How Costly Are Markups?

Supplementary Online Appendix

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

August 2022

This supplementary appendix is organized as follows. Appendix C provides additional empirical results used to explore the sensitivity of our results to key parameter values. Appendix D explains how we compute the steady state of our model and the transitional dynamics. Appendix E provides further details and quantitative results for two variations on our benchmark model with monopolistic competition: (i) where firm heterogeneity arises from persistence differences in quality across firms, and (ii) where we replace Kimball demand with symmetric translog demand. Appendix F analytically characterizes the aggregate markup with both Kimball demand and symmetric translog demand. These results also give us the mappings $\mathcal{M}(N)$ and $Z(N)$ that are crucial ingredients of our computational strategy. Appendix G derives value-added productivity in our model. Appendix H reports a simple formula for the welfare costs of markups in a static version of our model. Finally Appendix I analyzes the 'love for variety' effect in our model with Kimball demand.

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

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

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

C Additional empirical results

In this appendix we present additional empirical results that explore the sensitivity of our results to key parameter values.

Returns to scale. As discussed in Appendix B above, our estimates of firm-level markups $\mu_{it}(s)$ require an estimate of the elasticity of output with respect to labor $\alpha_t^l(s)$. In turn, to estimate this elasticity we need an estimate of the overall returns to scale (RTS) in production, i.e., $\text{RTS} := \alpha_t^k(s) + \alpha_t^l(s) + \alpha_t^x(s)$. Given the returns to scale, we can calculate firm-level markups and then estimate the key slope coefficient b from the within-sector relationship between markups and market shares.

To assess the sensitivity of our results to this assumption, Table C1 reports the estimated slope coefficient $b = \varepsilon/\bar{\sigma}$ for alternative values of the returns to scale. In particular, if we assume decreasing returns to scale a given input expenditure share implies proportionately lower output elasticities and hence lower levels of the implied markups. If we assume mildly decreasing returns to scale, $\text{RTS} = 0.95$ we find that the slope coefficient b barely changes. It gets slightly larger, rising from our benchmark 0.162 to 0.174, if we assume more strongly decreasing returns to scale, $\text{RTS} = 0.90$.

Sector heterogeneity. Our benchmark calibration takes the model at face-value and imposes a common slope coefficient $b(s) = b$. To assess the sensitivity of our results to this assumption, we provide alternative estimates of sector-specific $b(s)$ in two ways.

First, in Table C2 we report estimates of $b(s)$ with sectors selected by concentration ratios. In particular, we report $b(s)$ separately for sectors s with 4-firm concentration ratio (CR4) below and above 40%, a common threshold in the literature. The slope coefficients are very similar across sectors with different concentration levels. Second, in Table C3 we report a full set of $b(s)$ estimated separately for each 3-digit NAICS sector. For the main specification of interest, with sector × year and firm × sector fixed effects, we find the slope coefficient $b(s)$ range from a low of $b(s) = 0.081$ in Wood Product Manufacturing to a high of $b(s) = 0.242$ in Leather and Allied Product Manufacturing. Our benchmark estimate of $b = 0.162$ is almost exactly the midpoint of this range.

Table C1: Sensitivity to Returns to Scale

Dependent Variable	$\frac{1}{\mu_{it}(s)} + \log\left(1 - \frac{1}{\mu_{it}(s)}\right)$		
	RTS = 1.00	RTS = 0.95	RTS = 0.90
$\log \omega_{it}(s)$	0.162 (0.002)	0.162 (0.002)	0.174 (0.003)
Sector × Year FE	Y	Y	Y
Firm FE	Y	Y	Y
R^2	0.531	0.536	0.540
Observations	369,000	328,000	315,000

Sensitivity of estimated slope coefficent $b = \varepsilon/\bar{\sigma}$ to assumed returns to scale (RTS). Firm-level markups $\mu_{it}(s)$ constructed using data from the US Census of Manufactures from 1972 to 2012 as discussed in Appendix B. Benchmark specification assumes RTS $:= \alpha_t(s)^l + \alpha_t^k(s) + \alpha_t^x(s) = 1$. Estimated slope coefficient robust to RTS = 0.95 and RTS = 0.90. Standard errors clustered at the firm level. Number of observations drops with lower RTS because we exclude observations with $\mu_{it}(s) < 1$ so that the LHS of (59) is well-defined.

Table C2: Sensitivity to Sectoral Concentration

Dependent Variable	$\frac{1}{\mu_{it}(s)} + \log\left(1 - \frac{1}{\mu_{it}(s)}\right)$		
	Benchmark	CR4 $>$ 40%	CR4 $<$ 40%
$\log \omega_{it}(s)$	0.162 (0.002)	0.162 (0.009)	0.163 (0.002)
Sector × Year FE	Y	Y	Y
Firm FE	Y	Y	Y
R^2	0.531	0.536	0.530
Observations	369,000	21,000	348,000

Sensitivity of estimated slope coefficent $b = \varepsilon/\bar{\sigma}$ to sectoral concentration. Firm-level markups $\mu_{it}(s)$ constructed using data from the US Census of Manufactures from 1972 to 2012 as discussed in Appendix B. Sectors with four-firm concentration ration (CR4) $>$ 40% have almost identical b to sectors with less concentration. Standard errors clustered at the firm level.

