NBER WORKING PAPER SERIES

THE DARWINIAN RETURNS TO SCALE

David Baqaee
Emmanuel Farhi
Kunal Sangani

Working Paper 27139
http://www.nber.org/papers/w27139

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
May 2020, Revised August 2021

Emmanuel Farhi tragically passed away in July, 2020. Emmanuel was a one-in-a-lifetime collaborator and friend. We thank Cédric Duprez and Oleg Itskhoki for sharing their data. We thank Maria Voronina and Sihwan Yang for outstanding research assistance. We thank Pol Antras, Andrew Atkeson, Ariel Burstein, Elhanan Helpman, Chad Jones, Marc Melitz, and Simon Mongey for helpful comments. We acknowledge research financial support from the Ferrante fund at Harvard University and NSF grant #1947611. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

The Darwinian Returns to Scale
David Baqaee, Emmanuel Farhi, and Kunal Sangani
NBER Working Paper No. 27139
May 2020, Revised August 2021
JEL No. E0,E1,E23,O24,O4,O41

ABSTRACT

How does an increase in market size, say due to globalization, affect welfare? We study this question using a model with monopolistic competition, heterogeneous markups, and fixed costs. We characterize the change in welfare in the decentralized equilibrium, and decompose it into changes in technical efficiency and allocative efficiency. Allocative efficiency changes due to three different types of reallocations: (1) reallocations across firms with heterogeneous price elasticities due to increased entry, (2) reallocations due to the exit of marginally profitable firms, and (3) reallocations due to changes in firms' markups. Whereas the second and third effects have ambiguous implications for welfare, the first effect, which we call the Darwinian effect, always increases welfare regardless of the shape of demand curves. We non-parametrically calibrate residual demand curves with firm-level data from Belgian manufacturing firms and quantify our theoretical results. We find that mild increasing returns at the micro level can catalyze large increasing returns at the macro level. These aggregate gains are due to the Darwinian effect, which reallocates resources from low- to high-markup firms, and not the death of unproductive firms (2) or changes in markups (3). Our results suggest that a policy-maker can harness Darwinian reallocations in an economy with fixed resources by subsidizing firm entry.

David Baqaee
Department of Economics
University of California at Los Angeles
Bunche Hall
Los Angeles, CA 90095
and CEPR
and also NBER
baqaee@econ.ucla.edu

Emmanuel Farhi
Harvard University
*NA user is deceased

Kunal Sangani
Harvard University
Wyss Hall
20 N Harvard St
Boston, MA 02163
ksangani@g.harvard.edu

1 Introduction

Aggregate increasing returns to scale are at the core of some of the most fundamental issues in economics, ranging from the mechanics of growth, to the gains from trade, to the benefits from industrial and competition policy. Broadly speaking, there are two reasons why efficiency may increase as markets get larger. The first has to do with the technological features of production. If firms have increasing returns to scale, say due to fixed costs, then expanding the market will improve efficiency since fixed costs will be spread over a larger population. The second has to do with how resources are allocated in equilibrium. If competition intensifies in a bigger market, then perhaps this can reallocate resources in a way that improves aggregate efficiency. For example, Pavcnik (2002), Trefler (2004), and Mayer et al. (2014) document that as market size increases, resources are reallocated to high-performing firms and products.

In this paper, we propose a framework for decomposing these effects theoretically and quantitatively. We consider an economy with fixed entry and overhead costs, entry and exit, monopolistic competition, and heterogeneous markups. We argue that, to a large extent, increasing returns to scale at the aggregate level may reflect changes in allocative rather than technical efficiency. That is, a large share of the gains from an increase in market size—say due to immigration, fertility, or globalization (trade integration)—arise from how intensified competition reallocates resources across firms. Furthermore, we show that even mild increasing returns at the micro level (measured by the average ratio of marginal to average cost) can catalyze large increasing returns at the macro level. This reinforces the insight by Basu and Fernald (1997) that aggregation can exaggerate modest returns to scale at the micro level.

For tractability, models of monopolistic competition and entry often feature constant-elasticity-of-substitution (CES) demand. The classic reference is Melitz (2003), which is a workhorse model of reallocation. However, this model has an efficient equilibrium, so reallocations have no first-order effect on welfare (this is because the marginal social benefit of any input is equated across competing uses). Moreover, efficiency implies that micro- and macro-level returns to scale are the same, since, on the margin, allocating all incremental inputs to a single firm must yield the same aggregate return as the equilibrium allocation.

This simple elegance of CES demand comes at the expense of realism. CES demand imposes constant markups in both the cross-section and the time-series with complete pass-through of marginal costs into prices. In contrast, the data feature substantial heterogeneity in both markups and pass-throughs. Matching the empirical heterogeneity of markups and pass-throughs requires deviating from the CES benchmark. This, in turn, introduces distortions in the decentralized equilibrium and opens the door for endogenous reallocations triggered by a shock to affect welfare.[1]

[1]Of course, we are not the first to consider deviations from CES in models of free entry and monopolistic competition. Previous examples with inefficient equilibria include Krugman (1979), Mankiw and Whinston (1986), Venables (1985), Asplund and Nocke (2006), Melitz and Ottaviano (2008), Epifani and Gancia (2011), Zhelobodko et al. (2012), Edmond et al. (2018), Dhingra and Morrow (2019), Mrázová and Neary (2017), Mrázová and Neary (2019), Arkolakis et al. (2019),

We relax the restrictions of the CES demand system by using a generalized Kimball (1995) demand system, introduced by Matsuyama and Ushchev (2017). This demand system allows for the possibility that firms face different residual demand curves from one another, and allows each firm's desired markup and pass-through to vary flexibly as a function of its size.[2] Furthermore, this demand system is homothetic, which makes it relatively straightforward to embed our analysis of a single sector into a larger multi-sector general equilibrium model of the whole economy.

We characterize how welfare changes in response to an increase in market size. The response of welfare consists of a change in technical efficiency (i.e., an increase in welfare holding the allocation of resources across uses constant) and a change in allocative efficiency that arises due to endogenous reallocations. We show that changes in allocative efficiency can be further broken into three distinct channels, which correspond to firms' adjustments along three margins: the decision to enter, to exit the market, and to change markups. We call these three channels (1) the Darwinian effect, (2) the selection effect, and (3) the pro/anti-competitive effect.

The Darwinian effect (1) captures how firms with different price-elasticities are differentially affected by changes in the number of entrants. Each firm faces a demand curve, which pins down quantity as a function of the firm's price relative to an aggregate market-level price index. When the market expands and new firms enter, the aggregate price index falls, intensifying competition for all firms. Firms with more inelastic demand, however, are relatively insulated from changes in the aggregate price index, and hence expand relative to firms with elastic demand.

The markup of each firm is inversely related to its demand elasticity. Hence, the Darwinian effect reallocates resources from firms with elastic demand (and low markups) towards firms with inelastic demand (and high markups). From a social perspective, high-markup firms are too small relative to low-markup firms, and so this reallocation improves efficiency. We call this a *Darwinian* effect because a more competitive environment automatically selects and expands the "fittest" firms (those with the most inelastic demand).

Notably, this effect exists and is welfare-increasing regardless of the shape of demand curves, as long as there is non-trivial heterogeneity. In contrast, the selection and pro/anti-competitive effect, which have been studied in detail in previous work, have theoretically ambiguous effects on welfare.

The selection effect (2) results from the fact that, as the market expands, the minimum level of profitability a firm must have to survive can change. This mechanism only operates in models with overhead costs of production and is explored by Asplund and Nocke (2006), Melitz and Ottaviano (2008), Corcos et al. (2012), and Melitz and Redding (2015), among others. Unlike the Darwinian effect, whether or not the selection effect increases or reduces welfare is ambiguous. As pointed out by Dhingra and Morrow (2019), a toughening of the selection cut-off improves welfare only if the consumer surplus generated by the marginal firm relative to its sales is less than the average.

and Matsuyama and Ushchev (2020b). We discuss precisely how our approach and findings differ below.

[2]We also derive our results using other generalizations of CES preferences, which nest separable translog preferences and linear expenditure shares as special cases, in Appendix H. The results are similar both qualitatively and quantitatively.

Finally, the pro/anti-competitive effect (3) results from the fact that firms' desired markups may change as the market expands. Of the three channels, the pro/anti-competitive effect is the sole change in allocative efficiency arising in homogeneous firm models such as Krugman (1979). If firms have incomplete pass-through, as is the case considered by Krugman (1979), then as the price index falls due to an increase in market size, firms cut their desired markups (pro-competitive effect). Recent studies exploring the pro/anti-competitive effect include De Loecker et al. (2016), Feenstra and Weinstein (2017), Feenstra (2018), Arkolakis et al. (2019), and Matsuyama and Ushchev (2020b). We show that whether these changes in markups raise or lower welfare is also ambiguous.

These reallocative forces also have implications for policy. In particular, policy-makers can trigger these reallocations even in an economy with fixed resources by incentivizing entry. We show that a subsidy on firm entry costs can improve welfare even if, on the margin, entry is excessive. This is a consequence of the general theory of the second best (Lipsey and Lancaster, 1956)— since all optimality conditions cannot be satisfied, the second-best involves changing the amount of entry away from its first-best value. In our model, subsidizing entry above the first-best level can be desirable since entry triggers Darwinian reallocations that alleviate cross-sectional misallocation.

To quantify our theoretical results, we develop a strategy for taking the non-parametric model to data. Using cross-sectional firm-level information from Belgium on pass-throughs (from Amiti et al., 2019), we non-parametrically solve for the shape of the residual demand curve that can exactly rationalize the distributions of firm sales and pass-throughs. We then use our calibrated model to quantify the role reallocations play in aggregate returns to scale.

In our quantitative calibration, we find that changes in allocative efficiency are much more important than changes in technical efficiency in determining aggregate increasing returns to scale. They account for between 70% and 90% of the overall effect. As a result, mild increasing returns to scale at the microeconomic level can be associated with large increasing returns to scale at the aggregate level. Furthermore, we show that the selection and pro-competitive effects are either unimportant or harmful. Instead, the Darwinian mechanism that we isolate contributes the lion's share of the gains in allocative efficiency. In our quantitative calibration, we also find that these Darwinian reallocations concentrate a greater share of employment and sales in high-markup firms, tying the benefits of a market expansion to increases in concentration.[3]

We also relate our results for welfare to the behavior of real GDP. It is well-known that when the set of goods can change due to entry and exit, real GDP and welfare may not be the same (see e.g. Aghion et al., 2019a). In our model changes in real GDP are entirely driven by reductions in

[3]Baqaee and Farhi (2019) show that this type of reallocation—a reallocation from low-markup firms to high-markup firms—can explain a significant fraction of aggregate TFP growth in the US over the last two decades. De Loecker et al. (2020), Autor et al. (2020), and Aghion et al. (2019b) document a similar reallocation of market share to high-markup and high-productivity firms over time. This paper raises the possibility that increases in scale, perhaps driven by globalization, could be responsible for these reallocations.

markups, and do not depend on the reallocation effects that are so crucial for welfare. Quantitatively, we find that the elasticity of real GDP per capita to market size understates the elasticity of welfare to market size changes.

Many of the ideas that we develop regarding the response of the economy to changes in population apply to changes in other parameters and to other demand systems. In the appendix, we provide analytical comparative statics for changes in fixed costs of entry, fixed overhead costs, and the productivity distribution, as well as their decomposition into technical and allocative efficiency. We also show that the same intuitions can be rederived using other generalizations of CES preferences.

Related Literature. This paper builds on a large literature that considers how changes in market size affect entry, competition, and welfare. We adopt a framework with monopolistic competition and a representative consumer with a taste for variety, following Spence (1976) and Dixit and Stiglitz (1977).

The first analyses of how market size affect welfare assume that firms are homogeneous, such as Krugman (1979), Mankiw and Whinston (1986), Vives (1999), or Venables (1985). For example, Krugman (1979) shows that, in an economy with homogeneous firms, an increase in market size affects welfare through two channels: the entry of new varieties, and the decrease in markups as the relative share of each variety in total consumption falls. More recently, this line of research has been extended by Bilbiie et al. (2012) and Bilbiie et al. (2019) in a dynamic context, and by Matsuyama and Ushchev (2020b) for more general classes of homothetic preferences.

The heterogeneous firm case has been studied by Melitz (2003) when efficient, and by Asplund and Nocke (2006), Melitz and Ottaviano (2008), Epifani and Gancia (2011), Zhelobodko et al. (2012), Melitz and Redding (2015), Edmond et al. (2018), Dhingra and Morrow (2019), Mrázová and Neary (2017), Mrázová and Neary (2019), and Arkolakis et al. (2019) when inefficient. We highlight how our approach differs from a few of the most recent contributions in this literature.

Dhingra and Morrow (2019) decompose the gains from an increase in market size in an economy with heterogeneous firms compared to an economy with homogeneous firms under (non-homothetic) directly additive preferences. They show that certain restrictions on demand are sufficient for gains in a heterogeneous firm economy to be greater—namely, that markups are increasing in size (Marshall's second law of demand) and preferences are "aligned" (i.e., consumer surplus ratios associated with varieties are also increasing with quantity).[4] We provide a different decomposition focused instead on firms' margins of adjustment (entry, exit, and changes in markups). This allows us to isolate the Darwinian effect, which can be signed without restrictions on the shape of demand curves. By quantifying the model, we show that the Darwinian effect plays the dominant role.

