How Costly Are Markups?[*]

Chris Edmond[†] Virgiliu Midrigan[‡] Daniel Yi Xu[§]

First draft: July 2018. This draft: May 2021

Abstract

We study the welfare costs of markups in a dynamic model with heterogeneous firms and endogenously variable markups. Our general framework encompasses a range of popular market structures. We provide aggregation results showing how the macro implications of micro-level markup heterogeneity can be summarized by a few key statistics. We calibrate our model to match US Census of Manufactures firm-level data and find that the welfare costs of markups can be large. We decompose the costs of markups into three channels: (i) an aggregate markup that acts like a uniform output tax, (ii) misallocation of factors of production, and (iii) inefficient entry. Across all specifications, we find that the aggregate markup and misallocation channels account for the bulk of the costs of markups and that the entry channel is much less important. Subsidizing entry is not an effective tool in our model. While an increase in competition reduces incumbents' markups, it also reallocates market shares towards larger incumbent firms and the net effect is that the aggregate markup changes little.

Keywords: competition, concentration, misallocation, firm dynamics.

JEL classifications: D4, E2, L1, O4.

---

[*] We thank Greg Kaplan and five anonymous referees for valuable comments and suggestions. We are also particularly grateful for insightful feedback from our discussants Salomé Baslandze, Ariel Burstein, Jan Eeckhout, and Gino Gancia, and from Gauti Eggertsson, Emmanuel Farhi, Oleg Itskhoki, Pete Klenow and Iván Werning. We also thank participants at the Fall 2018 NBER EFG meeting, 2018 Bank of Italy-CEPR-EIEF conference on firm dynamics and economic growth and seminar participants at Columbia University, Duke University, ETH Zurich, the FRB New York, FRB St Louis, the Graduate Institute of Geneva, Keio University, the LSE, MIT, Northwestern, Notre Dame, NYU, Rochester University, UC Berkeley, USC, the University of Adelaide, the University of Michigan, and the University of Chicago for their comments. Edmond thanks the Australian Research Council for financial support under grant DP-150101857.

[†] University of Melbourne, cedmond@unimelb.edu.au.

[‡] New York University and NBER, virgiliu.midrigan@nyu.edu.

[§] Duke University and NBER, daniel.xu@duke.edu.

1 Introduction

How large are the welfare costs of product market distortions? What kinds of policies can best overcome these distortions? We answer these questions using a dynamic model with heterogeneous firms and endogenously variable markups. In our model, markups distort allocations through three channels. First, the *aggregate markup* acts like a uniform tax on all firms. Second, there is cross-sectional markup dispersion because larger firms face less competition and so charge higher markups. This markup dispersion gives rise to *misallocation* of factors of production. Third, there is *inefficient entry.* Our goal is to quantify these three channels using US micro data and to evaluate policies aimed at reducing the costs of markups.

We study economies in which, within a given sector, more productive firms are, in equilibrium, larger and face less elastic demand and so charge higher markups than less productive firms. As a consequence, changes in the environment that allow more productive firms to grow at the expense of less productive firms will be associated with an increase in the aggregate markup and a decline in the aggregate labor share. In this sense, our model is consistent with the reallocation of production from firms with relatively high measured labor shares to firms with relatively low measured labor shares (Autor, Dorn, Katz, Patterson and Van Reenen, 2020; Kehrig and Vincent, 2021) and the observation that firms with high markups have been getting larger, driving up the aggregate markup (Baqaee and Farhi, 2020).

Our general framework encompasses a range of popular market structures including (i) *monopolistic competition* with Kimball (1995) demand or symmetric translog demand as in Feenstra (2003), and (ii) *oligopolistic competition* with nested-CES demand as in Atkeson and Burstein (2008) and Edmond, Midrigan and Xu (2015). We consider settings where firms can differ in both productivity and quality and provide aggregation results showing that the macro implications of micro-level markup heterogeneity can be summarized by a few key statistics. One such result is that the aggregate markup, the 'wedge' in aggregate employment and investment decisions, is given by the *cost-weighted* average of firm-level markups.[1] By contrast the empirical literature on the macro implications of markup heterogeneity focuses on the *sales-weighted* average of firm-level markups.[2] We show that the sales-weighted average is the cost-weighted average plus a correction term reflecting the *variance* of markups. In this sense the sales-weighted average overstates the aggregate markup by including a term that reflects misallocation rather than the level of markups per se. Importantly, these aggregation results hold independent of the market structure details.

---

[1] Or the sales-weighted *harmonic* average, as in Edmond, Midrigan and Xu (2015) and Grassi (2017).

[2] For example, for the US economy De Loecker, Eeckhout and Unger (2020) estimate a sharply increasing sales-weighted average markup rising from about 1.2 in 1980 to about 1.6 in 2016. By contrast the cost-weighted average is lower and has risen by less, from about 1.1 to about 1.25. The difference reflects the increase in cross-sectional markup dispersion. We discuss these measures at length in Appendix A.

Regardless of market structure, we find that markups distort allocations through the three channels mentioned at the outset: the aggregate markup, misallocation due to markup dispersion, and inefficient entry. We show that the efficient allocation can be implemented by a specific nonlinear schedule of direct subsidies with two components, a *uniform component* that subsidizes all firms and that can be used to eliminate the aggregate markup, and a *size-dependent* component that jointly eliminates misallocation and the entry distortion.

We quantify the welfare costs of markups by asking how much the representative consumer would benefit if the economy transitioned from an initial steady state with markup distortions to the efficient steady state. Because eliminating the markup distortions entails a large increase in the capital stock, taking into account the cost of building up the capital stock is critical to correctly assess the welfare gains from such policies. We calibrate the initial steady state using US Census of Manufactures firm-level data from 1972 to 2012 to match levels of sales concentration and the firm-level relationship between labor shares and market shares observed in 6-digit NAICS industries, controlling for firm fixed effects and 6-digit NAICS industry-year effects to wash out other persistent sources of firm and industry heterogeneity.

Our calibration strategy makes use of the fact that, though the precise mapping depends on market structure, all versions of our model imply a simple firm-level relationship between markups and market shares. We use the estimated parameters from this relationship and observed market shares to calculate the markups implied by our model. We prefer to use the markups implied by our model rather than separately estimated markups based on production function elasticities because of the difficulties in identifying markups when only revenue data is available, as shown by Bond, Hashemi, Kaplan and Zoch (2020). Given the wide range of estimates of the aggregate markup for the US economy,[3] we report the costs of markups for a wide range of values for the aggregate markup, recalibrating the model each time.

We find that the welfare costs of markups can be large. Depending on the market structure and assumed level of the aggregate markup, the costs can be as high as 25% in consumption-equivalent terms. Overall, the costs tend to be lower if we assume monopolistic competition and target a low level of the aggregate markup, but are much higher if we assume oligopolistic competition or if we target a high aggregate markup.

We then turn to quantifying the relative importance of the three channels by which markups reduce welfare in our model. Across all specifications, we find that the aggregate markup and misallocation channels account for the bulk of the costs of markups and that the entry channel is much less important. That said, the relative importance of the aggregate markup and misallocation channels vary depending on the market structure and target for the aggregate markup. For example, the Kimball specification implies that the *share* of the total costs accounted for by the aggregate markup increases from 1/2 to 3/4 as we increase

---

[3] See e.g., Atkeson, Burstein and Chatzikonstantinou (2019), Barkai (2020), De Loecker, Eeckhout and Unger (2020), Gutiérrez and Phillippon (2017a,b), and Hall (2018) etc. Karabarbounis and Neiman (2019) demonstrate the difficulties in using movements in factor shares to infer movements in the aggregate markup.

the aggregate markup from 1.05 to 1.25. The balance of the costs are almost entirely due to misallocation, the losses from the entry distortion are negligible.

Although the losses from misallocation in our model can be sizeable, accounting for gross output TFP losses of around 1% to 3%, depending on the specification, they are small relative to standard estimates in the literature (Restuccia and Rogerson, 2008; Hsieh and Klenow, 2009). This is because we measure misallocation using the dispersion in marginal revenue products implied by the endogenous markup distribution in our model, i.e., that relatively small share of the dispersion in average revenue products systematically related to market shares. We *do not* attribute all variation in observed average revenue products to markups.

In representative firm models, subsidizing entry (or reducing barriers to entry) so as to increase competition is a powerful tool for reducing the aggregate markup and hence reducing the costs of markups (Bilbiie, Ghironi and Melitz, 2008, 2019). By contrast we find that, with heterogeneous firms, subsidizing entry is *not* a powerful tool. For all our specifications, we find that even large increases in the number of firms have small effects on the aggregate markup.[4] To understand this, recall that the aggregate markup is a cost-weighted average of firm-level markups. An increase in the number of firms has two effects on this weighted average. The direct effect is a reduction in the markup of each firm, due to a reduction in each firm's market share. But there is also an important compositional effect: small firms face more elastic demand and are more vulnerable to competition from entrants; large firms face less elastic demand and are less vulnerable. So when there is an increase in the number of firms, small, low markup firms contract by more than large, high markup firms and the resulting reallocation keeps the aggregate markup almost unchanged, despite the reduction in firm-level markups. In all our specifications, this offsetting compositional effect is almost as large as the direct effect so overall the aggregate markup falls by a small amount.[5]

The different specifications we consider each have their own strengths and weaknesses. The model with Kimball demand is more flexible than the model with symmetric translog demand and is better able to match our calibration targets. But the model with translog demand is more tractable than Kimball demand and leads to sharp analytic results. Both monopolistic competition models are simple computationally. The oligopoly model is computationally challenging but has richer empirical content. Though our aggregation results hold regardless of the assumed market structure, the oligopoly model makes a number of predictions that differ from the monopolistic competition models. First, we find larger amounts of markup dispersion and hence larger losses from misallocation in the oligopoly model than in

---

[4] There are however standard love-of-variety gains from increasing the number of firms.

[5] These offsetting direct and compositional effects are reminiscent of results in the trade literature, e.g., Bernard, Eaton, Jensen and Kortum (2003) and especially Arkolakis, Costinot, Donaldson and Rodríguez-Clare (2019). We derive analogous results for Kimball and translog demand but unlike in their analysis, we do not assume from the outset that the 'choke price' in either demand system is binding, since this is an equilibrium outcome. For the translog case, we also provide closed-form solutions for the aggregate markup and the cutoff productivity that pins down the cross-sectional distributions of markups and market shares.

either of the monopolistic competition models. Second, while the monopolistic competition models predict that there are *too few* firms in equilibrium, the oligopoly model predicts that there are *too many*. But since the entry margin is not a quantitatively important source of losses in any specification, this qualitative difference is not important.

Existing results on costs of markups. The starting point for discussion of the welfare costs of markups is Dixit and Stiglitz (1977), though the literature goes back to Lerner (1934). Recent work such as Zhelobodko, Kokovin, Parenti and Thisse (2012), Dhingra and Morrow (2019) and Behrens, Mion, Murata and Suedekum (2020) studies variable markups in static models with heterogeneous firms. By contrast, our model is dynamic. Like us, Bilbiie, Ghironi and Melitz (2008, 2019) study a dynamic model and quantify the costs of markups but they assume a representative firm. We find, however, that firm heterogeneity plays a crucial role in understanding the costs of markups. Other related work includes Atkeson and Burstein (2010, 2019) who provide a welfare analysis of innovation policies in firm dynamics models but who abstract from variable markups and Peters (2020) who studies innovation, firm dynamics, and variable markups but who does not evaluate the welfare costs of markups.

Markups and misallocation. In our model markups increase with firm size. This is one form of misallocation in the sense of Restuccia and Rogerson (2008), and Hsieh and Klenow (2009). We find that the productivity losses from this form of misallocation are on the order of 1 to 3%. We view these numbers as an upper bound on the gains from size-dependent subsidies since we attribute all of the systematic relationship between firm revenue productivity and firm size to market power, and not to, say, overhead costs as in Autor, Dorn, Katz, Patterson and Van Reenen (2020) and Bartelsman, Haltiwanger and Scarpetta (2013). Because of this we are likely somewhat overstating the true relationship between markups and firm size and overstating the losses from this form of misallocation.

It is important to recognize that we abstract from all other sources of markup variation that may cause misallocation. Firms may operate in different locations or sell different products in different sectors and charge different markups depending on the amount of competition they face in those different markets.[6] In principle policies that condition on location or other relevant market details may be able to address these forms of misallocation too. But implementing finely-tuned policies that condition on details of market conditions location-by-location seems challenging in practice. For this reason we have limited our analysis to size-dependent markup variation and we find that the productivity gains from eliminating misallocation due to size-dependent markup variation are likely no more than 1 to 3%.

---

[6] Rossi-Hansberg, Sarte and Trachter (2020) show that while aggregate US product-market concentration has been rising since the early 1990s, concentration in geographically-specific local markets has been falling.

2 Model

There is a representative consumer with preferences over final consumption and labor supply and who owns all the firms. The final good is produced by perfectly competitive firms using inputs from many sectors. Within each sector there are heterogeneous imperfectly competitive firms producing differentiated products using capital, labor and materials. Firms enter by paying a sunk cost in units of labor and then obtain a one-time productivity draw in a randomly allocated sector. Exit is random and there is no aggregate uncertainty. We focus on characterizing the steady state and transitional dynamics after a policy change.

2.1 Setup

A key feature of our analysis is a set of aggregation results that hold regardless of the details of market structure within each sector. We proceed in two steps, first explaining the basic setup and aggregate outcomes that hold independent of market structure within each sector and then turning to the remaining details where market structure matters.

Representative consumer. The representative consumer maximizes

$$\sum_{t=0}^{\infty} \beta^t \Big(\log C_t - \psi \frac{L_t^{1+\nu}}{1+\nu} \Big) \tag{1}$$

subject to the budget constraint

$$C_t + I_t = W_t L_t + R_t K_t + \Pi_t \tag{2}$$

where C_t denotes consumption of the numeraire final good, $I_t = K_{t+1} - (1-\delta)K_t$ denotes investment, K_t denotes physical capital, L_t denotes labor supply, W_t the real wage, R_t the rental rate of capital, and Π_t denotes aggregate profits net of the cost of creating new firms.

The representative consumer's labor supply satisfies

$$\psi C_t L_t^{\nu} = W_t \tag{3}$$

and their investment choice satisfies

$$1 = \beta \frac{C_t}{C_{t+1}} (R_{t+1} + 1 - \delta) \tag{4}$$

Since firms are owned by the representative consumer, they use the one-period discount factor $\beta C_t / C_{t+1}$ to discount future profit flows.

Final good producers. Let Y_t denote gross output of the final good. This can be used for consumption C_t, investment I_t, or as materials X_t, so that

$$C_t + I_t + X_t = Y_t \tag{5}$$

The use of the final good as materials gives the model a simple *roundabout* production structure, as in Jones (2011) and Baqaee and Farhi (2020).

The final good Y_t is produced by perfectly competitive firms using inputs $y_t(s)$ from a continuum of sectors

$$Y_t = \left(\int_0^1 y_t(s)^{\frac{\eta-1}{\eta}}\, ds\right)^{\frac{\eta}{\eta-1}} \tag{6}$$

where $\eta > 1$ is the elasticity of substitution *across* sectors $s \in [0,1]$. Let $p_t(s)$ denote the price index for sector s. Since the final good is the numeraire, these satisfy

$$1 = \left(\int_0^1 p_t(s)^{1-\eta}\, ds\right)^{\frac{1}{1-\eta}} \tag{7}$$

Within sectors. Within each sector there are imperfectly competitive firms producing differentiated goods. As discussed extensively below, we consider two market structures: *monopolistic competition* with a continuum of firms $i \in [0, n_t(s)]$ per sector, or *oligopolistic competition* with a finite number of firms $i = 1, \ldots, n_t(s)$ per sector. Except where noted, our results below hold for both cases.

Technology. Firms enter by paying a sunk cost κ in units of labor and then obtain a one-time productivity draw $z_i(s) \sim G(z)$ in a random sector s. A firm's *gross output* is then

$$y_{it}(s) = z_i(s)\left(\phi^{\frac{1}{\theta}}\, v_{it}(s)^{\frac{\theta-1}{\theta}} + (1-\phi)^{\frac{1}{\theta}}\, x_{it}(s)^{\frac{\theta-1}{\theta}}\right)^{\frac{\theta}{\theta-1}} \tag{8}$$

where $v_{it}(s)$ is the firm's *value-added*, a composite of physical capital and labor

$$v_{it} = k_{it}(s)^{\alpha} l_{it}(s)^{1-\alpha} \tag{9}$$

We impose a unit elasticity of substitution between capital and labor. The elasticity of substitution between value-added $v_{it}(s)$ and materials $x_{it}(s)$ is given by θ, and can be lower.

Input demands. Taking input prices as given, cost minimization gives the input demands

$$R_t k_{it}(s) = \alpha \left\{\left(\frac{R_t}{\alpha}\right)^{\alpha}\left(\frac{W_t}{1-\alpha}\right)^{1-\alpha}\right\} v_{it}(s) \tag{10}$$

$$W_t l_{it}(s) = (1-\alpha) \left\{\left(\frac{R_t}{\alpha}\right)^{\alpha}\left(\frac{W_t}{1-\alpha}\right)^{1-\alpha}\right\} v_{it}(s) \tag{11}$$

where the term in braces on the right is the price index for the value-added composite. In turn, demand for the value-added composite and demand for materials are given by

$$v_{it}(s) = \phi \left\{ \frac{\left(\frac{R_t}{\alpha}\right)^{\alpha} \left(\frac{W_t}{1-\alpha}\right)^{1-\alpha}}{\Omega_t} \right\}^{-\theta} \frac{y_{it}(s)}{z_i(s)} \tag{12}$$

and

$$x_{it}(s) = (1-\phi) \left\{ \frac{1}{\Omega_t} \right\}^{-\theta} \frac{y_{it}(s)}{z_i(s)} \tag{13}$$

where Ω_t is the input price index dual to the technologies in (8) and (9), namely

$$\Omega_t = \left(\phi \left\{ \left(\frac{R_t}{\alpha}\right)^{\alpha} \left(\frac{W_t}{1-\alpha}\right)^{1-\alpha} \right\}^{1-\theta} + (1-\phi) \right)^{\frac{1}{1-\theta}} \tag{14}$$

where materials have a relative price of 1 since they are in units of the numeraire. Notice that the capital/labor and value-added/materials ratios are common to all firms.

Marginal cost. These factor demands imply that a firm's *marginal cost* is given by

$$\frac{\Omega_t}{z_i(s)} \tag{15}$$

Profits and markups. A firm's profits are then given by

$$\pi_{it}(s) = p_{it}(s) y_{it}(s) - \frac{\Omega_t}{z_i(s)} y_{it}(s) \tag{16}$$

Firms maximize profits subject to the demand system they face, which depends on the market structure details. At the optimum a firm's price can be written as a markup $\mu_{it}(s)$ over marginal cost

$$p_{it}(s) = \mu_{it}(s) \frac{\Omega_t}{z_i(s)}, \qquad \mu_{it}(s) = \frac{\sigma_{it}(s)}{\sigma_{it}(s) - 1} \tag{17}$$

where $\sigma_{it}(s)$ denotes the (endogenous) *demand elasticity* facing firm i. Different demand systems imply different determinants of $\sigma_{it}(s)$ as discussed below. Profits can then be written in terms of markups and sales

$$\pi_{it}(s) = \left(1 - \frac{1}{\mu_{it}(s)}\right) p_{it}(s) y_{it}(s) \tag{18}$$

Labor shares. Combining a firm's labor demand from (11)-(12) with markup pricing (17), a firm's labor share can be written

$$\frac{W_t l_{it}(s)}{p_{it}(s) y_{it}(s)} = \frac{(1-\alpha)\zeta_t}{\mu_{it}(s)} \tag{19}$$

where ζ_t denotes the elasticity of output with respect to value-added

$$\zeta_t := \frac{\frac{\phi}{1-\phi}\left\{\left(\frac{R_t}{\alpha}\right)^{\alpha}\left(\frac{W_t}{1-\alpha}\right)^{1-\alpha}\right\}^{1-\theta}}{1 \; + \; \frac{\phi}{1-\phi}\left\{\left(\frac{R_t}{\alpha}\right)^{\alpha}\left(\frac{W_t}{1-\alpha}\right)^{1-\alpha}\right\}^{1-\theta}} \tag{20}$$

This elasticity is common to all firms but in general varies over time. All cross-sectional variation in labor shares is due to cross-sectional variation in markups $\mu_{it}(s)$.

We next briefly outline how the distribution of markups $\mu_{it}(s)$ affects productivity within and across sectors. We focus on aggregation results that obtain independent of within-sector market structure.

Aggregate productivity. Let $k_t(s)$, $l_t(s)$, and $x_t(s)$ denote sector-level capital, labor and materials. These are the integrals (or sums) of $k_{it}(s), l_{it}(s)$ and $x_{it}(s)$ over i within s. We can then write the gross output of sector s as

$$y_t(s) = z_t(s) F\left(k_t(s), l_t(s), x_t(s)\right) \tag{21}$$

where

$$F(k,l,x) = \left(\phi^{\frac{1}{\theta}}\left(k^{\alpha} l^{1-\alpha}\right)^{\frac{\theta-1}{\theta}} + (1-\phi)^{\frac{1}{\theta}} x^{\frac{\theta-1}{\theta}}\right)^{\frac{\theta}{\theta-1}} \tag{22}$$

and where sector-level productivity satisfies

$$z_t(s) = \left(\int_0^{n_t(s)} \frac{q_{it}(s)}{z_i(s)}\, di\right)^{-1} \tag{23}$$

where $q_{it}(s) := y_{it}(s)/y_t(s)$ denotes the relative size of firm i in sector s. The only difference having a finite number of firms makes is that the integral should be replaced by a finite sum.

Likewise, let $K_t, \tilde{L}_t$, and X_t denote aggregate capital, labor *used in production*, and materials. These are the integrals of $k_t(s), l_t(s)$ and $x_t(s)$ over $s \in [0,1]$. We then have aggregate gross output $Y_t = Z_t F(K_t, \tilde{L}_t, X_t)$ where aggregate productivity is given in the same way as sector productivity

$$Z_t = \left(\int_0^1 \frac{q_t(s)}{z_t(s)}\, ds\right)^{-1} \tag{24}$$

where $q_t(s) := y_t(s)/Y_t$ denotes the relative size of sector s.

Thus sector-level productivity $z_t(s)$ is a firm-size-weighted harmonic average of firm-level productivity $z_i(s)$ and aggregate productivity Z_t is a sector-size-weighted harmonic average of sector-level productivity. Sector-level productivity and aggregate productivity are affected by markups $\mu_{it}(s)$ through the effects of markups on the distribution of firm-size $q_{it}(s)$ within sectors and the distribution of sector-size $q_t(s)$ across sectors.

Aggregate markup. Let $\mu_t(s)$ denote the sector-level markup, implicitly defined by the sector-level labor share

$$\frac{W_t l_t(s)}{p_t(s)y_t(s)} = \frac{(1-\alpha)\zeta_t}{\mu_t(s)} \tag{25}$$

Combining the sector-level labor share with its firm-level counterpart (19) we can write the sales-share of firm i in sector s as

$$\frac{p_{it}(s)y_{it}(s)}{p_t(s)y_t(s)} = \frac{\mu_{it}(s)}{\mu_t(s)} \times \frac{l_{it}(s)}{l_t(s)} \tag{26}$$

Integrating both sides, the sector-level markup can be written *either* as an employment-weighted arithmetic average or a sales-weighted harmonic average of firm-level markups, as in Edmond, Midrigan and Xu (2015),

$$\mu_t(s) \;=\; \int_0^{n_t(s)} \mu_{it}(s)\,\frac{l_{it}(s)}{l_t(s)}\,di \;=\; \left(\int_0^{n_t(s)} \frac{1}{\mu_{it}(s)}\,\frac{p_{it}(s)y_{it}(s)}{p_t(s)y_t(s)}\,di\right)^{-1} \tag{27}$$

where, again, the only difference having a finite number of firms makes is that the integral should be replaced by a finite sum. From either of these and the expression for sector-level productivity $z_t(s)$ we see that the sector-level markup satisfies $p_t(s) = \mu_t(s)\Omega_t/z_t(s)$, i.e., the sector price index can be expressed as the sector-level markup over marginal cost.