Table C3: Sector-Specific Relationships Between Markups and Market Shares

	Dependent Variable	$\frac{1}{\mu_{it}(s)} + \log\left(1 - \frac{1}{\mu_{it}(s)}\right)$					
	$\log \omega_{it}(s)$	$b(s)$	s.e.	Obs.	$b(s)$	s.e.	Obs.
311	Food Manufacturing	0.058	0.002	36,500	0.179	0.009	22,000
312	Beverage and Tobacco Product Manufacturing	0.065	0.007	4,900	0.148	0.027	3,000
313	Textile Mills	0.067	0.006	6,500	0.202	0.022	3,900
314	Textile Product Mills	0.063	0.005	11,000	0.208	0.018	6,000
315	Apparel Manufacturing	0.097	0.003	31,500	0.112	0.010	12,500
316	Leather and Allied Product Manufacturing	0.038	0.008	4,100	0.242	0.030	2,400
321	Wood Product Manufacturing	0.108	0.003	31,000	0.081	0.011	19,000
322	Paper Manufacturing	0.037	0.005	9,800	0.153	0.019	6,300
323	Printing and Related Support Activities	0.035	0.002	71,000	0.098	0.008	4,3000
324	Petroleum and Coal Products Manufacturing	0.034	0.007	3,300	0.095	0.021	2,200
325	Chemical Manufacturing	0.058	0.003	19,000	0.135	0.009	12,000
326	Plastics and Rubber Products Manufacturing	0.070	0.003	27,000	0.135	0.010	17,000
327	Nonmetallic Mineral Product Manufacturing	0.046	0.003	32,500	0.119	0.009	22,000
331	Primary Metal Manufacturing	0.070	0.005	12,500	0.212	0.016	8,300
332	Fabricated Metal Product Manufacturing	0.086	0.002	115,000	0.157	0.006	74,500
333	Machinery Manufacturing	0.054	0.002	58,500	0.116	0.008	37,500
334	Computer and Electronic Product Manufacturing	0.043	0.002	27,500	0.123	0.009	14,000
335	Electrical Equipment, Appliance, and Component Manufacturing	0.055	0.004	13,000	0.154	0.014	7,900
336	Transportation Equipment Manufacturing	0.050	0.003	17,500	0.195	0.013	9,900
337	Furniture and Related Product Manufacturing	0.075	0.003	34,000	0.166	0.012	20,000
339	Miscellaneous Manufacturing	0.054	0.002	43,500	0.166	0.009	26,500
	Sector × Year FE		Y			Y	
	Firm × Sector FE					Y	

Relationship between firm-level markups $\mu_{it}(s)$ and 6-digit market shares $\omega_{it}(s)$ of firm i estimated separately for each 3-digit NAICS sector as shown. For each 3-digit sector we report the slope coefficient $b(s)$ from equation (59), standard error on the slope coefficient and number of observations. Two specifications are reported, one with sector × year FE only, the other with both sector × year and firm × sector FE. Standard errors clustered at the firm level.

Table C4: Log-Linear Markups and Market Shares

Dependent Variable	$\log \mu_{it}(s)$	
$\log \omega_{it}(s)$	0.031 (0.000)	0.072 (0.001)
Sector × Year FE	Y	Y
Firm FE		Y
R^2	0.098	0.647
Observations	609,000	369,000

Firm-level markups $\mu_{it}(s)$ constructed using data from the US Census of Manufactures from 1972 to 2012 as discussed in Appendix B. Market shares $\omega_{it}(s)$ of firm i within each 6-digit NAICS sector s. We include sector × year fixed effects to control for sector-specific shifts in the Kimball demand index $d_t(s)$. Our benchmark specification also includes firm fixed effects to contol for any time-invariant firm-specific component of demand. Standard errors clustered at the firm level.

Other distortions and a log-linear specification. Our model implies a non-linear relationship between markups and market size

$$\frac{1}{\mu_{it}(s)} + \log\left(1 - \frac{1}{\mu_{it}(s)}\right) = a(s) + a_i(s) + a_t(s) + b(s)\,\log \omega_{it}(s)$$

This non-linear relationship makes it difficult to use fixed effects to absorb persistent firm- or sector-level distortions that confound the measurement of markups in (60). To assess the impact of such distortions, we take a log-linear approximation to the LHS to write

$$f(\mu) := \frac{1}{\mu} + \log\left(1 - \frac{1}{\mu}\right) \approx f(\bar{\mu}) + f'(\bar{\mu})\bar{\mu}(\log \mu - \log \bar{\mu}) \tag{C1}$$

where $\bar{\mu} \geq 1$ is the point of approximation, a nuisance parameter. Up to this approximation, our model then implies that the true relationship between markups and market size is

$$f(\bar{\mu}) + f'(\bar{\mu})\bar{\mu}(\log \mu_{it}(s) - \log \bar{\mu}) = a(s) + a_i(s) + a_t(s) + b(s)\,\log \omega_{it}(s) \tag{C2}$$

or

$$\log \mu_{it}(s) = \tilde{a} + \tilde{a}(s) + \tilde{a}_i(s) + \tilde{a}_t(s) + \tilde{b}(s)\,\log \omega_{it}(s) \tag{C3}$$

where $\tilde{b}(s) = b(s)/(f'(\bar{\mu})\bar{\mu})$ etc where $f'(\bar{\mu})\bar{\mu} = 1/(\bar{\mu}(\bar{\mu}-1))$.