[4] Alternatively, gains in a heterogeneous firm economy are also greater than those in a homogeneous firm economy if markups and consumer surplus ratios are both decreasing with quantity instead, as long as the product of price elasticities and pass-throughs are increasing in quantity.

Relatedly, Mrázová and Neary (2019) show that, for demand systems satisfying Marshall's second law of demand, an increase in scale increases the profits of large firms, which they term the "Matthew Effect." While their focus on firm profits is different from our focus on consumer welfare, we show in our quantitative calibration that the Darwinian effect leads to a reallocation of employment and market share to high-markup firms. If markups are positively correlated with firm size, then increases in market size lead to an increase in market concentration consistent with Mrázová and Neary (2019).

Arkolakis et al. (2019) explore pro-competitive effects in an open economy with an export margin following shocks to iceberg trade costs. They find that pro-competitive effects on welfare are zero when preferences are homothetic and mildly reduce, rather than increase, welfare for important classes of non-homothetic preferences. Compared to Arkolakis et al. (2019), we avoid restrictions that make welfare invariant to the creation of varieties or make entry invariant to the shock. For example, in their model, the absence of fixed costs of accessing domestic and foreign markets means that the creation and destruction of "cut-off" goods has no first-order effects on welfare (i.e., the selection effect is always zero). Moreover, since the mass of firms that choose to enter is not affected by changes in iceberg costs in their model, the Darwinian effect is absent in their results. In our model, firms incur overhead costs to operate and the mass of entrants changes in response to changes in the size of the market; as a result, none of the three effects (Darwinian, selection, and pro-competitive) are generically zero following a change in market size. Nevertheless, our findings on the pro-competitive effects of scale accord with Arkolakis et al. (2019): in our calibration, we find that adjustments on the markup margin are small in magnitude and mildly reduce, rather than enhance, welfare.

Edmond et al. (2018) also consider monopolistically competitive economies with free entry and Kimball (1995) preferences. They compute the economy's distance from the Pareto-efficient frontier under a Klenow and Willis (2016) parameterization. Our focus is different since we study returns to scale in the decentralized equilibrium rather than the distance to the frontier.[5]

Finally, compared to previous work, we provide a new strategy for calibrating our non-parametric model. This approach allows us to quantify the importance of the Darwinian, selection, and pro-competitive channels. We find this approach offers significant advantages compared to calibrating an off-the-shelf functional form, since parametric specifications may mute important features of the data.[6]

Structure of the paper. The structure of the rest of the paper is as follows. Section 2 sets up the model and defines the equilibrium. Section 3 decomposes changes in welfare into changes in technical and allocative efficiency and introduces sufficient statistics in the data that we use to

[5]We provide a characterization of the economy's distance to the efficient frontier in Appendix F.

[6]We show in Appendix I that the popular Klenow and Willis (2016) specification of Kimball preferences is unable to match key features of our data, and hence quantitatively understates allocative efficiency effects compared to our benchmark results.

characterize our results. Section 4 shows how welfare responds to an increase in market size and what role reallocations play. Section 5 draws out the implications of these reallocations for how welfare responds to a tax or subsidy on entry. Section 6 introduces a calibration strategy allowing us to take the model to the data non-parametrically. Section 7 is a quantitative application. Section 8 summarizes extensions, and Section 9 concludes. The appendix contains all the proofs.

2 Model Setup

In this section, we specify the households' and firms' problems and define the equilibrium.

Households. There is a population of L identical consumers. Each consumer supplies one unit of labor and consumes different varieties of final goods indexed by a type θ. Consumers have homothetic preferences, with per-capita utility Y defined implicitly in money-metric terms by

$$\int_{\theta\in\Theta} \Upsilon_\theta(\frac{y_\theta}{Y})dF(\theta) = 1, \tag{1}$$

where y_θ is the per-capita consumption of variety θ, the function Υ_θ is increasing and concave with $\Upsilon_\theta(0) = 0$, the set Θ contains all potential varieties, and $dF(\theta)$ is the measure of varieties of type θ. We return to the definitions of Θ and $dF(\theta)$ with more precision when we discuss the firm side of the economy below.

These preferences, introduced by Matsuyama and Ushchev (2017), are a generalization of Kimball (1995) preferences, since the aggregator function Υ_θ is allowed to vary by variety. CES preferences are a special case of equation (1) when $\Upsilon_\theta(x) = \Upsilon(x) = x^{\frac{1-\sigma}{\sigma}}$. Matsuyama and Ushchev (2017) show that there are other ways one could generalize CES preferences while maintaining homotheticity and tractability. In Appendix H, we show that our theoretical and quantitative results are very similar if we use these alternatives.[7]

Consumers maximize their utility Y subject to the budget constraint

$$\int_{\theta\in\Theta} p_\theta y_\theta dF(\theta) = 1, \tag{2}$$

where p_θ is the price of variety θ. We normalize the nominal wage to one, so that each consumer's income is equal to one. This expression for the budget anticipates the fact that free entry forces profits to zero in equilibrium, so that wages are the sole source of household income.

Solving the household problem yields the per-capita inverse-demand curve for an individual variety θ,

$$\frac{p_\theta}{P} = \Upsilon'_\theta(\frac{y_\theta}{Y}), \tag{3}$$

[7]We have also derived similar versions of our results (available upon request) using non-homothetic separable preferences (as in Krugman, 1979 and Dhingra and Morrow, 2019).

where the *price aggregator* P and the *demand index* $\bar{\delta}$ are defined as

$$P = \frac{\bar{\delta}}{Y}, \qquad \text{and} \qquad \frac{1}{\bar{\delta}} = \int_{\theta \in \Theta} \Upsilon'_\theta(\frac{y_\theta}{Y}) \frac{y_\theta}{Y} dF(\theta). \tag{4}$$

Equation (3) demonstrates the appeal of these preferences — by choosing Υ_θ, we can generate residual demand curves of any desired (downward-sloping) shape for each variety. Furthermore, since Υ'_θ can vary by θ, different varieties can face different residual demand curves. Equation (3) also makes clear that the relative demand for a variety θ is determined by the ratio of its price, p_θ, to the price aggregator, P. Hence, the price aggregator P mediates competition between any given variety and all other available goods.

Note that the price aggregator P does not, in general, coincide with the ideal price index for the representative consumer, and hence deflating income by P does not yield welfare (except in the case of CES preferences).[8]

Firms. Each firm supplies a single variety and seeks to maximize profits under monopolistic competition similar to the production structure in Melitz (2003). To enter, firms incur a fixed entry cost of f_e units of labor. Upon entry, firms draw their type $\theta \in [0,1]$ from a distribution with density $g(\theta)$ and cumulative distribution function $G(\theta)$. Having drawn its type, each firm then decides whether to produce or to exit. Production requires paying an overhead cost of $f_{o,\theta}$ units of labor and a constant marginal cost of $1/A_\theta$ units of labor per unit of the good produced. Finally, the firm decides what price to set, taking as given its residual demand curve. We allow the firm's residual demand curve Υ'_θ, overhead cost $f_{o,\theta}$, and productivity A_θ to vary with the firm's type θ.

From (3), the price-elasticity of demand facing a variety of type θ, denoted σ_θ, is given by

$$\sigma_\theta(\frac{y}{Y}) = -\frac{\partial \log y_\theta}{\partial \log p_\theta} = \frac{\Upsilon'_\theta(\frac{y}{Y})}{-\frac{y}{Y}\Upsilon''_\theta(\frac{y}{Y})}. \tag{5}$$

Conditional on operating, a firm of type θ will set its price equal to a markup μ_θ times its marginal cost $1/A_\theta$. The profit-maximizing markup is given by the usual Lerner formula,

$$\mu_\theta(\frac{y}{Y}) = \frac{1}{1 - \frac{1}{\sigma_\theta(\frac{y}{Y})}}. \tag{6}$$

In the case of CES preferences, firms face identical price-elasticities of demand $\sigma_\theta = \sigma$, and hence have constant desired markups $\mu_\theta = \sigma/(\sigma-1)$ in the cross-section and time-series. The generalized preferences we consider instead allow firms' desired markups to vary with type θ and with each

[8] Let $e(\{p_\theta\}, Y)$ be the expenditure function of a household as a function of the price of all varieties p_θ and welfare Y, where the price of unavailable varieties is set to infinity. Since preferences are homothetic, we can write $e(\{p_\theta\}, Y) = P^Y Y$, where P^Y is the ideal price index. Changes in the price aggregator are $d \log P = d \log \bar{\delta} + d \log P^Y$. Since $\bar{\delta}$ is not, in general, a constant, the price aggregator and the ideal price index do not coincide. The exception is CES preferences, under which $\bar{\delta} = \sigma/(\sigma-1)$ is constant, and thus $P = P^Y$.

firm's position on the demand curve (y/Y).

To ensure that each firm's profit-maximizing price is unique, we assume restrictions on Υ_θ such that marginal revenue curves are strictly downward sloping.[9] Since y_θ is the per-capita output of the firm, the firm's total output is Ly_θ.

A firm of type θ chooses to produce if, and only if, its variable profits exceed the overhead cost of production, i.e.,

$$Lp_\theta y_\theta \left(1 - \frac{1}{\mu_\theta}\right) \geq f_{o,\theta}. \tag{7}$$

Denote the ratio of variable profits to overhead costs by

$$X_\theta = \frac{Lp_\theta y_\theta}{f_{o,\theta}} \left(1 - \frac{1}{\mu_\theta}\right), \tag{8}$$

and assume that firm types are ordered so that X_θ is strictly increasing and continuous in $\theta \in [0, 1]$. Furthermore, assume that X_θ varies smoothly in θ.[10] Define θ^* to be the infimum of the set $\{\theta \in [0, 1] : X_\theta \geq 1\}$. Firms with types $\theta \geq \theta^*$ decide to produce, since variable profits for these firms exceed overhead costs, and firms of type $\theta < \theta^*$ exit.

Free entry implies that firms enter until the expected variable profit minus overhead costs of any entering firm is equal to the fixed cost of entry:

$$\int_{\theta^*}^{1} \left[Lp_\theta y_\theta \left(1 - \frac{1}{\mu_\theta}\right) - f_{o,\theta}\right] g(\theta) d\theta \geq f_e. \tag{9}$$

The set of operating firms, and hence varieties available to the representative consumer, is $\{\theta \in [0, 1] : \theta \geq \theta^*\}$. The measure of firms of type θ is defined by the density $dF(\theta) = Mg(\theta)\mathbf{1}_{(\theta \geq \theta^*)} d\theta$, where M is the mass of entrants and $\mathbf{1}$ is an indicator function.

Equilibrium. In equilibrium, consumers maximize utility taking prices as given; firms maximize profits taking prices other than their own and consumer welfare as given; and markets clear. That is, an equilibrium is determined by equations (1), (2), (3), (4), (6), (7), and (9).

Notation. Denote the sales share density by

$$\lambda_\theta = (1 - G(\theta^*)) M p_\theta y_\theta. \tag{10}$$

[9] In terms of primitives, we assume that $x\Upsilon_\theta'''(x) < -2\Upsilon_\theta''(x)$ for all x and all θ.

[10] In terms of primitives, this means that firms are ordered in such a way that $\frac{-\sigma_\theta}{\rho_\theta} \frac{\partial \log \mu_\theta}{\partial \theta} + \left(\frac{\sigma_\theta}{\rho_\theta} - 1\right) \frac{\partial \log A_\theta}{\partial \theta} - \frac{\partial \log f_{o,\theta}}{\partial \theta} \geq 0$ where ρ_θ is the pass-through function defined in terms of primitives by (13).

This is a density because it is always non-negative and integrates to one.[11] For some variable z_θ, define the sales-weighted average by

$$\mathbb{E}_\lambda[z_\theta] = \int_{\theta^*}^{\infty} \lambda_\theta z_\theta \frac{g(\theta)}{1 - G(\theta^*)} d\theta. \tag{11}$$

Similarly, denote the sales-weighted covariance of any two variables x_θ and z_θ by

$$Cov_\lambda[x_\theta, z_\theta] = \mathbb{E}_\lambda[x_\theta z_\theta] - \mathbb{E}_\lambda[x_\theta]\mathbb{E}_\lambda[z_\theta].$$

3 Central Concepts

In this section, we introduce some central concepts that will guide our analysis. First, we define how welfare changes can be decomposed into changes in technical and allocative efficiency. Second, we introduce statistics related to the shape of the demand curve that help make sense of equilibrium reallocations. Third, we discuss how welfare is determined in terms of some intuitive, but endogenous, variables. We build on the definitions in this section to prove our main results in Sections 4 and 5.

3.1 Changes in Technical and Allocative Efficiency

To understand the drivers of changes in welfare, it is useful to decompose changes in welfare into those driven by technical and allocative efficiency changes. Changes in technical efficiency capture the direct impact of the shock, holding the allocation of resources constant. Changes in allocative efficiency capture the indirect impact of the shock resulting from the endogenous beneficial (or harmful) reallocations that are triggered by the shock.[12]

Following Baqaee and Farhi (2019), we define the exogenous technology vector $\mathcal{A} = (L, f_e, \{f_{o,\theta}\}, \{A_\theta\})$ and the endogenous allocation vector $\mathcal{X} = (l_e, \{l_{o,\theta}\}, \{l_\theta\})$. The technology vector $\mathcal{A}$ captures the primitive parameters defining the production possibilities of the economy. The allocation vector $\mathcal{X}$, on the other hand, describes the fractions of labor allocated to the following activities: entry, overhead, and variable production of varieties of each type θ. Together, $\mathcal{A}$ and $\mathcal{X}$ entirely describe any feasible allocation. Let $\mathcal{Y}(\mathcal{A}, \mathcal{X})$ be the associated level of consumer welfare. Our analysis

[11] Since M is the mass of entrants and θ^* is the selection cut-off, $(1 - G(\theta^*))M$ is the mass of surviving firms and this integrates to one from the budget constraint (2).