Likewise, let $\mathcal{M}_t$ denote the aggregate, economy-wide markup. Following the same steps, this can be written either as an employment-weighted arithmetic average or a sales-weighted harmonic average of sector-level markups

$$\mathcal{M}_t \;=\; \int_0^1 \mu_t(s)\,\frac{l_t(s)}{\tilde{L}_t}\,ds \;=\; \left(\int_0^1 \frac{1}{\mu_t(s)}\,\frac{p_t(s)y_t(s)}{Y_t}\,ds\right)^{-1} \tag{28}$$

The aggregate markup satisfies $1 = \mathcal{M}_t\Omega_t/Z_t$, i.e., the aggregate price level (normalized to one) is the aggregate markup over aggregate marginal cost. We discuss these and related measures of average markups in more detail in Appendix A.

Markup dispersion and productivity. To see how markup dispersion affects productivity, observe from (6) that sector size $q_t(s) = y_t(s)/Y_t$ satisfies $q_t(s) = p_t(s)^{-\eta}$ and since $p_t(s) = \mu_t(s)\Omega_t/z_t(s)$ and $1 = \mathcal{M}_t\Omega_t/Z_t$ we can write

$$q_t(s) = \left(\frac{\mu_t(s)}{\mathcal{M}_t}\,\frac{Z_t}{z_t(s)}\right)^{-\eta} \tag{29}$$

Plugging this into our expressions for aggregate productivity and solving for Z_t we obtain

$$Z_t = \left(\int_0^1 \left(\frac{\mu_t(s)}{\mathcal{M}_t}\right)^{-\eta} z_t(s)^{\eta-1}\,ds\right)^{\frac{1}{\eta-1}} \tag{30}$$

In turn, sector-productivity $z_t(s)$ and markups $\mu_t(s)$ depend on the distribution of firm-level productivity $z_i(s)$ and markups $\mu_{it}(s)$ within sector s — but the details of this layer of aggregation *do* depend on the within-sector market structure.

2.2 Role of market structure

In this section we explain how the details of within-sector market structure matter. First, the market structure matters for the determining the relative size distribution $q_{it}(s) = y_{it}(s)/y_t(s)$ within each sector s. That said, *taking* $n_t(s)$ *as given*, we can cover a range of popular specifications in a unified way, as explained below. Second, and more substantively, the market structure matters for the entry problem that determines $n_t(s)$. The entry problem is simple with monopolistic competition but more involved with oligopolistic competition.[7]

Relative size distribution. Taking $n_t(s)$ as given, the relative size distribution $q_{it}(s)$ within sector s is pinned down by the static markup-pricing condition (17). To cover alternative specifications in a unified way, we write this as

$$f(q) = \frac{\sigma(q)}{\sigma(q)-1}\,\frac{A_t(s)}{z_i(s)}, \qquad A_t(s) := \frac{\Omega_t}{p_t(s)d_t(s)} \tag{31}$$

where the function $f(q)$ is proportional to the inverse demand curve, $\sigma(q)$ is the associated demand elasticity, with markup $\mu(q) = \sigma(q)/(\sigma(q)-1)$, and where $p_t(s)$ is the price index for sector s and $d_t(s)$ is a demand index that depends on the market structure.[8] Let $q(z\,;A)$ denote the solution to $p(q) = \mu(q)A/z$ for arbitrary $A > 0$. We then pick the specific value of A that satisfies the within-sector aggregator. For example:

(i) Monopolistic competition with Kimball demand. Let sector s consist of a mass $n_t(s) > 0$ firms and let sector output be given implicitly by the Kimball aggregator

$$\int_0^{n_t(s)} \Upsilon\Big(\frac{y_{it}(s)}{y_t(s)}\Big)\,di = 1 \tag{32}$$

where $\Upsilon(q)$ is strictly increasing and strictly concave. For this specification inverse demand $f(q)$ and the demand elasticity $\sigma(q)$ are given by

$$f(q) = \Upsilon'(q) \qquad \text{and} \qquad \sigma(q) = -\frac{\Upsilon'(q)}{\Upsilon''(q)q} \tag{33}$$

The associated demand index $d_t(s)$ is given by

$$d_t(s) = \left(\int_0^{n_t(s)} \Upsilon'(q_{it}(s))q_{it}(s)\,di\right)^{-1} \tag{34}$$

The scalar $A_t(s) := \Omega_t/p_t(s)d_t(s)$ is then pinned down by satisfying the Kimball aggregator and thus depends on the mass of firms $n_t(s)$.

---

[7] In our model with oligopoly, the *potential* number of firms per sector $n_t(s)$ is endogenous. This problem is challenging because potential entrants anticipate their impact on a sector, and the distribution of sectoral configurations is a very high-dimensional object. By contrast in Atkeson and Burstein (2008), Edmond, Midrigan and Xu (2015), and De Loecker, Eeckhout and Mongey (2021), the potential number of firms is static and exogenous, with firms simply deciding whether to operate or not.

[8] In this notation, a firm of size $q_{it}(s)$ has price $p_{it}(s) = f(q_{it}(s)) \times p_t(s)d_t(s)$.

(ii) Oligopolistic competition with CES demand. Let sector s consist of a finite $n_t(s) \in \mathbb{N}$ firms and let sector output be given by the CES aggregator

$$\sum_{i=1}^{n_t(s)} \Upsilon\Big(\frac{y_{it}(s)}{y_t(s)}\Big) = 1, \qquad \Upsilon(q) = q^{\frac{\gamma-1}{\gamma}} \tag{35}$$

where $\gamma > \eta > 1$ denotes the elasticity of substitution within sector s. Relative to the Kimball specification we have a finite number of firms, hence genuine strategic interactions, but restrict the kernel of the aggregator $\Upsilon(q)$ to be a power function. For this specification inverse demand $f(q)$ is given by

$$f(q) = \Upsilon'(q) = \frac{\gamma-1}{\gamma}\, q^{-\frac{1}{\gamma}} \tag{36}$$

while the demand index is simply

$$d_t(s) = \left(\sum_{i=1}^{n_t(s)} \Upsilon'(q_{it}(s))q_{it}(s)\right)^{-1} = \frac{\gamma}{\gamma-1} \tag{37}$$

Depending on whether competition is in quantities or prices, the demand elasticity facing a firm of size q is given by

$$\sigma(q) = \begin{cases} \left(\dfrac{1}{\eta}q^{\frac{\gamma-1}{\gamma}} + \dfrac{1}{\gamma}(1-q^{\frac{\gamma-1}{\gamma}})\right)^{-1} & \text{[Cournot competition]} \\ \\ \eta q^{\frac{\gamma-1}{\gamma}} + \gamma(1-q^{\frac{\gamma-1}{\gamma}}) & \text{[Bertrand competition]} \end{cases} \tag{38}$$

where $q^{\frac{\gamma-1}{\gamma}}$ is the sales share of a firm of size q, equal to the kernel of the aggregator $\Upsilon(q)$ in the CES case but not in general. The scalar $A_t(s) := \Omega_t/p_t(s)d_t(s)$ is pinned down by satisfying the CES aggregator and thus depends on $n_t(s)$.

With the relative size distribution $q_{it}(s)$ solved for in this way, we then know the distribution of markups $\mu_{it}(s) = \mu(q_{it}(s))$ and hence can compute sector-level productivity $z_t(s)$ and markups $\mu_t(s)$ and then aggregate productivity Z_t and the aggregate markup $\mathcal{M}_t$.

Entry and exit. Firms enter by paying a sunk cost κ in units of labor and then obtain a one-time productivity draw $z_i(s) \sim G(z)$ in a randomly allocated sector $s \in [0,1]$. Let $N_t = \int_0^1 n_t(s)\,ds$ denote the aggregate mass of firms and let $M_t = \int_0^1 m_t(s)\,ds$ denote the aggregate mass of entrants. With a continuum of sectors, entry per sector $m_t(s)$ is IID Poisson with rate parameter M_t.[9] Firms operate in their sector, obtaining a stream of profits $\pi_{it}(s)$, until they are hit with an IID exit shock, which happens with probability φ per period. The aggregate mass of firms thus evolves according to

$$N_{t+1} = (1-\varphi)N_t + M_t \tag{39}$$

---

[9] With a finite number of sectors S, entry per sector $m_t(s)$ would be IID Binomial with number of trials M_tS and success per trial $1/S$. Taking $S \to \infty$ this converges to a Poisson with rate parameter M_t.

Free entry condition. Now consider the decision problem of a potential entrant. In all versions of our model, entry occurs to the point at which ex ante expected discounted profits are offset by the sunk cost

$$\kappa W_t \geq \beta \sum_{j=1}^{\infty} (\beta(1-\varphi))^{j-1} \frac{C_t}{C_{t+j}} \int_0^1 \bar{\pi}_{t+j}(s)\, ds \tag{40}$$

with strict equality whenever $M_t > 0$ and where $\bar{\pi}_t(s)$ denotes expected profits conditional on operating in sector s. Where these market structures differ is in how these expected profits are calculated. Under *monopolistic competition*, with a continuum $[0, n_t(s)]$ of firms per sector, the entry of any individual firm i has no effect on sector-level variables. But under *oligopolistic competition*, with a finite $n_t(s) \in \mathbb{N}$ firms per sector, the entry of a new firm has non-negligible effects on post-entry sector-level variables. Specifically:

(i) MONOPOLISTIC COMPETITION. Let $\pi_t(z_i, s) := \pi_{it}(s)$ denote the ex post profits of an individual firm with productivity draw z_i in sector s. In the monopolistic competition case, the expected profits conditional on operating in sector s are equal to the average profits of the incumbent firms in that sector

$$\bar{\pi}_t(s) = \int \pi_t(z_i, s)\, dG(z_i) \tag{41}$$

(ii) OLIGOPOLISTIC COMPETITION. Let $\boldsymbol{z}(s)$ denote a sector-specific vector

$$\boldsymbol{z}(s) = (\, z_1(s)\, ,\, z_2(s)\, ,\, \ldots\, ,\, z_{n_t(s)}(s)\,) \tag{42}$$

of $n_t(s)$ independent draws from $G(z)$. Let $\pi_t(z_i, \boldsymbol{z}(s))$ denote the ex post profits of an individual firm with productivity z_i in a sector with $n_t(s)$ other firms with productivities $\boldsymbol{z}(s)$. The free-entry condition is again given by equation (40) but now the expected profits conditional on operating in sector s are given by

$$\bar{\pi}_t(s) = \iint \pi_t(z_i, \boldsymbol{z}(s))\, dG_{n_t(s)}(\boldsymbol{z}(s))\, dG(z_i) \tag{43}$$

where $G_{n_t(s)}(\boldsymbol{z}(s)) = G(z_1) \times G(z_2) \times \cdots \times G(z_{n_t(s)}(s))$ denotes the joint distribution of the vector $\boldsymbol{z}(s)$. In the oligopolistic competition case, the *expected* profits from entering sector s are no longer equal to the *average* profits of those that do operate in sector s. There are two reasons for this. First, even if firms were identical, an entrant of non-negligible size would reduce the market shares of incumbents, tending to decrease expected profits. Second, sectors are heterogeneous, even two sectors with the same $n_t(s)$ will have different samples $\boldsymbol{z}(s)$, and, given this heterogeneity, Jensen's inequality can push expected profits *above* average profits.

2.3 Equilibrium

Given an initial mass of firms $n_0(s)$ per sector and an aggregate capital stock K_0, an *equilibrium* is (i) a sequence of firm prices $p_{it}(s)$ and allocations $y_{it}(s)$, $k_{it}(s)$, $l_{it}(s)$, $x_{it}(s)$ and (ii) aggregate gross output Y_t, consumption C_t, investment I_t, materials X_t, labor L_t, wage rate W_t, rental rate R_t, and mass of entrants M_t such that firms and consumers optimize and the labor, capital and goods markets all clear. In particular

$$L_t = \iint l_{it}(s)\, di\, ds + \kappa M_t \tag{44}$$

$$K_t = \iint k_{it}(s)\, di\, ds \tag{45}$$

$$X_t = \iint x_{it}(s)\, di\, ds \tag{46}$$

(or the equivalent finite sums over i in the case of oligopolistic competition). Note that κM_t denotes labor used in the entry of new firms.

Solving the model. We discuss the solution method in Appendix B. The key to solving the model it to recognize that aggregate markups $\mathcal{M}_t$, aggregate productivity Z_t and aggregate expected profits $\bar{\Pi}_t := \int_0^1 \bar{\pi}_t(s)\, ds$, are given by *time-invariant* functions of the aggregate mass of firms N_t, independent of all other aggregate variables, say $\mathcal{M}_t = \mathcal{M}(N_t)$, $Z_t = Z(N_t)$, and $\bar{\Pi}_t = \Pi(N_t)$. These functions summarize all the implications of market structure for aggregate outcomes. We solve the model by interpolating these functions and then use the remaining conditions, i.e., the production functions, input choices, optimality conditions of the representative consumer, and our aggregation results to simultaneously determine $Y_t, C_t, I_t, X_t, L_t, W_t, R_t, M_t$ given the state variables N_t and K_t.

3 Efficient allocation

In this section we derive the efficient allocation in our economy by considering the problem of a benevolent planner who faces the same technological and resource constraints as in the decentralized economy. Comparing the efficient allocation chosen by the planner to the decentralized allocation reveals three channels through which markups distort outcomes in the decentralized economy: (i) the aggregate markup acts like a uniform output tax, (ii) markup dispersion gives rise to misallocation of factors of production, and (iii) markups distort the entry margin.

3.1 Planner's problem

The planner chooses how many varieties to create, how to allocate inputs, consumption, investment, and employment so as to maximize the representative consumer's utility taking as given the resource constraints for capital, labor and goods and the production functions for individual varieties. To facilitate comparisons with the decentralized equilibrium, the planner *cannot direct* the creation of new varieties towards specific sectors. We use asterisks to denote variables in the planner's problem.

The planner's problem has two parts: (i) a static allocation problem that determines aggregate productivity, and (ii) a dynamic problem that determines aggregate investment in new varieties, aggregate investment in physical capital, and aggregate employment. The link between the two parts is that the aggregate productivity solving the static allocation problem is a function of the stock of varieties, $Z_t^* = Z(N_t^*)$, which the planner internalizes when choosing how many varieties to create.

Dynamic problem. Starting with the dynamic problem, just as in the decentralized problem, we can use the resource constraints for capital, labor and goods and the production functions for individual varieties to derive the aggregate production function (22). We can then write the the planner's problem as maximizing

$$\sum_{t=0}^{\infty} \beta^t \Big(\log C_t^* - \psi \frac{\big(\tilde{L}_t^* + \kappa(N_{t+1}^* - (1-\varphi)N_t^*)\big)^{1+\nu}}{1+\nu} \Big) \tag{47}$$

subject to the resource constraint for goods,

$$C_t^* + K_{t+1}^* + X_t^* = Z(N_t^*)F(K_t^*, \tilde{L}_t^*, X_t^*) + (1-\delta)K_t^* \tag{48}$$

taking as given the function $Z(N_t^*)$ implied by the static allocation problem. The initial conditions for this problem are the mass of varieties N_0 and capital stock K_0.

The planner's optimality conditions for consumption, investment, and employment are standard. The shadow wage is equated to the marginal product of labor

$$\psi C_t^* \tilde{L}_t^{*\,\nu} = Z_t^* F_{L,t}^* \tag{49}$$

while the marginal product of capital satisfies

$$1 = \beta \frac{C_t^*}{C_{t+1}^*} \Big(Z_{t+1}^* F_{K,t+1}^* + 1 - \delta \Big) \tag{50}$$

and the marginal product of materials is simply $Z_t^* F_{X,t}^* = 1$. Comparing these conditions with their decentralized counterparts, we see that the aggregate markup $\mathcal{M}_t$ acts like a uniform output tax, reducing the overall scale of production and hence reducing the use of all inputs relative to the planner's problem.

Planner's choice of varieties. Now consider the planner's choice of varieties N_{t+1}^*. Letting $W_t^* = \psi C_t^* \tilde{L}_t^{*\,\nu}$ denote the shadow wage, we can write the first order condition

$$\kappa W_t^* = \beta(1-\varphi)\kappa W_{t+1}^* + \beta \frac{C_t^*}{C_{t+1}^*}\left(\frac{dZ_{t+1}^*}{dN_{t+1}^*}\frac{N_{t+1}^*}{Z_{t+1}^*}\right)\frac{Y_{t+1}^*}{N_{t+1}^*} \tag{51}$$

Iterating forward this gives

$$\kappa W_t^* = \beta \sum_{j=1}^{\infty}(\beta(1-\varphi))^{j-1}\frac{C_t^*}{C_{t+j}^*}\left(\frac{dZ_{t+j}^*}{dN_{t+j}^*}\frac{N_{t+j}^*}{Z_{t+j}^*}\right)\frac{Y_{t+j}^*}{N_{t+j}^*} \tag{52}$$

This is the planner's counterpart to the free-entry condition in the decentralized problem. In the decentralized problem, a firm's incentive to enter is given by its expected discounted profits, which depend on its markup and sales. By contrast, the planner's incentive to create new varieties depends on the elasticity of aggregate productivity with respect to the mass of firms — and this depends on the solution to the static allocation problem.

Static allocation problem. Now consider the problem of maximizing aggregate productivity Z_t^* taking as given $n_t(s)$. The allocation of activity across sectors $q_t^*(s) = y_t^*(s)/Y_t^*$ is given by $q_t^*(s) = (z_t^*(s)/Z_t^*)^\eta$ so that in terms of sector-level productivity, aggregate productivity is $Z_t^* = (\int_0^1 z_t^*(s)^{\eta-1}\,ds)^{1/(\eta-1)}$, i.e., as in (30) but with no dispersion in sector-level markups. In turn, the allocation of activity within sectors $q_{it}^*(s) = y_{it}^*(s)/y_t^*(s)$ is given by

$$\Upsilon'(q_{it}^*(s))d_t^*(s) = \frac{z_t^*(s)}{z_i(s)} \tag{53}$$

where $d_t^*(s)$ is the planner's demand index, the counterpart of (34) or (37). In other words, at the optimum the planner's shadow value of a variety is simply the planner's marginal cost of producing it. This optimality condition holds for both our monopolistic competition model with Kimball demand and our oligopolistic competition model with CES demand. As in the decentralized problem, the scalar $z_t^*(s)/d_t^*(s)$ is pinned down by satisfying the within-sector aggregator. In our oligopolistic competition model with CES demand this gives sector-level productivity $z_t^*(s) = (\sum_{i=1}^{n_t(s)} z_{it}^*(s)^{\gamma-1})^{1/(\gamma-1)}$ with constant demand index $d_t^*(s) = \frac{\gamma}{\gamma-1}$.

There is *misallocation* in the sense of Hsieh and Klenow (2009) whenever there is variation in marginal revenue products across firms, i.e., when the equilibrium $q_{it}(s)$ does not coincide with the planner's $q_{it}^*(s)$. This happens whenever markups $\mu_{it}(s)$ vary across firms.

Value of an additional variety. Now consider the value to the planner of an additional variety. Abstracting from any integer constraints on $n_t(s)$, an application of the envelope theorem gives

$$\frac{dZ_t^*}{dn_t(s)}\frac{n_t(s)}{Z_t^*} = \left(d_t^*(s) - 1\right) q_t^*(s)\,\frac{Z_t^*}{z_t^*(s)} \tag{54}$$

To interpret this condition, we use the planner's demand index to write

$$d_t^*(s) - 1 = \int_0^{n_t(s)} \left(\epsilon_{it}^*(s) - 1\right) p_{it}^*(s) q_{it}^*(s)\, di \tag{55}$$

(or the equivalent finite sum in the case of oligopolistic competition), where we define

$$\epsilon_{it}^*(s) := \frac{\Upsilon(q_{it}^*(s))}{\Upsilon'(q_{it}^*(s))q_{it}^*(s)}, \qquad \text{and} \qquad p_{it}^*(s) := \Upsilon'(q_{it}^*(s))d_t^*(s) \tag{56}$$

The term $\epsilon_{it}^*(s)$ is the *inverse elasticity* of the within-sector aggregator $\Upsilon(q)$ evaluated at the planner's allocation for a particular variety $q_{it}^*(s)$. The term $p_{it}^*(s)$ is the social value of an additional unit of that variety, i.e., the planner's counterpart to the market price.

Comparing the free-entry condition in the decentralized equilibrium to the planner's entry condition, we recover an important insight of Bilbiie, Ghironi and Melitz (2008, 2019), Zhelobodko, Kokovin, Parenti and Thisse (2012) and Dhingra and Morrow (2019), namely that the planner's incentives to create new varieties are determined by the inverse elasticity $\epsilon_{it}^*(s)$ of the aggregator while the incentives for new firms to enter are determined by their markups $\mu_{it}(s)$. Whether there is too much or too little entry compared to the planner's allocation is in general ambiguous and depends on precise details of the parameterization.

To summarize, variable markups distort outcomes in the decentralized economy through three channels: (i) the aggregate markup $\mathcal{M}_t$ acts like a uniform output tax, (ii) markup dispersion $\mu_{it}(s)$ gives rise to misallocation of factors of production, and (iii) markups distort the entry margin.

4 Quantifying the model

In this section we outline our parameterization and calibration strategy and our model's implications for the cross-sectional distribution of markups. We then calculate the aggregate productivity losses due to misallocation.

4.1 Benchmark parameterization

Kimball demand. To this point we have stressed aggregation results that hold regardless of the details of market structure within each sector. But to quantify the model we need to take a stand on demand and market structure. For our benchmark model we assume *monopolistic competition with Kimball demand*, as in (32) above. In particular, we assume the Kimball aggregator has the functional form introduced by Klenow and Willis (2016).

This specification implies that inverse demand curves are given by[10]

$$\Upsilon'(q) = \frac{\bar{\sigma}-1}{\bar{\sigma}} \exp\left(\frac{1-q^{\,\varepsilon/\bar{\sigma}}}{\varepsilon}\right), \qquad \bar{\sigma} > 1 \tag{57}$$

which in turn implies that the demand elasticity $\sigma(q)$ is log-linear in relative size

$$\sigma(q) := -\frac{\Upsilon'(q)}{\Upsilon''(q)q} = \bar{\sigma}\, q^{-\varepsilon/\bar{\sigma}} \tag{58}$$

The parameter $\varepsilon/\bar{\sigma}$ is the elasticity of the demand elasticity with respect to relative size and is often known as the *super-elasticity*. If $\varepsilon = 0$ we have the constant demand elasticity $\sigma(q) = \bar{\sigma}$. If $\varepsilon > 0$, relatively large firms will face less elastic demand and charge high markups. If $\varepsilon < 0$, relatively large firms will face more elastic demand and charge low markups.

Productivity distribution. For parsimony and as is standard in the literature we assume that the distribution of productivity $G(z)$ is Pareto with tail parameter ξ.