To see the advantage of this log-linear specification, suppose we have measured markups

$$\hat{\mu}_{it}(s) = \frac{p_{it}(s)y_{it}(s)}{W_t l_{it}(s)}\,\hat{\alpha}_t^l(s) \tag{C4}$$

But suppose the measured markups $\hat{\mu}_{it}(s)$ confound the true markup $\mu_{it}(s)$ and a multiplicative wedge

$$\hat{\mu}_{it}(s) = \mu_{it}(s) \exp(\tau_i(s) + \tau_t(s)) \tag{C5}$$

Then a regression of log measured markups on log market share is equivalent to

$$\log \hat{\mu}_{it}(s) = \tilde{a} + \tilde{a}(s) + (\tilde{a}_i(s) + \tau_i(s)) + (\tilde{a}_t(s) + \tau_t(s)) + \tilde{b}(s) \log \omega_{it}(s) \tag{C6}$$

So the persistent firm-level distortion $\tau_i(s)$ is absorbed by the firm fixed effects and the persistent sector-time distortion $\tau_t(s)$ is absorbed by the sector-time fixed effects. In this log-linear specification the slope coefficient $\tilde{b}(s)$ no longer has a structural interpretation, i.e., is not the super-elasticity, but is related to the super-elasticity via $\tilde{b}(s) = \bar{\mu}(\bar{\mu} - 1)b(s)$ where $\bar{\mu} \geq 1$ is the approximation point.

We report the results from estimating this log-linear specification in Table C4. When we impose a common slope coefficient $\tilde{b}(s) = \tilde{b}$ as in our benchmark model we find a tightly estimated $\tilde{b} = 0.072$ with standard error 0.001 clustered at the firm level. To intepret this magnitude, if we set $\bar{\mu} = 1.2$ then the implied super-elasticity is $b = \tilde{b}/(\bar{\mu}(\bar{\mu} - 1)) = 0.072/((1.2)(0.2)) = 0.3$, somewhat higher than in our benchmark model. That said, this implied value for the super-elasticity is almost exactly the same as the super-elasticity we estimate by indirect inference in an extension of our model where a firm fixed effect is required to control for permanent differences in quality across firms, see Appendix E below.

To summarize, even without imposing the additional structure from the Kimball demand system, we see clearly that markups positively covary with market shares, both within firms over time and across firms at a point in time. Another advantage of this log-linear specification is that the elasticity $\alpha_t^l(s)$ is also absorbed by the sector-time effects. In this sense, this exercise also serves to demonstrate that our results are not driven by the estimates of $\alpha_t^l(s)$ used to construct $\mu_{it}(s)$.

Estimates based on Taiwanese product-level data. As a further 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 labor input expenditure shares as in (B10) and estimate the slope coefficient b in (59) 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 comovement 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 Computational details

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

D.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

$$\mathcal{M}_t = \mathcal{M}(N_t), \qquad \text{and} \qquad Z_t = Z(N_t) \tag{D1}$$

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

$$q(z;A) = \operatorname*{argmax}_{q\geq 0} \left[\Upsilon'(q)q - \frac{A}{z}q\right] \tag{D2}$$

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, 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

$$\left[\Upsilon'(q(z;A)) - \mu(q(z;A))\frac{A}{z}\right]q(z;A) = 0 \tag{D3}$$

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

$$N\int \Upsilon(q(z;A))\,dG(z) = 1 \tag{D4}$$

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

$$\mathcal{M}(N) = \left(\frac{\displaystyle\int \frac{1}{\mu(q(z;A(N)))}\Upsilon'(q(z;A(N)))q(z;A(N))\,dG(z)}{\displaystyle\int \Upsilon'(q(z;A(N)))q(z;A(N))\,dG(z)}\right)^{-1} \tag{D5}$$

and

$$Z(N) = \left(N \int \frac{q(z; A(N))}{z} \, dG(z)\right)^{-1} \tag{D6}$$

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

$$\kappa W_t = \beta \sum_{j=1}^{\infty} (\beta(1-\varphi))^{j-1} \frac{C_t}{C_{t+j}} \left(1 - \frac{1}{\mathcal{M}_{t+j}}\right) \frac{Y_{t+j}}{N_{t+j}} \tag{D7}$$

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.

D.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

$$\mu(z_i, s) = \frac{1}{1 - \left(\frac{1}{\eta}\omega(z_i, z) + \frac{1}{\gamma}(1 - \omega(z_i, s))\right)} \tag{D8}$$

$$\omega(z_i, s) = \frac{\mu(z_i, s)^{1-\gamma} \, z_i^{\gamma-1}}{\sum_{i=1}^{n(s)} \mu(z_i, s)^{1-\gamma} \, z_i^{\gamma-1}} \tag{D9}$$

for each firm $i = 1, 2, \ldots, n(s)$ 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 $\boldsymbol{z}(s) = (z_1(s), z_2(s), \ldots, z_{n(s)}(s))$ the entrant understands that the configuration of productivity will become

$$\boldsymbol{z}'(s) = (\boldsymbol{z}(s), z) \tag{D10}$$

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

$$\hat{\Pi}(N) = \int \left(\int_0^1 \left(1 - \frac{1}{\mu(z, (\boldsymbol{z}(s), z))} \right) \omega(z, (\boldsymbol{z}(s), z))\, \bar{\omega}(\boldsymbol{z}(s), z)\, ds \right) dG(z) \tag{D11}$$

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

$$\kappa W_t \geq \beta \frac{C_t}{C_{t+1}} Q_{t+1} \tag{D12}$$

where

$$Q_t = \hat{\Pi}(N_t) Y_t + \beta(1 - \varphi) \frac{C_t}{C_{t+1}} Q_{t+1} \tag{D13}$$

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 (D11) 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).