[12] Our notion of allocative efficiency compares changes in welfare due to reallocations against the benchmark where the allocation of resources is held constant. A different notion of allocative efficiency that is also sometimes used in the literature measures changes in the distance to the efficient frontier. Changes in that measure of allocative efficiency depend both on whether reallocations are beneficial/harmful and how far the efficient frontier moves due to changes in technology. See Baqaee and Farhi (2019) for a discussion.

decomposes changes in welfare into changes in technical and allocative efficiency as

$$d\log Y = \underbrace{\frac{\partial \log \mathcal{Y}}{\partial \log \mathcal{A}} d\log \mathcal{A}}_{\text{technical efficiency}} + \underbrace{\frac{\partial \log \mathcal{Y}}{\partial \mathcal{X}} d\mathcal{X}}_{\text{allocative efficiency}}. \tag{12}$$

At the efficient allocation, the envelope theorem implies that changes in allocative efficiency are zero to a first-order. Inefficiencies in the initial allocation of resources open the door for reallocations to have first-order effects on welfare. Hence, in the general case, our model will feature changes in both technical and allocative efficiency following a shock to market size.

3.2 Pass-Throughs and Consumer Surplus Ratios

In this section, we introduce two statistics related to the shape of demand curves that we use to characterize changes in consumer welfare. We define the pass-through of a variety as the elasticity of its price to its marginal cost. A firm's pass-through can be expressed as a function of primitives,

$$\rho_\theta(\tfrac{y}{Y}) = \frac{\partial \log p_\theta}{\partial \log mc_\theta} = 1 + \frac{\partial \log \mu_\theta}{\partial \log mc_\theta} = \frac{1}{1 + \frac{\frac{y}{Y}\mu'_\theta(\frac{y}{Y})}{\mu_\theta(\frac{y}{Y})}\sigma_\theta(\frac{y}{Y})}, \tag{13}$$

where the price-elasticity of demand and markup functions are given by (5) and (6). Under CES preferences, firms' desired markups are constant, and hence firms exhibit "complete pass-through" ($\rho_\theta = 1$). In general, however, firms' desired markups may vary with size. For example, if a firm's desired markup is increasing in its size, the firm will exhibit "incomplete pass-through" ($\mu'_\theta(\frac{y}{Y}) > 0$ and thus $\rho_\theta < 1$).[13]

We denote the consumer surplus per unit sales for a variety by δ_θ. More precisely, δ_θ is the ratio of consumption-equivalent utility from a marginal variety, $\bar{\delta}\Upsilon_\theta(\frac{y}{Y})$, to its per-capita sales:

$$\delta_\theta(\tfrac{y}{Y}) = \frac{\bar{\delta}\Upsilon_\theta(\frac{y}{Y})}{p_\theta y_\theta} = \frac{\Upsilon_\theta(\frac{y}{Y})}{\frac{y}{Y}\Upsilon'_\theta(\frac{y}{Y})}. \tag{14}$$

Figure 1 gives a visual representation of δ_θ as the ratio of consumer surplus $A + B$ to revenues A. Naturally, the consumer surplus ratio $\delta_\theta \geq 1$ for all θ. In a CES model, δ_θ measures the "love-of-variety" effect. In this model, this love-of-variety effect can vary both by type θ and relative size y_θ/Y.

By integrating over Equation (14), we can show that the demand index in (4) is simply the sales-weighted average of this consumer surplus ratio,

$$\mathbb{E}_\lambda[\delta_\theta] = \bar{\delta}. \tag{15}$$

[13]This is sometimes referred to as Marshall's second law of demand (see Melitz, 2018 for more information).

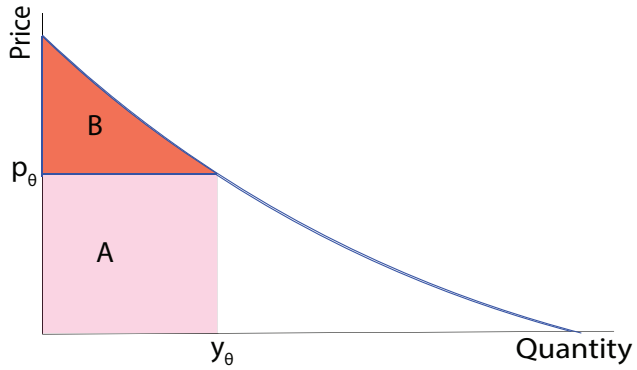


Figure 1: Graphical illustration of δ_θ as the area under the residual demand curve divided by revenues. That is $\delta_\theta = (A + B)/A \geq 1$.

3.3 Welfare

We are interested in how welfare (per capita) responds to changes in market size. The change in consumer welfare, measured using either the equivalent or compensating variation, is given by

$$d\log Y = \underbrace{\left(\mathbb{E}_\lambda[\delta_\theta] - 1\right) d\log M}_{\substack{\text{Consumer surplus} \\ \text{from entry of new varieties}}} \underbrace{- \left(\delta_{\theta^*} - 1\right) \lambda_{\theta^*} \frac{g(\theta^*)}{1 - G(\theta^*)} d\theta^*}_{\substack{\text{Consumer surplus (loss)} \\ \text{from exit of varieties } d\theta^*}} \underbrace{- \mathbb{E}_\lambda \left[d\log \mu_\theta\right]}_{\substack{\text{Marginal surplus} \\ \text{from price changes}}} . \tag{16}$$

Intuitively, welfare changes $d\log Y$ incorporate the consumer surplus brought about by the entry of new varieties $d\log M$ or destroyed by the exit of varieties $d\theta^*$ via the first two terms on the right-hand side of (16). The final term captures how marginal changes in prices of non-entering and non-exiting goods affect the consumer.

As pointed out by Aghion et al. (2019a), statistical agencies calculate real GDP using the change in prices for continuing varieties present before and after a change. Hence, the change in measured real GDP per capita in our model is equal to the last summand in (16)—that is, the change in nominal income per capita deflated by the change in prices for continuing varieties weighted by their market shares,[14]

$$d\log Q = -\mathbb{E}_\lambda[d\log p_\theta] = -\mathbb{E}_\lambda \left[d\log \mu_\theta\right]. \tag{17}$$

Since we assume that labor is the only factor of production, changes in real GDP per capita are also equal to changes in aggregate TFP. This extends the observation made by Jaimovich and Floetotto

[14]In principle, changes in real GDP can either be defined using a quantity index or a price index. In practice, real GDP is measured using the GDP deflator, so we use the price index definition. We include a discussion of the quantity index in Appendix J.

(2008) who show that, in a model with entry and variable markups, variations in markups affect aggregate productivity. The key point is that changes in real GDP per capita and aggregate TFP do not generically coincide with changes in welfare if $\delta_\theta \neq 1$.

It is worth discussing a couple cases in which consumer welfare changes and real GDP changes do coincide. If the model did not allow for the creation and destruction of varieties, then the first two terms of (16) would be zero, and changes in consumer welfare would equal changes in real GDP per capita. Consumer welfare changes and real GDP changes also coincide in models of entry featuring no fixed costs and demand curves with choke prices. If new goods enter smoothly from the choke price, then $\delta_\theta = 1$ for all entrants, and the first two terms are zero.[15]

4 Changes in Market Size

In this section, we characterize how an increase in L affects welfare. Following Krugman (1979), one can think of this as capturing the effect of trade integration of symmetric economies. Suppose we have N countries with identical tastes and technologies, with populations $L_1, L_2, ..., L_N$. The market equilibrium if these N countries trade freely is the same as the market equilibrium in a single, closed economy with size $L_1 + L_2 + ... + L_N$; hence, comparative statics of the equilibrium with respect to L can be interpreted as the effect of opening to trade with symmetric foreign markets.

As noted by Helpman and Krugman (1985), reallocations associated with increased competition can mitigate or exacerbate existing distortions. We first describe the distortions in the initial equilibrium. Then, we show how the reallocations caused by a market expansion interact with these pre-existing distortions to affect allocative efficiency.

4.1 Sources of Inefficiency

An allocation is inefficient if welfare can be increased by reallocating labor between entry, overhead, and variable production while keeping the total amount of labor fixed. There are three margins along which the allocation can be inefficient in this model: (1) entry can be excessive or insufficient; (2) selection can be too tough or too weak; (3) the cross-sectional allocation of labor across variable production may be distorted. We discuss these three different kinds of inefficiency in turn and show that each can be characterized with simple conditions on the sufficient statistics presented in Section 3.2.

In what follows, we define *local* efficiency for each margin. That is, whether a marginal reallocation along some dimension improves or decreases welfare. This is distinct from global

[15] When new varieties enter smoothly from the choke price, rather than across the type distribution, the first term will rely on δ_{θ^*} rather than $\mathbb{E}_\lambda[\delta_\theta]$. This discussion applies, for example, to Arkolakis et al. (2019): In their model, there are no fixed costs of exporting and export quantities vary smoothly from zero (at the choke price), so the response of real GDP per capita and welfare to a change in iceberg trade costs coincide.

efficiency which compares the allocation to the first-best allocation. These local notions of efficiency are the ones that are relevant for reallocation effects in the decentralized equilibrium.

Entry efficiency. Consider a marginal reallocation that reduces variable production labor and increases entry and overhead labor, keeping the selection cut-off and the relative allocation of labor across varieties constant. If this perturbation raises welfare, we say that entry is insufficient. If the opposite holds, we say that entry is excessive.

Lemma 1 (Excessive/Insufficient Entry)**.** *Entry is insufficient if, and only if,*

$$\mathbb{E}_\lambda \left[\mu_\theta^{-1}\right]^{-1} < \mathbb{E}_\lambda[\delta_\theta]. \tag{18}$$

If this inequality is reversed, entry is excessive.

In words, there is too little entry if the harmonic (sales-weighted) average of firm markups is less than the sales-weighted average consumer surplus ratio. Intuitively, entrants respond to average markups (since markups determine profits), but the value of entry for consumers depends on the average consumer surplus ratio. In a CES model, (18) holds as an equality and so the CES model has efficient entry.

Selection efficiency. We say that selection is too weak if marginally increasing the selection cutoff—and reallocating the labor from those newly exiting varieties proportionately to entry, overhead, and variable production—increases welfare.

Lemma 2 (Tough/Weak Selection)**.** *Selection is too weak if, and only if,*

$$\delta_{\theta^*} < \mathbb{E}_\lambda[\delta_\theta]. \tag{19}$$

If this inequality is reversed, selection is too tough.

Suppose that the selection cutoff θ^* increases. If the consumer surplus associated with the marginal variety δ_{θ^*} is lower than the average $\mathbb{E}_\lambda[\delta_\theta]$, the welfare associated with new varieties created from the freed-up labor outweighs the welfare loss from the exiting varieties. Since the increase in the selection cut-off is welfare-improving, in this case, we say that selection was initially too weak.

Crucially, note that if the inequality in (19) is reversed, then an increase in the selection cut-off $d\theta^* > 0$ reduces efficiency and welfare. Therefore, tougher selection and the death of marginally profitable firms is not, ipso facto, evidence that efficiency is rising. In a CES model, (19) holds as an equality and so the CES model has efficient entry.

Relative production efficiency. Finally, we say that the amount of variable labor dedicated to the production of one variety is too high compared to another if, on the margin, welfare increases when variable labor is reallocated from the former to the latter.

Lemma 3 (Cross-section misallocation)**.** *Variable labor of variety θ' is too high compared to that of variety θ if, and only if,*

$$\mu_{\theta'} < \mu_{\theta}. \tag{20}$$

Intuitively, firms with higher markups are inefficiently small in the cross-section compared to firms with lower markups. Hence, reallocating labor from a low-markup firm to a high-markup firm increases allocative efficiency. Crucially, it is a comparison of markups μ_θ, and not productivities A_θ, that determines whether or not one firm should be larger than another from a social perspective. If markups happen to be positively associated with productivity, then an expansion of more productive firms increases welfare, but this is only because "high productivity" proxies for "high markup."[16]

In a CES model, (20) holds as an equality and so the CES model has an efficient cross-sectional allocation of resources.

4.2 Welfare and Shocks to Market Size

We characterize the change in welfare following an exogenous change in market size.