Calibration strategy. We assign values to a number of conventional macro parameters that are held constant through all our quantitative exercises. We calibrate the parameters of the demand system and the productivity distribution to match facts on the amount of sales concentration and the relationship between markups and market shares in US data.

Assigned parameters. We assume that a period is one year and set the discount factor $\beta = 0.96$ and depreciation rate $\delta = 0.06$. We set the exit rate to $\varphi = 0.04$ to match the employment share of exiting firms, as in Boar and Midrigan (2020). We set the elasticity of value-added to capital $\alpha = 1/3$ and set the elasticity of substitution between value-added and materials to $\theta = 0.5$, both conventional values. Preferences (1) are homothetic and consistent with balanced growth. We set the inverse of the Frisch elasticity of labor supply to $\nu = 1$. We normalize the disutility from labor supply ψ and the entry cost κ to achieve a steady-state output of $Y = 1$ and a steady-state total mass of firms $N = 1$ for our benchmark economy. We report these parameter choices in Panel A of Table 1.

Calibrated parameters. The level and dispersion of markups in our benchmark model depend crucially on three underlying parameters: (i) the Pareto tail parameter ξ, (ii) the super-elasticity $\varepsilon/\bar{\sigma}$ that determines the sensitivity of a firm's demand elasticity to its relative

---

[10] The aggregator $\Upsilon(q)$ itself is given by

$$\Upsilon(q) = 1 + (\bar{\sigma}-1)\exp\left(\frac{1}{\varepsilon}\right)\varepsilon^{\frac{\bar{\sigma}}{\varepsilon}-1}\left[\Gamma\left(\frac{\bar{\sigma}}{\varepsilon},\frac{1}{\varepsilon}\right) - \Gamma\left(\frac{\bar{\sigma}}{\varepsilon},\frac{q^{\varepsilon/\bar{\sigma}}}{\varepsilon}\right)\right]$$

where $\Gamma(s,x) := \int_x^\infty t^{s-1}e^{-t}dt$ denotes the upper incomplete Gamma function.

Table 1: Parameterization

Panel A: Assigned Parameters

β	discount factor	0.96
δ	depreciation rate	0.06
φ	exit rate	0.04
α	elasticity of value-added to capital	1/3
ν	elasticity of labor supply	1
θ	elasticity of substitution between value-added and materials	0.5

Panel B: Calibrated Parameters

			low	medium	high
ξ	Pareto tail		20.70	6.84	4.07
$\bar{\sigma}$	demand elasticity		29.10	10.86	7.21
$\varepsilon/\bar{\sigma}$	super-elasticity		0.16	0.16	0.16
ϕ	weight on value-added		0.51	0.43	0.33
calibration targets		*data*			
$\mathcal{M}$	aggregate markup		1.05	1.15	1.25
	top 5% sales share	0.57	0.57	0.57	0.57
	materials share	0.45	0.45	0.45	0.45
$\hat{b}$	regression coefficient	0.16	0.16	0.16	0.16

Panel A reports assigned parameters held constant through all our quantitative exercises. Panel B reports calibrated parameters for our benchmark model with monopolistic competition and Kimball demand. We report three cases, *low*, *medium*, and *high*, which refer to alternative targets for the level of the aggregate markup, $\mathcal{M} = 1.05, 1.15$, and 1.25 respectively. For each $\mathcal{M}$ we calibrate the Pareto tail ξ, demand elasticity $\bar{\sigma}$, super-elasticity $\varepsilon/\bar{\sigma}$ and weight on value-added ϕ to match the targets shown in Panel B. For each model we choose the super-elasticity $\varepsilon/\bar{\sigma}$ so that the slope coefficient b from equation (59) in the model matches the estimated slope coefficient $\hat{b}$. See the text for more details.

size, and (iii) the ‘average’ demand elasticity $\bar{\sigma}$. Intuitively, the Pareto tail parameter ξ is pinned down by the amount of concentration in the distribution of firm size, the super-elasticity $\varepsilon/\bar{\sigma}$ is pinned down by the cross-sectional relationship between markups and market shares, and $\bar{\sigma}$ is pinned down by the overall level of markups. Specifically we target:

(i) Sales concentration. The Pareto tail parameter ξ is pinned down by our target for sales concentration. We target the average sales share of the top 5% of firms (by industry market share) in 6-digit NAICS industries. For 2012 US manufacturing, the top 5% of firms account for 57% of sales.

(ii) Relationship between markups and market shares. The super-elasticity $\varepsilon/\bar{\sigma}$ is pinned down by the relationship between markups and market shares in our model. In particular, within a given industry markups are a function of relative size, $\mu_{it}(s) = \mu(q_{it}(s))$, and market shares $\omega_{it}(s) := p_{it}(s)q_{it}(s)$ are a function of relative size, $\omega_{it}(s) \sim \Upsilon'(q_{it}(s))q_{it}(s)$ up to a constant. Eliminating $q_{it}(s)$ between these expressions gives

$$\frac{1}{\mu_{it}(s)} + \log\left(1 - \frac{1}{\mu_{it}(s)}\right) = \text{constant} \ + \ b \log \omega_{it}(s), \qquad b = \varepsilon/\bar{\sigma} \tag{59}$$

The LHS is a strictly increasing function of the markup $\mu_{it}(s)$. Thus the slope coefficient b on the RHS is an estimate of the strength of the within-industry relationship between market shares and markups.[11] As discussed below, we estimate this regression on firm-level data from the US Census of Manufactures and obtain a precisely estimated $\hat{b} = 0.16$. In our benchmark model, the slope coefficient b in this regression *is* the super-elasticity so for our benchmark model we set $\varepsilon/\bar{\sigma} = \hat{b} = 0.16$. The super-elasticity $\varepsilon/\bar{\sigma}$ is specific to Kimball demand. In other versions of our model with different demand systems we use indirect inference, choosing parameters so that the slope coefficient in the model matches the estimated slope coefficient $\hat{b} = 0.16$ in this regression.

(iii) Aggregate Markup. The average elasticity $\bar{\sigma}$ is pinned down by our target for the aggregate markup $\mathcal{M}$. The recent literature on markups in the US economy provides a wide range of estimates for $\mathcal{M}$.[12] Given this range of estimates, rather than commit to a single target for the aggregate markup, for our benchmark model we recalibrate $\bar{\sigma}$ (jointly, with our other parameters) for $\mathcal{M}$ ranging from 1.05 to 1.35.

Finally, we calibrate the weight ϕ on value-added in the gross-output production function by targeting a materials share of 45% for the US economy in 2012. For each $\mathcal{M}$ we calibrate this parameter jointly with the three key parameters $\xi, \varepsilon/\bar{\sigma}$, and $\bar{\sigma}$ as discussed above.

---

[11] This is similar to how we estimated the within-industry relationship between market shares and markups in Edmond, Midrigan and Xu (2015) but adapted to the Kimball demand system used here.

[12] See e.g., Atkeson, Burstein and Chatzikonstantinou (2019), Barkai (2020), De Loecker, Eeckhout and Unger (2020), Gutiérrez and Phillippon (2017a,b), and Hall (2018) etc. Basu (2019) surveys this literature.

Regression details. As discussed in Appendix C, to implement the regression (59), we construct $1/\mu_{it}(s)$ using data from the US Census of Manufactures from 1972 to 2012. In particular, the firm-level labor share (19) implies[13]

$$\frac{1}{\mu_{it}(s)} = \frac{W_t l_{it}(s)}{p_{it}(s) y_{it}(s)} \times \frac{1}{(1-\alpha)\zeta_t} \tag{60}$$

Cost minimization implies that, for each firm,

$$(1-\alpha)\zeta_t = \frac{W_t l_{it}(s)}{W_t l_{it}(s) + R_t k_{it}(s) + x_{it}(s)} \tag{61}$$

Our key assumption is that the elasticity $(1-\alpha)\zeta_t$ is *common to all firms within an industry*. We estimate this elasticity by aggregating (61) over firms within each 6-digit NAICS industry.[14] We allow this elasticity to vary over time by constructing it for each Census year. We then have an estimate of $(1-\alpha)\zeta_t$ that we can plug back into (60) to construct $1/\mu_{it}(s)$.

Our simple model abstracts from other sources of persistent firm-level and industry-level heterogeneity, such as variation in capital shares or returns to scale, and from other distortions (implicit or explicit taxes, etc) that may drive a *wedge* between cost shares and $1/\mu_{it}(s)$. Given this, in our empirical analysis we control for firm fixed effects in addition to the 6-digit NAICS industry-year effects. We obtain an estimated slope coefficient $\hat{b} = 0.162$ with standard error 0.002 clustered at the firm level.

An alternative to this simple cost share approach would be to estimate sector-specific production functions. But recent work by Bond, Hashemi, Kaplan and Zoch (2020) demonstrates that in the presence of variable markups it is not possible to consistently estimate output elasticities when only revenue data is available. Using the simple labor cost share approach also makes our results easier to compare to recent empirical work, such as Autor, Dorn, Katz, Patterson and Van Reenen (2020) and De Loecker, Eeckhout and Unger (2020), that also report such measures.

Model fit. Panel B of Table 1 reports the parameter values that minimize our objective function for three values of the aggregate markup $\mathcal{M} = 1.05$ ('low'), $\mathcal{M} = 1.15$ ('medium') and $\mathcal{M} = 1.25$ ('high'). To match a low level of markups $\mathcal{M} = 1.05$ while targeting a top 5% sales share of 0.573 requires a high average demand elasticity $\bar{\sigma} = 29.1$ and a thin-tailed productivity distribution $\xi = 20.7$. To match a high level of markups $\mathcal{M} = 1.25$ while targeting the same top 5% sales share requires a much lower average demand elasticity $\bar{\sigma} = 7.21$ and a fatter-tailed productivity distribution $\xi = 4.07$. Though our estimate of $\varepsilon/\bar{\sigma} = 0.162$ is much lower than typically assumed in macro studies that attempt to match

---

[13] To deal with multi-establishment firms we construct $1/\mu_{eit}(s)$ for each establishment e and then aggregate to $1/\mu_{it}(s)$ using the labor cost share of each establishment e of firm i.

[14] These calculations also use industry-specific user costs of capital from the NBER-CES and BLS.

Table 2: Markup Distribution

	low	medium	high
cost-weighted distribution of markups			
aggregate markup, $\mathcal{M}$	1.05	1.15	1.25
p25 markup	1.04	1.11	1.17
p50 markup	1.05	1.14	1.23
p75 markup	1.06	1.18	1.31
p90 markup	1.07	1.23	1.40
p99 markup	1.11	1.35	1.63
productivity losses from misallocation			
$\frac{Z^*-Z}{Z} \times 100$	0.28	0.97	1.86

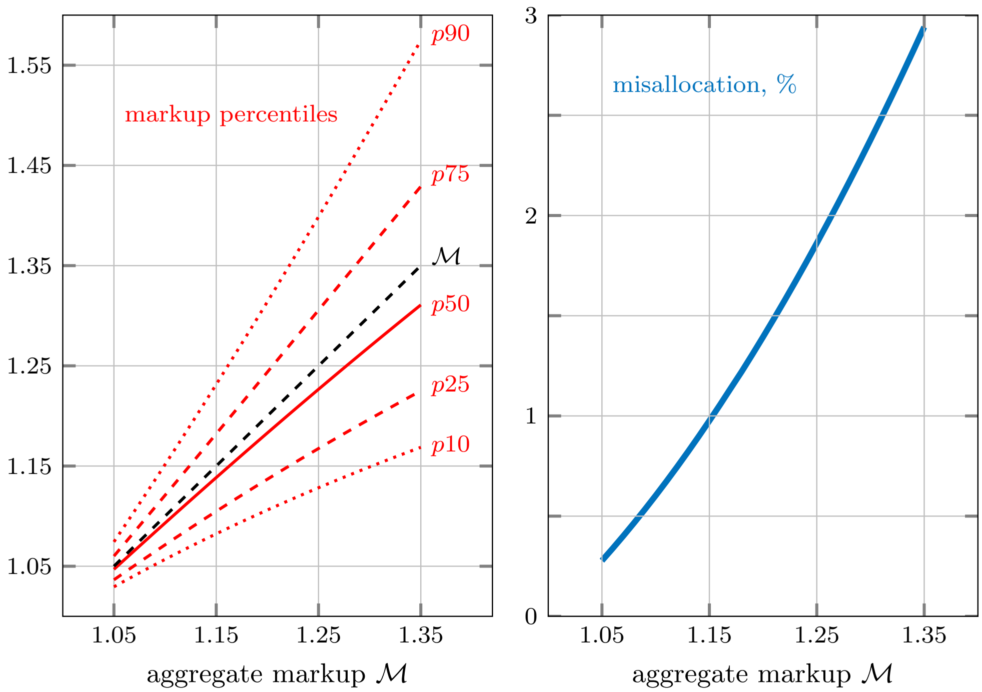
the response of prices to changes in monetary policy or exchange rates, it is in line with the micro estimates surveyed by Klenow and Willis (2016). In Appendix C we find an almost identical super-elasticity $\varepsilon/\bar{\sigma} = 0.16$ best fits the relationship between markups and market shares in the Taiwanese manufacturing firms studied by Edmond, Midrigan and Xu (2015).

4.2 Markups and misallocation

Markup distribution. Table 2 reports the cost-weighted percentiles of the steady-state distribution of markups in our model for the same three values of the aggregate markup $\mathcal{M} = 1.05$ ('low'), $\mathcal{M} = 1.15$ ('medium') and $\mathcal{M} = 1.25$ ('high'). As we target higher levels of the aggregate markup $\mathcal{M}$ the model implies more markup dispersion. This occurs because as we target higher $\mathcal{M}$, requiring a lower average demand elasticity $\bar{\sigma}$, we need a fatter-tailed productivity distribution to hold the top 5% sales share unchanged. In turn, a fatter-tailed productivity distribution creates more large firms who charge large markups, increasing markup dispersion. We illustrate this in Figure 1 using a fine grid for $\mathcal{M}$.

Misallocation. The markup dispersion generated by our model implies that there are aggregate productivity losses due to misallocation. We measure the amount of misallocation by comparing aggregate productivity Z in the steady state of our benchmark economy to the level of aggregate productivity Z^* that could be achieved by a planner facing the same technology and resource constraints who could reallocate factors of production across producers. We illustrate this in Figure 2, which compares the relative size $q(z)$ and employment $l(z)$ of a

Figure 1: Markup Distribution and Misallocation in Benchmark Model



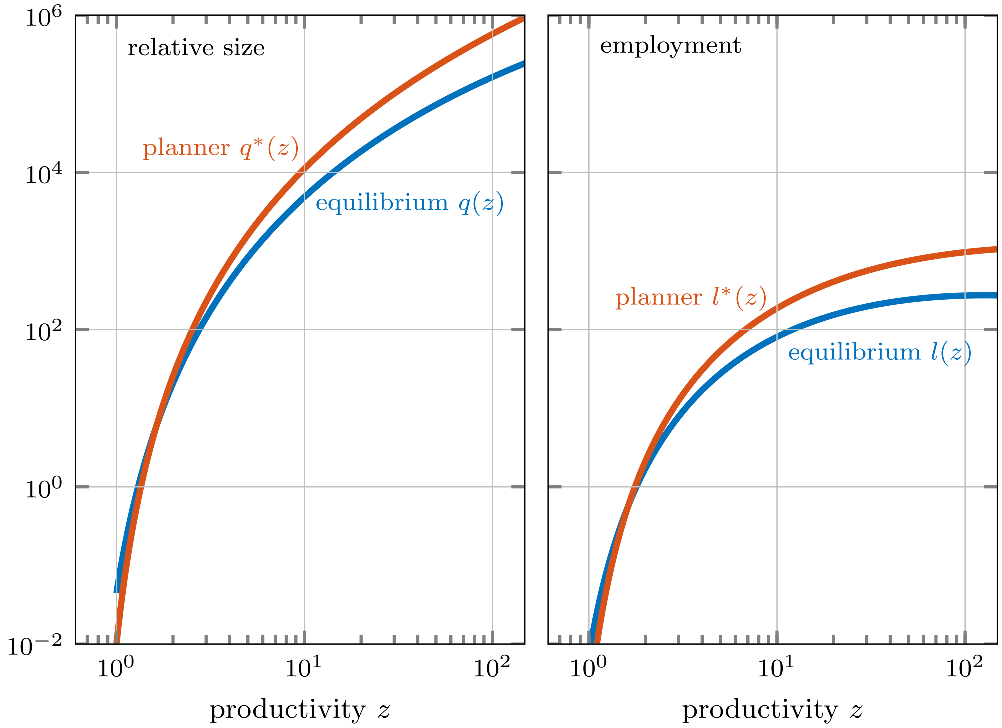
The left panel shows cost-weighted percentiles of the markup distribution in our benchmark model with monopolistic competition and Kimball demand for a range of targets for the aggregate markup $\mathcal{M}$. For each $\mathcal{M}$ we recalibrate the Pareto tail ξ, demand elasticity $\bar{\sigma}$, super-elasticity $\varepsilon/\bar{\sigma}$ and weight on value-added ϕ to match our calibration targets as given in Table 1. The right panel shows the implied amount of misallocation for each $\mathcal{M}$.

firm with productivity z in the decentralized equilibrium to the planner's counterparts $q^*(z)$ and $l^*(z)$. More productive firms have higher markups and produce and employ too little compared to the planner's allocation. Less productive firms produce and employ too much compared to the planner's allocation. As shown in the right panel of Figure 1, aggregate productivity Z in our benchmark economy is on the order of 1% or 2% below the level of aggregate productivity Z^* that could be achieved by a planner.

Notice in Figure 2 that the planner's allocation is not log-linear in productivity, as it would be with CES demand. The extra concavity reflects strongly diminishing marginal productivity as the relative size q increases. If misallocation losses were calculated assuming a constant demand elasticity $\bar{\sigma}$ rather than variable demand elasticities $\sigma(q) = \bar{\sigma} q^{-\varepsilon/\bar{\sigma}}$ we would find higher misallocation (for a given amount of dispersion in marginal revenue products) because we would overstate the gains from reallocating factors from small, less productive firms to large, more productive firms.

Comparison with Baqaee and Farhi (2020). In related work, Baqaee and Farhi (2020) calculate that the aggregate productivity gains from eliminating all markups are about 20%, an order of magnitude larger than in our model. Why do they find much larger losses from

Figure 2: Equilibrium and Planner Allocations



The left panel shows the equilibrium relative size $q(z)$ and the planner's relative size $q^*(z)$ as functions of productivity for our benchmark economy with $\mathcal{M} = 1.15$. The right panel shows the equilibrium employment $l(z)$ and the planner's employment $l^*(z)$ for the same economy. More productive firms have higher markups and produce too little and employ too little compared to the planner's allocation. Less productive firms produce too much and employ too much compared to the planner's allocation. In this figure aggregate employment in the decentralized equilibrium is the same as aggregate employment for the planner. Our measure of *misallocation* is the aggregate output loss implied by the equilibrium allocation relative to the planner's allocation.

misallocation? There are two points of difference. First, we report gross output losses rather than value-added losses. As is well-understood, a high materials share substantially amplifies value-added TFP losses relative to gross output TFP losses. Second, for 'markup dispersion' they use the *total amount of dispersion in average revenue products* estimated in the literature (e.g., as in De Loecker, Eeckhout and Unger, 2020; Gutiérrez and Phillippon, 2017b), whereas we use only *that small fraction of dispersion in marginal revenue products explained by market shares.* That is, they use the word 'markup' to refer to all the Hsieh and Klenow (2009) distortions in average revenue products, whereas we use the word 'markup' to refer to the markup distribution implied by our model, calibrated to the US firm size distribution. Because they use the total dispersion, they find much larger losses from misallocation.[15]

---

[15] See Eslava and Haltiwanger (2020) who study the life-cycle of Colombian manufacturing plants and find that markup variation plays only a small role in accounting for variation in average revenue products.

5 How costly are markups?

We now present our main results on the welfare costs of markups. We first quantify the total welfare costs of markups in our benchmark economy for a range of values for the aggregate markup $\mathcal{M}$. We then show how the efficient allocation can be implemented by a specific nonlinear schedule of size-dependent subsidies and show how to isolate aspects of this policy to quantify the relative magnitudes of the different markup channels. We also study simple entry subsidies that *indirectly* affect markup distortions through the amount of competition.

We measure the welfare costs of markups by asking how much the representative consumer would benefit from implementing the efficient allocation that eliminates all markup distortions, taking the transitional dynamics into account. We find that the total welfare costs of markups can be large. For example, for an economy with aggregate markup $\mathcal{M} = 1.15$, implementing the efficient allocation results in a consumption-equivalent welfare gain of about 8.7%, rising to 23.6% for an economy with $\mathcal{M} = 1.25$. We find that a uniform output subsidy that offsets the aggregate markup alone goes a long way towards achieving full efficiency.

5.1 Welfare cost of markups

We first compare the distorted steady state in our decentralized equilibrium to that chosen by a planner, then calculate the welfare gains from implementing the efficient steady state taking the transitional dynamics into account.

Steady state comparisons. The first six columns of Table 3 report the percentage change in consumption C, gross output Y, employment L, mass of firms N, physical capital K, and aggregate productivity Z from the initial distorted steady state to the efficient steady state for each of three values of the aggregate markup $\mathcal{M} = 1.05$, $\mathcal{M} = 1.15$, and $\mathcal{M} = 1.25$. The efficient steady state features higher consumption, higher output, and employment. Aggregate productivity is higher, both because of the elimination of misallocation and because of the increase in product variety, i.e., increase in the mass of firms N.[16]

Welfare gains from implementing efficient allocation. The last column of Table 3 reports the welfare gains for the representative consumer in consumption-equivalent units including the transition, i.e., these take into account the deferred increase in consumption as investment in physical capital and product variety accumulates over time. These dynamics also take into account the time path of employment. We find that if the aggregate markup is low, $\mathcal{M} = 1.05$, the representative consumer needs to be compensated with an additional 1.35% consumption per period in order to be indifferent between the initial distorted steady state and the transition to the efficient steady state. This increases to 8.69% consumption

---

[16] We discuss the effects of variety on aggregate productivity in more detail in Appendix F.