E Monopolistic competition extensions

In this appendix 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.

E.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

$$N_t \int z\,\Upsilon\Big(\frac{y_t(z)}{Y_t}\Big)\,dG(z) = 1 \tag{E1}$$

This implies the inverse demand curve

$$p_t(z) = z\,\Upsilon'(q_t(z))\,D_t \tag{E2}$$

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

$$D_t = \left(N_t \int z\,\Upsilon'(q_t(z))q_t(z)\,dG(z)\right)^{-1} \tag{E3}$$

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

$$z\,\Upsilon'(q_t(z)) = \frac{\sigma(q_t(z))}{\sigma(q_t(z))-1}\,A_t, \qquad A_t := \frac{\Omega_t}{D_t} \tag{E4}$$

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

$$\frac{1}{\mu_t(z)} + \log\left(1 - \frac{1}{\mu_t(z)}\right) = a \ + \ b \, \log \omega_t(z) \ - \ b \, \log z, \qquad b = \frac{\varepsilon}{\bar{\sigma}} \tag{E5}$$

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 E1 reports the parameters for the quality model when we target an aggregate markup of $\mathcal{M} = 1.15$. The quality model fits the data 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. This value of the super-elasticity is almost exactly the same as we infer in a log-linear approximation to (59) where we can use firm fixed effects to control for persistent quality differences, see Appendix C above.

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 E2. For the quality model calibrated to an aggregate markup of $\mathcal{M} = 1.15$ the aggregate productivity losses due to misallocation are 1.75%, 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 E3, 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.

Discussion. In the quality specification used here a firm's product *does not directly* affect the firm's production function. By contrast, in the literature it is standard to assume that higher-quality firms need higher-quality inputs in production, e.g., as in Fieler, Eslava and Xu (2018), Jaimovich, Rebelo and Wong (2019), and Verhoogen (2008). To the extent that quality affects production in a Hicks-neutral way, this is without loss of generality. For

Table E1: Parameterization, Extensions

calibration targets		*data*	*quality*	*translog*	*benchmark*
$\mathcal{M}$	aggregate markup	1.1 $\sim$ 1.4	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
parameter values					
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

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 E2: Markup Dispersion and Productivity Losses, 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
aggregate productivity losses, %			
gross output	1.75	2.81	0.97
value-added	4.20	6.16	2.71
value-added, $\mathcal{M} = 1$	3.35	5.33	1.85

Cost-weighted steady state distribution of markups and aggregate productivity losses for various monopolistic competition models. For brevity we focus on the case $\mathcal{M} = 1.15$. Gross output aggregate productivity loss is $(Z - Z^*)/Z^* \times 100$, and similarly for the value-added aggregate productivity loss. To isolate the effect of misallocation on value-added aggregate productivity we also report the value-added aggregate productivity loss with the same amount of markup dispersion but holding $\mathcal{M} = 1$ to eliminate the distortion between value-added and materials, see text for details.

Table E3: 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	59.6	44.5	18.0	20.1	100.4	4.1	8.67
uniform subsidy	51.8	35.8	17.0	9.5	88.5	1.5	5.90
size-dependent subsidy	5.3	6.2	1.0	8.3	6.6	2.3	2.87
entry subsidy	6.3	7.4	2.4	20.0	8.1	3.0	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.*

example, if firms with quality z have production function $y = a(z)F(k, l, x)$ we can rescale quality $\tilde{z} := z/a(z)$ and use (E4) to solve for relative size $q_t(\tilde{z})$ and hence markups $\mu_t(\tilde{z})$ in terms of the rescaled quality $\tilde{z}$. But if quality affects production through the use of specialized capital, labor or materials in a factor-biased (non-Hicks-neutral) way, there would be genuine interactions between a firm's pricing decisions and input choice that make the model more complex. Our results focus on the simple Hicks-neutral setup which is suficient for our purposes, i.e., demonstrating the effects of breaking the one-to-one link between size and markups.

E.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 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

$$\begin{aligned}\log(P_tY_t) \;=\; \log Y_t \;+\; \frac{1}{2\bar{\sigma}N_t} \;+\; \int \log p_t(z)\, dG(z) \\ +\; \frac{\bar{\sigma}N_t}{2}\left(\left(\int \log p_t(z)\, dG(z)\right)^2 - \int \log p_t(z)^2\, dG(z)\right)\end{aligned} \tag{E6}$$

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

$$\omega_t(z) := \frac{p_t(z)y_t(z)}{P_tY_t} = \frac{d\log(P_tY_t)}{d\log p_t(z)} = \bar{\sigma}\log\Big(\frac{p_t^*}{p_t(z)}\Big), \qquad p_t(z) < p_t^* \tag{E7}$$

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

$$\log p_t^* := \frac{1}{2\bar{\sigma}N_t} + \int \log p_t(z)\, dG(z) \tag{E8}$$

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

$$y_t(z) = \bar{\sigma}\log\Big(\frac{p_t^*}{p_t(z)}\Big)\frac{P_tY_t}{p_t(z)}, \qquad p_t(z) < p_t^* \tag{E9}$$

