarte, P.-D. and Trachter, N. (2024), 'Plants in space', *Journal of Political Economy* **132**(3), 867–909.

Sandström, M. (2020), 'Can a rise of intangible capital explain an increase in markups?', *Working paper* .

Weiss, J. (2020), 'Intangible investment and market concentration', *Working paper* .

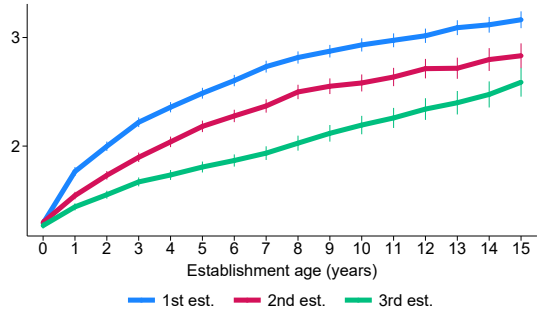
Zhelobodko, E., Kokovin, S., Parenti, M. and Thisse, J.-F. (2012), 'Monopolistic competition: Beyond the constant elasticity of substitution', *Econometrica* **80**(6), 2765–2784.

A Additional Empirical Results

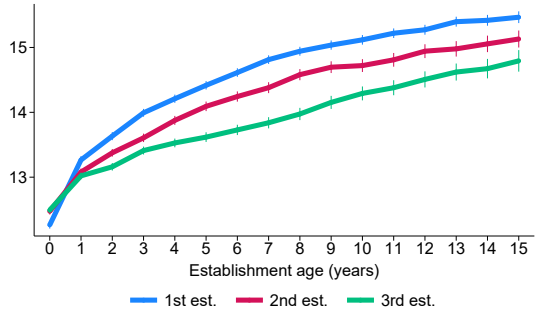
This section contains additional results related to our empirical findings. Figure 9 shows the size of successive establishments at firms with at least n establishments for $n \in \{3, 4\}$. We see the same pattern as we saw for firms with at least 5 establishments: each successive establishment at a firm tends to be smaller, conditional on establishment age. The patterns are similar for $n > 5$ (not shown), but the confidence intervals become large and the results are less clear. This makes sense because only 0.6% of our firm-year observations have more than 5 establishments, so the sample size becomes small.

Figure 9: Size of successive establishments

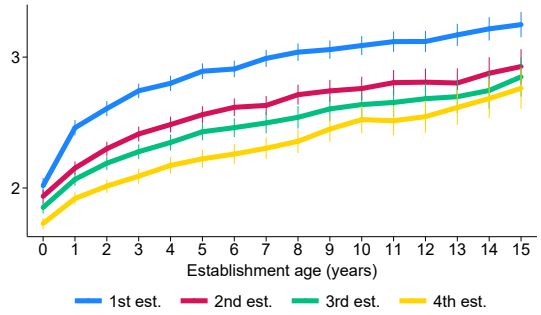
(a) Log employment ($\geq$ 3 establishments)



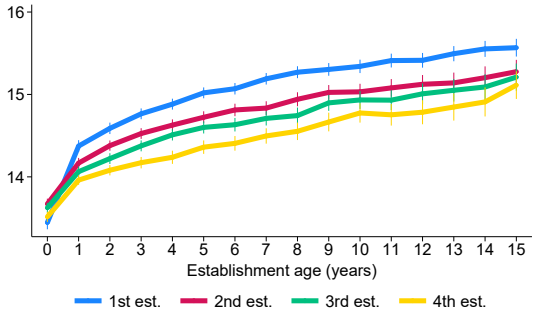
(b) Log wage bill ($\geq$ 3 establishments)



(c) Log employment ($\geq$ 4 establishments)



(d) Log wage bill ($\geq$ 4 establishments)



Average log employment (number of employees) and wage bill (in 2017 SEK) for firms' first $n \in \{3, 4\}$ establishments as functions of establishment age, conditional on survival. Averages are computed across firms with at least n establishments. The vertical bars are 95% confidence intervals.

Table 7 reports the first stage results of our instrumental variable regression listed in column 6 of Table 3 (from Section 2.4 on size-dependent markups). In the first stage, we predict a firm's log sales per establishment using the firm's previous year log sales per establishment, current log number of establishments, as well as fixed effects for the firm's

5-digit industry crossed with the year. This follows from the idea that a firm chooses its number of establishments ahead of time, but then its sales include noise.