Table 3: Implications of Alternative Policies, Benchmark Model

		steady state comparisons, %						
		Y	C	L	N	K	Z	welfare, %
$\mathcal{M} = 1.05$	efficient	15.3	11.0	6.0	12.3	24.1	0.9	1.35
	uniform subsidy	13.7	9.1	5.7	3.4	22.0	0.2	0.65
	size-dependent subsidy	1.4	1.7	0.3	8.1	1.8	0.7	0.72
	entry subsidy	1.1	1.3	0.4	11.2	1.4	0.6	0.07
$\mathcal{M} = 1.15$	efficient	59.7	44.6	18.0	20.1	100.6	4.1	8.69
	uniform subsidy	51.9	35.9	17.0	9.5	88.7	1.5	5.92
	size-dependent subsidy	5.3	6.2	1.0	8.3	6.6	2.3	2.87
	entry subsidy	6.3	7.5	2.4	20.1	8.2	3.1	0.56
$\mathcal{M} = 1.25$	efficient	134.0	102.3	30.1	26.7	246.1	8.9	23.62
	uniform subsidy	112.7	79.7	28.2	15.0	208.2	3.9	17.34
	size-dependent subsidy	10.8	12.5	1.8	8.1	13.6	4.1	6.25
	entry subsidy	17.4	20.4	6.1	29.3	23.0	7.4	1.96

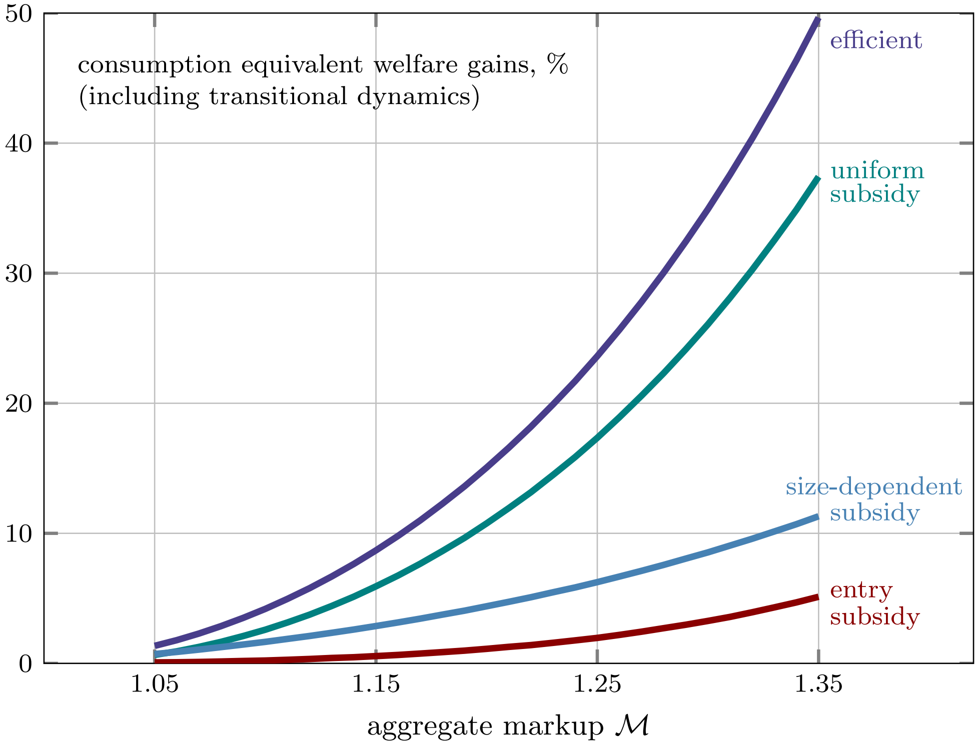
The first six columns report the percentage change from the initial distorted steady state to the new steady state. The last column reports the consumption equivalent welfare gains (including transitional dynamics). For each $\mathcal{M}$ we recalibrate the Pareto tail ξ, demand elasticity $\bar{\sigma}$, super-elasticity $\varepsilon/\bar{\sigma}$ and weight on value-added ϕ. The alternative policies are (i): the *efficient allocation*, where all markups are removed, (ii) a *uniform subsidy* that eliminates the aggregate markup, (iii) *size-dependent subsidies* that eliminate misallocation and the entry distortion, and (iv) the uniform *entry subsidy* that leads to the largest welfare gain.

per period if the aggregate markup is $\mathcal{M} = 1.15$ and to 23.62% consumption per period if the aggregate markup is $\mathcal{M} = 1.25$. The welfare gains are higher when we target higher $\mathcal{M}$. Indeed the gains are *convex* in $\mathcal{M}$. As we target higher $\mathcal{M}$ for the benchmark economy, both the level of markups and the amount of markup dispersion increase. We illustrate this convexity in Figure 3 using a fine grid for $\mathcal{M}$ with the upper bound extended to 1.35.

5.2 Implementing the efficient allocation

We now show how the efficient allocation can be implemented by a specific nonlinear schedule of size-dependent subsidies. This policy removes the aggregate markup distortion, removes markup dispersion (and hence misallocation), and removes the entry distortion. We then show how to isolate different aspects of this policy to quantify the relative magnitudes of the different markup channels. This policy is financed by lump-sum taxes on the representative consumer. We view these calculations as a device for isolating the role of each distortion. The actual consequences of such a policy would of course be much more complex in economies with heterogeneous consumers and other frictions (see Boar and Midrigan, 2020, for example).

Figure 3: Welfare Gains from Alternative Policies



Consumption equivalent welfare gains (including transitional dynamics), from the initial distorted steady state to the new steady state for a range of targets for the aggregate markup $\mathcal{M}$. For each $\mathcal{M}$ we recalibrate the Pareto tail ξ, demand elasticity $\bar{\sigma}$, super-elasticity $\varepsilon/\bar{\sigma}$ and weight on value-added ϕ. The alternative policies are (i): the *efficient allocation*, where all markups are removed, (ii) a *uniform subsidy* that eliminates the aggregate markup, (iii) *size-dependent subsidies* that eliminate misallocation and the entry distortion, and (iv) the uniform *entry subsidy* that leads to the largest welfare gain.

Direct policy intervention to remove markup distortions. In the decentralized equilibrium, the profits of a firm with productivity z facing Kimball demand can be written

$$\pi_t(z) = \Big[\Upsilon'(q_t(z))q_t(z)D_t - \frac{\Omega_t}{z}\,q_t(z)\Big]Y_t \tag{62}$$

where D_t denotes the Kimball demand index from (34) above.[17] Now suppose that firms are paid a size-dependent subsidy $T_t(q)$ given by

$$T_t(q) = \Big[\Upsilon(q) - \Upsilon'(q)q\Big]D_tY_t \tag{63}$$

This policy takes away revenues in proportion to $\Upsilon'(q)q$ and returns revenues in proportion to $\Upsilon(q)$ which will then induce firms to price at marginal cost. In particular, given the subsidy $T_t(q)$, a firm has net profits $\hat{\pi}_t(z) := \pi_t(z) + T_t(q_t(z))$ which simplifies to

$$\hat{\pi}_t(z) = \Big[\Upsilon(q_t(z))D_t - \frac{\Omega_t}{z}\,q_t(z)\Big]Y_t \tag{64}$$

---

[17] In our benchmark economy, sectors $s \in [0,1]$ are ex post identical and we have $d_t(s) = D_t$, $y_t(s) = Y_t$, $p_t(s) = 1$, $z_t(s) = Z_t$, $n_t(s) = N_t$ etc.

This leads to the optimal price

$$p_t(z) = \Upsilon'(q_t(z))D_t = \frac{\Omega_t}{z} \tag{65}$$

In other words, this policy induces firms to price at marginal cost with firm-level wedge $\mu_t(z) = 1$. Hence the aggregate wedge is also $\mathcal{M}_t = 1$. Given this, net profits are equal to the transfer $\hat{\pi}_t(z) = T_t(q_t(z))$ and so the free entry condition becomes

$$\begin{aligned}\kappa W_t &= \beta \sum_{j=1}^{\infty} (\beta(1-\varphi))^{j-1} \frac{C_t}{C_{t+j}} \int \Big[\Upsilon(q_{t+j}(z)) - \Upsilon'(q_{t+j}(z))q_{t+j}(z)\Big] D_{t+j} Y_{t+j}\, dG(z) \\ &= \beta \sum_{j=1}^{\infty} (\beta(1-\varphi))^{j-1} \frac{C_t}{C_{t+j}} \left(D_{t+j} - 1\right) \frac{Y_{t+j}}{N_{t+j}} \end{aligned} \tag{66}$$

where the second line follows using the definitions of the Kimball aggregator (32) and its demand index (34). To see how the free-entry condition under this policy compares to the planner's entry condition, use (54) to write the planner's elasticity of aggregate productivity with respect to new varieties

$$\frac{dZ_t^*}{dN_t^*} \frac{N_t^*}{Z_t^*} = (D_t^* - 1) \tag{67}$$

Plugging this elasticity into the planner's entry condition (52) we see that the free-entry condition under the policy $T_t(q)$ coincides with the planner's entry condition, i.e., this policy also eliminates the entry distortion. We next show how to use a generalization of this policy to isolate and quantify the relative importance of each channel.

5.3 Decomposing the implementation.

The nonlinear schedule $T_t(q)$ directly implements the efficient allocation. To study each channel in isolation, it is helpful to generalize this to

$$T_t(q) = \Big[a_0\, \Upsilon(q) + a_1\, \Upsilon'(q)q\Big] D_t Y_t \tag{68}$$

We can then recover the main cases of interest by setting the policy parameters a_0, a_1 appropriately. There are three main cases of interest: (i) setting $a_0 = 1$ and $a_1 = -1$ implements the *efficient allocation* as discussed above, (ii) setting $a_0 = 0$ and $a_1 = \chi > 0$ implements a *uniform subsidy* that leaves the dispersion in marginal revenue products unchanged but drives the aggregate wedge down to $\mathcal{M}/(1+\chi)$, while (iii) setting $a_0 = 1/(1+\chi)$ and $a_1 = -1$ implements *size-dependent subsidies* that eliminate the dispersion in marginal revenue products while leaving an aggregate wedge equal to $1+\chi$.

Uniform subsidy. Setting $a_0 = 0, a_1 = \chi$ implements a *uniform subsidy* giving net profits

$$\hat{\pi}_t(z) = \Big[(1+\chi)\Upsilon'(q_t(z))q_t(z)D_t - \frac{\Omega_t}{z}\, q_t(z)\Big] Y_t \tag{69}$$

which leads firms to set the price

$$p_t(z) = \frac{\mu_t(z)}{1+\chi}\,\frac{\Omega_t}{z} \tag{70}$$

where $\mu_t(z)$ is the benchmark markup of a firm with productivity z. This subsidy induces firms to produce more and to use more of each input, driving the wedge between price and marginal cost down to $\mu_t(z)/(1+\chi)$ and driving the aggregate wedge in the optimality conditions of the representative firm down to $\mathcal{M}_t/(1+\chi)$. Thus by setting $\chi = \mathcal{M} - 1$ for the initial distorted steady state we can put in motion a transition to a new steady state where the aggregate wedge has been eliminated. But this uniform subsidy has no effect on relative markups and so leaves steady state misallocation unchanged. This subsidy affects the entry condition but generally leaves it distorted.

Table 3 reports the effect of introducing the uniform subsidy on steady state outcomes for three levels of the aggregate markup $\mathcal{M}$. Figure 3 reports the effect on welfare, including the transitional dynamics, for a fine grid of $\mathcal{M}$. For all levels of $\mathcal{M}$, the uniform subsidy accounts for a large share of the potential welfare gains. For example, if the aggregate markup is low, $\mathcal{M} = 1.05$, the uniform subsidy increases gross output by 13.7%, consumption by 9.1%, and employment by 5.7%. These increases are only slightly smaller than those from implementing the efficient allocation. If the aggregate markup is higher, $\mathcal{M} = 1.15$ or $\mathcal{M} = 1.25$, the uniform subsidy delivers larger increases because the economy is more distorted to begin with. The uniform subsidy delivers less of an increase to aggregate productivity Z and the mass of firms N because these reflect the continued presence of misallocation and a distorted entry margin. Notice that as we increase $\mathcal{M}$, not only are the welfare gains from the uniform subsidy larger, they are also larger as a share of the total gains. For example, if $\mathcal{M} = 1.05$ the uniform subsidy accounts for about one-half of the total welfare gains (0.65% out of 1.35%), rising to nearly three-quarters of the total welfare gains if $\mathcal{M} = 1.25$ (17.34% out of 23.62%).

Size-dependent subsidies. Setting $a_0 = 1/(1+\chi)$ and $a_1 = -1$ implements *size-dependent subsidies* that drives the wedge between price and marginal cost down to $\mu_t(z)/(1+\chi) = 1$ for each firm but leaves the aggregate wedge in the optimality conditions of the representative firm equal to $1 + \chi$. Thus by setting $\chi = \mathcal{M} - 1$ for the initial distorted steady state we can put in motion a transition to a new steady state where the the aggregate wedge remains $\mathcal{M}$ but where the marginal revenue product of factors are equated across firms, i.e., a new steady state where there is no misallocation, and where the entry distortion is partly offset.

Table 3 shows that such subsidies have a more modest impact than the uniform subsidy. If the aggregate markup is low, $\mathcal{M} = 1.05$, these size-dependent subsidies increase gross output by 1.4%, consumption by 1.7%, and employment by 0.3%, noticeably less than the impact of the uniform subsidy. Where these polices have more success is on aggregate productivity Z which now increases by 0.7% when misallocation is eliminated as opposed to the 0.2% gain from the uniform subsidy driven by love-of-variety effects. If the aggregate markup is higher, $\mathcal{M} = 1.15$ or $\mathcal{M} = 1.25$, the amount of markup dispersion in the benchmark economy is larger and so the level of misallocation is also higher. In terms of the *share* of the total gains, the size-dependent subsidies account for about one-half if $\mathcal{M} = 1.05$ (0.72% out of 1.35%), falling to about one-quarter if $\mathcal{M} = 1.25$ (6.25% out of 23.62%).

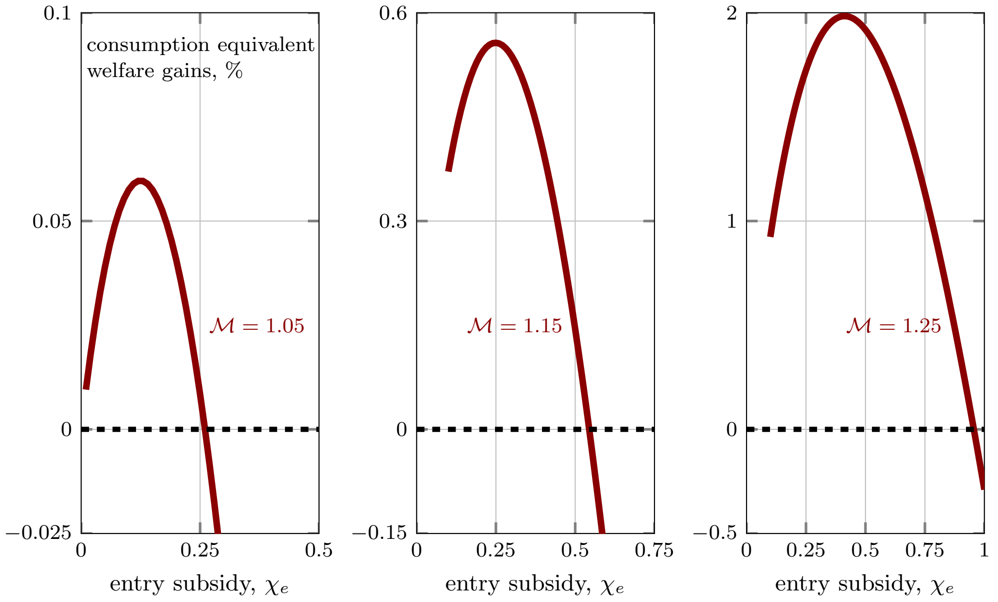
The direct intervention $T_t(q)$ eliminates all markup distortions when both the uniform subsidy component and the size-dependent component are switched on. If only one or other of these components is switched on, entry generally remains distorted as well. We next evaluate the extent to which indirect interventions in the product market, such as those which encourage entry and competition, can reduce markup distortions.

5.4 Subsidizing entry

A policy intervention like $T_t(q)$ reduces markup distortions *directly*, i.e., markups act like a tax on production so subsidizing production reduces the distortion. We now contrast such direct policies with a more *indirect* policy for reducing markup distortions — subsidizing entry, to increase the amount of competition.

Optimal entry subsidy. Consider the introduction of uniform entry subsidy χ_e that reduces the sunk entry cost from κ to $\kappa/(1+\chi_e)$. We calculate the gains from such a policy for many values of χ_e and report in Table 3 the impact of the *optimal* entry subsidy that delivers the largest total welfare gain. If the aggregate markup is low, $\mathcal{M} = 1.05$, we find that the optimal entry subsidy is $\chi_e = 0.13$, as shown in the left panel of Figure 4. This delivers a relatively large 11.2% increase in the mass of firms N but has a more modest effect on economic activity, increasing gross output by 1.1%, consumption by 1.3%, and employment by 0.4%. Aggregate productivity increases by 0.6%, reflecting the increase in variety. But these increases in activity do not lead to substantial welfare gains, due to the cost of creating new varieties incurred during the transition. The gains from the optimal entry subsidy are 0.07%, about one-twentieth of the total gains available (0.07% out of 1.35%). If the aggregate markup is higher, $\mathcal{M} = 1.15$, the optimal entry subsidy increases to $\chi_e = 0.25$. This delivers a 20.1% increase in the mass of firms N but still entry only accounts for just over one-twentieth of the total gains (0.56% out 8.69%). If $\mathcal{M} = 1.25$, the optimal entry subsidy increases to $\chi = 0.41$, delivering a 29.3% increase in the mass of firms N but still entry accounts for less than one-tenth of the total gains (1.96% out of 23.62%).

Figure 4: Optimal Entry Subsidy



Each panel shows the consumption equivalent welfare gains (including transitional dynamics) as a function of the entry subsidy χ_e for different levels of the aggregate markup $\mathcal{M}$. The welfare gains for entry subsidies reported in Table 3 are for the *optimal* entry subsidies, i.e., for the peaks of these curves. In our benchmark calibration there is insufficient entry in the initial distorted steady state so the optimal entry subsidy is positive. But entry subsidies that are too large lead to welfare losses.

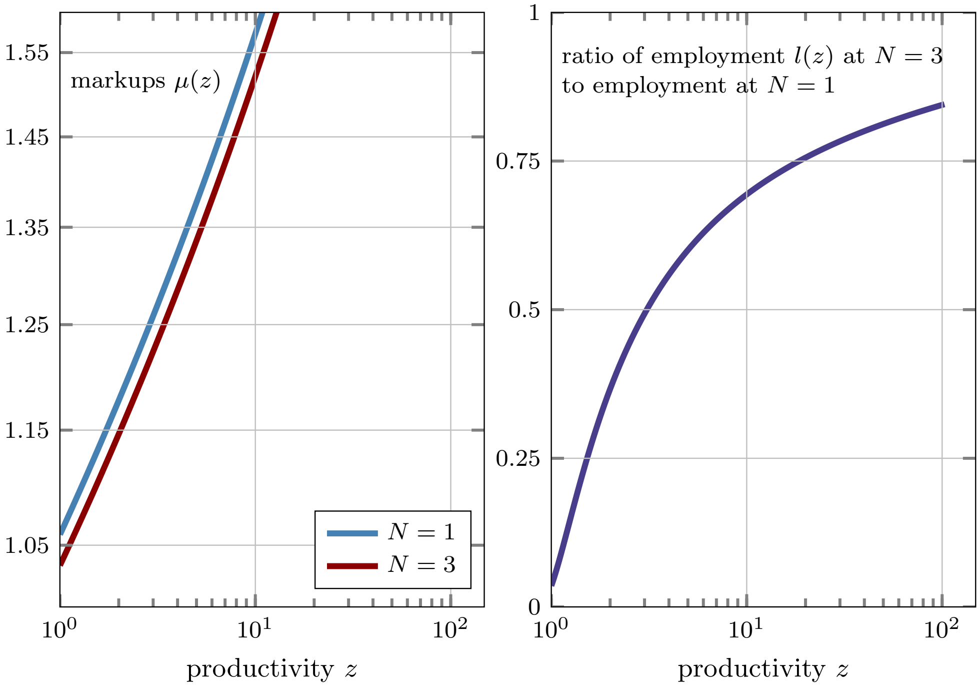
Why are the gains from subsidizing entry so low? The gains from entry are low because increasing the number of firms has tiny effects on both the aggregate markup and on misallocation. In this sense, subsidizing entry is too blunt a tool to deal with product market distortions. For example, if the benchmark economy has $\mathcal{M} = 1.05$ the optimal entry subsidy delivers an 11.2% increase in the mass of firms N but the aggregate markup falls by only about 0.02% to $\mathcal{M} = 1.0498$. Similarly if the benchmark economy has $\mathcal{M} = 1.15$, the optimal entry subsidy delivers a 20.1% increase in the mass of firms N but the aggregate markup hardly changes, falling to $\mathcal{M} = 1.149$. Entry subsidies do deliver increases in aggregate productivity, but these are due to love-of-variety effects, not due to a reduction in misallocation.

The result that *more competition* does not decrease the aggregate markup may appear counterintuitive but is, in fact, a robust result in a large class of models in the international trade literature which have shown that the removal of trade costs (which subjects domestic producers to more competition) leaves the markup distribution unchanged.[18] To understand this result, recall that the aggregate markup is a cost-weighted average of firm-level markups. An increase in the number of firms has two effects on this weighted average. The direct effect

---

[18] See Bernard, Eaton, Jensen and Kortum (2003) and Arkolakis, Costinot, Donaldson and Rodríguez-Clare (2019) who show that the markup distribution is invariant to changes in trade costs in models where variable markups arise due to limit pricing and monopolistic competition with non-CES demand, respectively.

Figure 5: Effect of Entry Subsidy on Markups



The left panel shows steady-state markups $\mu(z)$ for an economy with mass of firms $N = 1$ and an entry subsidy chosen to triple the mass of firms to $N = 3$. The right panel shows the ratio of employment $l(z)$ at $N = 3$ to employment at $N = 1$. Small, low markup firms contract by more than large, high markup firms so that high markup firms get relatively more weight in the aggregate markup calculation. Because of this, the aggregate markup hardly changes. In this example, the aggregate markup barely changes, from $\mathcal{M} = 1.150$ to $\mathcal{M} = 1.146$, even though the mass of firms triples.

is a reduction in the relative size q and hence a reduction in the markups $\mu(q)$ of each firm. But there is also an important compositional effect. Recall that in our model, small firms face more elastic demand. This makes them more vulnerable to competition from entrants. By contrast large firms face less elastic demand and are less vulnerable to competition from entrants. An entry subsidy that increases the number of firms causes small, low markup firms to contract by more than large, high markup firms and the resulting reallocation means high markup firms get relatively more weight in the aggregate markup calculation. In our model, this offsetting compositional effect is almost exactly as large as the direct effect so that overall the aggregate markup falls by a negligible amount. We develop this argument more formally in Appendix E.

We illustrate the two offsetting effects in Figure 5. For visual clarity, we consider an extreme parameterization in which we make the entry subsidy large enough to *triple* the number of firms. Notice in the left panel that markups fall for all firms when the number of firms increases. But the right panel shows that the largest, most productive firms shrink by much less than the smallest, least productive firms. We show below that similar results are obtained with other market structures.

6 Monopolistic competition extensions

In this section we consider two variations on our benchmark model: (i) where we retain Kimball demand but where firm heterogeneity arises from differences in *quality* (demand shifters) rather than differences in productivity, and (ii) where we replace Kimball demand with symmetric *translog demand*. For both these variations we retain the assumption of monopolistic competition. We present results for our model with oligopolistic competition and a finite number of firms per sector in the following section.

6.1 Heterogeneity in quality

In our benchmark model, markups are pinned down entirely by market shares. We now consider an extension where differences in quality imply differences in demand schedules across firms, breaking the tight link between markups and market shares in our benchmark.

Setup. Let $z \sim G(z)$ denote the *quality* of a firm's product and write the Kimball aggregator

$$N_t \int z\, \Upsilon\Big(\frac{y_t(z)}{Y_t}\Big)\, dG(z) = 1 \tag{71}$$

This implies the inverse demand curve

$$p_t(z) = z\, \Upsilon'(q_t(z))\, D_t \tag{72}$$

where as before $q_t(z) = y_t(z)/Y_t$ denotes a firm's relative size and D_t denotes the Kimball demand index, now given by

$$D_t = \left(N_t \int z\, \Upsilon'(q_t(z)) q_t(z)\, dG(z)\right)^{-1} \tag{73}$$

Firms have the same technology as in our benchmark model except that now all firms have the same productivity which we normalize to 1. Thus all firms have marginal cost Ω_t given by the same index of factor prices (14) and we can write the static markup condition

$$z\, \Upsilon'(q_t(z)) = \frac{\sigma(q_t(z))}{\sigma(q_t(z)) - 1}\, A_t, \qquad A_t := \frac{\Omega_t}{D_t} \tag{74}$$

where as before $\sigma(q) = \bar{\sigma} q^{-\varepsilon/\bar{\sigma}}$ denotes the demand elasticity of a firm of size q. Conditional on a given A_t this static markup condition pins down the cross-sectional distribution of relative size $q_t(z)$ and hence markups $\mu_t(z) = \mu(q_t(z))$, just as in the benchmark model.