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

$$\sigma(\rho) = \frac{1+\log\left(\frac{1}{\rho}\right)}{\log\left(\frac{1}{\rho}\right)}, \qquad \mu(\rho) = 1 + \log\Big(\frac{1}{\rho}\Big) \tag{E10}$$

We can then write the static markup-pricing condition

$$\rho_t(z) = \frac{\sigma(\rho_t(z))}{\sigma(\rho_t(z))-1}\,\frac{z_t^*}{z}, \qquad z_t^* := \frac{\Omega_t}{p_t^*} \tag{E11}$$

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 (E10) we can write

$$\mu_t(z) = 1 + \frac{1}{\bar{\sigma}}\,\omega_t(z) \tag{E12}$$

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

$$\mu + \log \mu = 1 + \log\left(\frac{z}{z_t^*}\right), \qquad z > z_t^* \tag{E13}$$

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 $\rho_t(z)$.

Calibration. As is clear from our analytic expressions for z_t^* and $\mathcal{M}_t$ in the main text, the translog model is less flexible than our Kimball benchmark. In particular, 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.[32] 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 reproducing 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 E1, the translog model reproduces an aggregate markup of $\mathcal{M} = 1.15$ but implies too little sales concentration (a top 5% sales share of 0.21, as opposed to 0.57 in the data) and too strong a relationship between markups and market shares (regression coefficient $b = 0.43$ as opposed to $\hat{b} = 0.16$ 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 E2. For our translog model calibrated to an aggregate markup of $\mathcal{M} = 1.15$ the aggregate productivity losses due to misallocation are 2.81%, 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 E3, 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.

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

F 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.

F.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

$$\mathcal{M} = \frac{\displaystyle\int_1^\infty \mu(q(z))\frac{q(z)}{z}\,dG(z)}{\displaystyle\int_1^\infty \frac{q(z)}{z}\,dG(z)} \tag{F1}$$

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

$$\Upsilon'(q) = \mu(q)\frac{A}{z} \tag{F2}$$

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

$$N\int_1^\infty \Upsilon(q(z/A))\,dG(z) = 1 \tag{F3}$$

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

$$\mathcal{M}(N) = \frac{\displaystyle\int_{1/A(N)}^\infty \mu(q(\hat{z}))\frac{q(\hat{z})}{\hat{z}}\,dG(\hat{z})}{\displaystyle\int_{1/A(N)}^\infty \frac{q(\hat{z})}{\hat{z}}\,dG(\hat{z})} \tag{F4}$$

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

$$\mathcal{M}'(N) = (\mu_{min} - \mathcal{M}) \times \frac{q_{min}\, g_{min}}{\displaystyle\int_{1/A(N)}^{\infty} \frac{q(\hat{z})}{\hat{z}}\, dG(\hat{z})} \times \frac{A'(N)}{A(N)} \leq 0 \tag{F5}$$

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

$$z^*(N) = \max\left[\, 1\ ,\ \frac{\bar{\sigma}}{\bar{\sigma} - 1}\, e^{-\frac{1}{\varepsilon}}\, A(N)\, \right] \tag{F6}$$

where $A(N)$ solves the Kimball aggregator (F3) 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 (F5), 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

$$Z(N) = \Big(NA(N)^{-\xi-1}\int_{1/A(N)}^{\infty}\frac{q(\hat{z})}{\hat{z}}\,dG(\hat{z})\Big)^{-1} \tag{F7}$$

which likewise determines $Z(N)$ given the $A(N)$ which solves the Kimball aggregator (F3).

F.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

$$\mu + \log\mu = 1 + \log(z/z^*), \qquad z > z^* \tag{F8}$$

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

$$\log p^* = \frac{1}{\bar{\sigma}N} + \int \log p(z)\,dG(z) \tag{F9}$$

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

$$(1-G(z^*))\log z^* = -\Big(\frac{1}{\bar{\sigma}N} + \int_{z^*}^{\infty}\log\big(\frac{\mu(z)}{z}\big)\,dG(z)\Big) \tag{F10}$$

To simplify this we need to calculate the integral on the RHS. Using (F8) to rewrite the integrand, we get

$$\begin{aligned}\int_{z^*}^{\infty}\log\big(\frac{\mu(z)}{z}\big)\,dG(z) &= \int_{z^*}^{\infty}\big(1-\mu(z)-\log z^*\big)\,dG(z)\\ &= \big(1-\ln z^*\big)(1-G(z^*)) - \int_{z^*}^{\infty}\mu(z)\,g(z)\,dz\\ &= \big(1-\ln z^*\big)(1-G(z^*)) - \int_{1}^{\infty}\mu\,g(z(\mu))z'(\mu)\,d\mu\\ &= \big(1-\ln z^*\big)(1-G(z^*)) - \xi\int_{1}^{\infty}(1+\mu)\,\big\{\,z^*\mu e^{\mu-1}\,\big\}^{-\xi}\,d\mu\\ &= \big(1-\ln z^*\big)(1-G(z^*)) - (1-G(z^*))\xi\int_{1}^{\infty}(1+\mu)\mu^{-\xi}\,e^{-\xi(\mu-1)}\,d\mu\end{aligned} \tag{F11}$$