Theorem 1 (Welfare Effect of Change in Market Size)**.** *In response to changes in population $d\log L$, changes in consumer welfare are given by*

$$d\log Y = \underbrace{\left(\mathbb{E}_\lambda[\delta_\theta] - 1\right) d\log L}_{\textit{technical efficiency}} + \underbrace{\frac{\xi^\epsilon + \xi^{\theta^*} + \xi^\mu}{1 - \xi^\epsilon - \xi^{\theta^*} - \xi^\mu}\left(\mathbb{E}_\lambda[\delta_\theta]\right) d\log L}_{\textit{allocative efficiency}}, \tag{21}$$

where

$$\begin{aligned}
(\textit{Darwinian Effect}) \quad \xi^\epsilon &= \left(\mathbb{E}_\lambda[\delta_\theta] - 1\right) Cov_\lambda\left[\sigma_\theta, \frac{1}{\mu_\theta}\right], \\
(\textit{Selection Effect}) \quad \xi^{\theta^*} &= \left(\mathbb{E}_\lambda[\delta_\theta] - \delta_{\theta^*}\right) \lambda_{\theta^*} \gamma_{\theta^*} \left(\mathbb{E}_\lambda\left[\frac{\sigma_{\theta^*}}{\sigma_\theta}\right] - 1\right), \\
(\textit{Pro/Anti-competitive Effect}) \quad \xi^\mu &= \mathbb{E}_\lambda\left[\left(1 - \rho_\theta\right)\sigma_\theta\left(1 - \frac{\mathbb{E}_\lambda[\delta_\theta]}{\mu_\theta}\right)\right]\mathbb{E}_\lambda\left[\frac{1}{\sigma_\theta}\right],
\end{aligned}$$

and $\gamma_{\theta^*} > 0$ is the hazard rate of the profitability distribution X_θ at the selection cut-off. In terms of primitives, this is

$$\frac{1}{\gamma_{\theta^*}} = \frac{1 - G(\theta^*)}{g(\theta^*)}\left[\left.\frac{\partial \log X_\theta}{\partial \theta}\right|_{\theta^*}\right] = \frac{1 - G(\theta^*)}{g(\theta^*)}\left[\frac{-\sigma_\theta}{\rho_\theta}\frac{\partial \log \mu_\theta}{\partial \theta} + \left(\frac{\sigma_\theta}{\rho_\theta} - 1\right)\frac{\partial \log A_\theta}{\partial \theta} - \left.\frac{\partial \log f_{o,\theta}}{\partial \theta}\right|_{\theta^*}\right].$$

[16] In general, the productivity level is irrelevant for whether a given reallocation improves or worsens efficiency. This contrasts with statistical decompositions, for example Olley and Pakes (1996), which consider a reallocation towards firms with higher *levels* of productivity A_θ as an indicator of an improvement in efficiency. For more detail, see the discussion in Baqaee and Farhi (2019).

Equation (21) decomposes the change in welfare into a technical and allocative efficiency effect according to the definition in Section 3.1. We start by discussing the technical efficiency term before discussing the allocative efficiency term.

The first term in Equation (21) captures the changes in technical efficiency that arise due to an increase in market size, holding fixed the proportional allocation of resources across uses (entry, overhead, and variable production). Because the fraction of labor allocated to entry is held fixed, the increase in population implies a proportional increase in entry. This has two offsetting effects. First, the new varieties increase consumer welfare by $\mathbb{E}_\lambda[\delta_\theta] d \log L$, since the consumer's surplus associated with the new varieties will average $\mathbb{E}_\lambda[\delta_\theta]$. On the other hand, the increase in the number of varieties reduces the per-capita consumption of existing varieties by $d \log L$. The net effect balances these two offsetting effects. Since $\delta_\theta \geq 1$, the technical efficiency term is always positive. In a CES model, this is the only effect.

The second term in (21) captures how changes in the allocation of resources contribute to welfare. Each of ξ^ϵ, ξ^{θ^*}, and ξ^μ relates to a particular type of reallocation. Throughout, we assume that $\xi^\epsilon + \xi^\mu + \xi^{\theta^*} < 1$, which guarantees that the equilibrium exists and is locally unique. We provide sufficient conditions that guarantee this in Appendix D.1.[17]

One way of thinking about the decomposition of the allocative efficiency effect into ξ^ϵ, ξ^{θ^*}, and ξ^μ is that the general equilibrium response can be analyzed as a series of three successive allocations, each of which allows firms to adjust along a greater number of margins.[18] In the first restricted allocation, we allow free entry, but hold markups and the selection cutoff constant (i.e., μ_θ and θ^* are fixed using implicit taxes). The change in welfare in this allocation is the same as in Theorem 1, but setting $\xi^{\theta^*} = \xi^\mu = 0$. In the second allocation, firms can also change their decision to operate but still cannot alter their markups. The change in welfare in this allocation is equal to Theorem 1, but setting $\xi^\mu = 0$. Finally, the third allocation allows firms to adjust on all three margins: entry, exit, and choice of markup.

To fix ideas, we consider three special cases, each of which isolates and focuses on the intuition for a different margin of adjustment.

[17] As discussed in Appendix D.1, since $\xi^\epsilon + \xi^\mu + \xi^{\theta^*}$ is a function of $\{\{\lambda_\theta, \delta_\theta, \mu_\theta, \rho_\theta, X_\theta\}_{\theta \in \Theta}, \theta^*\}$, and any feasible collection $\{\{\lambda_\theta, \delta_\theta, \mu_\theta, \rho_\theta, X_\theta\}_{\theta \in \Theta}, \theta^*\}$ can be rationalized via some collection of primitives $\{\Upsilon_\theta, A_\theta, f_{o,\theta}\}$, then as long as $0 \leq \xi^\epsilon + \xi^\mu + \xi^{\theta^*} < 1$, we can use the inverse function theorem to show that an equilibrium exists and is locally unique. This requirement is satisfied in our calibrated application.

[18] The decomposition in Theorem 1 is different to the one provided by Dhingra and Morrow (2019). We focus on how welfare is affected by different margins of adjustment. Dhingra and Morrow (2019) instead decompose gains from an increase in market size into those present in homogeneous versus heterogeneous firm models. The quantity reallocations they isolate, for example, group together Darwinian effects with effects due to heterogeneous pass-throughs, and cannot be signed without assumptions on the shape of demand.

4.2.1 Darwinian Effect

To isolate the role of the Darwinian effect, consider an economy in which there are no overhead costs ($f_{o,\theta} = 0$) so that $\theta^* = 0$. Furthermore, assume that preferences are given by

$$\int_{\theta^*}^{\infty} \left(\frac{y_\theta}{Y}\right)^{\frac{\sigma_\theta - 1}{\sigma_\theta}} dF(\theta) = 1, \tag{22}$$

which is a special case of (1) with $\Upsilon_\theta(x) = x^{(\sigma_\theta - 1)/\sigma_\theta}$.[19]

In this example, markups can vary in the cross-section of firms because $\mu_\theta = \frac{\sigma_\theta}{\sigma_\theta - 1}$, but markups do not vary in time-series because pass-though is complete ($\rho_\theta = 1$). The fact that markups do not change means that $\xi^\mu = 0$, and the fact that there are no overhead costs means that $\xi^{\theta^*} = 0$. Hence, we have the following.

Corollary 1 (Darwinian Effect)**.** *When preferences are given by* (22) *and overhead costs are zero, the change in welfare from an increase in market size is given by*

$$d\log Y = \underbrace{\Big(\mathbb{E}_\lambda[\delta_\theta] - 1\Big) d\log L}_{\textit{technical efficiency}} + \underbrace{\frac{\xi^\epsilon}{1 - \xi^\epsilon}\Big(\mathbb{E}_\lambda[\delta_\theta]\Big) d\log L}_{\textit{allocative efficiency}} \geq 0. \tag{23}$$

Changes in allocative efficiency are strictly positive ($\xi^\epsilon > 0$) as long as there is any heterogeneity in markups (and therefore price-elasticities):

$$\xi^\epsilon = (\mathbb{E}_\lambda[\delta_\theta] - 1) Cov_\lambda\left[\sigma_\theta, \frac{1}{\mu_\theta}\right] = -(\mathbb{E}_\lambda[\delta_\theta] - 1) Cov_\lambda\left[\sigma_\theta, \frac{1}{\sigma_\theta}\right] \geq 0. \tag{24}$$

In other words, the Darwinian effect is unambiguously positive regardless of the shape of demand curves and does not depend on whether entry is excessive or insufficient.

To understand this effect, note that the change in the relative per-capita quantity of each variety depends on the price-elasticity of demand and its price relative to the price index:

$$d\log(\frac{y_\theta}{Y}) = -\sigma_\theta(d\log p_\theta - d\log P) = \sigma_\theta d\log P.$$

The second equality follows from the fact that in this example $d\log p_\theta = d\log \mu_\theta = 0$. Now consider how an increase in market size affects demand for this firm. The increase in market size, and the entry of new firms, causes the price aggregator to fall $d\log P < 0$. The reduction in the price aggregator triggers bigger reductions in per-capita quantities for firms that face more elastic demand. The result is that low-markup firms (who have high price-elasticities of demand) shrink more than high-markup firms (who have low price-elasticities). By Lemma 3, high markup

[19] These preferences were introduced by Matsuyama and Ushchev (2020a). They refer to these as "constant-price-elasticity" preferences. When the σ_θ parameter is uniform across firm types, this collapses to CES.

firms were initially too small relative to the efficient allocation, so this reallocation reduces relative productive inefficiencies and improves welfare. We call this a *Darwinian* effect because a more competitive environment, from a reduction in the price index, selects and shifts resources towards the "fittest" firms (those with the most inelastic demand). The multiplier $(\mathbb{E}_\lambda[\delta_\theta] - 1)$ in (24) appears because the reallocations caused by the Darwinian effect save on labor, and these extra resources are funneled into additional entry.

4.2.2 Selection Effect

We now relax the assumption of zero overhead costs, while retaining the constant markups and complete pass-throughs of the previous example. As a result, we reintroduce a source of allocative efficiency changes due to changes in the selection cut-off (ξ^{θ^*}), but continue to hold $\xi^\mu = 0$.

Corollary 2 (Darwinian and Selection Effect)**.** *When preferences are given by* (22) *and overhead costs are nonzero, the change in welfare from an increase in market size is given by*

$$d\log Y = \underbrace{\left(\mathbb{E}_\lambda[\delta_\theta] - 1\right) d\log L}_{\textit{technical efficiency}} + \underbrace{\frac{\xi^\epsilon + \xi^{\theta^*}}{1 - \xi^\epsilon - \xi^{\theta^*}} \left(\mathbb{E}_\lambda[\delta_\theta]\right) d\log L}_{\textit{allocative efficiency}}. \tag{25}$$

Whilst the Darwinian effect is always positive, changes in the selection cut-off will only increase welfare if

$$\xi^{\theta^*} = \left(\mathbb{E}_\lambda[\delta_\theta] - \delta_{\theta^*}\right) \lambda_{\theta^*} \gamma_{\theta^*} \left(\mathbb{E}_\lambda\left[\frac{\sigma_{\theta^*}}{\sigma_\theta}\right] - 1\right) \geq 0.$$

This happens, for example, if consumer surplus ratio at the cut-off δ_{θ^*} is lower than average $\mathbb{E}_\lambda[\delta_\theta]$, and the price elasticity σ_{θ^*} is higher than average $\mathbb{E}_\lambda[\sigma_\theta]$. The second condition ensures that the selection cut-off increases in response to an increase in market size since the marginal firms are more exposed to competition than the average firm, and the first condition ensures that the death of marginal firms is beneficial since selection was too weak to begin with.

An important implication is that increases in the selection cutoff, $d\theta^* > 0$, are not, on their own, evidence of an improvement in allocative efficiency. Increases in selection due to intensifying competition are only socially desirable if allocating labor to the marginal firm provides households with less consumer surplus than reallocating that labor to entry and other surviving firms. Indeed, in our quantitative application in Section 7, we find that increases in the selection cut-off are welfare-reducing.

4.2.3 Pro/Anti-Competitive Effect

In our third and final example, we turn off the Darwinian and selection effects by considering an economy with homogeneous firms. In this example, reallocations are driven purely by the fact that firms change their markups in response to changes in market size.

Corollary 3 (Pro/Anti-competitive effect)**.** *Suppose that all varieties face the same residual demand curve* $\Upsilon'_\theta = \Upsilon'$*, overhead cost* $f_{o,\theta} = f_o$*, and productivity* $A_\theta = 1$*. The change in welfare from an increase in market size is given by*

$$d\log Y = \underbrace{(\delta - 1) d\log L}_{\text{technical efficiency}} + \underbrace{\delta \frac{\xi^\mu}{1-\xi^\mu} d\log L}_{\text{allocative efficiency}}. \tag{26}$$

When firms are homogeneous, $\xi^\epsilon = \xi^{\theta^*} = 0$, and ξ^μ simplifies to

$$\xi^\mu = (1-\rho)\left(1 - \frac{\delta}{\mu}\right). \tag{27}$$

If firms exhibit incomplete pass-through ($\rho < 1$), the allocative effects of markup adjustments are welfare-enhancing if, and only if, there is initially too much entry ($\mu > \delta$). Intuitively, the increase in market size causes the price index to fall, and this causes markups to decrease if $\rho < 1$. A reduction in markups deters entry, which is beneficial if entry was excessive to begin with.

The literature typically refers to the idea that markups may fall with market size as the *pro-competitive effect* of scale. In this example, the pro-competitive effect is captured entirely by $\rho < 1$: markups decrease since the *per-capita* consumption of each firm's output decreases in response to entry. As (27) makes clear, the welfare impact of these pro-competitive effects then depends on the initial efficiency of entry.[20,21]

4.3 Real GDP and Shocks to Market Size

We end this section by considering changes in real GDP. As we discuss in Section 3.3, changes in welfare and real GDP per capita do not generically coincide when we allow for firm entry and exit. Proposition 1 characterizes the change in real GDP per capita following a change in market size.