Table 7: Size-dependent markups: first stage IV

	Log sales per establishment$_t$
Log sales per establishment$_{t-1}$	0.878
	(0.006)
Log number of establishments$_t$	0.013
	(0.003)
Only multi-establishment firms	✓
Industry-year fixed effects	✓
R^2	0.860
Number of observations	86,515

Results from regressing a firm's log sales per establishment on the firm's previous year log sales per establishment, current log number of establishments, and fixed effects for the firm's 5-digit industry crossed with the year. We only use firm-year observations with more than one establishment. Standard errors (in parentheses) are clustered at the firm level.

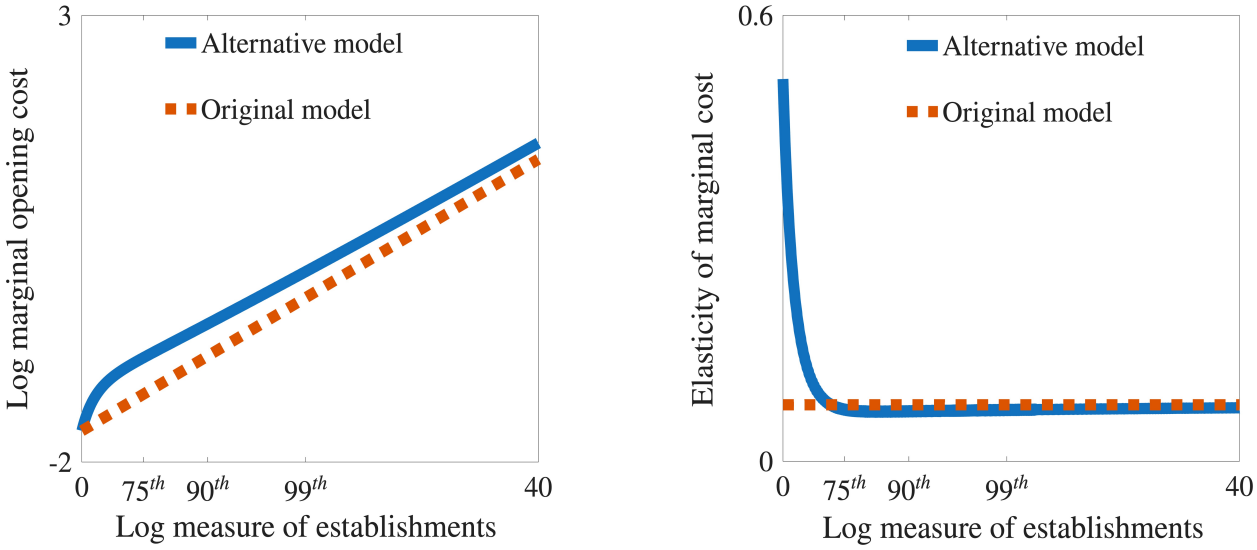
The estimated coefficient on the lag of log sales per establishment is 0.878. Since this is close to 1, it indicates that sales per establishment is highly persistent and strongly autocorrelated, rather than being driven by random noise. The F-statistic testing the null hypothesis that lagged sales per establishment and current number of establishments are irrelevant is 11,712.54, meaning that the instruments are strong.

B Alternative Model Details

Formally, the alternative model from Section 5.1 is as follows. Let $F(\cdot, i)$ be the cdf for the establishment quality distribution across firm i's establishments in the competitive equilibrium of our original model. Then, suppose that each establishment firm i opens has quality drawn from cdf $F(\cdot, i)$, regardless of how many establishments firm i opens. Next, set the establishment opening cost function and the inelastic labor supply so that in the competitive equilibrium of the alternative model, each firm chooses the same set of establishments and production at each establishment as in the competitive equilibrium of the original model.

Specifically, as in the original model, the establishment opening cost function is the same for all firms and is equal to 0 if a firm does not open any establishments beyond its initial unit measure. Then, for each $N \geq 1$, let $\hat{\varphi}(N)$ be the unique baseline quality of a firm that opens N establishments in the competitive equilibrium of the original model. In the alternative model, we set the marginal establishment opening cost at N establishments equal to the profits per establishment of a firm with baseline quality $\hat{\varphi}(N)$. Finally, we change the inelastic labor supply so that given the implied labor used for opening establishments, the remaining labor for production is the same as in the original model.

Figure 10: Establishment Opening Cost Functions



Left panel: a firm's log marginal establishment opening cost, in the alternative and original models, as functions of its log measure of establishments. Right panel: the elasticity of a firm's marginal establishment opening cost with respect to its measure of establishments, in the alternative and original models, as functions of its log measure of establishments. In both panels, m^{th} denotes the log measures of establishments such that firms earning m% of sales in the competitive equilibrium (of either model) have fewer establishments.