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 (F8) 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 (F10), cancelling common terms and simplifying then gives

$$\frac{1}{1 - G(z^*)} = z^{*\,\xi} = \bar{\sigma} N (I(\xi) - 1) \tag{F12}$$

where $I(\xi)$ is the constant

$$I(\xi) := \xi \int_1^\infty (1 + \mu)\mu^{-\xi}\, e^{-\xi(\mu-1)}\, d\mu \tag{F13}$$

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

$$I(\xi) = 1 + e^{\xi} E_{\xi}(\xi), \qquad E_n(x) := \int_1^\infty \frac{e^{-xt}}{t^n}\, dt \tag{F14}$$

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

$$\boxed{z^*(N) = \max\Big[\, 1 \,,\; \bar{\sigma} N\, e^{\xi} E_{\xi}(\xi) \,\Big]^{1/\xi}} \tag{F15}$$

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

$$\mathcal{M}^{-1} = N \int_1^\infty \frac{\omega(z)}{\mu(z)}\, dG(z) = \bar{\sigma} N \int_1^\infty \frac{\mu(z) - 1}{\mu(z)}\, dG(z) = \bar{\sigma} N \int_{z^*}^\infty \frac{\mu(z) - 1}{\mu(z)}\, dG(z) \tag{F16}$$

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

$$\mathcal{M}^{-1} = \bar{\sigma} N (1 - G(z^*)) \left\{ 1 - \xi \int_1^\infty (1 + \mu)\, \mu^{-2-\xi}\, e^{-\xi(\mu-1)}\, d\mu \right\} \tag{F17}$$

which we can again write in terms of generalized exponential integrals

$$\mathcal{M}^{-1} = \bar{\sigma} N (1 - G(z^*)) \Big(1 - \xi e^{\xi} [E_{\xi+1}(\xi) + E_{\xi+2}(\xi)] \Big) \tag{F18}$$

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

(F12) and (F14) 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

$$\mathcal{M} = \frac{e^{\xi} E_{\xi}(\xi)}{1 - \xi e^{\xi}[E_{\xi+1}(\xi) + E_{\xi+2}(\xi)]} \tag{F19}$$

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

$$E_n(x) = x^{n-1}\Gamma(1 - n, x), \qquad \Gamma(s, x) := \int_x^{\infty} t^{s-1} e^{-t}\, dt \tag{F20}$$

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

$$\Gamma(s + 1, x) = s\Gamma(s, x) + x^s e^{-x} \tag{F21}$$

Using these properties to collect terms and simplify

$$\mathcal{M} = \frac{e^{\xi} E_{\xi}(\xi)}{1 - \xi e^{\xi}[E_{\xi+1}(\xi) + E_{\xi+2}(\xi)]} = \frac{e^{\xi}\xi^{\xi-1}\Gamma(1 - \xi, \xi)}{e^{\xi}\xi^{\xi+1}\Gamma\big(-(1 + \xi), \xi\big)} = \xi^{-2}\,\frac{\Gamma\big(+1 - \xi, \xi\big)}{\Gamma\big(-1 - \xi, \xi\big)} \tag{F22}$$

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

$$\frac{\Gamma(s + 2, x)}{\Gamma(s, x)} = s(s + 1) + (s + 1 + x)\frac{x^s e^{-x}}{\Gamma(s, x)} \tag{F23}$$

which follows from iterating forward twice using our recursion (F21). Evaluating this at $s = -(1 + \xi)$ and $x = \xi$ and simplifying

$$\frac{\Gamma(+1 - \xi, \xi)}{\Gamma(-1 - \xi, \xi)} = -(1+\xi)[-(1+\xi)+1] + [-(1+\xi)+(1+\xi)]\frac{\xi^{-(1+\xi)}e^{-\xi}}{\Gamma(-(1+\xi), \xi)} = \xi(1+\xi) \tag{F24}$$

Hence we get the very simple expression for the aggregate markup

$$\boxed{\mathcal{M} = 1 + \frac{1}{\xi}, \qquad \text{if } z^* > 1} \tag{F25}$$

To take the second case, if instead $z^* = 1$, so that $G(z^*) = 0$, then from (F18), (F22) and (F24) we have

$$\mathcal{M} = \left(1 + \frac{1}{\xi}\right) \times \left(\bar{\sigma}N\, e^{\xi} E_{\xi}(\xi)\right)^{-1}, \qquad \text{if } z^* = 1 \tag{F26}$$

Collecting these cases together, we conclude that

$$\boxed{\mathcal{M}(N) = \left(1 + \frac{1}{\xi}\right) \times \left(\max\Big[\, 1\ ,\ \bar{\sigma}N\, e^{\xi} E_{\xi}(\xi)\ \Big]\right)^{-1}} \tag{F27}$$

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.