Proposition 1 (Real GDP Effect of Change in Market Size)**.** *In response to changes in population* $d\log L$*, changes in real GDP per capita are*

$$d\log Q = \mathbb{E}_\lambda\left[1 - \rho_\theta\right] \mathbb{E}_\lambda\left[\frac{1}{\sigma_\theta}\right] (d\log Y + d\log L), \tag{28}$$

where $d\log Y$ *is given by Theorem 1.*

[20]This discussion is closely related to the contemporaneous findings from Matsuyama and Ushchev (2020b), who show that if entry is globally pro-competitive, then entry is excessive in models with homogeneous firms.

[21]Corollary 3 abstracts from firm heterogeneity. If firms are heterogeneous and pass-through is incomplete, whether or not $\xi^\mu > 0$ does not hinge purely on whether or not entry is excessive or insufficient. This can be seen by inspecting Theorem 1. With heterogeneous firms, changes in markups also change the cross-sectional distribution of resources. Whether or not these reallocations are beneficial or harmful is in general ambiguous even if we know that entry is excessive.

An increase in population leads to a reduction in markups under incomplete pass-through $\rho_\theta < 1$, and this pro-competitive effect reduces the price of continuing varieties. Since real GDP depends on the change in the price of continuing varieties, this can cause real GDP per capita to rise $d \log Q = -\mathbb{E}_\lambda[d \log p_\theta]$. If pass-through is complete, then real GDP per capita does not change as the market expands. Hence, reallocation affects welfare and real GDP through very different channels. This also applies to aggregate productivity, as measured by national income accounts, which in this model coincides with real GDP per capita.

5 Policy Interventions

In this section, we consider the implications of our results for policy. Section 4.1 discussed the three margins along which the decentralized allocation can be distorted—entry inefficiency, selection inefficiency, and relative production inefficiencies. The policy that obtains the first-best allocation eliminates all three margins of distortion. Achieving the first-best requires at least as many policy instruments as there are firm types, since at the minimum this policy must correct for all pairwise mismatches in markups across any two firm types θ and θ'. Moreover, the planner also needs to regulate selection by comparing consumer surplus at the cut-off against the average. Whereas such extensive interventions in the market are impracticable, regulating entry is, in comparison, straightforward.[22]

In this section, we consider how a marginal entry tax affects welfare, and show that an entry tax can trigger similar reallocative forces as those in Section 4. The tax on entry, τ, modifies the free entry condition given in (9), so that each entering firm now pays $(1+\tau)f_e$ units of labor upon entry:

$$\int_{\theta^*}^{\infty} \left[\left(1 - \frac{1}{\mu_\theta} \right) p_\theta y_\theta w L - f_{o,\theta} \right] g(\theta) d\theta = (1+\tau) f_e. \tag{29}$$

Government revenues from this tax are rebated lump-sum to households.

For brevity, we include additional details of how these changes affect the system of equilibrium conditions in Appendix E and continue now to the welfare result. Proposition 2 characterizes the response of welfare to a tax on entry, starting from the point where entry is untaxed.

Proposition 2 (Welfare Effect of an Entry Tax)**.** *Suppose entry is initially untaxed (unsubsidized). The response of welfare to a marginal tax on entry is given by*

$$d \log Y = \left(1 - \frac{\mathbb{E}_\lambda \left[\delta_\theta \right] / \mathbb{E}_\lambda \left[\mu_\theta^{-1} \right]^{-1} + \left(\mathbb{E}_\lambda \left[\delta_\theta \right] - \delta_{\theta^*} \right) \lambda_{\theta^*} \gamma_{\theta^*}}{1 - \xi^\epsilon - \xi^{\theta^*} - \xi^\mu} \right) \psi_e \, d\tau, \tag{30}$$

where $\psi_e = f_e / \left(f_e + (1 - G(\theta^*)) \, \mathbb{E}\left[f_{o,\theta} \right] \right)$ is the entry cost share of all fixed costs, and ξ^ϵ, ξ^{θ^*}, and ξ^μ are as

[22] For more discussion of first-best policy, see Appendix F.1, where we characterize the policy that achieves first-best. We use this optimal policy to estimate the distance of the decentralized equilibrium to the efficient frontier.

defined in Theorem 1.

Whether an entry tax increases welfare depends on the sign of the term in parentheses in (30). This term is more likely to be positive—and an entry tax is more likely to be welfare-enhancing—if entry is excessive ($\mathbb{E}_\lambda[\delta_\theta] < \mathbb{E}_\lambda[\mu_\theta^{-1}]^{-1}$), if selection is too tough ($\mathbb{E}_\lambda[\delta_\theta] < \delta_{\theta^*}$), or if the beneficial reallocations from entry given by ξ^ϵ, ξ^{θ^*}, and ξ^μ are small.

An immediate implication of Proposition 2 is that excessive entry (as defined in Lemma 1) is not a sufficient condition for an entry tax to be welfare-increasing. For example, if the beneficial reallocations from entry ($\xi^\epsilon + \xi^{\theta^*} + \xi^\mu$) are sufficiently large, then attempting to correct for excessive entry with an entry tax will actually be welfare-reducing because the economy loses the beneficial cross-sectional reallocations associated with entry.

We illustrate this intuition by briefly discussing the welfare effect of the entry tax in the three special cases from Section 4.

Darwinian effect. Consider again the economy in Section 4.2.1, where there are no overhead costs and preferences are given by (22). In this example, the entry tax has no effect on firms' markups or on selection.

Corollary 4 (Darwinian Effect)**.** *When preferences are given by* (22) *and overhead costs are zero, the change in welfare from a marginal tax on entry is positive if, and only if,*

$$\mathbb{E}_\lambda\left[\delta_\theta\right] < \left(1 - \xi^\epsilon\right) \mathbb{E}_\lambda\left[\mu_\theta^{-1}\right]^{-1}. \tag{31}$$

Note that this condition is more stringent than the condition for excessive entry in Lemma 1, since $\xi^\epsilon > 0$ in any economy with heterogeneous markups. Intuitively, since entry alleviates relative production inefficiencies due to Darwinian reallocations, the welfare impact of an entry tax may be negative if the loss of those Darwinian reallocations outweighs the benefits of moving closer to the efficient level of entry.

Selection effect. Suppose we retain complete pass-through preferences, but now allow for nonzero overhead costs, as in Section 4.2.2. The economy now features both Darwinian and selection effects, but pro-/anti-competitive effects are still absent.

Corollary 5 (Darwinian and Selection Effect)**.** *When preferences are given by* (22) *and overhead costs are nonzero, the change in welfare from a marginal tax on entry is positive if, and only if,*

$$\mathbb{E}_\lambda\left[\delta_\theta\right] < \left(1 - \xi^\epsilon - \left(\mathbb{E}_\lambda\left[\delta_\theta\right] - \delta_{\theta^*}\right) \lambda_{\theta^*} \gamma_{\theta^*} \mathbb{E}_\lambda\left[\frac{\sigma_{\theta^*}}{\sigma_\theta}\right]\right) \mathbb{E}_\lambda\left[\mu_\theta^{-1}\right]^{-1}. \tag{32}$$

This condition is more stringent than the condition in Corollary 4 if selection is too weak ($\delta_{\theta^*} < \mathbb{E}_\lambda[\delta_\theta]$), and less stringent if selection is too tough. Intuitively, an entry tax decreases selection, which is only beneficial if the initial level of selection was too tough.

Pro/anti-competitive effect. Finally, consider an economy with homogeneous firms, as in Section 4.2.3. In this economy, entry has no Darwinian or selection effects, since firms are identical.

Corollary 6 (Pro/Anti-Competitive Effect)**.** *Suppose that all varieties face the same residual demand curve* $\Upsilon'_\theta = \Upsilon'$*, overhead cost* $f_{o,\theta} = f_o$*, and productivity* $A_\theta = 1$*. The change in welfare from a marginal tax on entry is positive if, and only if, entry is excessive:*

$$\delta < \mu. \tag{33}$$

Without firm heterogeneity, the entry margin is the sole source of potential inefficiency. As a result, the change in welfare following an entry tax depends only on whether entry is initially excessive or insufficient as in Lemma 1.

6 Calibration Strategy

In this section, we discuss how we take the theory to the data. We first describe our non-parametric calibration procedure. We then implement it using Belgian data and show how the primitives can be derived from the data. In Section 7, we use the calibrated model to perform quantitative experiments.

6.1 Non-Parametric Calibration Approach

The model has many degrees of freedom, so to calibrate the model, we impose two restrictions: (1) firms face identical overhead costs $f_{o,\theta} = f_o$, and (2) the aggregators Υ_θ take the form,

$$\Upsilon_\theta(\frac{y_\theta}{Y}) = \Upsilon(B_\theta \frac{y_\theta}{Y}). \tag{34}$$

Hence, firms differ in their productivities A_θ and taste shifters B_θ. These taste shifters are equivalent to price reductions, but they are unobservable. Allowing for taste-shifters is important since, in practice, two firms that charge the same price in the data can have very different sales. The presence of taste-shifters allow us to accommodate this possibility.[23]

Given our assumption that overhead costs are the same for all firms, we can identify a firm's type from its position in the sales distribution. We rank firms by sales and assign their type to be the fraction of firms with less sales. We will take two objects as data: (1) the density of sales shares λ_θ, and (2) the distribution of pass-throughs ρ_θ. As we will show, the pass-through function is a third-order differential equation in the Kimball aggregator, and can be used to calibrate the model up to boundary conditions. For boundary conditions, we need to take a stand on the average levels

[23] If there were no taste-shifters, then one could identify the residual demand curve by simply plotting price against quantity in the cross-section. In practice, this is infeasible because the prices firms report are not directly comparable to one-another.

of first and second derivatives, i.e. on the average markup and the average consumer surplus ratio (these will be constants of integration). We will present our estimates for different values of these variables.

Markups and consumer surplus ratios. In the cross-section, markups μ_θ and sales λ_θ must solve two differential equations,

$$\frac{d\log\lambda_\theta}{d\theta} = \frac{\rho_\theta}{\mu_\theta - 1}\frac{d\log\left(A_\theta B_\theta\right)}{d\theta}, \tag{35}$$

$$\frac{d\log\mu_\theta}{d\theta} = \left(1-\rho_\theta\right)\frac{d\log\left(A_\theta B_\theta\right)}{d\theta}. \tag{36}$$

Differences in consumer tastes B_θ are isomorphic to difference in productivities A_θ, and we do not identify them separately. For simplicity, we refer to $(A_\theta B_\theta)$ as a variety's productivity.

The intuition for the first differential equation (35) is that, compared to a firm of type θ, a firm with type $\theta + d\theta$ has higher productivity $d\log\left(A_\theta B_\theta\right)/d\theta$, lower "taste-adjusted" price $d\log p_\theta/d\theta = \rho_\theta d\log\left(A_\theta B_\theta\right)/d\theta$, and thus higher sales $d\log\lambda_\theta/d\theta = (\sigma_\theta - 1)d\log p_\theta/d\theta$. The second differential equation (36) comes from the fact that the relationship of desired markups to productivity is $d\log\mu_\theta/d\log\left(A_\theta B_\theta\right) = 1-\rho_\theta$.

Combining the two equations yields

$$\frac{d\log\mu_\theta}{d\theta} = (\mu_\theta - 1)\frac{1-\rho_\theta}{\rho_\theta}\frac{d\log\lambda_\theta}{d\theta}. \tag{37}$$

Given sales shares λ_θ and pass-throughs ρ_θ, this differential equation allows us to recover markups μ_θ up to a constant μ_{θ^*}. We choose the initial value $\mu_{\theta^*} \geq 1$ to match a given value of the (harmonic) sales-weighted average markup $\bar{\mu} = \mathbb{E}_\lambda[\mu_\theta^{-1}]^{-1}$.

Either of the two differential equations for sales shares or markups then allows us to recover $A_\theta B_\theta$ up to a constant $A_{\theta^*}B_{\theta^*}$, which we normalize to 1.

Finally, we recover consumer surplus ratios using the differential equation

$$\frac{d\log\delta_\theta}{d\theta} = \frac{\mu_\theta - \delta_\theta}{\delta_\theta}\frac{d\log\lambda_\theta}{d\theta}, \tag{38}$$

with the initial condition δ_{θ^*} chosen to match a given value of the average consumer surplus ratio $\mathbb{E}_\lambda[\delta_\theta] = \bar{\delta}$.

Fixed costs and selection cut-off. The information so far does not reveal the cut-off value θ^*, so calibrating this number requires outside information. To calibrate the marginal type θ^*, we step slightly outside the model and imagine that new firms operate for one year before they choose to shut down; after the first year, firms face a constant, exogenous death rate. Hence, the difference between the probability of exit in the first year versus later years identifies θ^*. Conditional on θ^*,

we can back out the fixed costs using the free-entry condition

$$\frac{f_e}{L} + (1 - G(\theta^*))\frac{f_o}{L} = \frac{1}{M}\mathbb{E}\left[\lambda_\theta\left(1 - \frac{1}{\mu_\theta}\right)\right], \tag{39}$$

and the selection condition

$$(1 - G(\theta^*))\frac{f_o}{L} = \frac{1}{M}\lambda_{\theta^*}\left(1 - \frac{1}{\mu_{\theta^*}}\right), \tag{40}$$

where the total population L and the mass of firms M can be normalized to 1. This completes what we need to calibrate the model.