Relationship between markups and market shares. Where the quality interpretation substantively changes the analysis is in the implied relationship between markups and market shares used in our calibration strategy. In particular, market shares $\omega_t(z) := p_t(z)q_t(z)$ are now given by $\omega_t(z) \sim z\Upsilon'(q_t(z))q_t(z)$ and so depend not just on $q_t(z)$ as in our benchmark but also on quality z. Eliminating $q_t(z)$ to write the relationship between markups and market shares now gives

$$\frac{1}{\mu_t(z)} + \log\left(1 - \frac{1}{\mu_t(z)}\right) = \text{constant} \ + \ b\,\log\omega_t(z) \ - \ b\,\log z \qquad b = \varepsilon/\bar{\sigma} \tag{75}$$

Unlike our benchmark model, cross-sectional variation in market shares is no longer a sufficient statistic for the effect of variation in z. In our benchmark, we interpreted the estimated $\hat{b}$ as a direct estimate of $\varepsilon/\bar{\sigma}$. But in this extension, since the market share is negatively correlated with the empirically *unobserved* quality z, the linear regression coefficient is no longer a consistent estimate of $\varepsilon/\bar{\sigma}$. In recalibrating the model, we use indirect inference to pin down $\varepsilon/\bar{\sigma}$, increasing the value of $\varepsilon/\bar{\sigma}$ until the coefficient in the model b equals its counterpart in the data, $\hat{b} = 0.162$, jointly with our other calibration targets.

Calibration. Table 4 reports the parameters for the quality model when we target an aggregate markup of $\mathcal{M} = 1.15$. The quality model fits the data just as well as our benchmark. The most important difference is that the super-elasticity needs to be substantially higher than in our benchmark, $\varepsilon/\bar{\sigma} = 0.304$ as opposed to 0.162. With $\varepsilon/\bar{\sigma} = 0.304$ the regression coefficient b in the quality model matches its counterpart $\hat{b}$ in the data.

Results. Given the substantially higher super-elasticity, $\varepsilon/\bar{\sigma} = 0.304$, for a given aggregate markup $\mathcal{M}$ the quality model implies more markup dispersion, especially in the upper tail. This leads to larger losses from misallocation, as shown in Table 5. For the quality model calibrated to an aggregate markup of $\mathcal{M} = 1.15$ the aggregate productivity losses due to misallocation are 1.77%, as opposed to 0.97% for our benchmark model with $\mathcal{M} = 1.15$. Because of the larger amount of misallocation in the initial distorted steady state, the total welfare costs are larger than in our benchmark and the gains from size-dependent policies that eliminate misallocation and the entry distortion are both larger in absolute terms and larger as a share of the total than in our benchmark. That said, as reported in Table 6, we continue to find that a uniform output subsidy alone can go more than half way to achieving full efficiency. As in our benchmark, the gains from the optimal entry subsidy are still an order of magnitude smaller than the gains from other policies.

6.2 Translog demand

We now consider a version of our model where we replace Kimball demand with symmetric *translog* demand as in Feenstra (2003). For this version of the model we revert to our

Table 4: Parameterization, Extensions

			quality	*translog*	*benchmark*
ξ	Pareto tail		7.69	6.67	6.84
$\bar{\sigma}$	demand elasticity		12.60	20*	10.86
$\varepsilon/\bar{\sigma}$	super-elasticity		0.30	–	0.16
ϕ	weight on value-added		0.42	0.44	0.43
calibration targets					
$\mathcal{M}$	aggregate markup	1.15	1.15	1.15	1.15
	top 5% sales share	0.57	0.57	0.21	0.57
	materials share	0.45	0.45	0.45	0.45
$\hat{b}$	regression coefficient	0.16	0.16	0.43	0.16

The calibrated parameters for our monopolistic competition extensions. For our *quality* model with Kimball demand we calibrate the Pareto tail ξ, demand elasticity $\bar{\sigma}$, super-elasticity $\varepsilon/\bar{\sigma}$ and weight on value-added ϕ to match the targets shown, the same as for our benchmark model but here for brevity we focus on the case $\mathcal{M} = 1.15$. Our *translog* model has effectively one less parameter and so fits the data less well, see text for more details. All other parameters are assigned as in Panel A of Table 1.

Table 5: Markup Distribution, Extensions

	quality	*translog*	*benchmark*
cost-weighted distribution of markups			
aggregate markup, $\mathcal{M}$	1.15	1.15	1.15
p25 markup	1.09	1.07	1.11
p50 markup	1.13	1.12	1.14
p75 markup	1.19	1.20	1.18
p90 markup	1.26	1.30	1.23
p99 markup	1.43	1.53	1.35
productivity losses from misallocation			
$\frac{Z^*-Z}{Z} \times 100$	1.77	2.89	0.97

Table 6: Implications of Alternative Policies, Extensions

	steady state comparisons, %						
	Y	C	L	N	K	Z	welfare, %
quality							
efficient	68.4	54.3	19.1	30.7	114.0	6.8	11.55
uniform subsidy	52.4	36.8	16.9	10.6	89.2	1.9	6.44
size-dependent subsidy	10.8	12.8	2.0	16.9	13.5	4.7	5.58
entry subsidy	10.3	12.3	3.6	31.1	13.2	5.0	1.22
translog							
efficient	61.6	46.4	16.8	8.6	103.1	4.2	13.43
uniform subsidy	51.3	35.4	17.0	9.9	87.9	1.4	5.67
size-dependent subsidy	7.5	8.6	0.1	-1.1	9.0	2.7	7.47
entry subsidy	2.7	3.2	1.1	9.5	3.4	1.4	0.14
benchmark ($\mathcal{M} = 1.15$)							
efficient	59.7	44.6	18.0	20.1	100.6	4.1	8.69
uniform subsidy	51.9	35.9	17.0	9.5	88.7	1.5	5.92
size-dependent subsidy	5.3	6.2	1.0	8.3	6.6	2.3	2.87
entry subsidy	6.3	7.5	2.4	20.1	8.2	3.1	0.56

The first six columns report the percentage change from the initial distorted steady state with $\mathcal{M} = 1.15$ to the new steady state. The last column reports the consumption equivalent welfare gains (including transitional dynamics). The alternative policies are (i): the *efficient allocation*, where all markups are removed, (ii) a *uniform subsidy* that eliminates the aggregate markup, (iii) *size-dependent subsidies* that eliminate misallocation and the entry distortion, and (iv) the optimal *entry subsidy*.

benchmark setting where firm heterogeneity arises from differences in productivity.

Setup. Let the technology for final good producers be given by a symmetric translog expenditure (cost) function which we write

$$\begin{aligned}\log(P_tY_t) \;=\; & \log Y_t \;+\; \frac{1}{2\bar{\sigma}N_t} \;+\; \int \log p_t(z)\,dG(z) \\ & +\; \frac{\bar{\sigma}N_t}{2}\left(\left(\int \log p_t(z)\,dG(z)\right)^2 - \int \log p_t(z)^2\,dG(z)\right) \end{aligned} \tag{76}$$

From Shephard's lemma, the market share $\omega_t(z)$ of a firm with productivity z is given by

$$\omega_t(z) := \frac{p_t(z)y_t(z)}{P_tY_t} = \frac{d\log(P_tY_t)}{d\log p_t(z)} = \bar{\sigma}\log\left(\frac{p_t^*}{p_t(z)}\right), \qquad p_t(z) < p_t^* \tag{77}$$

where any price $p_t(z)$ larger than the *choke price* p_t^* given by

$$\log p_t^* := \frac{1}{2\bar{\sigma} N_t} + \int \log p_t(z)\, dG(z) \tag{78}$$

will lead to zero sales. We can then write the residual demand curve

$$y_t(z) = \bar{\sigma} \log\Big(\frac{p_t^*}{p_t(z)}\Big) \frac{P_t Y_t}{p_t(z)}, \qquad p_t(z) < p_t^* \tag{79}$$

Let $\rho_t(z) := p_t(z)/p_t^*$ denote a firm's relative price and let $\omega(\rho) = \bar{\sigma}\log(1/\rho)$ denote the market share and $y(\rho) \sim \omega(\rho)/\rho$ the residual demand for a firm with relative price $\rho \leq 1$. Let $\sigma(\rho)$ and $\mu(\rho)$ denote the associated demand elasticity and markup. These are given by

$$\sigma(\rho) = \frac{1 + \log\left(\frac{1}{\rho}\right)}{\log\left(\frac{1}{\rho}\right)}, \qquad \mu(\rho) = 1 + \log\Big(\frac{1}{\rho}\Big) \tag{80}$$

We can then write the static markup-pricing condition

$$\rho_t(z) = \frac{\sigma(\rho_t(z))}{\sigma(\rho_t(z)) - 1}\, \frac{z_t^*}{z}, \qquad z_t^* := \frac{\Omega_t}{p_t^*} \tag{81}$$

where z_t^* is the cutoff productivity such that $p_t^* = \Omega_t/z_t^*$, i.e., the cutoff firm with productivity z_t^* has price equal to its marginal cost Ω_t/z_t^*. Conditional on z_t^* this static markup condition pins down the cross-sectional distribution of relative prices $\rho_t(z)$ and hence markups $\mu_t(z) = \mu(\rho_t(z))$, just as in the benchmark model.

Markups and market shares. This translog specification implies a *linear* relationship between markups and market shares. From (80) we can write

$$\mu_t(z) = 1 + \frac{1}{\bar{\sigma}}\, \omega_t(z) \tag{82}$$

As in our benchmark model, firms with higher market shares have higher markups. With translog demand, the strength of this relationship is governed by $1/\bar{\sigma}$.

Markups. Inverting $\mu(\rho)$ to write $\rho(\mu) = e^{1-\mu}$ we can write the static markup condition

$$\mu + \log \mu = 1 + \log\Big(\frac{z}{z_t^*}\Big), \qquad z > z_t^* \tag{83}$$

which implicitly determines the markup $\mu_t(z)$, strictly increasing in z. Notice that the productivity cutoff z_t^* is the *only* aggregate variable that matters for the cross-sectional distribution of markups — and hence the only aggregate variable that matters for the the cross-sectional distributions of market shares $\omega_t(z)$ and relative prices $\rho_t(z)$.

To this point, our analysis of the translog model has restated standard results in the trade literature, familiar from Feenstra (2003), Rodriguez-Lopez (2011), and Arkolakis, Costinot, Donaldson and Rodríguez-Clare (2010, 2019) among others. We next show that given a Pareto distribution of firm-level productivity $G(z)$ we can solve explicitly for the cutoff productivity z_t^* and then aggregate markup $\mathcal{M}_t$. Though closely related to these existing papers, to the best of our knowledge, the following results are novel and may be of some independent interest to researchers working with translog demand and Pareto distributions.

Solving for the cutoff z_t^*. As shown in Appendix E, using $z_t^* = \Omega_t/p_t^*$, the definition of the choke price p_t^* in (78), the static markup condition (83), and the assumption that $G(z)$ is Pareto, the cutoff z_t^* is given by

$$z_t^* = \max\Big[\, 1 \,,\; \bar{\sigma} N_t \, e^{\xi} E_{\xi}(\xi) \,\Big]^{1/\xi} \tag{84}$$

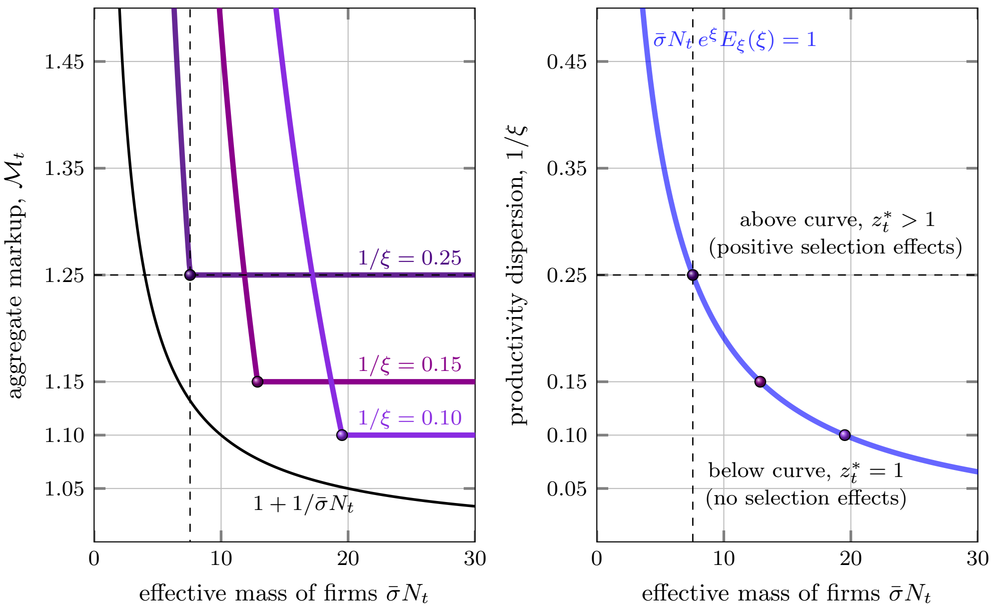
where $E_n(x) := \int_1^\infty t^{-n} e^{-xt}\, dt$ denotes the generalized exponential integral. Since the mass of firms N_t is a state variable (is predetermined), this determines z_t^* and from (83) we then know the entire distribution of markups, market shares and relative prices given N_t. The constant $e^{\xi} E_{\xi}(\xi)$ depends only on the Pareto tail parameter $\xi > 1$ and is strictly decreasing in ξ, i.e., increasing in productivity dispersion $1/\xi$. If either the 'effective' mass of firms $\bar{\sigma} N_t$ is sufficiently low or productivity dispersion $1/\xi$ is sufficiently low we have $z_t^* = 1$, meaning that there are no selection effects and all firms operate. But if either $\bar{\sigma} N_t$ is sufficiently high or productivity dispersion $1/\xi$ is sufficiently high we have $z_t^* > 1$, meaning that there are positive selection effects. Intuitively, when demand is more elastic, or when the mass of firms is larger, or when productivity is more dispersed, there is more competitive pressure and selection effects are stronger, increasing the cutoff z_t^*. The right panel of Figure 6 illustrates, showing how the locus $\bar{\sigma} N_t \, e^{\xi} E_{\xi}(\xi) = 1$ partitions the parameter space into the regions where $z_t^* = 1$ (below the curve) and $z_t^* > 1$ (above the curve).

Solving for the aggregate markup $\mathcal{M}_t$. Our assumption that $G(z)$ is Pareto also implies a simple solution for $\mathcal{M}_t$. As shown in Appendix E, using the fact that the aggregate markup can be written as a harmonic weighted average of firm-level markups, the linear relationship between market shares and markups (82), the static markup condition (83), and our solution for the cutoff productivity z_t^*, the aggregate markup $\mathcal{M}_t$ is given by

$$\mathcal{M}_t = \Big(1 + \frac{1}{\xi}\Big) \times \Big(\max\Big[\, 1 \,,\; \bar{\sigma} N_t \, e^{\xi} E_{\xi}(\xi) \,\Big]\Big)^{-1} \tag{85}$$

Since the mass of firms N_t is a state variable, this determines $\mathcal{M}_t$. Now observe from (84) that if $\bar{\sigma} N_t \, e^{\xi} E_{\xi}(\xi) \leq 1$, implying $z_t^* = 1$, then the aggregate markup is strictly decreasing in N_t with an elasticity of -1. But whenever $\bar{\sigma} N_t \, e^{\xi} E_{\xi}(\xi) > 1$, i.e., whenever there are positive

Figure 6: Aggregate Markup with Translog Demand



Left panel shows the solution for the aggregate markup $\mathcal{M}_t$ with translog demand as a function of the effective mass of firms $\bar{\sigma}N_t$ for various levels of the Pareto tail ξ. Right panel shows how the parameter space is partitioned into regions where there are positive selection effects, $z_t^* > 1$, or no selection effects, $z_t^* = 1$. Whenever there are positive selection effects, the aggregate markup is constant at $\mathcal{M}_t = 1 + 1/\xi$. By contrast with identical firms, the aggregate markup would be given by $1 + 1/\bar{\sigma}N_t$ as shown. With firm heterogeneity, the aggregate markup is decreasing only in $\bar{\sigma}N_t$ only if there are no selection effects, $z_t^* = 1$.

selection effects, $z_t^* > 1$, then the aggregate markup is *constant* at the specific value

$$\mathcal{M}_t = 1 + \frac{1}{\xi}, \qquad \text{whenever } z_t^* > 1 \tag{86}$$

So whenever there are positive selection effects, $z_t^* > 1$, e.g., $\bar{\sigma}$ or productivity dispersion $1/\xi$ is sufficiently high, then increases in the mass of firms N_t have *no effect* on the aggregate markup $\mathcal{M}_t$. Instead, increases in N_t are absorbed by increases in the cutoff z_t^*, i.e., by stronger selection effects. This analytic result reinforces the lesson from our benchmark model with Kimball demand where we found numerically that the aggregate markup is extremely insensitive to changes in N_t.[19] The reason is the same: whenever $z_t^* > 1$, an increase in N_t increases z_t^* thereby directly reducing all firm-level markups $\mu_t(z)$ according to (83). But low markup firms contract by more than large, high markup firms and the resulting reallocation means high markup firms get relatively more weight in the aggregate markup calculation. In the translog case, so long as parameters are such that $z_t^* > 1$, this offsetting compositional effect is exactly as large as the direct effect so that overall the aggregate markup is unchanged.

---

[19] As discussed in Appendix E, the model with Kimball demand is qualitatively similar to translog demand in that for Kimball demand the aggregate markup $\mathcal{M}_t$ is also invariant to N_t *if* there are positive selection effects. But in our benchmark calibration of the Kimball model, there are no selection effects and changes in N_t *do* change $\mathcal{M}_t$ albeit by negligible amounts.

Role of heterogeneity. Firm heterogeneity is essential to this result. If by contrast all firms were identical, as in say Bilbiie, Ghironi and Melitz (2008, 2019), each firm would have market share $1/N_t$ and the aggregate markup would be $\mathcal{M}_t = 1 + 1/\bar{\sigma}N_t$ and would always be decreasing in N_t. In the representative firm setting, there is the direct effect of an increase in N_t on firm-level markups but this effect is the same for all firms so there is no offsetting compositional effect. In this sense, accounting for the role of firm heterogeneity is crucial for understanding the welfare effects of changes in the mass of firms N_t.[20]

Calibration. The translog model is less flexible than our Kimball benchmark. Whenever there are positive selection effects, $z_t^* > 1$, the Pareto tail ξ is pinned down by our target for the aggregate markup $\mathcal{M} = 1 + 1/\xi$. Moreover the parameter $\bar{\sigma}$ always enters in the form $\bar{\sigma}N$ and so is not separately identified.[21] In this sense, the translog model only has two key parameters to work with, not the three parameters of our Kimball benchmark. Given this, it is not surprising that the translog model does less well in hitting our calibration targets. The translog model cannot simultaneously hit our aggregate markup target, sales concentration target, and regression coefficient $\hat{b}$. As reported in Table 4, the translog model reproduces an aggregate markup of $\mathcal{M} = 1.15$ but implies too little sales concentration (a top 5% sales share of 0.207, as opposed to 0.573 in the data) and too strong a relationship between markups and market shares (regression coefficient $b = 0.434$ as opposed to $\hat{b} = 0.162$ in the data).

Results. As with the quality differences model, the translog model implies considerably more markup dispersion, especially in the upper tail. This again leads to larger losses from misallocation relative to our benchmark model, as shown in Table 5. For our translog model calibrated to an aggregate markup of $\mathcal{M} = 1.15$ the aggregate productivity losses due to misallocation are 1.89%, as opposed to 0.97% for our benchmark model with $\mathcal{M} = 1.15$. Because of the larger amount of misallocation in the initial distorted steady state, the total welfare costs are larger than in our benchmark and the gains from size-dependent policies that eliminate misallocation and the entry distortion are both larger in absolute terms and larger as a share of the total than in our benchmark. Indeed, as shown in Table 6, this effect is even stronger than in the quality model so now we find that the size-dependent policies have a larger effect than the uniform output subsidy. Again we find that the gains from the optimal entry subsidy are much, much smaller than the gains from other policies.

---

[20] Rodriguez-Lopez (2011) derives a related result, solving for the *average* markup $\int \mu_t(z)\, dG(z)$ with translog demand and Pareto productivity and shows that this depends only on the Pareto tail ξ. Also related, Arkolakis, Costinot, Donaldson and Rodríguez-Clare (2019) show that with translog demand and Pareto productivity the univariate distribution of markups $\text{Prob}[\mu' \leq \mu]$ depends only on the Pareto tail ξ. Our key analytic contribution is to explicitly compute the *aggregate* markup, the sales-weighted harmonic average $\mathcal{M}_t = \left(N_t \int (\omega_t(z)/\mu_t(z)) dG(z)\right)^{-1}$, which, as we have stressed throughout, is the key wedge in the optimality conditions of the representative firm.

[21] Recall that we choose the sunk entry cost κ to normalize $N = 1$ in the initial distorted steady state.

7 Oligopolistic competition

We now present calculations based on an alternative model featuring oligopolistic competition rather than monopolistic competition. Our goal in this section is to assess to what extent our quantitative results are sensitive to the specific market structure we used in our benchmark.

Setup. Let there be $n_t(s) \in \mathbb{N}$ firms per sector with IID productivity draws $z_i(s) \sim G(z)$. Let the within-sector aggregator be $\Upsilon(q) = q^{\frac{\gamma-1}{\gamma}}$ for $\gamma > \eta > 1$ so that the model has the nested-CES structure used by Atkeson and Burstein (2008) and Edmond, Midrigan and Xu (2015). For our quantitative work we assume *Cournot competition* so that, as in (38) above, the demand elasticity of a firm is given by the sales-weighted harmonic average

$$\sigma_{it}(s) = \left(\frac{1}{\eta}\,\omega_{it}(s) + \frac{1}{\gamma}\,(1-\omega_{it}(s)\right)^{-1} \tag{87}$$

where $\omega_{it}(s) = q_{it}(s)^{\frac{\gamma-1}{\gamma}}$ denotes the market share of firm i in sector s.[22] As stressed at length above, this oligopoly model is encompassed by our general framework except that for the free-entry condition (40) expected profits are given by (43). In practice however, solving this oligopoly model with a forward-looking free-entry condition endogenously determining the number of firms is challenging.[23] This is because there are many firms that each have non-negligible effects on sector-level outcomes, outcomes in any given sector are a function of the vector $\boldsymbol{z}(s) = (z_1(s), z_2(s), \ldots, z_{n_t(s)}(s))$ of productivities. Because of the finite number of firms, sectors are heterogeneous and we cannot invoke the law of large numbers to compute expected profits. And because $n_t(s)$ is typically large,[24] we need to compute high-dimensional integrals with respect to the joint distribution $G_{n_t(s)}(\boldsymbol{z}(s))$ of $\boldsymbol{z}(s)$. In principle, the heterogeneity across sectors creates incentives for firms to *direct* entry towards more profitable sectors. But to simplify the problem computationally, we assume that entry is *random*, that firms can not direct entry in this way. We discuss these issues in more detail in Appendix B.