Figure 10 plots the marginal establishment opening cost function and its elasticity (with respect to measure of establishments) in the original and alternative models. In the original model, each multi-establishment firm's equilibrium marginal establishment opening cost must equal the profits at its lowest quality establishment. By assumption, the elasticity of the marginal establishment opening cost with respect to measure of establishments is constant at $\theta > 0$. In the alternative model, each firm's equilibrium marginal establishment opening cost must equal the *average* profits across its establishments. As a result, the elas-

ticity of the marginal establishment opening cost with respect to measure of establishments is higher than θ for a low measure of establishments, slightly lower than θ for a high measure of establishments, and converges to θ as the measure of establishments goes to infinity.

C Proofs

C.1 Proof of Theorem 1

Before we prove Theorem 1, it is useful to state and prove the following lemma.

Lemma 1. *Suppose the demand elasticity $\sigma(q)$ is strictly decreasing in relative output q and there exists a $\bar{q} > 0$ such that $\sigma(q) > 1$ for all $q \leq \bar{q}$. Then, for all $q \in (0, \bar{q})$,*

$$\frac{\Upsilon(q)}{q\Upsilon'(q)} < \frac{\sigma(q)}{\sigma(q) - 1} \tag{9}$$

and $\frac{\Upsilon(q)}{q\Upsilon'(q)}$ is strictly increasing in q.

Proof. To prove the lemma, we first show that inequality (9) holds weakly in the limit as q goes to 0. Then, we show that if (9) does not hold for some $q \in (0, \bar{q})$, then it cannot hold weakly as q goes to 0. This is a contradiction, so it proves that (9) must hold for all $q \in (0, \bar{q})$. Finally, we show that $\frac{\Upsilon(q)}{q\Upsilon'(q)}$ is strictly increasing in q.

First, we show that $\lim_{q\to 0} \frac{\sigma(q)}{\sigma(q)-1}$ exists. By assumption, $\sigma(q)$ is strictly greater than 1 for all $q \in (0, \bar{q})$ and is strictly decreasing in q. Therefore, $\frac{\sigma(q)}{\sigma(q)-1}$ is strictly greater than 1 for all $q \in (0, \bar{q}]$ and is strictly *increasing* in q. It follows that as q goes to 0, $\frac{\sigma(q)}{\sigma(q)-1}$ is strictly falling and is bounded below by 1. Thus, $\lim_{q\to 0} \frac{\sigma(q)}{\sigma(q)-1}$ exists.

Next, we show that $\lim_{q\to 0} \frac{\Upsilon(q)}{q\Upsilon'(q)}$ exists and is less than or equal to $\lim_{q\to 0} \frac{\sigma(q)}{\sigma(q)-1}$. The derivative of $q\Upsilon'(q)$ with respect to q is $\Upsilon'(q) + q\Upsilon''(q)$, which equals $\Upsilon'(q)\frac{\sigma(q)-1}{\sigma(q)}$. Since $\sigma(q)$ is strictly greater than 1 for all $q \in (0, \bar{q})$, it follows that $q\Upsilon'(q)$ is strictly increasing in q for all $q \in (0, \bar{q})$. Therefore, as q goes to 0, $q\Upsilon'(q)$ is strictly falling and bounded below by 0. Thus, $\lim_{q\to 0} q\Upsilon'(q)$ exists and is a weakly positive real number. Now, there are two cases. For case 1, suppose $\lim_{q\to 0} q\Upsilon'(q) > 0$. Then, since $\Upsilon(0) = 0$, it follows that $\lim_{q\to 0} \frac{\Upsilon(q)}{q\Upsilon'(q)} = 0$. For case 2, suppose $\lim_{q\to 0} q\Upsilon'(q) = 0$. Then, since $\Upsilon(0) = 0$, it follows that both the numerator and denominator of $\frac{\Upsilon(q)}{q\Upsilon'(q)}$ go to 0 as q goes to 0. To use L'Hôpital's rule, take the derivative of

the numerator over the derivative of the denominator to get $\frac{\Upsilon'(q)}{\Upsilon'(q)+q\Upsilon''(q)}$, which equals $\frac{\sigma(q)}{\sigma(q)-1}$. Since $\lim_{q\to 0}\frac{\sigma(q)}{\sigma(q)-1}$ exists, it then follows from L'Hôpital's rule that $\lim_{q\to 0}\frac{\Upsilon(q)}{q\Upsilon'(q)} = \lim_{q\to 0}\frac{\sigma(q)}{\sigma(q)-1}$. Hence, in either case, $\lim_{q\to 0}\frac{\Upsilon(q)}{q\Upsilon'(q)}$ exists and is less than or equal to $\lim_{q\to 0}\frac{\sigma(q)}{\sigma(q)-1}$.