In principle, one could alternatively use estimates of markups μ_θ or consumer surplus ratios δ_θ in conjunction with sales λ_θ to calibrate the model. We instead rely on pass-throughs, since estimating markups is notoriously difficult (typically requiring a fully-specified model of demand or production function estimation), and since estimating δ_θ would require experimental data tracing out individual demand curves. The downside is that calibrating the model using ρ_θ requires outside information to pin down boundary conditions $\bar{\mu}$ and $\bar{\delta}$.

6.2 Calibration Implementation

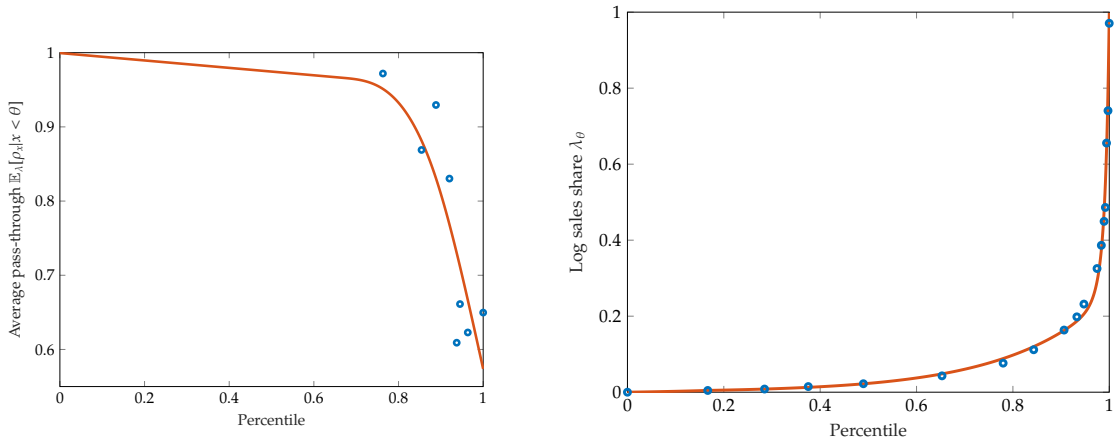
In this section, we implement the calibration procedure described above using estimates of the firm-level pass-throughs and the distribution of firms sales. We refer readers interested in a more detailed description of our data sources to Appendix A.

Data sources. To calibrate the model, we need data on pass-throughs ρ_θ, firm sales λ_θ, and the selection cut-off θ^*. For ρ_θ, we use estimates of pass-throughs by firm size for manufacturing firms in Belgium from Amiti et al. (2019). They use annual administrative firm-product level data (Prodcom) from 1995-2007, which contains information on prices and sales, collected by Statistics Belgium. Using exchange rate shocks as instruments for changes in marginal cost, they are able to control for the portion of price changes due to competitors' prices, and hence identify the partial equilibrium pass-through by firm size (under assumptions consistent with our model). Their estimates are shown in Figure 2a.[24]

Prodcom does not sample very small firms (firms must have sales greater than 1 million euros to be included). Therefore, we merge their estimates of the pass-through function ρ_θ (as a function of size) with the sales distribution λ_θ for the universe of Belgian manufacturing firms (from VAT declarations). The cumulative sales share distribution is shown in Figure 2b.

For firms that are smaller than the smallest firms in Prodcom, we interpolate their pass-through in such a way that the smallest firm has pass-through equal to one, since Amiti et al. (2019) find

[24] Appendix A in Amiti et al. (2019) provides evidence that differences in pass-through across small and large firms are not driven by confounders (e.g., exporters or multinationals).



(a) Average cumulated pass-through (b) Cumulative sales share distributions.

Figure 2: The blue dots are data showing the average sales-weighted pass-through and cumulated sales share for firms smaller than the percentile given by the x-axis. The solid red line is a fitted spline.

that the average pass-through for the smallest 75% of firms in Prodcom is already 0.97.[25]

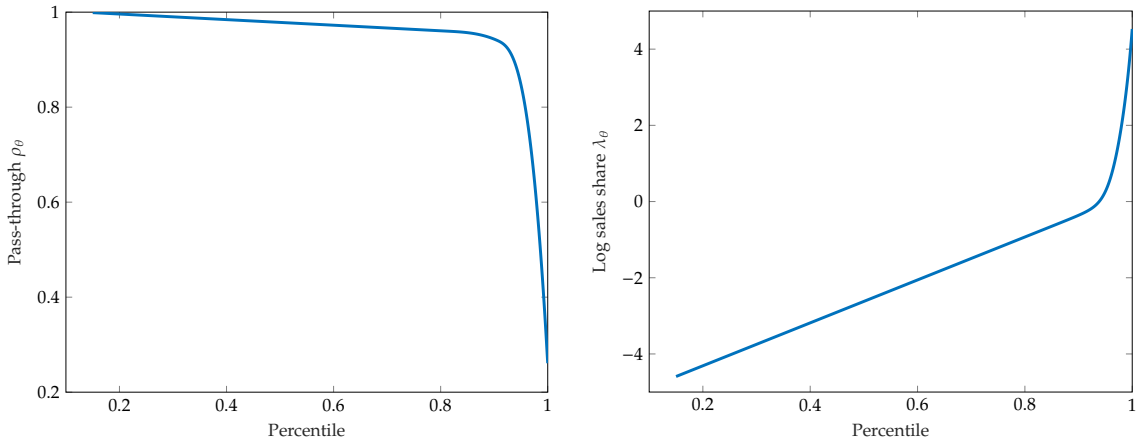
To calibrate θ^*, we fit a quasi-hyperbolic process to firm exit probability by age as reported by Pugsley et al. (2018). We find $\theta^* = 0.15$.

Boundary conditions. Our results require taking a stand on two boundary conditions: the average consumer surplus ratio $\bar{\delta}$ and the (harmonic) average markup $\bar{\mu}$. We focus on two benchmark calibrations of $\bar{\delta}$: (1) efficient entry $\bar{\delta} = \bar{\mu}$ (see Lemma 1), and (2) efficient selection $\bar{\delta} = \delta_{\theta^*}$ (see Lemma 2). We consider two different values for the average markup $\bar{\mu} = 1.045$ and $\bar{\mu} = 1.090$, which are chosen so that $d\log Y/d\log L \approx 0.13$ under the first assumption, and $d\log Y/d\log L \approx 0.30$ under the second assumption. An aggregate scale elasticity $d\log Y/d\log L \in [0.13, 0.3]$ is broadly in line with the literature.[26] In Appendix C, we vary both boundary conditions along a 2-dimensional grid and show that the two benchmark cases we focus on are representative of broader patterns.

Calibrated statistics. Figures 3a and 3b display pass-throughs ρ_θ and log sales $\log\lambda_\theta$ as a function of firm type θ. These are derived by differentiating the splines in Figures 2 (see Appendix A for more details). Figure 3a shows that pass-throughs decrease from 1 for the smallest firms to about 0.3 for the largest firms. Figure 3b shows that sales are initially increasing exponentially (linear in logs), but become super-exponential towards the end reflecting a high degree of

[25] In mapping the model to the data, we assume that products sold by the same firm are perfect substitutes, so each firm is a different variety. We could alternatively assume that each product is a separate variety. Appendix B provides results using this assumption. The computed elasticities are different, but the overall message does not change.

[26] For context, in a CES model, $d\log Y/d\log L = 0.13$ corresponds to setting an elasticity of substitution around 8 whilst $d\log Y/d\log L = 0.3$ corresponds to an elasticity of substitution around 4.



(a) Pass-through ρ_θ (b) Log sales share density $\log \lambda_\theta$

Figure 3: Pass-throughs and sales share density as a function of firm type θ.

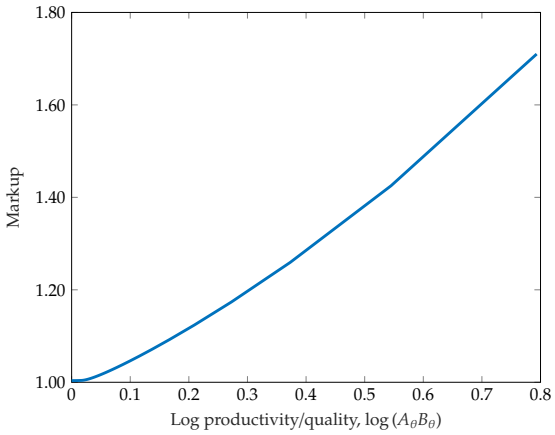
concentration in the tail.

The results from solving the differential equations are shown in Figure 4. Figure 4a shows that markups μ_θ are increasing and convex in log productivity/quality, $\log(A_\theta B_\theta)$. For brevity, we only show graphs of the estimates for $\bar{\mu} = 1.090$ but the patterns are similar for the other case (though of course, markups are lower when we set $\bar{\mu} = 1.045$). The net markup ranges from close to zero for the smallest firms to almost 80% for the very largest firms. Figure 4b shows the log productivity distribution. As with the sales density, the productivity density is also initially exponential, and becomes super exponential in the tail. Since price elasticities are decreasing in θ, productivity has to change by more than sales in the cross-section to allow firms to get large. Figures 4d and 4c show the consumer surplus ratio δ for the efficient-selection case ($\delta_{\theta^*} = \bar{\delta}$) and the efficient-entry case ($\bar{\mu} = \bar{\delta}$).

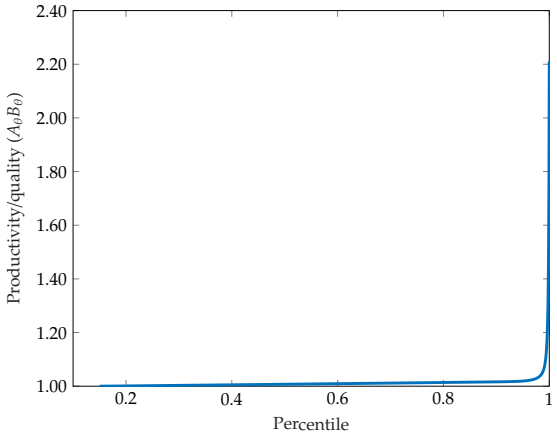
Finally, Figure 5 plots the inverse residual demand curve in linear and log-log terms. Figure 5a shows that our estimate has a distinctly non-isoelastic shape, indicating substantial departures from CES.

7 Quantitative Results

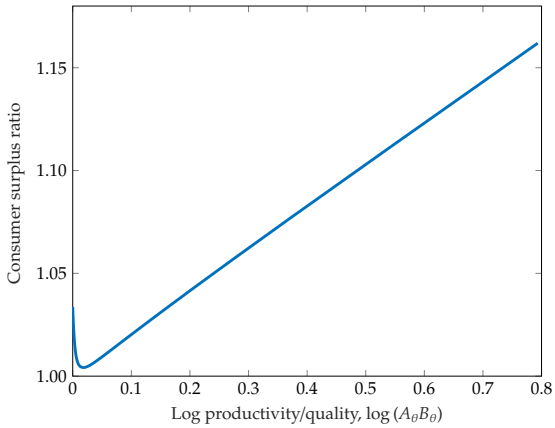
In this section, we compute how welfare changes in response to an increase in market size and in response to a tax on entry, using the calibrated Belgian data. For both exercises, we decompose welfare gains into technical and allocative efficiency, and further decompose allocative efficiency changes into the Darwinian, selection, and pro-competitive margins. As extensions, we compare macro returns to scale (at the aggregate level) to micro returns to scale, show that our local approximations provide a good guide to the nonlinear response of the model, and illustrate how increases in market size increase industrial concentration.



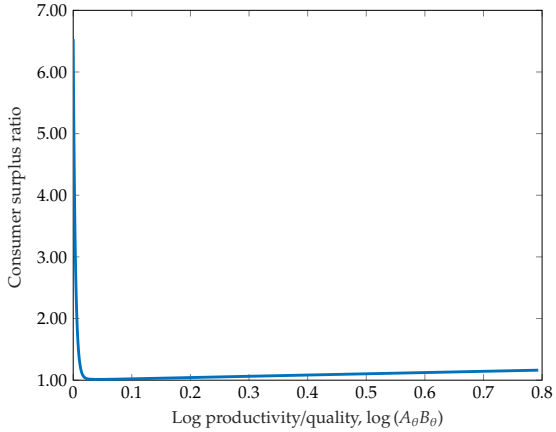
(a) Markup μ_θ



(b) Log productivity/quality $(A_\theta B_\theta)$

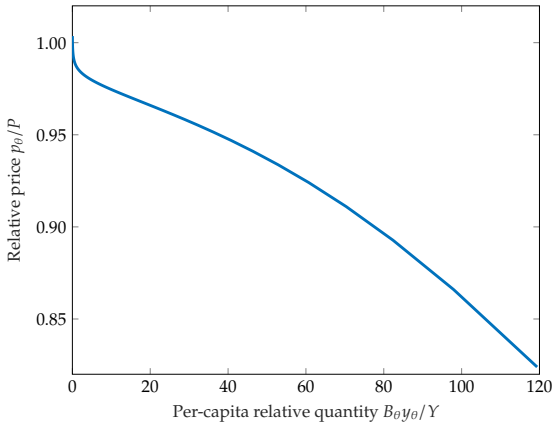


(c) Consumer surplus ratio δ_θ (efficient selection).