Relationship between markups and market shares. This nested-CES specification implies that the inverse markup is linear decreasing in the market share

$$\frac{1}{\mu_{it}(s)} = \left(1 - \frac{1}{\gamma}\right) - \left(\frac{1}{\eta} - \frac{1}{\gamma}\right)\,\omega_{it}(s) \tag{88}$$

As in our benchmark model, firms with higher market shares have higher markups. Here, the strength of this relationship is governed by the gap between the between-sector elasticity of

---

[22] In this oligopoly model, sectors are ex post heterogeneous so we put back dependence on s in the notation.

[23] Other applications of this oligopoly setup, e.g., Atkeson and Burstein (2008), Edmond, Midrigan and Xu (2015), and De Loecker, Eeckhout and Mongey (2021), treat the number of potential producers as *exogenous*.

[24] In our calibration, there are on average about 360 firms per sector.

substitution η and the within-sector elasticity of substitution $\gamma > \eta$. Multiplying both sides of (88) by $\omega_{it}(s)$ and summing over all firms i within sector s gives

$$\frac{1}{\mu_t(s)} = \left(1 - \frac{1}{\gamma}\right) - \left(\frac{1}{\eta} - \frac{1}{\gamma}\right) \sum_{i=1}^{n_t(s)} \omega_{it}(s)^2 \tag{89}$$

The model predicts a linear decreasing relationship between the sector-level inverse markup $1/\mu_t(s)$ and the sector's *Herfindahl-Hirschman index* (HHI) of sales concentration. From (25), the sector-level labor share is proportional to the inverse markup, $W_t l_t(s)/p_t(s)y_t(s) = (1-\alpha)\zeta_t/\mu_t(s)$. Motivated by this, in calibrating the oligopoly model we use indirect inference to pin down the gap between γ and η, choosing parameters so that our model reproduces the $\hat{b} = -0.21$ slope coefficient in a regression of the change over time of sector-level labor shares on the change in sector-level HHIs, as in Autor, Dorn, Katz, Patterson and Van Reenen (2020), jointly with our other calibration targets.

Calibration. We calibrate the oligopoly model targeting measures of concentration for 4-digit industries in the 2012 US Census of Manufactures as reported by Autor, Dorn, Katz, Patterson and Van Reenen (2020). In particular, we target their top 4 sales share (CR4) of 0.43, top 20 sales share (CR20) of 0.73, and the slope in a regression of the change over time in sector-level labor shares on the change in sector-level HHIs of $\hat{b} = -0.21$.[25] We also target an aggregate markup of $\mathcal{M} = 1.15$ and a materials share of 0.45. Intuitively, the two measures of sales concentration pin down the Pareto tail ξ, which controls the amount of productivity dispersion, and the sunk entry cost κ. The aggregate markup then pins down γ, while the slope coefficient pins down the gap between γ and η. As shown in Table 7, the oligopoly model does well at hitting our calibration targets. Relative to the benchmark model with Kimball demand and monopolistic competition, the oligopoly model requires less productivity dispersion, $\xi = 8.51$ as opposed to $\xi = 6.84$. On average, there is a relatively large number of firms per sector, $N = 359$, but most of these firms are very small.

Results. Table 8 reports the cost-weighted percentiles of the steady-state distribution of firm-level markups $\mu_{it}(s)$ and sector-level markups $\mu_t(s)$. The distribution of sector-level markups alone is as dispersed as the unconditional markup distribution in our benchmark model with monopolistic competition calibrated to the same aggregate markup $\mathcal{M} = 1.15$ (for which sectors are identical).[26] The unconditional distribution of markups in the oligopoly model is considerably more dispersed than in our benchmark, especially in the upper tail. This leads to larger losses from misallocation, 3.3% up from 0.97% in the benchmark. As

---

[25] The CR4 and CR20 are read off Panel A of Figure 4 while the regression coefficient $\hat{b}$ is from Table 2 baseline column (3) in Autor, Dorn, Katz, Patterson and Van Reenen (2020).

[26] This amount of dispersion in sector-level markups is however less costly, because of the low elasticity of substitution η between sectors. The amount of markup dispersion within sectors is more important.

Table 7: Parameterization, Oligopoly

ξ	Pareto tail		8.51
γ	elasticity of substitution within sectors		12.76
η	elasticity of substitution between sectors		1.35
N	average number of firms per sector		359
ϕ	weight on value-added		0.28
calibration targets		*data*	*model*
$\mathcal{M}$	aggregate markup	1.15	1.15
CR4	top 4 sales share	0.43	0.43
CR20	top 20 sales share	0.73	0.72
	materials share	0.45	0.45
$\hat{b}$	regression coefficient	-0.21	-0.21

The calibrated parameters for our oligopoly model. We calibrate the Pareto tail ξ, the within- and between-sector elasticities of substitution γ and η, the sunk entry cost κ, and weight on value-added ϕ to match the targets shown. For brevity we focus on the case $\mathcal{M} = 1.15$. In practice, we choose the average number of firms N per sector and back out the sunk cost κ that rationalizes N. The cross-sectional regression is of the change over time in sector-level labor shares on the change in sector-level HHIs, as discussed in the text. All other parameters are assigned as in Panel A of Table 1.

Table 8: Markup Distribution, Oligopoly

	oligopoly		*benchmark*
	sectoral	unconditional	unconditional
cost-weighted distribution of markups			
aggregate markup, $\mathcal{M}$	1.15	1.15	1.15
p25 markup	1.12	1.09	1.11
p50 markup	1.14	1.11	1.14
p75 markup	1.16	1.17	1.18
p90 markup	1.21	1.27	1.23
p99 markup	1.36	1.56	1.35
productivity losses from misallocation			
$\frac{Z^{*}-Z}{Z} \times 100$		3.30	0.97

Table 9: Implications of Alternative Policies, Oligopoly

	steady state comparisons, %						
	Y	C	L	N	K	Z	welfare, %
oligopoly							
efficient	56.4	40.4	15.1	−9.1	94.8	2.1	14.65
uniform subsidy	50.5	34.5	17.0	9.8	86.7	1.1	5.13
size-dependent subsidy	4.1	4.4	−1.8	−18.3	4.5	0.8	8.67
entry subsidy	−2.1	−2.5	−1.0	−8.7	−2.7	−1.1	0.12
benchmark ($\mathcal{M} = 1.15$)							
efficient	59.7	44.6	18.0	20.1	100.6	4.1	8.69
uniform subsidy	51.9	35.9	17.0	9.5	88.7	1.5	5.92
size-dependent subsidy	5.3	6.2	1.0	8.3	6.6	2.3	2.87
entry subsidy	6.3	7.5	2.4	20.1	8.2	3.1	0.56

The first six columns report the percentage change from the initial distorted steady state with $\mathcal{M} = 1.15$ to the new steady state. The last column reports the consumption equivalent welfare gains (including transitional dynamics). The alternative policies are (i): the *efficient allocation*, where all markups are removed, (ii) a *uniform subsidy* that eliminates the aggregate markup, (iii) *size-dependent subsidies* that eliminate misallocation and the entry distortion, and (iv) the optimal *entry subsidy*, which here is a tax, since the initial steady state has too many firms.

shown in Table 9, in other respects the oligopoly model implies remarkably similar long-run changes in gross output, consumption, employment, and capital as our benchmark. A notable difference is that in our oligopoly model the initial distorted steady state features *too many* firms and transitioning to the efficient steady state involves reducing the average number of firms N by about 9.1%. Because of the larger amount of misallocation, the oligopoly model implies substantially larger costs of markups, 14.65% in consumption-equivalent terms, up from 8.69% for our benchmark. The gains from size-dependent subsidies that eliminate misallocation and the entry distortion are 8.67% for the oligopoly model, up from 2.87% for our benchmark. The gains from a uniform subsidy that eliminates the aggregate markup distortion are similar to our benchmark, 5.13% down slightly from 5.92%, but are correspondingly a smaller share of the total. Again, the gains from the optimal entry subsidy are much, much smaller than the gains from other policies.[27]

---

[27] Since the initial steady state has too many firms, the optimal entry subsidy is a *tax*, of about $\chi_e = -0.1$.

There are two important caveats regarding these results. First, in the oligopoly model, subsidies to eliminate misallocation would have to be both sector- and size-dependent, as opposed to just size-dependent as they are in our benchmark model with monopolistic competition. Second, the losses from misallocation may be lower if entry could be directed to specific sectors. It remains an open question and an important direction for future research to assess how much misallocation would be reduced if firms could direct entry.

8 Conclusion

We study the welfare costs of product market distortions in a dynamic model with heterogeneous firms and endogenously variable markups. Our model encompasses several popular market structures and we provide aggregation results showing how the macro implications of micro-level markup heterogeneity can be summarized by a few key statistics. We calibrate our model to match levels of sales concentration and the firm-level relationship between labor shares and market shares observed in 6-digit US Census of Manufactures data. We find that the welfare costs of markups can be large. Depending on the market structure and assumed level of the aggregate markup, the representative consumer can gain as much as 25% in consumption-equivalent terms if all markup distortions are eliminated, once transitional dynamics are taken into account.

In our model markups reduce welfare because the aggregate markup distortion acts like a uniform output tax, reducing employment and investment by all firms, because markup variation across firms causes misallocation of factors of production, and because there is an inefficient rate of entry due to the misalignment between private and social incentives to create new firms. Across all specifications, we robustly find that the aggregate markup and misallocation channels account for the bulk of the costs of markups and that the entry channel is much less important.

Although we focus on the normative side of our model, our results also have clear empirical implications. One simple but important finding is that the overall level of markups is best measured as a *cost-weighted* average of firm-level markups. This is the relevant 'wedge' in aggregate employment and investment decisions. By contrast a *sales-weighted* average of firm-level markups, as used in the empirical literature, overstates the rise in the overall level of market power. In addition, our results provide two reasons to be skeptical of explanations for the simultaneous rise in concentration and markups that focus on increasing barriers to entry. First, in our model increasing barriers to entry *reduce concentration*, because the resulting lack of competition makes it easier for small firms to survive. Second, in our model changes in entry have negligible effects on the overall level of markups because entry is associated with a reallocation of production towards high productivity, high markup firms.

To keep our model tractable enough that we can aggregate cross-sectional outcomes and study transitional dynamics for a broad range of alternative market structures, we have abstracted from a number of considerations that might play an important role in the development of a more complete account of the macroeconomic implications of product market distortions. First, while markups in our model are a return to sunk investments, there are no positive spillovers from such investment to the stock of knowledge in the economy at large and hence no implications for endogenous growth. But as emphasized by Atkeson, Burstein and Chatzikonstantinou (2019), in the endogenous growth models they survey, a higher markup acts like a uniform subsidy to innovation and is welfare-improving, the quantitative details depending sensitively on the specification of the technology for research. In principle, these effects could be large. That said, in endogenous growth models with variable markups, such as Peters (2020), the interactions between entry, aggregate innovation and misallocation are more subtle with the overall effects on growth ambiguous. An important challenge for future work in this area is to provide detailed evidence on technologies for research and the magnitudes of spillovers that can be used to refine such models to help quantify the relative importance of these growth effects and the level effects of markups emphasized in this paper.

Second, we have made the assumption, standard in the literature, that the underlying sources of firm size differences are fundamental differences in productivity or quality. Because of this, large firms with high markups represent a lost opportunity — they should be even larger, not smaller, but charge lower prices. But if large firms are large not because they are more productive or because their products are higher quality but instead because they receive special tax breaks, or have political connections that help them evade antitrust actions or other forms of regulation, then such firms may well be too large, not too small. Another important challenge for future work in this area is to build models that blend political connections, as in Akcigit, Baslandze and Lotti (2018), with endogenous product market distortions so that we can quantitatively evaluate size-dependent policy interventions when both fundamental and non-fundamental sources of firm size are operative.

Finally, to keep the analysis focused, we have abstracted from tax wedges and other distortions that affect aggregate labor supply and capital accumulation decisions. For standard second best reasons, such distortions may either amplify or mitigate the costs of product market distortions. Quantifying the interactions between these different types of distortions also seems a natural topic for ongoing research.

Appendix

A Cost-weighted vs. sales-weighted average markups

In this appendix we derive an exact relationship between a *cost-weighted* average markup $\mathcal{M}$ and a *sales-weighted* average markup $\tilde{\mathcal{M}}$. The key result is

$$\boxed{\frac{\tilde{\mathcal{M}} - \mathcal{M}}{\mathcal{M}} = \mathrm{Var}[\hat{\mu}_i]}$$

where $\mathrm{Var}[\hat{\mu}_i]$ is a measure of the cross-sectional dispersion in the idiosyncratic component in markups, $\hat{\mu}_i := \mu_i/\mathcal{M}$. This derivation makes no assumptions about demand or market structure but makes one key assumption about technology, specifically, that all firms within a given industry have the same *cost elasticity.*

Notation. Consider an industry with $i = 1, 2, \ldots, n$ firms. Let p_i, y_i, μ_i and c_i denote respectively a firm's price, output, markup, and *total variable costs.*

Cost elasticity assumption. Let $\vartheta > 0$ denote a firm's *cost elasticity*, that is, the elasticity of total variable costs with respect to output

$$\vartheta := \frac{\partial \log c}{\partial \log y} = \frac{\partial c}{\partial y}\frac{y}{c} = \frac{\text{marginal cost}}{\text{average cost}} \tag{A1}$$

Our key assumption is that the cost elasticity ϑ is common to all firms within a given industry, $\vartheta_i = \vartheta$. In other words, all firms within a given industry have the same *returns to scale*, but this may be either increasing, constant, or decreasing at the industry level. Marginal costs are then given by $\vartheta c_i/y_i$. Importantly we do not put any restrictions on marginal costs, these can vary arbitrarily across firms within the industry.

Aggregate markup. Given the assumption that all firms within a given industry have the same cost elasticity ϑ, it is straightforward to show that the industry aggregate markup, that is, the ratio of industry price to industry marginal cost, is given by a cost-weighted average of firm-level markups (equivalently, a sales-weighted *harmonic* average). Following the same steps as in the main text, since prices p_i are a markup μ_i over marginal cost $\vartheta c_i/y_i$ we have revenues $p_i y_i = \vartheta \mu_i c_i$ so if we are to write $\mathcal{M}$ as the 'wedge' between industry revenue $PY := \sum_i p_i y_i$ and industry costs $\vartheta \sum_i c_i$ (i.e., so that $\mathcal{M}$ is the ratio of the industry price level to industry marginal costs), then

$$\mathcal{M} = \sum_{i=1}^{n} \mu_i\, \omega_i, \qquad \omega_i := \frac{c_i}{\sum_i c_i} \tag{A2}$$

where in slight abuse of notation we now use ω_i to denote the *cost-weights.* Notice that this derivation makes no assumptions about the demand system or market structure that generates the markups μ_i.

Relationship between cost-weighted and sales-weighted averages. By contrast, the applied literature on markups has emphasized sales-weighted averages, which can be written

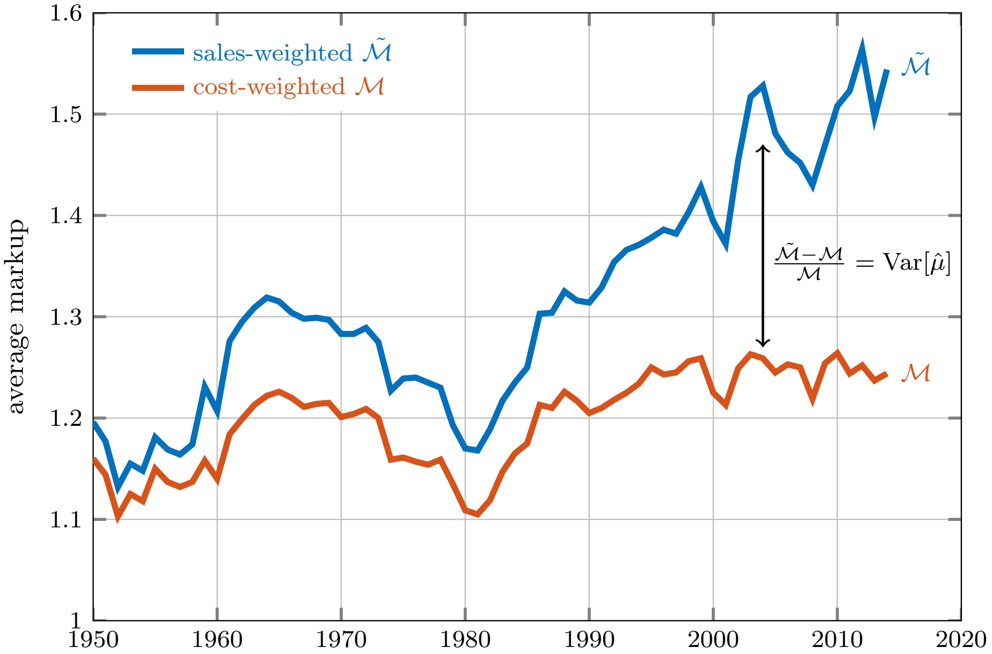
$$\tilde{\mathcal{M}} = \sum_{i=1}^{n} \mu_i\, \tilde{\omega}_i, \qquad \tilde{\omega}_i := \frac{p_i y_i}{\sum_i p_i y_i} \tag{A3}$$

We will now show that the sales-weighted average $\tilde{\mathcal{M}}$ can be decomposed into the cost-weighted average $\mathcal{M}$ plus a term that reflects the cross-sectional dispersion in markups.

Let $\mathbb{E}[\cdot]$ denote averages with respect to the cost weights so that $\mathcal{M} = \mathbb{E}[\mu_i]$. Then we can write the sales-weighted average as

$$\tilde{\mathcal{M}} = \sum_{i=1}^{n} \mu_i\, \tilde{\omega}_i = \sum_{i=1}^{n} \mu_i\, \frac{\tilde{\omega}_i}{\omega_i}\, \omega_i = \mathbb{E}\Big[\, \mu_i\, \frac{\tilde{\omega}_i}{\omega_i}\, \Big] \tag{A4}$$

Figure A1: Cost-Weighted vs. Sales-Weighted Average Markups, Compustat



The sales-weighted average $\tilde{\mathcal{M}}$ of firm-level markups in Compustat data, as in De Loecker, Eeckhout and Unger (2020), and the cost-weighted average of firm-level markups $\mathcal{M}$. The former is higher and has increased by a larger amount. The proportional difference between the two averages reflects the cross-sectional dispersion in markups, which has been increasing.

Expanding the expectation of the product into the covariance plus the product of the expectations then gives

$$\begin{aligned}\tilde{\mathcal{M}} = \mathbb{E}\Big[\,\mu_i\,\frac{\tilde{\omega}_i}{\omega_i}\,\Big] &= \mathrm{Cov}\Big[\,\mu_i\,,\,\frac{\tilde{\omega}_i}{\omega_i}\,\Big] + \mathbb{E}\big[\mu_i\big]\,\mathbb{E}\Big[\,\frac{\tilde{\omega}_i}{\omega_i}\,\Big] \\ &= \mathrm{Cov}\Big[\,\mu_i\,,\,\frac{\tilde{\omega}_i}{\omega_i}\,\Big] + \mathcal{M} \end{aligned} \tag{A5}$$

since $\mathcal{M} = \mathbb{E}[\mu_i]$ and $\mathbb{E}[\frac{\tilde{\omega}_i}{\omega_i}] = \sum_i \tilde{\omega}_i = 1$. In short, the *absolute difference* between the sales-weighted and cost-weighted average markups is given by the covariance of the markups μ_i and the relative weights $\tilde{\omega}_i/\omega_i$.

But under the assumption of a common cost elasticity ϑ the relative weights are *proportional to the markups themselves*

$$\frac{\tilde{\omega}_i}{\omega_i} = \frac{p_i y_i}{c_i}\frac{\sum_i c_i}{\sum_i p_i y_i} = \frac{\mu_i \vartheta \frac{c_i}{y_i} y_i}{c_i}\frac{\sum_i c_i}{\sum_i p_i y_i} = \frac{\mu_i}{\mathcal{M}} \tag{A6}$$

where the last equality follows because $\mathcal{M}$ is the 'wedge' between industry revenue $\sum_i p_i y_i$ and industry costs $\vartheta \sum_i c_i$. In short, we can write

$$\mathrm{Cov}\Big[\,\mu_i\,,\,\frac{\tilde{\omega}_i}{\omega_i}\,\Big] = \mathrm{Cov}\Big[\,\mu_i\,,\,\mu_i\frac{1}{\mathcal{M}}\,\Big] = \frac{1}{\mathcal{M}}\mathrm{Var}[\mu_i] \tag{A7}$$

And hence our key decomposition can be written

$$\tilde{\mathcal{M}} = \mathcal{M} + \frac{1}{\mathcal{M}}\mathrm{Var}[\mu_i] \tag{A8}$$

That is, the sales-weighted average can be expressed as the cost-weighted average plus a term that reflects the cross-sectional dispersion in markups.

Multiplicative decomposition. A slightly more intuitive version of this decomposition obtains if we decompose the markups μ_i multiplicatively into the common $\mathcal{M}$ component and an idiosyncratic component $\hat{\mu}_i$ with mean normalized to one

$$\hat{\mu}_i := \mu_i / \mathcal{M} \tag{A9}$$

Then $\text{Var}[\mu_i] = \mathcal{M}^2 \text{Var}[\hat{\mu}_i]$ and we can write

$$\frac{\tilde{\mathcal{M}} - \mathcal{M}}{\mathcal{M}} = \text{Var}[\hat{\mu}_i] \tag{A10}$$

That is, the percentage difference between the sales-weighted average and the cost-weighted average is given by the cross-sectional variance of the idiosyncratic component $\hat{\mu}_i$.

Hence $\tilde{\mathcal{M}} \geq \mathcal{M}$ with equality only if there is no markup dispersion. The statistic $\tilde{\mathcal{M}}$ can rise over time either due to increasing $\mathcal{M}$ or increasing $\text{Var}[\hat{\mu}_i]$ or both. The statistic $\tilde{\mathcal{M}}$ can be rising even if $\mathcal{M}$ is constant. Indeed $\tilde{\mathcal{M}}$ can be rising even if $\mathcal{M}$ is falling if the increase in dispersion $\text{Var}[\hat{\mu}_i]$ is large enough.

Compustat example. To get a quantitative sense of the difference between the cost-weighted average $\mathcal{M}$ and the sales-weighted average $\tilde{\mathcal{M}}$, we compute these statistics using publicly available Compustat data for the US economy. We follow the approach of De Loecker, Eeckhout and Unger (2020) using the ratio of sales to the cost of goods sold, scaled by estimates (at the 2-digit industry level) of the output elasticity of the production function from Karabarbounis and Neiman (2019). We show the results in Figure A1.[28] Clearly the sales weighted average $\tilde{\mathcal{M}}$ is higher and has risen by substantially more than the cost-weighted average $\mathcal{M}$. The additional increase in $\tilde{\mathcal{M}}$ reflects the increasing dispersion of markups.

Although researchers may not always have reliable data on total variable costs, under the assumption that all firms within a given industry share the same cost elasticity ϑ, the cost-weighted *arithmetic* average is equivalent to the sales-weighted *harmonic* average, which can of course be computed if the sales-weighted arithmetic average can.

B Computational details

In this appendix we outline how we compute the steady state of the model and the transitional dynamics.