Now, suppose $\frac{\Upsilon(q_0)}{q_0\Upsilon'(q_0)} \geq \frac{\sigma(q_0)}{\sigma(q_0)-1}$ for some $q_0 \in (0,\bar{q})$. Then, the derivative of $\frac{\Upsilon(q)}{q\Upsilon'(q)}$ with respect to q must be weakly negative at $q = q_0$ because

$$\frac{\partial\frac{\Upsilon(q)}{q\Upsilon'(q)}}{\partial q} = \frac{1}{q} - \frac{\Upsilon(q)(\Upsilon'(q)+q\Upsilon''(q))}{(q\Upsilon'(q))^2} = \frac{1}{q}\left(1 - \frac{\Upsilon(q)}{q\Upsilon'(q)}\frac{\sigma(q)-1}{\sigma(q)}\right). \tag{10}$$

It follows that if we decrease q below q_0, then $\frac{\Upsilon(q)}{q\Upsilon'(q)}$ weakly rises. Moreover, by assumption, $\frac{\sigma(q)-1}{\sigma(q)}$ is strictly positive at $q = q_0$ and strictly rises if we decrease q. Therefore, if we decrease q below q_0, then the derivative of $\frac{\Upsilon(q)}{q\Upsilon'(q)}$ with respect to q becomes strictly negative. This same argument shows that for all $q \in (0, q_0)$, the derivative of $\frac{\Upsilon(q)}{q\Upsilon'(q)}$ with respect to q is strictly negative. Moreover, by assumption, for all $q \in (0, q_0)$, $\frac{\sigma(q)}{\sigma(q)-1}$ is strictly increasing in q. Thus, since $\frac{\Upsilon(q_0)}{q_0\Upsilon'(q_0)} \geq \frac{\sigma(q_0)}{\sigma(q_0)-1}$, it follows that $\lim_{q\to 0}\frac{\Upsilon(q)}{q\Upsilon'(q)} > \lim_{q\to 0}\frac{\sigma(q)}{\sigma(q)-1}$. This is a contradiction because we saw earlier that the opposite inequality weakly holds. It follows that for all $q \in (0,\bar{q})$, inequality (9) holds strictly. Finally, from (10), it follows that for all $q \in (0,\bar{q})$, the derivative of $\frac{\Upsilon(q)}{q\Upsilon'(q)}$ with respect to q is strictly positive, so $\frac{\Upsilon(q)}{q\Upsilon'(q)}$ is strictly increasing in q. This completes the proof of the lemma. ■

We can now prove Theorem 1.

Proof. Suppose a planner has an infinitesimal quantity Δ of labor beyond the inelastic labor supply $\bar{L}$. First, suppose they use this labor to increase production at firm i's establishment n, holding fixed the measure of establishments at all firms from the competitive equilibrium, and holding fixed production at all other establishments from the competitive equilibrium. Then, household consumption (final good output) rises by $V_{prod}(i,n)\Delta$, where $V_{prod}(i,n)$ must be such that the Kimball aggregator constraint (3) continues to hold:

$$0 = \varphi(i,n)\Upsilon'(q(i,n))\frac{1}{Y}\Delta - \left(\int_0^1\int_0^{N(j)}\varphi(j,m)\Upsilon'(q(j,m))q(j,m)\frac{1}{Y}\right)dmdjV_{prod}(i,n)\Delta,$$

where the first term on the right-hand side is the effect of the increase in relative output at firm i's establishment n and the second term is the effect of the resulting decrease in relative

output at all other establishments due to the increase in final good output. It follows that

$$V_{prod}(i,n) = \varphi(i,n)\Upsilon'(q(i,n))D = p(i,n) = \frac{\sigma(q(i,n))}{\sigma(q(i,n))-1}W, \tag{11}$$

which follows from the derivation of the demand curve (4), where D is the demand index, and from expression (6) for an establishment's competitive equilibrium markup.