G Value-added productivity

In this appendix we derive value-added aggregate productivity in our model. To begin with, recall that aggregate value added is given by

$$\text{GDP} = Y - X \tag{G1}$$

And recall that we can write the aggregate production function for gross output

$$Y = Z\left[\phi^{\frac{1}{\theta}}\left(K^{\alpha}\tilde{L}^{1-\alpha}\right)^{\frac{\theta-1}{\theta}} + (1-\phi)^{\frac{1}{\theta}}X^{\frac{\theta-1}{\theta}}\right]^{\frac{\theta}{\theta-1}} \tag{G2}$$

G.1 Planner's value-added aggregate productivity

To calculate the amount the planner can produce with given K and $\tilde{L}$, we choose materials X^* to maximize

$$\text{GDP}^* = Z^*\left[\phi^{\frac{1}{\theta}}\left(K^{\alpha}\tilde{L}^{1-\alpha}\right)^{\frac{\theta-1}{\theta}} + (1-\phi)^{\frac{1}{\theta}}X^{*\frac{\theta-1}{\theta}}\right]^{\frac{\theta}{\theta-1}} - X^* \tag{G3}$$

The first order condition for this problem is

$$(1-\phi)^{\frac{1}{\theta}}\,Z^*\left[\phi^{\frac{1}{\theta}}\left(K^{\alpha}\tilde{L}^{1-\alpha}\right)^{\frac{\theta-1}{\theta}} + (1-\phi)^{\frac{1}{\theta}}X^{*\frac{\theta-1}{\theta}}\right]^{\frac{1}{\theta-1}}\left(X^*\right)^{-\frac{1}{\theta}} = 1 \tag{G4}$$

or equivalently

$$X^* = (1-\phi)\left(Z^*\right)^{\theta-1}Y^* \tag{G5}$$

We can then eliminate materials X^* from the objective to get

$$Y^* = Z^*\left[\phi^{\frac{1}{\theta}}\left(K^{\alpha}\tilde{L}^{1-\alpha}\right)^{\frac{\theta-1}{\theta}} + (1-\phi)\left(\left(Z^*\right)^{\theta-1}Y^*\right)^{\frac{\theta-1}{\theta}}\right]^{\frac{\theta}{\theta-1}} \tag{G6}$$

which implicitly determines the planner's gross output Y^* in terms of the given K and $\tilde{L}$ and their gross output productivity Z^*. Solving for the planner's gross output Y^* we get

$$Y^* = \phi^{\frac{1}{\theta-1}}\frac{Z^*}{\left(1-(1-\phi)\left(Z^*\right)^{(\theta-1)}\right)^{\frac{\theta}{\theta-1}}}\left(K^{\alpha}\tilde{L}^{1-\alpha}\right) \tag{G7}$$

which implies that the planner's aggregate value-added is

$$\begin{aligned}\text{GDP}^* = Y^* - X^* &= \left(1-(1-\phi)\left(Z^*\right)^{\theta-1}\right)Y^* \\ &= \phi^{\frac{1}{\theta-1}}\frac{\left(1-(1-\phi)\left(Z^*\right)^{\theta-1}\right)}{\left(1-(1-\phi)\left(Z^*\right)^{\theta-1}\right)^{\frac{\theta}{\theta-1}}} \times Z^*\left(K^{\alpha}\tilde{L}^{1-\alpha}\right)\end{aligned} \tag{G8}$$

So the planner's value-added aggregate productivity is

$$Z^*_{\text{value-added}} = \phi^{\frac{1}{\theta-1}}\,\frac{\left(1-(1-\phi)\,Z^{*\theta-1}\right)}{\left(1-(1-\phi)Z^{*\theta-1}\right)^{\frac{\theta}{\theta-1}}}\,Z^* \tag{G9}$$

G.2 Decentralized value-added aggregate productivity

For the decentralized economy, aggregate value-added is given by

$$\text{GDP} = Y - X = \left(1 - (1-\phi)\left(\frac{1}{\Omega}\right)^{1-\theta}\frac{1}{\mathcal{M}}\right)Y \tag{G10}$$

Moreover we know that $\Omega = Z/\mathcal{M}$ so we can write materials

$$X = (1-\phi)\left(\frac{Z}{\mathcal{M}}\right)^{\theta-1}\frac{Y}{\mathcal{M}} = (1-\phi)\,Z^{\theta-1}\mathcal{M}^{-\theta}Y \tag{G11}$$

Using this to eliminate materials X from the aggregate production function for gross output (G2) we have

$$Y^{\frac{\theta-1}{\theta}} = \left[\phi^{\frac{1}{\theta}}\left(ZK^{\alpha}\tilde{L}^{1-\alpha}\right)^{\frac{\theta-1}{\theta}} + (1-\phi)Z^{\theta-1}\mathcal{M}^{1-\theta}Y^{\frac{\theta-1}{\theta}}\right] \tag{G12}$$

Solving for gross output Y we get

$$Y = \phi^{\frac{1}{\theta-1}}\frac{Z}{(1-(1-\phi)Z^{\theta-1}\mathcal{M}^{1-\theta})^{\frac{\theta}{\theta-1}}}K^{\alpha}\tilde{L}^{1-\alpha} \tag{G13}$$

which implies that aggregate value-added in the decentralized economy is

$$\text{GDP} = \left(1-(1-\phi)\,Z^{\theta-1}\mathcal{M}^{-\theta}\right)Y = \phi^{\frac{1}{\theta-1}}\frac{\left(1-(1-\phi)\,Z^{\theta-1}\mathcal{M}^{-\theta}\right)}{(1-(1-\phi)^{\theta-1}\mathcal{M}^{1-\theta})^{\frac{\theta}{\theta-1}}} \times ZK^{\alpha}\tilde{L}^{1-\alpha} \tag{G14}$$

So value-added aggregate productivity in the decentralized economy is

$$Z_{\text{value-added}} = \phi^{\frac{1}{\theta-1}}\,\frac{\left(1-(1-\phi)\,Z^{\theta-1}\mathcal{M}^{-\theta}\right)}{(1-(1-\phi)Z^{\theta-1}\mathcal{M}^{1-\theta})^{\frac{\theta}{\theta-1}}}\,Z \tag{G15}$$

Comparing the levels of value-added aggregate productivity in the decentralized economy to its counterpart from the planner's problem, we see that value-added aggregate productivity is distorted both because markup dispersion makes Z too low relative to the planner's Z^* and because the aggregate markup $\mathcal{M}$ leads to an inefficient use of materials.