B.1 Monopolistic competition

We first use our aggregation results to calculate the aggregate markup $\mathcal{M}_t$ and aggregate productivity Z_t. In our monopolistic competition model, sectors are identical and these are time-invariant functions of the aggregate mass of producers N_t, say

$$\mathcal{M}_t = \mathcal{M}(N_t), \qquad \text{and} \qquad Z_t = Z(N_t) \tag{B1}$$

Calculating these objects requires solving for firm-level markups. To be concrete we illustrate using our Kimball specification. For this specification we can write the problem of a firm with productivity z as choosing relative output

$$q(z; A) = \underset{q \geq 0}{\text{argmax}} \left[\Upsilon'(q)q - \frac{A}{z} q \right] \tag{B2}$$

where $A > 0$ is a scalar that summarizes the aggregate conditions faced by an individual firm, including the overall amount of competition, as determined by the demand index D and the unit cost of production Ω,

---

[28] See also Figure II, Panel B in De Loecker, Eeckhout and Unger (2020).

as determined by the equilibrium wage and rental rate. Solving this problem for an arbitrary A gives the relative quantity $q(z; A)$, which satisfies the complementary slackness condition

$$\Big[\Upsilon'(q(z;A)) - \mu(q(z;A))\frac{A}{z}\Big]q(z;A) = 0 \tag{B3}$$

where $\mu(q) = \sigma(q)/(\sigma(q) - 1)$ is the markup of a firm of size q and where for our Kimball specification $\sigma(q) = \bar{\sigma}q^{-\varepsilon/\bar{\sigma}}$. The equilibrium value of A is then pinned down by satisfying the Kimball aggregator

$$N \int \Upsilon(q(z;A))\, dG(z) = 1 \tag{B4}$$

We then have $A(N)$ for any arbitrary mass of producers $N > 0$. This mapping is time-invariant because the distribution $G(z)$ is time-invariant.

To implement this, we discretize $G(z)$ using Gauss-Legendre quadrature with 5000 grid points and obtain $q(z; A)$ using a non-linear solver for each of these grid points. We then use another non-linear solver to find the equilibrium $A(N)$ that satisfies the Kimball aggregator. With the optimal relative output $q(z; A(N))$ and markups $\mu(q(z; A(N))$ in hand, we can calculate the aggregate markup $\mathcal{M}(N)$ and aggregate productivity $Z(N)$ using our aggregation results

$$\mathcal{M}(N) = \left(\frac{\displaystyle\int \frac{1}{\mu(q(z;A(N)))}\,\Upsilon'(q(z;A(N)))q(z;A(N))\, dG(z)}{\displaystyle\int \Upsilon'(q(z;A(N)))q(z;A(N))\, dG(z)}\right)^{-1} \tag{B5}$$

and

$$Z(N) = \left(N \int \frac{q(z;A(N))}{z}\, dG(z)\right)^{-1} \tag{B6}$$

We interpolate the functions $\mathcal{M}(N)$ and $Z(N)$ using Chebyshev polynomials and solve the resulting system of equations that characterize the steady state and transition dynamics using the perfect foresight solver in DYNARE. The advantage of the model with monopolistic competition is that the free-entry condition can be written as

$$\kappa W_t = \beta \sum_{j=1}^{\infty} (\beta(1-\varphi))^{j-1}\, \frac{C_t}{C_{t+j}} \left(1 - \frac{1}{\mathcal{M}_{t+j}}\right) \frac{Y_{t+j}}{N_{t+j}} \tag{B7}$$

and is therefore straightforward to evaluate alongside the other equilibrium conditions. We use a similar approach to solve for the efficient allocations, replacing the decentralized equilibrium conditions with the first-order conditions that characterize the planner's allocations.

B.2 Oligopolistic competition

With oligopolistic competition, the distribution of productivity is no longer sector- and time-invariant. Rather, each sector s is characterized by a productivity vector $\boldsymbol{z}(s) = (z_1(s), z_2(s), \ldots, z_{n(s)}(s))$ of the $n(s)$ firms in that sector. Notice here that $\boldsymbol{z}(s)$ varies across sectors both because the number of firms varies and because, with a finite number of firms, the exact configuration of productivity draws also varies even for two sectors with the same number of firms.

Let $\lambda(\boldsymbol{z})$ denote the distribution of productivity vectors $\boldsymbol{z}$ across sectors. For a given $\lambda(\boldsymbol{z})$, we can solve for the aggregate markup and aggregate productivity by first calculating the within-industry equilibrium for each $\boldsymbol{z}$. For example, when firms compete in quantities, we solve the following system of $2n(s)$ equations

$$\mu(z_i, s) = \frac{1}{1 - \left(\frac{1}{\eta}\omega(z_i, z) + \frac{1}{\gamma}(1 - \omega(z_i, s))\right)} \tag{B8}$$

$$\omega(z_i, s) = \frac{\mu(z_i, s)^{1-\gamma}\, z_i^{\gamma-1}}{\sum_{i=1}^{n(s)} \mu(z_i, s)^{1-\gamma}\, z_i^{\gamma-1}} \tag{B9}$$

for each firm $i = 1, 2, \dots, n(s)$ in each sector s. We can then use the resulting distribution of markups and relative size within and across sectors and the aggregation results in the main text to calculate the aggregate markup and aggregate productivity. Our assumption that entry is random, not directed at individual sectors, allows us to write these aggregate variables as functions of the average number of firms per sector, $N = \int_0^1 n(s)\,ds$, just as in the model with monopolistic competition.

Now consider the free-entry condition. To evaluate this condition, we need to recognize that a potential entrant understands that, because there are a finite number of firms, its entry will change the equilibrium in the sector it enters. If an entrant is assigned to sector s with existing productivity distribution $\boldsymbol{z}(s) = (z_1(s), z_2(s), \dots, z_{n(s)}(s))$ the entrant understands that the configuration of productivity will become

$$\boldsymbol{z}'(s) = (\boldsymbol{z}(s), z) \tag{B10}$$

where z is the entrant's productivity, independently drawn from $G(z)$.

To implement this, we solve for the industry equilibrium for every sector and every possible draw of z. In practice we have more than 300 firms per sector, it infeasible to use tensor-based Gaussian quadrature to approximate the distribution $\lambda(\boldsymbol{z})$ across sectors. Instead, we use Monte-Carlo methods to approximate $\lambda(\boldsymbol{z})$ across 100,000 sectors (we also verify that our answers do not change when we increase the number of sectors further). We again use Gauss-Legendre quadrature to approximate the univariate distribution $G(z)$.

Let $\hat{\Pi}(N)$ denote a firm's expected profits per period (scaled by aggregate output) from entering and drawing productivity z from $G(z)$ and being assigned to a random sector s with initial productivity configuration $\boldsymbol{z}(s)$, that is

$$\hat{\Pi}(N) = \int \left(\int_0^1 \left(1 - \frac{1}{\mu(z, (\boldsymbol{z}(s), z))} \right) \omega(z, (\boldsymbol{z}(s), z))\, \bar{\omega}(\boldsymbol{z}(s), z)\, ds \right) dG(z) \tag{B11}$$

where $\mu(z, (\boldsymbol{z}(s), z))$ and $\omega(z, (\boldsymbol{z}(s), z))$ denote the markup and market share of an individual firm with productivity z in a sector with productivity configuration $\boldsymbol{z}'(s) = (\boldsymbol{z}(s), z)$ and where $\bar{\omega}(\boldsymbol{z}(s), z)$ denotes the associated market share of sector s to which the firm is assigned. Because firms are randomly assigned, these expected profits depend only on the average number of firms N, not the entire productivity distribution.

The free-entry condition can then be written as

$$\kappa W_t \geq \beta \frac{C_t}{C_{t+1}} Q_{t+1} \tag{B12}$$

where

$$Q_t = \hat{\Pi}(N_t) Y_t + \beta(1 - \varphi) \frac{C_t}{C_{t+1}} Q_{t+1} \tag{B13}$$

As with the monopolistic competition case, we use Chebyshev polynomials to approximate the time-invariant functions $\hat{\Pi}(N)$, $\mathcal{M}(N)$ and $Z(N)$, which then allows us to use standard methods to characterize the equilibrium transition dynamics.

Computing the function $\hat{\Pi}(N)$ is the key step and is extremely time consuming, because doing so requires resolving the industry equilibrium for every sector the firm may be assigned to for every possible realization of its own productivity draw. But this step only has to be done once. Our assumption that entry is random is key to making even this feasible. If instead firms can direct their entry to individual sectors, one can no longer interchange the order of integration used to calculate $\hat{\Pi}(N)$ from (B11) and we would need to characterize the equilibrium law of motion for the vector $\boldsymbol{z}_{t+1}(s)$ given the current vector $\boldsymbol{z}_t(s)$ and the individual entry decisions, as well as how a firm's profits vary with both its own and its competitors' productivity, $\pi(z, (\boldsymbol{z}_t(s), z))$, in order to compute the expected present value of profits from entering a sector with a given vector of $\boldsymbol{z}_t(s)$ of incumbents' productivities. Because these are very high-dimensional objects, computing this alternative model would require resorting to a dimensionality-reduction approximation in the spirit of Krusell and Smith (1998).

C Data and empirical analysis

We use data from the US Census of Manufactures from 1972 to 2012. We focus on the Census of Manufactures for two reasons: (i) it has higher-quality input data relative to other sectors, such as Services, and (ii) the vast majority of manufacturing goods are easily transportable and not limited to local markets.

We start with the value of shipments $p_{eit}(s)y_{eit}(s)$ and total salaries/wages $W_t l_{eit}(s)$ for each establishment e of firm i in industry s. In the case of a single-establishment firm i in industry s, we have

$$\mu_{it}(s) = \mu_{eit}(s) = \frac{p_{eit}(s)y_{eit}(s)}{W_t l_{eit}(s)}\zeta_t(s) \tag{C1}$$

where $\zeta_t(s)$ summarizes output elasticities specific to industry s. In the slightly more complicated case of multi-establishment firms, we first calculate an establishment e specific markup in industry s and then aggregate over the establishments e of firm i to get

$$\mu_{it}(s) = \zeta_t(s) \sum_{e \in i} \mu_{eit}(s) \frac{W_t l_{eit}(s)}{\sum_{e' \in i} W_t l_{e'it}(s)} \tag{C2}$$

We trim the outliers by winsorizing μ_{eit} at the top and bottom 5 percentile of each Census year.

The empirical literature has proposed various strategies for recovering the output elasticities $\zeta_t(s)$ specific to industry s. In principle, one could estimate industry-specific production functions to recover these elasticities. However, recently Bond, Hashemi, Kaplan and Zoch (2020) have shown that in the presence of variable markups it is not possible to consistently estimate output elasticities when only revenue data is available. Given this, we follow an alternative approach, more in the spirit of growth accounting, where we use the firm's cost minimization conditions to write, for each establishment e and firm i

$$\zeta_t(s) = \frac{W_t l_{eit}(s)}{W_t l_{eit}(s) + R_t k_{eit}(s) + x_{eit}(s)} \tag{C3}$$

Because of measurement error at the establishment level, we take averages within industry s. Following Foster, Grim and Haltiwanger (2016), we estimate $\hat{\zeta}_t(s)$ using the cost-weighted average of labor cost shares of establishments in each industry s. We construct $\hat{\zeta}_t(s)$ separately for each Census year.

With $\hat{\zeta}_t(s)$ in hand, we can construct $\hat{\mu}_{it}(s)$ from (C2) and estimate key parameters of the model using each specifications' implications for the relationship between markups and market shares. For example, the Kimball version of the model implies

$$\frac{1}{\mu_{it}(s)} + \log\left(1 + \frac{1}{\mu_{it}(s)}\right) = a + b \log \omega_{it}(s) \tag{C4}$$

Our model abstracts from other sources of persistent firm-level and industry-level heterogeneity, such as variations in capital shares or returns to scale, and from other distortions (implicit or explicit taxes, etc) that may drive a wedge between cost shares and $1/\mu_{it}(s)$. As a result, in our empirical analysis, we control for firm fixed effects in addition to the 6-digit NAICS industry-year effects. The industry-year effects control for any short-run fluctuation of industry demand or input costs. We obtain an estimated slope coefficient $\hat{b} = 0.162$ with standard error 0.002 clustered at the firm level.

Estimates based on Taiwanese product-level data. As a robustness check, we have also estimated the slope coefficient b using a rich product-level panel dataset from Taiwanese manufacturing that we previously studied in Edmond, Midrigan and Xu (2015). The Taiwanese data is more detailed than the US Census data and allows us to control for any product-year specific effects. We again construct markups using the inverse of cost shares and estimate (C4) in two ways. In the first approach we exploit the cross-sectional variation of producers within a given product category by including product-year fixed effects. This gives an estimate of $\hat{b} = 0.15$ that is tightly estimated with a standard error of 0.002. In the second approach we exploit the panel structure of the data and include a producer fixed effect, thus using the time-series co-movement of a producer's sales and their markups to estimate b. This gives an estimate of $\hat{b} = 0.16$ with a standard error of 0.007, almost identical to our benchmark estimate $\hat{b} = 0.162$ from the US Census data.

D Static welfare calculation

In this appendix we derive a simple formula for the welfare losses from markups in a steady state version of our model. Suppose that the representative consumer has preferences

$$U(C,L)=\frac{C^{1-\sigma}}{1-\sigma}-\frac{L^{1+\nu}}{1+\nu} \tag{D1}$$

Suppose also that labor is the only factor of production[29] and that there is a representative firm with production function $Y = ZL$. Markups distort allocations by reducing aggregate productivity Z and by introducing a wedge $\mathcal{M}$ between the wage and marginal product of labor, $W = Z/\mathcal{M}$. Labor supply is given by $C^\sigma L^\nu = W = Z/\mathcal{M}$. Using goods market clearing $C = Y = ZL$, employment and consumption in the distorted allocation are given by

$$L=\mathcal{M}^{-\frac{1}{\sigma+\nu}}\,Z^{\frac{1-\sigma}{\sigma+\nu}},\qquad \text{and}\qquad C=\mathcal{M}^{-\frac{1}{\sigma+\nu}}\,Z^{\frac{1+\nu}{\sigma+\nu}} \tag{D2}$$

The associated level of utility is

$$U(C,L)=\left(\frac{1}{1-\sigma}-\frac{1}{1+\nu}\frac{1}{\mathcal{M}}\right)\mathcal{M}^{-\frac{1-\sigma}{\sigma+\nu}}\,Z^{\frac{(1+\nu)(1-\sigma)}{\sigma+\nu}} \tag{D3}$$

Similarly, the level of utility in the efficient allocation is

$$U(C^*,L^*)=\left(\frac{1}{1-\sigma}-\frac{1}{1+\nu}\right)Z^{*\,\frac{(1+\nu)(1-\sigma)}{\sigma+\nu}} \tag{D4}$$

Let $\mathcal{W}$ denote the level of consumption solving $U(\mathcal{W},0) = U(C,L)$ for the distorted allocation, namely

$$\mathcal{W}=\left(1-\frac{1-\sigma}{1+\nu}\frac{1}{\mathcal{M}}\right)^{\frac{1}{1-\sigma}}\mathcal{M}^{-\frac{1}{\sigma+\nu}}\,Z^{\frac{1+\nu}{\sigma+\nu}} \tag{D5}$$

Similarly, let $\mathcal{W}^*$ denote the level of consumption solving $U(\mathcal{W}^*,0) = U(C^*,L^*)$ for the efficient allocation

$$\mathcal{W}^*=\left(1-\frac{1-\sigma}{1+\nu}\right)^{\frac{1}{1-\sigma}}Z^{*\,\frac{1+\nu}{\sigma+\nu}} \tag{D6}$$

Hence the consumption-equivalent losses from markups can be written

$$\frac{\mathcal{W}}{\mathcal{W}^*}=\left(\frac{\left(1-\frac{1-\sigma}{1+\nu}\frac{1}{\mathcal{M}}\right)}{\left(1-\frac{1-\sigma}{1+\nu}\right)}\right)^{\frac{1}{1-\sigma}}\left(\frac{Z}{Z^*}\right)^{\frac{1+\nu}{\sigma+\nu}}\mathcal{M}^{-\frac{1}{\sigma+\nu}} \tag{D7}$$

With logarithmic utility, $\sigma \to 1$, as in the main text, this simplifies to

$$\frac{\mathcal{W}}{\mathcal{W}^*}=\left(\frac{Z}{Z^*}\right)\mathcal{M}^{-\frac{1}{\sigma+\nu}} \tag{D8}$$

To illustrate, if misallocation reduces aggregate productivity to $Z/Z^* = 0.99$ and the aggregate markup is $\mathcal{M} = 1.15$ with $\nu = 1$ as in our benchmark model, then this static formula implies $\mathcal{W}/\mathcal{W}^* = 0.9232$, a welfare loss of -7.68% in consumption-equivalent terms.

---

[29] A steady-state calculation including capital would overstate the costs of markups because it would ignore the deferred consumption required to build up the efficient capital stock.

E Aggregate markup analytics

In this appendix we characterize analytically the time-invariant function $\mathcal{M}(N)$ mapping the mass of firms into the aggregate markup for our two monopolistic competition specifications: (i) Kimball demand, and (ii) symmetric translog demand. This time invariant function, along with its counterpart for aggregate productivity $Z(N)$, plays a crucial role in solving our model.

The results below are in the spirit of results in Arkolakis, Costinot, Donaldson and Rodríguez-Clare (2019), but unlike in their analysis, we do not assume from the outset that the choke price in either demand system is binding, since this is an equilibrium outcome. In addition, for the translog case we provide a closed-form solution for the aggregate markup that may be of some independent interest.

E.1 Kimball demand

First observe that a firm's employment is proportional to its relative size scaled by productivity, $l(z) \sim q(z)/z$, so we can write the aggregate markup as the cost-weighted average

$$\mathcal{M} = \frac{\displaystyle\int_1^\infty \mu(q(z))\frac{q(z)}{z}\,dG(z)}{\displaystyle\int_1^\infty \frac{q(z)}{z}\,dG(z)} \tag{E1}$$

With Kimball demand, a firm's relative size $q(z)$ is pinned down by the static markup pricing condition

$$\Upsilon'(q) = \mu(q)\frac{A}{z} \tag{E2}$$

where $A > 0$ is an endogenous aggregate variable that depends on the demand index and the unit costs of production. Hence a firm's optimal size $q(z; A)$ is a function only of the ratio z/A and we can write $q(z/A)$. Plugging this back into the Kimball aggregator gives

$$N\int_1^\infty \Upsilon(q(z/A))\,dG(z) = 1 \tag{E3}$$

This implicitly determines $A(N)$. Since $q(z/A)$ is increasing in z/A for each z, from the implicit function theorem we obtain that $A'(N) > 0$, i.e., that a larger mass of firms N makes the market more competitive and shrinks the relative size of each firm $q(z/A(N))$.

We can then use a change of variables $\hat{z} = z/A$ and the assumption that $G(z)$ is Pareto to write the aggregate markup as a function of N via $A(N)$, namely

$$\mathcal{M}(N) = \frac{\displaystyle\int_{1/A(N)}^\infty \mu(q(\hat{z}))\frac{q(\hat{z})}{\hat{z}}\,dG(\hat{z})}{\displaystyle\int_{1/A(N)}^\infty \frac{q(\hat{z})}{\hat{z}}\,dG(\hat{z})} \tag{E4}$$

Hence changes in the number of competitors, summarized by changes in $A(N)$, only change the aggregate markup through their effect on the markups of the smallest firms. A direct calculation then gives

$$\mathcal{M}'(N) = (\mu_{min} - \mathcal{M}) \times \frac{q_{min}\,g_{min}}{\displaystyle\int_{1/A(N)}^\infty \frac{q(\hat{z})}{\hat{z}}\,dG(\hat{z})} \times \frac{A'(N)}{A(N)} \leq 0 \tag{E5}$$

where $\mu_{min} = \mu(1/A)$ and $q_{min} = q(1/A)$ are shorthand for the markups and relative size of the smallest type of firm, which has density in the population $g_{min} = g(1/A)$. A larger mass of firms N makes the market more competitive, increasing $A(N)$, and since the markups of the smallest firms are smaller than the markup of the average, $\mu_{min} \leq \mathcal{M}$, the aggregate markup falls.

Cutoff productivity $z^*(N)$**.** This derivation implicitly assumes that all firms have interior solutions to $\Upsilon'(q) = \mu(q)A(N)/z$ pinning down their relative size. But if $A(N)$ is sufficiently large, i.e., if N is sufficiently large, then firms with low productivity are at a corner solution and produce nothing. In particular, there is a cutoff productivity z^* satisfying $\Upsilon'(0) = A(N)/z^*$ such that all firms with $z \leq z^*$ have relative size $q = 0$. Since $G(z)$ is bounded below by 1 and $\Upsilon'(0) = (\bar{\sigma} - 1)e^{1/\varepsilon}/\bar{\sigma}$ from (57), we can write this cutoff

$$z^*(N) = \max \left[1 \, , \, \frac{\bar{\sigma}}{\bar{\sigma} - 1} \, e^{-\frac{1}{\varepsilon}} \, A(N) \right] \tag{E6}$$

where $A(N)$ solves the Kimball aggregator (E3) and is strictly increasing in N. In short if the mass of firms N is sufficiently small, then $z^* = 1$ and there are no selection effects. But if N is sufficiently large, then $z^* > 1$ and there are positive selection effects which become stronger the larger is N.

Now observe that if $z^* = 1$, then $q_{min} > 0$ so that, from (E5), for sufficiently small N the aggregate markup $\mathcal{M}(N)$ is strictly decreasing in N. But if $z^* > 1$, i.e., the choke price is binding, then $q_{min} = 0$ and the aggregate markup $\mathcal{M}(N)$ is invariant to N. In other words, for small N, increases in N are absorbed by a decline in the aggregate markup with no change in selectivity, but for larger N, increases in N are absorbed by an increase in selectivity with no further change in the aggregate markup. This latter case echoes Arkolakis, Costinot, Donaldson and Rodríguez-Clare (2019), but here we see that whether or not the choke price is binding is determined by $A(N)$, which then varies over time as the mass of firm evolves.

In our benchmark calibration of the Kimball model, there are no selection effects, $z^* = 1$, but the smallest firms of size q_{min} are tiny so the effects of changes in N on the aggregate markup are likewise tiny.

Aggregate productivity $Z(N)$**.** Similarly, aggregate productivity $Z(N)$ is a time-invariant function of the mass of firms. Following the same steps as for the aggregate markup, we can write

$$Z(N) = \left(N A(N)^{-\xi - 1} \int_{1/A(N)}^{\infty} \frac{q(\hat{z})}{\hat{z}} \, dG(\hat{z}) \right)^{-1} \tag{E7}$$

which likewise determines $Z(N)$ given the $A(N)$ which solves the Kimball aggregator (E3).

E.2 Translog demand

Symmetric translog demand is sufficiently tractable that we can obtain a closed form solution for $\mathcal{M}(N)$. The qualitative properties are essentially the same as for the Kimball specification.