Next, suppose the planner uses the additional Δ units of labor to open establishments at firm i and produce the competitive equilibrium quantity, $y(i,N(i))$, at those establishments, holding fixed the measure of establishments at all other firms and production at all other establishments from the competitive equilibrium. The planner opens $\frac{\Delta}{\kappa N(i)^\theta + y(i,N(i))}$ establishments at firm i because each establishment uses $\kappa N(i)^\theta$ units of labor for the marginal opening cost and $y(i,N(i))$ units of labor for production. Then, household consumption (final good output) must rise by $V_{estab}(i)\Delta$ where $V_{estab}(i)$ is such that the Kimball aggregator constraint (3) continues to hold:

$$0 = \frac{\varphi(i,N(i))\Upsilon(q(i,N(i)))\Delta}{\kappa N(i)^\theta + y(i,N(i))} - \left(\int_0^1 \int_0^{N(j)} \varphi(j,m)\Upsilon'(q(j,m))q(j,m)\frac{1}{Y}dmdj\right) V_{estab}(i)\Delta,$$

where the right-hand side of the first line is the effect of the increase in establishments at firm i and the second line is the effect of the resulting decrease in relative output at all other establishments due to the increase in final good output. It follows that

$$V_{estab}(i) = \frac{\varphi(i,N(i))\Upsilon(q(i,N(i)))Y}{\kappa N(i)^\theta + y(i,N(i))}D = \frac{p(i,N(i))y(i,N(i))}{W(\kappa N(i)^\theta + y(i,N(i)))}\frac{\Upsilon(q(i,N(i)))}{q(i,N(i))\Upsilon'(q(i,N(i)))}W,$$

which follows from the derivation of the demand curve (4), where D is the demand index. The first ratio after the second equal sign is revenue at firm i's marginal establishment relative to cost (including the establishment opening cost). If firm i is multi-establishment, then this ratio is 1 (otherwise the firm would open/close establishments), so

$$V_{estab}(i) = \frac{\Upsilon(q(i,N(i)))}{q(i,N(i))\Upsilon'(q(i,N(i)))}W. \tag{12}$$

On the other hand, if firm i is single-establishment, then

$$V_{estab}(i) = \frac{\sigma(i,1)/(\sigma(i,1)-1)}{\kappa/y(i,1)+1}\frac{\Upsilon(q(i,1))}{q(i,1)\Upsilon'(q(i,1))}W \leq \frac{\Upsilon(q(i,1))}{q(i,1)\Upsilon'(q(i,1))}W, \tag{13}$$

where we use expression (6) for an establishment's competitive equilibrium markup and where the inequality follows because otherwise the firm would open establishments.

Now, suppose the demand elasticity $\sigma(q)$ is constant in an establishment's relative output q at $\sigma > 1$ (it must be greater than 1 by Assumption 1). It follows from (11) that for all firms i and establishments n, the marginal social value of production is constant at $V_{prod}(i,n) = W\sigma/(\sigma-1)$. Moreover, since $\sigma = -\Upsilon'(q)/(q\Upsilon''(q))$ by definition, it follows that for all q, $\Upsilon'(q) = Aq^{-1/\sigma}$ for some $A > 0$. Along with $\Upsilon(0) = 0$, this implies that for all q, $\Upsilon(q) = A\frac{\sigma}{\sigma-1}q^{(\sigma-1)/\sigma}$. Thus, it follows from (12) that for all multi-establishment firms i, the marginal social value of opening establishments is constant at $V_{estab}(i) = W\sigma/(\sigma-1)$, which is equal to the constant value of all $V_{prod}(i,n)$. It follows from (13) that for all single-establishment firms i, $V_{estab}(i)$ is weakly lower. This completes the proof of the first statement of the theorem.

On the other hand, suppose the demand elasticity $\sigma(q)$ is strictly decreasing in an establishment's relative output q and is strictly greater than 1 for all $q < \bar{q}$ (the latter must be the case by Assumption 1). Since $\sigma(q)/(\sigma(q)-1)$ is strictly decreasing in $\sigma(q)$, it follows from (11) that $V_{prod}(i,n)$ is strictly increasing in the relative output of firm i's establishment n, $q(i,n)$. Moreover, it follows from (12) and Lemma 1 that if firm i is multi-establishment, then $V_{estab}(i)$ is strictly increasing in the relative output of firm i's marginal establishment, $q(i,N(i))$. Then since $\sigma(i,1)$ is strictly decreasing in $y(i,1)$, it follows from (13) and Lemma 1 that if firm i is single-establishment, then $V_{estab}(i)$ is strictly increasing in $q(i,N(i))$. Finally, it follows from (11), (12), (13), and Lemma 1 that for all firms i, $V_{estab}(i) < V_{prod}(j,n)$ if $q(i,N(i)) \leq q(j,n)$. This completes the proof of the theorem. ■