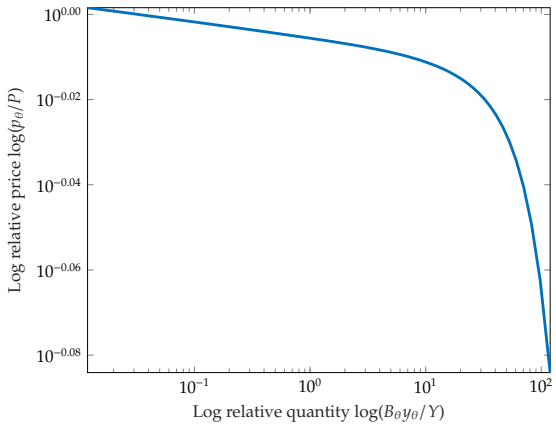


(d) Consumer surplus ratio δ_θ (efficient entry).

Figure 4: Markups and consumer surplus ratios with $\bar{\mu} = 1.090$.



(a) Inverse residual demand curve



(b) Log-log inverse residual demand curve

Figure 5: Residual demand curve (price against quality-adjusted quantity) for the efficient-selection case with $\bar{\mu} = 1.09$. The results for the efficient-entry case are similar.

Welfare effect of a market expansion. Table 1 reports the elasticity of consumer welfare to market size, following Theorem 1. The response of welfare is decomposed into changes due to technical efficiency and allocative efficiency,

$$\mathrm{d}\log Y = \mathrm{d}\log Y^{tech} + \mathrm{d}\log Y^{alloc}.$$

The table further decomposes the allocative effect by into the Darwinian, selection, and pro-competitive channels. We denote welfare under the Darwinian effect $\mathrm{d}\log Y^{\epsilon}$ only (holding fixed θ^* and markups μ_θ); welfare allowing the Darwinian and selection effect $\mathrm{d}\log Y^{\epsilon,\theta^*}$ (holding fixed markups μ_θ); and welfare when all three margins can adjust $\mathrm{d}\log Y^{\epsilon,\theta^*,\mu} = \mathrm{d}\log Y$.

	$\bar{\mu} = 1.045$		$\bar{\mu} = 1.090$	
	$\bar{\delta} = \delta_{\theta^*}$	$\bar{\delta} = \bar{\mu}$	$\bar{\delta} = \delta_{\theta^*}$	$\bar{\delta} = \bar{\mu}$
Welfare: $\mathrm{d}\log Y$	0.130	0.145	0.293	0.323
Technical efficiency: $\mathrm{d}\log Y^{tech}$	0.017	0.045	0.034	0.090
Allocative efficiency: $\mathrm{d}\log Y^{alloc}$	0.114	0.100	0.260	0.233
Darwinian effect: $\mathrm{d}\log Y^{\epsilon} - \mathrm{d}\log Y^{tech}$	0.117	0.408	0.272	1.396
Selection effect: $\mathrm{d}\log Y^{\epsilon,\theta^*} - \mathrm{d}\log Y^{\epsilon}$	0.000	-0.251	0.000	-1.006
Pro-competitive effect: $\mathrm{d}\log Y^{\epsilon,\theta^*,\mu} - \mathrm{d}\log Y^{\epsilon,\theta^*}$	-0.004	-0.057	-0.012	-0.157
Real GDP per capita	0.024	0.024	0.051	0.052

Table 1: The elasticity of welfare and real GDP per capita to population following Theorem 1 and Proposition 1.

When discussing the results, we focus on the case with $\bar{\mu} = 1.045$, but similar comments apply to the case where $\bar{\mu} = 1.090$. We start by discussing the case with efficient entry first ($\bar{\delta} = \delta_{\theta^*}$). By construction, the elasticity of consumer welfare to population is 0.13. Only around a tenth of the overall effect is due to the technical efficiency effect $\bar{\delta} - 1 = 0.017$. Changes in allocative efficiency 0.114 account for around nine-tenths of the overall effect. An increase in market size therefore brings about considerable improvements in allocative efficiency, and these improvements are about nine times larger than direct gains from technical efficiency.

The change in allocative efficiency from the Darwinian effect is large and positive at 0.117. The selection and pro-competitive effects are insignificant in comparison. The change in allocative efficiency from the selection effect is zero by construction, since the surplus associated with exiting varieties is equal to the average consumer surplus. Finally, the change in allocative efficiency from the pro-competitive effect is slightly negative at -0.004. This number includes the effects of an overall reduction in markups and entry, which is beneficial since entry is initially too high ($\bar{\mu} > \bar{\delta}$), and a reallocation effect between high-markup and low-markup firms that depends on the relative

pass-throughs and elasticities of firms. In principle, the overall effect is ambiguous in sign, and here we find that detrimental reallocation effects dominate the beneficial reduction in markups, leading to an overall reduction in welfare from the pro-competitive channel.[27]

The elasticity of real GDP per capita is much smaller than the elasticity of welfare to market size at 0.024. As discussed earlier, this difference is a consequence of the fact that the welfare benefits of new goods are not reflected in changes in real GDP.[28]

Next, consider the case with efficient entry. The elasticity of welfare with respect to population shocks is now slightly higher at 0.145. The technical efficiency effect is now 0.045, reflecting the fact that $\bar{\delta}$ is calibrated to equal $\bar{\mu} = 1.045$. The allocative efficiency effect is still much more important than the technical efficiency effect at 0.100.

The Darwinian effect is now much larger at 0.408. The main reason for the increase is because $\mathbb{E}_\lambda[\delta_\theta] - 1$ is now 0.045 instead of 0.017. This implies that entry is more valuable than it was before. Since the labor saved by the Darwinian effect is funneled into more entry, this makes the Darwinian effect more beneficial as well. The selection effect from the adjustment of the exit cut-off is now non-zero and negative at -0.27. The reason for this can be seen from inspecting Figure 4d, which shows that the consumer surplus ratio at the cut-off is much higher than average. Hence, as the cut-off increases in response to toughening competition, socially valuable small firms are forced to exit. Finally, the pro-competitive effect from the reduction in markups is still negative and larger in magnitude at -0.057. The reason the pro-competitive effect is now more negative is because entry was initially excessive in the efficient-selection case, so the overall reductions in markups had a beneficial effect on the entry efficiency. Since we are now imposing entry efficiency, this effect no longer operates, and the overall contribution of changing markups to welfare is more negative.

The response of real GDP per capita is basically unchanged at 0.024, since in both specifications, the average reduction in markups for existing firms is roughly the same.

How important can selection be? An important theme in the literature has been to emphasize the role of the selection margin (increases in the productivity/quality cut-off) as a driver of

[27]The effect of reductions in markups is complicated by cross-sectional misallocation. Since pass-throughs are below one for all firms, all firms cut their markups in response to entry. However, large firms cut their markups by more than small firms since their pass-through is lower. This pushes in the direction of reallocations towards large firms. However, the small firms have much more elastic demand curves, and this pushes in the direction of reallocations towards small firms. In our quantitative model, the fact that the pro-competitive effect is harmful implies that the elasticity effect dominates the pass-through effect. That is, the pro-competitive effect exacerbates cross-sectional misallocation because it reallocates resources from high-markup firms towards low-markup firms.

[28]The large gap between the welfare and real GDP effect should be interpreted with caution, because it is sensitive to a dimension of the problem, namely dynamics, which we have abstracted from. The reason is that real GDP, while it misses the consumer surplus created immediately upon entry of a new variety, captures all the post-entry productivity gains for this variety. Everything else equal, if new varieties enter small and grow larger over time by improving their productivity, as would be realistic if varieties were identified with firms, there would be less of a difference between welfare and real GDP. By contrast, if new varieties enter large, as would be realistic if varieties where products, then there would be bigger difference between welfare and real GDP.

productivity and welfare gains. However, in our baseline results, the selection margin is either neutral (when $\delta_{\theta^*} = \bar{\delta}$) or is deleterious (when $\bar{\delta} = \bar{\mu}$). One may wonder how robust this finding is and how it depends on our choice of boundary conditions.

To answer this question, we consider a third possibility for the initial conditions. We try setting $\delta_{\theta^*} = 1$, which implies that the residual demand curve for infra-marginal firms is perfectly horizontal. In other words, the marginal firms produce no excess consumer surplus for the household. This maximizes the importance of the selection margin for welfare, conditional on our choice of $\bar{\mu}$. The results, however, are quantitatively very similar to those in Table 1.

Specifically, when $\bar{\mu} = 1.045$, we find the overall effect on welfare is still 0.130 with a technical efficiency effect of 0.016 and an allocative efficiency effect of 0.114. The Darwinian effect still accounts for the bulk of changes in allocative efficiency at 0.116. The contribution of the selection effect to welfare is now positive, but still close to insignificant at 0.001. The pro-competitive effect is still negative and of similar magnitude to before at -0.003. Similarly, when $\bar{\mu} = 1.09$, the welfare effect is 0.293 with a technical efficiency effect of 0.033 and an allocative efficiency effect of 0.260. Once again, the overwhelming force is the Darwinian effect at 0.269, with a negligible contribution from the selection effect (0.003) and a small, negative pro-competitive effect (-0.012).

These results suggest that the small role played by the selection margin is not an anomaly resulting from our choice of initial conditions. For robustness to our choice of $\bar{\mu}$, see the robustness exercise in Appendix C.

How important is heterogeneity? To emphasize the interaction of heterogeneity and inefficiency, we compare our model to a model with homogeneous firms, calibrated to have a pass-through equal to the average (sales-weighted) pass-through and a markup equal to the harmonic average. Table 2 shows the results.

	$\bar{\mu} = 1.045$		$\bar{\mu} = 1.090$	
	$\bar{\delta} = \delta_{\theta^*}$	$\bar{\delta} = \bar{\mu}$	$\bar{\delta} = \delta_{\theta^*}$	$\bar{\delta} = \bar{\mu}$
Welfare: $\mathrm{d}\log Y$	0.030	0.045	0.060	0.090
Technical efficiency: $\mathrm{d}\log Y^{tech}$	0.017	0.045	0.034	0.090
Allocative efficiency: $\mathrm{d}\log Y^{alloc}$	0.013	0.000	0.026	0.000
Real GDP per capita	0.021	0.022	0.042	0.043

Table 2: The elasticity of welfare and real GDP per capita to market size in an economy with homogeneous firms.

The most striking difference is that the elasticity of welfare to market size is much smaller, because changes in allocative efficiency are an order of magnitude smaller due to the absence of the Darwinian effect. In a model with homogeneous firms (see Section 4.2.3), the sole source of inefficiency comes from excessive or insufficient entry. Thus, when entry is assumed to be efficient

(the second and fourth columns), there are no changes in allocative efficiency at all. Even when entry is not efficient, the changes in allocative efficiency are fairly small. Those beneficial changes in allocative efficiency are solely due to pro-competitive effects.

Are there larger increasing returns at the macro vs. micro levels? The micro return to scale for a surviving type θ is the ratio of average cost to marginal cost minus one $ac_\theta/mc_\theta - 1$, so that 0 corresponds to constant returns to scale. The average cost is $ac_\theta = [f_e/(1-G(\theta^*))+f_o+Ly_\theta/A_\theta]/(Ly_\theta)$. The marginal cost is $mc_\theta = 1/A_\theta$. The harmonic average across surviving producers of the micro return to scale is equal to $1/\mathbb{E}[1/(ac_\theta/mc_\theta - 1)] = \bar{\mu} - 1$.[29]

Hence average micro technological increasing returns to scale are 0.045 when $\bar{\mu} = 1.045$ and 0.090 when $\bar{\mu} = 1.090$. Increasing returns at the aggregate level are much larger: between 0.130 and 0.145 in the former case and between 0.293 and 0.323 in the latter case. This means that even small technological increasing returns at the micro level can give rise to large increasing returns to scale at the aggregate level. Once again, the interaction of inefficiency and heterogeneity is key. If the economy were efficient, macro and micro returns would be identical. In an economy with homogeneous firms, the difference between macro and micro returns is much smaller.

Nonlinear response. One might worry that the reallocative effects in Table 1 could peter out quickly if we kept increasing the size of the market. Since the model is calibrated globally, we can solve the model for large shocks and thereby analyze potential nonlinearities.[30] The results in Table 3 and Figure 6 below show that the forces identified for small shocks by Proposition 1 continue to apply for large shocks.