Cutoff productivity $z^*(N)$**.** We first characterize the cutoff productivity z^* as a function of the mass of firms N. With symmetric translog demand, firm-level markups $\mu(z)$ implicitly solve

$$\mu + \log \mu = 1 + \log(z/z^*), \qquad z > z^* \tag{E8}$$

with $\mu(z) = 1$ for all $z \leq z^*$ where z^* is the cutoff productivity dual to the choke price

$$\log p^* = \frac{1}{\bar{\sigma} N} + \int \log p(z) \, dG(z) \tag{E9}$$

Using $p^* z^* = \Omega$ and $p(z) = \mu(z)\Omega/z$ we can rewrite the choke price as a condition on the cutoff productivity

$$(1 - G(z^*)) \log z^* = -\Big(\frac{1}{\bar{\sigma} N} + \int_{z^*}^{\infty} \log \big(\frac{\mu(z)}{z} \big) \, dG(z) \Big) \tag{E10}$$

To simplify this we need to calculate the integral on the RHS. Using (E8) to rewrite the integrand, we get

$$
\begin{aligned}
\int_{z^*}^{\infty} \log\big(\frac{\mu(z)}{z}\big)\, dG(z) &= \int_{z^*}^{\infty} \big(1-\mu(z)-\log z^*\big)\, dG(z) \\
&= \big(1-\ln z^*\big)(1-G(z^*)) - \int_{z^*}^{\infty} \mu(z)\, g(z)\, dz \\
&= \big(1-\ln z^*\big)(1-G(z^*)) - \int_{1}^{\infty} \mu\, g(z(\mu)) z'(\mu)\, d\mu \\
&= \big(1-\ln z^*\big)(1-G(z^*)) - \xi \int_{1}^{\infty} (1+\mu)\left\{\, z^*\mu e^{\mu-1} \,\right\}^{-\xi} d\mu \\
&= \big(1-\ln z^*\big)(1-G(z^*)) - (1-G(z^*))\xi \int_{1}^{\infty} (1+\mu)\mu^{-\xi}\, e^{-\xi(\mu-1)}\, d\mu
\end{aligned}
\tag{E11}
$$

where the third line changes the variable of integration from z to μ using $\mu(z^*)=1$ and we then use the inverse $z(\mu)=z^*\mu e^{\mu-1}$ implied by (E8) and its derivative $z'(\mu)$ and the Pareto density $g(z)=\xi z^{-\xi-1}$ recognizing that ${z^*}^{-\xi}=1-G(z^*)$. Substituting this formula for the integral back into (E10), cancelling common terms and simplifying then gives

$$\frac{1}{1-G(z^*)} = {z^*}^{\xi} = \bar{\sigma} N (I(\xi)-1) \tag{E12}$$

where $I(\xi)$ is the constant

$$I(\xi) := \xi \int_{1}^{\infty} (1+\mu)\mu^{-\xi}\, e^{-\xi(\mu-1)}\, d\mu \tag{E13}$$

which depends only on the Pareto tail ξ. To simplify this further, note that we can write $I(\xi)$ in terms of the generalized exponential integral

$$I(\xi) = 1 + e^{\xi} E_{\xi}(\xi), \qquad E_n(x) := \int_{1}^{\infty} \frac{e^{-xt}}{t^n}\, dt \tag{E14}$$

Since the distribution $G(z)$ is bounded below by 1, our solution for the cutoff productivity is

$$\boxed{z^*(N) = \max\Big[\, 1\,,\ \bar{\sigma} N\, e^{\xi} E_{\xi}(\xi)\, \Big]^{1/\xi}} \tag{E15}$$

As with the Kimball specification, if the mass of firms N is sufficiently small, then $z^*=1$ and there are no selection effects. But if N is sufficiently large, then $z^*>1$ and there are positive selection effects which become stronger the larger is N.

Aggregate markup $\mathcal{M}(N)$. For the translog case, begin by writing the aggregate markup as the sales-weighted *harmonic* average and use the translog's linear relationship between markups and market shares

$$\mathcal{M}^{-1} = N \int_{1}^{\infty} \frac{\omega(z)}{\mu(z)}\, dG(z) = \bar{\sigma} N \int_{1}^{\infty} \frac{\mu(z)-1}{\mu(z)}\, dG(z) = \bar{\sigma} N \int_{z^*}^{\infty} \frac{\mu(z)-1}{\mu(z)}\, dG(z) \tag{E16}$$

Changing variables from z to μ and following the same steps as in the derivation of the cutoff z^* gives

$$\mathcal{M}^{-1} = \bar{\sigma} N (1-G(z^*)) \left\{ 1 - \xi \int_{1}^{\infty} (1+\mu)\, \mu^{-2-\xi}\, e^{-\xi(\mu-1)}\, d\mu \right\} \tag{E17}$$

which we can again write in terms of generalized exponential integrals

$$\mathcal{M}^{-1} = \bar{\sigma} N (1-G(z^*)) \Big(1 - \xi e^{\xi} [E_{\xi+1}(\xi) + E_{\xi+2}(\xi)] \Big) \tag{E18}$$

Since we know $z^*(N)$, this implicitly gives $\mathcal{M}(N)$ too.

But we can say more than this. To simplify further, we consider the cases $z^* > 1$ and $z^* = 1$ in turn. To take the first case, if N is sufficiently high, such that $z^* > 1$, then from (E12) and (E14) we have $1 = \bar{\sigma}N(1 - G(z^*))\, e^{\xi} E_{\xi}(\xi)$ so we can eliminate the multiplicative term $\bar{\sigma}N(1 - G(z^*))$ to get

$$\mathcal{M} = \frac{e^{\xi} E_{\xi}(\xi)}{1 - \xi e^{\xi}[E_{\xi+1}(\xi) + E_{\xi+2}(\xi)]} \tag{E19}$$

This is a constant, independent of N. Although it looks complicated, it simplifies nicely. To do so, we first rewrite the exponential integrals in terms of upper incomplete gamma functions, using the standard result

$$E_n(x) = x^{n-1}\Gamma(1-n, x), \qquad \Gamma(s, x) := \int_x^{\infty} t^{s-1} e^{-t}\, dt \tag{E20}$$

and then use the standard recursion formula for upper incomplete gamma functions

$$\Gamma(s+1, x) = s\Gamma(s, x) + x^s e^{-x} \tag{E21}$$

Using these properties to collect terms and simplify

$$\mathcal{M} = \frac{e^{\xi} E_{\xi}(\xi)}{1 - \xi e^{\xi}[E_{\xi+1}(\xi) + E_{\xi+2}(\xi)]} = \frac{e^{\xi}\xi^{\xi-1}\Gamma(1-\xi, \xi)}{e^{\xi}\xi^{\xi+1}\Gamma\big(-(1+\xi), \xi\big)} = \xi^{-2}\,\frac{\Gamma\big(+1-\xi, \xi\big)}{\Gamma\big(-1-\xi, \xi\big)} \tag{E22}$$

Now note that the ratio of gamma functions on the RHS is of the form

$$\frac{\Gamma(s+2, x)}{\Gamma(s, x)} = s(s+1) + (s+1+x)\frac{x^s e^{-x}}{\Gamma(s, x)} \tag{E23}$$

which follows from iterating forward twice using our recursion (E21). Evaluating this at $s = -(1+\xi)$ and $x = \xi$ and simplifying

$$\frac{\Gamma(+1-\xi, \xi)}{\Gamma(-1-\xi, \xi)} = -(1+\xi)[-(1+\xi)+1] + [-(1+\xi)+(1+\xi)]\frac{\xi^{-(1+\xi)}e^{-\xi}}{\Gamma(-(1+\xi), \xi)} = \xi(1+\xi) \tag{E24}$$

Hence we get the very simple expression for the aggregate markup

$$\boxed{\mathcal{M} = 1 + \frac{1}{\xi}, \qquad \text{if } z^* > 1} \tag{E25}$$

To take the second case, if instead $z^* = 1$, so that $G(z^*) = 0$, then from (E18), (E22) and (E24) we have

$$\mathcal{M} = \Big(1 + \frac{1}{\xi}\Big) \times \Big(\bar{\sigma}N\, e^{\xi} E_{\xi}(\xi)\Big)^{-1}, \qquad \text{if } z^* = 1 \tag{E26}$$

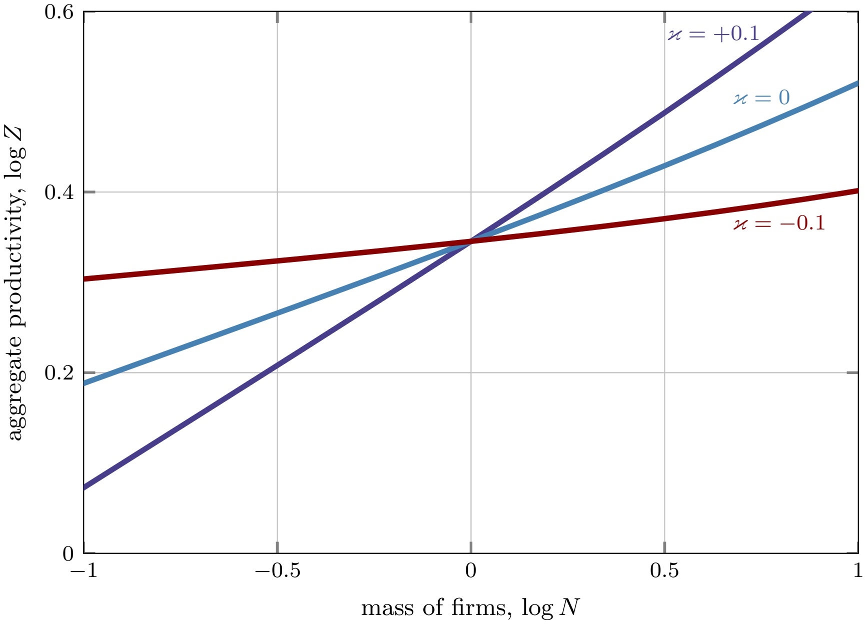
Collecting these cases together, we conclude that

$$\boxed{\mathcal{M}(N) = \Big(1 + \frac{1}{\xi}\Big) \times \Big(\max\Big[\, 1\, ,\ \bar{\sigma}N\, e^{\xi} E_{\xi}(\xi)\,\Big]\Big)^{-1}} \tag{E27}$$

since $z^* = 1$ if $\bar{\sigma}N\, e^{\xi} E_{\xi}(\xi) \leq 1$ and $z^* > 1$ otherwise.

Qualitatively, this is essentially the same as with the Kimball specification. For sufficiently small N the aggregate markup $\mathcal{M}(N)$ is strictly decreasing in N. But if $z^* > 1$, i.e., the choke price is binding, the aggregate markup $\mathcal{M}(N) = 1 + 1/\xi$ is invariant to N and depends only on the amount of productivity dispersion $1/\xi$. Just as with the Kimball specification, for small N, increases in N are absorbed by a decline in the aggregate markup with no change in selectivity, but for larger N, increases in N are absorbed by an increase in selectivity with no further change in the aggregate markup.

Figure F1: Love For Variety Effects



Aggregate productivity $\log Z$ as a function of the mass of varieties $\log N$. The parameter $\varkappa$ controls the strength of the variety effect: $\varkappa = 0$ is our benchmark model, $\varkappa = -0.1$ has a much weaker variety effect, $\varkappa = +0.1$ has a much stronger variety effect.

F Love for variety

Our model with variable markups has a 'love for variety' effect, an increase in N increases aggregate productivity $Z(N)$ because of the concavity of the technology in each individual variety, as in a CES model. In a CES model with identical firms and demand elasticity $\bar{\sigma} > 1$ we would have $Z(N) = N^{\frac{1}{\bar{\sigma}-1}}$, log-linear in N.

To assess the variety effect in our benchmark model, we modify the Kimball aggregator to

$$N^{1+\varkappa} \int_1^\infty \Upsilon(q(z))\, dG(z) = 1 \tag{F1}$$

where $\varkappa$ parameterizes the strength of the variety effect. If $\varkappa = 0$, we have our benchmark model. If $\varkappa < 0$, there is a weaker variety effect, if $\varkappa > 0$ there is a stronger variety effect.

Figure F1 plots $\log Z$ as a function of $\log N$ for a weaker variety effect, $\varkappa = -0.1$ and a stronger variety effect $\varkappa = +0.1$. To interpret these parameter values, recall that in the CES special case, $\Upsilon(q) = q^{\frac{\bar{\sigma}-1}{\bar{\sigma}}}$ we would remove the variety effect altogether by setting $\varkappa = -1/\bar{\sigma}$. In the CES case, this would make $Z(N)$ invariant to N. For our benchmark model calibrated to $\mathcal{M} = 1.15$ we have $\bar{\sigma} = 10.86$, this would require $\varkappa = -0.0921$, say -0.1 in round numbers. Figure F1 shows that $\varkappa = -0.1$ significantly reduces the variety effect but does not eliminate it entirely. With variable markups, and hence a higher profit share, it takes a more negative $\varkappa$ to eliminate the variety effect. In particular, we need a value of $\varkappa$ consistent with the calibrated profit share, something like $\varkappa \approx -(\mathcal{M}-1)/\mathcal{M} = -0.13$.

Table F1 reports the welfare costs of markups under various alternative policy scenarios for $\varkappa = -0.1$ and $\varkappa = +0.1$. With $\varkappa = -0.1$ the planner wants many fewer varieties, but with $\varkappa = +0.1$ the planner wants many more varieties. Notice that regardless of the sign of $\varkappa$, the welfare costs are larger than in our benchmark model. This reflects the additional inefficiency due to entry externalities in the market equilibrium. The value of $\varkappa$ does not affect the amount of misallocation, which remains 0.97% of gross output TFP, as in our benchmark model with $\mathcal{M} = 1.15$. Nonetheless, the welfare gains from size-dependent subsidies are considerably larger. This is because the size-dependent subsidies correct *both* misallocation and

Table F1: Implications of Alternative Policies, Variety Effects

		steady state comparisons, %						
		Y	C	L	N	K	Z	welfare, %
$\varkappa = -0.1$	efficient	42.4	24.4	9.4	-66.9	72.4	-3.7	17.48
	uniform subsidy	47.9	31.8	17.0	10.0	82.9	0.4	3.70
	size-dependent subsidy	-3.9	-5.5	-7.2	-69.8	-6.1	-4.1	11.55
	entry subsidy	-7.5	-8.6	-7.8	-66.8	-11.0	-4.5	8.66
$\varkappa = \ \ 0.0$	efficient	59.7	44.6	18.0	20.1	100.6	4.1	8.69
	uniform subsidy	51.9	35.9	17.0	9.5	88.7	1.5	5.92
	size-dependent subsidy	5.3	6.2	1.0	8.3	6.6	2.3	2.87
	entry subsidy	6.3	7.5	2.4	20.1	8.2	3.1	0.56
$\varkappa = +0.1$	efficient	115.7	108.5	25.3	90.0	189.0	21.5	20.20
	uniform subsidy	55.3	39.6	16.9	9.1	93.8	2.5	8.01
	size-dependent subsidy	41.7	50.4	8.5	71.3	53.7	17.9	13.74
	entry subsidy	47.9	57.9	11.1	91.7	62.7	20.6	11.56

The first six columns report the percentage change from the initial distorted steady state with $\mathcal{M} = 1.15$ to the new steady state. The last column reports the consumption equivalent welfare gains (including transitional dynamics). The parameter $\varkappa$ controls the strength of the variety effect: $\varkappa = 0$ is our benchmark model, $\varkappa = -0.1$ has a much weaker variety effect, $\varkappa = +0.1$ has a much stronger variety effect. The alternative policies are (i): the *efficient allocation*, where all markups are removed, (ii) a *uniform subsidy* that eliminates the aggregate markup, (iii) *size-dependent subsidies* that eliminate misallocation and the entry distortion, and (iv) the uniform *entry subsidy* that leads to the largest welfare gain. Regardless of $\varkappa$ the amount of misallocation is *the same* as in our benchmark. But there are now larger welfare gains because of a more distorted entry margin.

the entry distortion and the entry distortion here is larger than in our benchmark. Although these channels are not perfectly additive, by comparing the gains from the full set of size-dependent subsidies to the gains from the optimal uniform entry subsidy, one can see that the welfare gains from correcting the misallocation distortion are approximately 2.5% in all cases — larger than than the gross output TFP loss because of the standard multiplier effect from intermediates.

References

Akcigit, Ufuk, Salomé Baslandze, and Francesca Lotti, "Connecting to Power: Political Connections, Innovation, and Firm Dynamics," October 2018. NBER Working Paper 25136.

Arkolakis, Costas, Arnaud Costinot, Dave Donaldson, and Andrés Rodríguez-Clare, "Gains From Trade under Monopolistic Competition: A Simple Example with Translog Expenditure Functions and Pareto Distributions of Firm-Level Productivity," September 2010. Yale University Working Paper.

_ , _ , _ , and _ , "The Elusive Pro-Competitive Effects of Trade," *Review of Economic Studies*, January 2019, *86* (1), 46–80.

Atkeson, Andrew and Ariel Burstein, "Pricing-to-Market, Trade Costs, and International Relative Prices," *American Economic Review*, 2008, *98* (5), 1998–2031.

_ and _ , "Innovation, Firm Dynamics, and International Trade," *Journal of Political Economy*, June 2010, *118* (3), 433–484.

_ and _ , "The Aggregate Implications of Innovation Policy," *Journal of Political Economy*, 2019, *127* (6), 2625–2683.

_ , _ , and Manolis Chatzikonstantinou, "Transitional Dynamics in Aggregate Models of Innovative Investment," *Annual Review of Economics*, August 2019, *11*, 273–301.

Autor, David, David Dorn, Lawrence F. Katz, Christina Patterson, and John Van Reenen, "The Fall of the Labor Share and the Rise of Superstar Firms," *Quarterly Journal of Economics*, May 2020, *135* (2), 645–709.

Baqaee, David Rezza and Emmanuel Farhi, "Productivity and Misallocation in General Equilibrium," *Quarterly Journal of Economics*, February 2020, *135* (1), 105–163.

Barkai, Simcha, "Declining Labor and Capital Shares," *Journal of Finance*, 2020, *75* (5), 2421–2463.

Bartelsman, Eric, John Haltiwanger, and Stefano Scarpetta, "Cross-Country Differences in Productivity: The Role of Allocation and Selection," *American Economy Review*, 2013, *103* (1), 305–334.

Basu, Susanto, "Are Price-Cost Markups Rising in the United States? A Discussion of the Evidence," *Journal of Economic Perspectives*, 2019, *33* (3), 3–22.

Behrens, Kristian, Giordano Mion, Yasusada Murata, and Jens Suedekum, "Quantifying the Gap between Equilibrium and Optimum Under Monopolistic Competition," *Quarterly Journal of Economics*, November 2020, *134* (4), 2299–2360.

Bernard, Andrew B., Jonathan Eaton, J. Bradford Jensen, and Samuel Kortum, "Plants and Productivity in International Trade," *American Economic Review*, September 2003, *93* (4), 1268–1290.

Bilbiie, Florin O., Fabio Ghironi, and Marc J. Melitz, "Monopoly Power and Endogenous Product Variety: Distortions and Remedies," October 2008. NBER Working Paper 14383.

_ , _ , and _ , "Monopoly Power and Endogenous Product Variety: Distortions and Remedies," *American Economic Journal: Macroeconomics*, October 2019, *11* (4), 140–174.

Boar, Corina and Virgiliu Midrigan, "Markups and Inequality," November 2020. NBER Working Paper 25952.

Bond, Steve, Arshia Hashemi, Greg Kaplan, and Piotr Zoch, "Some Unpleasant Markup Arithmetic: Production Function Elasticities and their Estimation from Production Data," April 2020. NBER Working Paper 27002.

De Loecker, Jan, Jan Eeckhout, and Gabriel Unger, "The Rise of Market Power and the Macroeconomic Implications," *Quarterly Journal of Economics*, May 2020, *135* (2), 561–644.

_ , _ , and Simon Mongey, "Quantifying Market Power and Business Dynamism in the Macroeconomy," April 2021. KU Leuven Working Paper.

Dhingra, Swati and John Morrow, "Monopolistic Competition and Optimum Product Diversity Under Firm Heterogeneity," *Journal of Political Economy*, February 2019, *127* (1), 196–232.

Dixit, Avinash K. and Joseph E. Stiglitz, "Monopolistic Competition and Optimum Product Diversity," *American Economic Review*, 1977, *67* (3), 297–308.

Edmond, Chris, Virgiliu Midrigan, and Daniel Yi Xu, "Competition, Markups, and the Gains from International Trade," *American Economic Review*, October 2015, *105* (10), 3183–3221.

Eslava, Marcela and John C. Haltiwanger, "The Size and Life-Cycle Growth of Plants: The Role of Productivity, Demand and Wedges," May 2020. NBER Working Paper 27184.

Feenstra, Robert C., "A Homothetic Utility Function for Monopolistic Competition Models, Without Constant Price Elasticity," *Economics Letters*, January 2003, *78* (1), 79–86.

Foster, Lucia, Cheryl Grim, and John Haltiwanger, "Reallocation in the Great Recession: Cleansing or Not?," *Journal of Labor Economics*, January 2016, *34* (S1), S293–S331.

Grassi, Basile, "IO in I-O: Size, Industrial Organization, and the Input-Output Network Make a Firm Structurally Important," December 2017. Bocconi University Working Paper.

Gutiérrez, Germán and Thomas Phillippon, "Declining Competition and Investment in the US," November 2017. NYU Stern Working Paper.

_ and _ , "Investment-Less Growth: An Empirical Investigation," *Brooking Papers on Economic Activity*, Fall 2017, pp. 89–169.

Hall, Robert E., "New Evidence on the Markup of Prices over Marginal Costs and the Role of Mega-Firms in the US Economy," May 2018. NBER Working Paper.

Hsieh, Chang-Tai and Peter J. Klenow, "Misallocation and Manufacturing TFP in China and India," *Quarterly Journal of Economics*, November 2009, *124* (4), 1403–1448.

Jones, Charles I., "Intermediate Goods and Weak Links in the Theory of Economic Development," *American Economic Journal: Macroeconomics*, 2011, *3* (2), 1–28.

Karabarbounis, Loukas and Brent Neiman, "Accounting for Factorless Income," in Martin Eichenbaum and Jonathan A. Parker, eds., *NBER Macroeconomics Annual*, University of Chicago Press, June 2019.

Kehrig, Matthias and Nicolas Vincent, "The Micro-Level Anatomy of the Aggregate Labor Share Decline," *Quarterly Journal of Economics*, May 2021, *136* (2), 1031–1087.

Kimball, Miles S., "The Quantitative Analytics of the Basic Neomonetarist Model," *Journal of Money, Credit, and Banking*, 1995, *27* (4, Part 2), 1241–1277.

Klenow, Peter J. and Jonathan L. Willis, "Real Rigidities and Nominal Price Changes," *Economica*, July 2016, *83*, 443–472.

Krusell, Per and Anthony A. Smith, "Income and Wealth Heterogeneity in the Macroeconomy," *Journal of Political Economy*, October 1998, *106* (5), 867–896.

Lerner, Abba P., "The Concept of Monopoly and the Measurement of Monopoly Power," *Review of Economic Studies*, 1934, *1* (3), 157–175.

Peters, Michael, "Heterogeneous Mark-Ups, Growth and Endogenous Misallocation," *Econometrica*, September 2020, *88* (5), 2037–2073.

Restuccia, Diego and Richard Rogerson, "Policy Distortions and Aggregate Productivity with Heterogeneous Establishments," *Review of Economic Dynamics*, October 2008, *11* (4), 707–720.

Rodriguez-Lopez, Jose Antonio, "Prices and Exchange Rates: A Theory of Disconnect," *Review of Economic Studies*, July 2011, *78* (3), 1135–1177.

Rossi-Hansberg, Esteban, Pierre-Daniel Sarte, and Nicholas Trachter, "Diverging Trends in National and Local Concentration," in Martin Eichenbaum and Erik Hurst, eds., *NBER Macroeconomics Annual*, University of Chicago Press, 2020.

Zhelobodko, Evgeny, Sergey Kokovin, Mathieu Parenti, and Jacques-François Thisse, "Monopolistic Competition: Beyond the Constant Elasticity of Substitution," *Econometrica*, November 2012, *80* (6), 2765–2784.