H 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

$$U(C,L)=\frac{C^{1-\sigma}}{1-\sigma}-\frac{L^{1+\nu}}{1+\nu} \tag{H1}$$

Suppose also that labor is the only factor of production[33] 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

$$L=\mathcal{M}^{-\frac{1}{\sigma+\nu}}\,Z^{\frac{1-\sigma}{\sigma+\nu}},\qquad \text{and}\qquad C=\mathcal{M}^{-\frac{1}{\sigma+\nu}}\,Z^{\frac{1+\nu}{\sigma+\nu}} \tag{H2}$$

The associated level of utility is

$$U(C,L)=\left(\frac{1}{1-\sigma}-\frac{1}{1+\nu}\frac{1}{\mathcal{M}}\right)\mathcal{M}^{-\frac{1-\sigma}{\sigma+\nu}}\,Z^{\frac{(1+\nu)(1-\sigma)}{\sigma+\nu}} \tag{H3}$$

Similarly, the level of utility in the efficient allocation is

$$U(C^*,L^*)=\left(\frac{1}{1-\sigma}-\frac{1}{1+\nu}\right)Z^{*\,\frac{(1+\nu)(1-\sigma)}{\sigma+\nu}} \tag{H4}$$

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

$$\mathcal{W}=\left(1-\frac{1-\sigma}{1+\nu}\frac{1}{\mathcal{M}}\right)^{\frac{1}{1-\sigma}}\mathcal{M}^{-\frac{1}{\sigma+\nu}}\,Z^{\frac{1+\nu}{\sigma+\nu}} \tag{H5}$$

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

$$\mathcal{W}^*=\left(1-\frac{1-\sigma}{1+\nu}\right)^{\frac{1}{1-\sigma}}Z^{*\,\frac{1+\nu}{\sigma+\nu}} \tag{H6}$$

Hence the consumption-equivalent losses from markups can be written

$$\frac{\mathcal{W}}{\mathcal{W}^*}=\left(\frac{\left(1-\frac{1-\sigma}{1+\nu}\frac{1}{\mathcal{M}}\right)}{\left(1-\frac{1-\sigma}{1+\nu}\right)}\right)^{\frac{1}{1-\sigma}}\left(\frac{Z}{Z^*}\right)^{\frac{1+\nu}{\sigma+\nu}}\mathcal{M}^{-\frac{1}{\sigma+\nu}} \tag{H7}$$

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

$$\frac{\mathcal{W}}{\mathcal{W}^*}=\left(\frac{Z}{Z^*}\right)\mathcal{M}^{-\frac{1}{1+\nu}} \tag{H8}$$

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.

[33] 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.

I 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

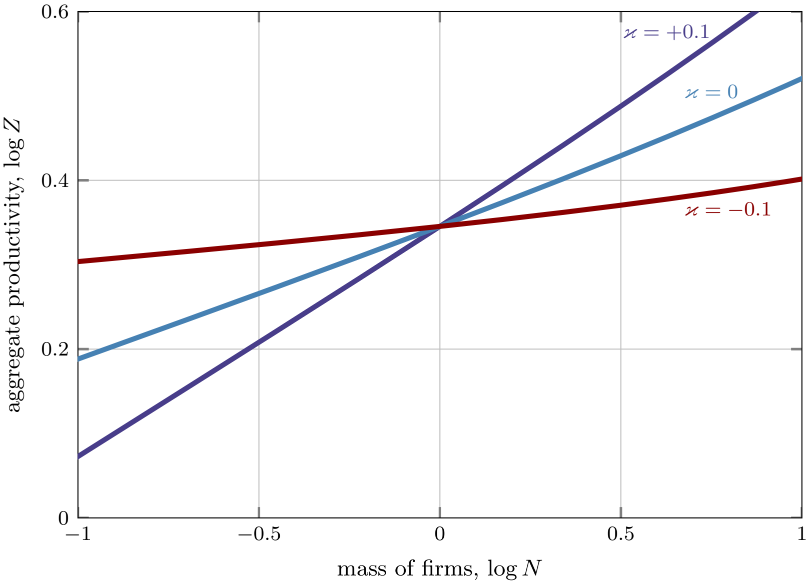
$$N^{1+\varkappa} \int_1^{\infty} \Upsilon(q(z))\, dG(z) = 1 \tag{I1}$$

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 I1 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 I1 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 I1 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 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 between 2 and 3% in all cases — larger than than the gross output TFP loss because of the standard multiplier effect from intermediates.

Figure I1: 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.

Table I1: 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	59.6	44.5	18.0	20.1	100.4	4.1	8.67
	uniform subsidy	51.8	35.8	17.0	9.5	88.5	1.5	5.90
	size-dependent subsidy	5.3	6.2	1.0	8.3	6.6	2.3	2.87
	entry subsidy	6.3	7.4	2.4	20.0	8.1	3.0	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.