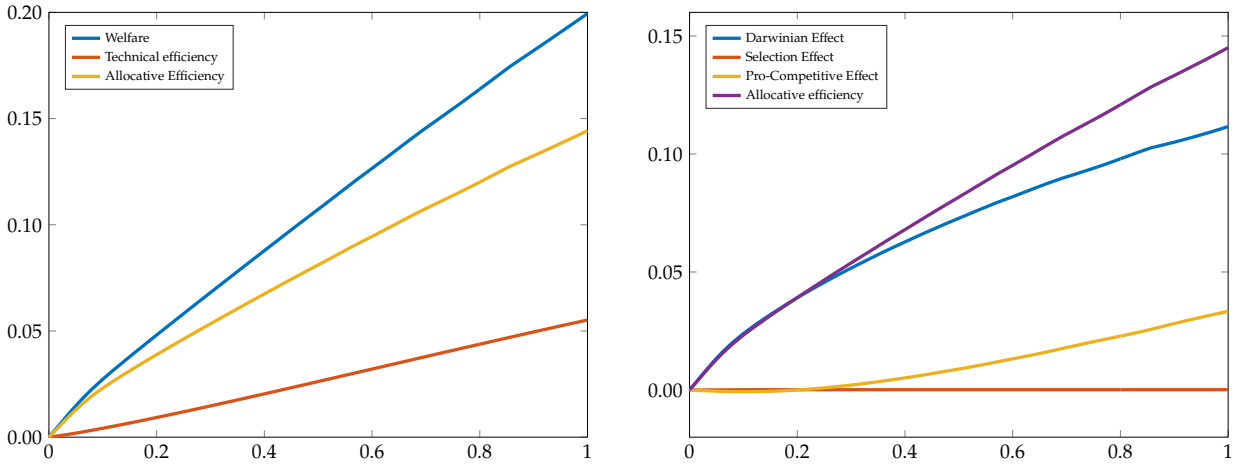
	$\bar{\mu} = 1.045$		$\bar{\mu} = 1.090$	
	$\bar{\delta} = \delta_{\theta^*}$	$\bar{\delta} = \bar{\mu}$	$\bar{\delta} = \delta_{\theta^*}$	$\bar{\delta} = \bar{\mu}$
Welfare: $\Delta \log Y$	0.100	0.099	0.215	0.216
Technical efficiency: $\Delta \log Y^{tech}$	0.025	0.048	0.052	0.098
Allocative efficiency: $\Delta \log Y^{alloc}$	0.075	0.051	0.162	0.117
Darwinian effect: $\Delta \log Y^{\epsilon} - \Delta \log Y^{tech}$	0.066	0.107	0.145	0.272
Selection effect: $\Delta \log Y^{\epsilon,\theta^*} - \Delta \log Y^{\epsilon}$	0.000	-0.065	0.000	-0.176
Pro-competitive effect: $\Delta \log Y^{\epsilon,\theta^*,\mu} - \Delta \log Y^{\epsilon,\theta^*}$	0.008	0.008	0.017	0.021
Real GDP per capita	0.025	0.024	0.054	0.051

Table 3: The average elasticity of welfare and real GDP per capita to population for a large shock $\Delta \log L = 0.5$.

[29] From $ac_\theta = [(l_e + l_o)/(1 - G(\theta^*)) + l_\theta]/(A_\theta l_\theta)$ and $mc_\theta = 1/A_\theta$, we have $ac_\theta/mc_\theta - 1 = [(l_e + l_o)/(1 - G(\theta^*))]/l_\theta$ and hence $1/(ac_\theta/mc_\theta - 1) = (1 - G(\theta^*)l_\theta/(l_e + l_o)$. The result follows since $(1 - G(\theta^*))l_\theta = \lambda_\theta/\mu_\theta$ and $l_e + l_o = 1 - 1/\bar{\mu}$.

[30] We do this by numerically solving the system of ordinary differential equations in Appendix D.

Table 3 reports the average (rather than the marginal) elasticity of welfare to a 0.5 log point increase in population (a roughly 68% increase). The magnitude of and the decomposition of the average effects are similar to those for the marginal effects reported in Table 1. Although the model is far from being log-linear, the qualitative conclusions are unchanged.



(a) Welfare: technical and allocative efficiency as functions of $\Delta \log L$.

(b) Allocative efficiency: adjustments of the different margins as functions of $\Delta \log L$.

Figure 6: Decomposition of changes in welfare and allocative efficiency following Proposition 1, obtained by separately computing each term in the decomposition and integrating (cumulating) the changes. The model is calibrated to have efficient selection and $\bar{\mu} = 1.09$ at the initial point.

Figure 6 shows cumulated changes in welfare and each channel for the calibration with efficient selection $\bar{\delta} = \delta_{\theta^*}$ and $\bar{\mu} = 1.09$ (column 3). The first panel shows that even though their relative importance decreases slightly with the size of the shock, changes in allocative efficiency continue to dwarf changes in technical efficiency even for large shocks. The second panel shows that as the population grows, changes in allocative efficiency due to the pro-competitive channel start to account for a non-trivial part of overall changes in allocative efficiency. This happens because as we increase population, the harmonic average of markups increases due to the Darwinian effect. This means that entry becomes more excessive, and hence that reallocations triggered by individual markup reductions improve allocative efficiency more.

Implications for industrial concentration. Figure 7 shows the Lorenz curve for the distribution of sales as the market size increases. This graph shows the proportion of sales accounted for by firms up to a given centile of the size distribution. The figure shows that concentration rises as the market expands. This is primarily due to the Darwinian effect, which causes large, high-markup firms to expand relative to small, low-markup firms. Hence, in our quantitative application, increases in market size, from say globalization, raise welfare at the aggregate level via reallocations that also increase industrial concentration. Baqaee and Farhi (2019) find that

allocative efficiency in the U.S. economy improved from 1997-2015 due to a reallocation of market share of high-markup firms; we speculate that Darwinian reallocations from globalization may have contributed to this trend.

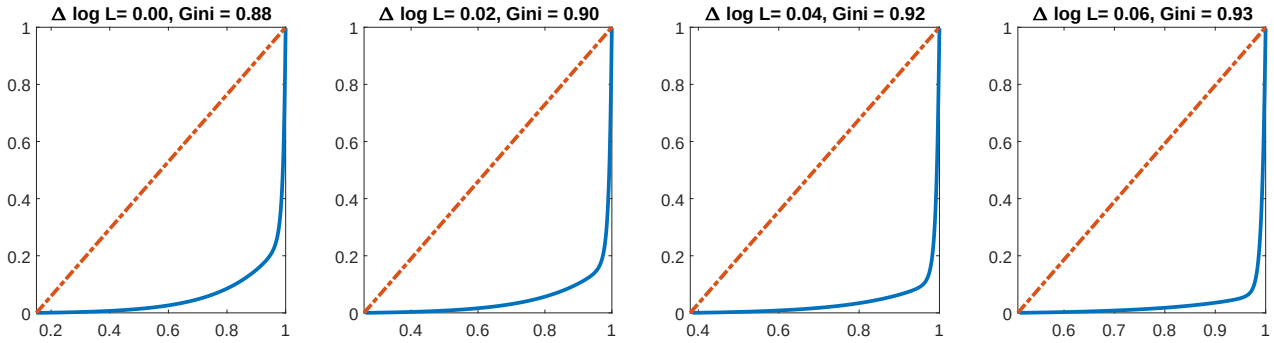


Figure 7: Each panel depicts the Lorenz curve for the sales distribution for different values of the market size parameter L. The dotted red line indicates the line of perfect equality (i.e. homogenous firms). The Gini coefficient, which is a measure of inequality, is also reported.

Welfare effect on an entry tax. Table 4 shows the effect of an entry tax on welfare, following Proposition 2. Note that the technology available to the economy is fixed, so all changes in welfare arise from changes in allocative efficiency. We again decompose the welfare change into the Darwinian, selection, and pro-competitive effects, where $\mathrm{d}\log Y^{\epsilon}$ holds fixed θ^* and markups μ_θ, $\mathrm{d}\log Y^{\epsilon,\theta^*}$ holds fixed only markups μ_θ, and $\mathrm{d}\log Y^{\epsilon,\theta^*,\mu} = \mathrm{d}\log Y$ allows all three margins to adjust. The last row of the table re-computes the welfare effect of an entry tax in a model with homogeneous firms calibrated to have a pass-through equal to the average sales-weighted pass-through and a markup equal to the harmonic average.

	$\bar{\mu} = 1.045$		$\bar{\mu} = 1.090$	
	$\bar{\delta} = \delta_{\theta^*}$	$\bar{\delta} = \bar{\mu}$	$\bar{\delta} = \delta_{\theta^*}$	$\bar{\delta} = \bar{\mu}$
Welfare: $\mathrm{d}\log Y$	-0.082	-0.089	-0.187	-0.198
Darwinian effect: $\mathrm{d}\log Y^{\epsilon} - \mathrm{d}\log Y^{tech}$	-0.085	-0.391	-0.199	-1.283
Selection effect: $\mathrm{d}\log Y^{\epsilon,\theta^*} - \mathrm{d}\log Y^{\epsilon}$	0.000	0.248	0.000	0.943
Pro-competitive effect: $\mathrm{d}\log Y^{\epsilon,\theta^*,\mu} - \mathrm{d}\log Y^{\epsilon,\theta^*}$	0.003	0.055	0.012	0.142
Welfare with homogeneous firms: $\mathrm{d}\log Y^{homog}$	0.014	0.000	0.028	0.000

Table 4: Welfare effect of an entry tax, following Proposition 2.

For all four choices of the boundary conditions, we find that the entry tax is welfare-reducing. Since the tax reduces entry, the Darwinian effect operates in reverse, loosening competition,

reallocating to low-markup firms, and thus exacerbating existing relative production inefficiencies. The reduction in entry has beneficial pro-competitive effects and selection effects (when selection is inefficient), but the losses due to Darwinian reallocations outweigh these benefits. In contrast, when firm heterogeneity is excluded from the model, the entry tax is beneficial or has no effect (when entry is efficient).

These results suggest that a social planner can increase welfare by enacting an entry subsidy. Notably, the Darwinian effects that constitute the entire gains from an entry subsidy are absent in a model with homogeneous firms. Thus, ignoring firm heterogeneity would lead us to recommend a tax (rather than a subsidy) on firm entry.

8 Extensions

Before concluding, we describe some extensions of the basic framework.

Multi-sector economies. Although our analysis uses a single sector model, embedding this structure into a larger multi-sector structure is relatively straightforward since preferences are homothetic. For example, suppose that consumers have Cobb-Douglas preferences over sectors

$$U = \prod_{\mathcal{I}} Y_{\mathcal{I}}^{\beta_{\mathcal{I}}},$$

where $\mathcal{I}$ indexes different sectors and each sector's output is implicitly pinned down by

$$\int_{\Theta_{\mathcal{I}}} \Upsilon_\theta(\frac{y_{\theta,\mathcal{I}}}{Y_{\mathcal{I}}}) dF_{\mathcal{I}}(\theta) = 1.$$

In this case, it is straightforward to show that changes in welfare are simply a weighted-average of changes in sectoral output

$$\frac{d\log U}{d\log L} = \sum_{\mathcal{I}} \beta_{\mathcal{I}} \frac{d\log Y_{\mathcal{I}}}{d\log L},$$

where the change in sectoral output $d\log Y_{\mathcal{I}}/d\log L$ is given by a sector-specific version of Theorem 1.[31]

Optimal policy and distance to the efficient frontier. In the main text of the paper, we have focused exclusively on comparative statics of the second-best equilibrium. For completeness, in Appendix F.1, we provide an analytical characterization of optimal policy. We also provide an analytical second-order approximation of the distance to the efficient frontier which decomposes the contributions of the different margins of inefficiency (entry, selection, and relative production)

[31] The same logic can also be extended to more complicated sectoral models since, due to homotheticity, we can still break the problem into two blocks: within and across sectors.

to the overall amount of misallocation. In Appendix F.2, we compute the distance to the efficient frontier in our calibrated model. There, we quantify the extent of misallocation in the decentralized equilibrium compared to the first-best allocation. We find the number to be somewhere between 2.5% and 6.8% in Belgium depending on the boundary condition. Therefore, there can be large changes in allocative efficiency even though the decentralized economy is not too far from efficiency. This appendix also helps cement that idea that when we increase the size of the market, the frontier also moves. Therefore, changes in allocative efficiency due to reallocation are fundamentally different to changes in the distance from the frontier. Reallocations can boost welfare on the margin, even as the distance with the efficient frontier widens.

Other demand systems. In the main text, we focus on generalized Kimball preferences. This is a class of preferences highlighted by Matsuyama and Ushchev (2017) as being both flexible and tractable. In Appendix H, we show that our theoretical results, our calibration strategy, and quantitative application are very similar under the other alternatives Matsuyama and Ushchev (2017) point out.

Other shocks. In the main text of the paper, we have focused exclusively on shocks to population. In Appendix G, we provide comparative statics with respect to other parameters, like productivity or fixed costs.

9 Conclusion

In this paper, we analyze the origins of aggregate increasing returns to scale. We decompose the overall effect of a market expansion into changes in technical and allocative efficiency and quantify our model using a non-parametric calibration exercise.

We find that changes in allocative efficiency, due to the reallocation of resources, are a more important source of welfare gains from increases in scale than changes in technical efficiency. Quantitatively, the most important reallocation is a composition effect that shifts resources from firms with low markups towards those high markups, which we call the Darwinian effect. This effect is distinct from changes in the marginal profitability cut-off and changes in markups. In fact, increases in the cut-off and reductions in markups play only minor roles in comparison. Furthermore, we find that a planner may want to subsidize entry, even if entry is excessive compared to the first best, to take advantage of Darwinian reallocations.

References

Aghion, Philippe, Antonin Bergeaud, Timo Boppart, Peter J. Klenow, and Huiyu Li, "Missing growth from creative destruction," *American Economic Review*, 2019, *109* (8), 2795–2822.

_ , _ , _ , _ , **and** _ , "A Theory of Falling Growth and Rising Rents," Technical Report 26448, National Bureau of Economic Research 2019.

Amiti, Mary, Oleg Itskhoki, and Jozef Konings, "International Shocks, Variable Markups, and Domestic Prices," *The Review of Economic Studies*, 2019, *86* (6), 2356–2402.

Arkolakis, Costas, Arnaud Costinot, Dave Donaldson, and Andrés Rodríguez-Clare, "The elusive pro-competitive effects of trade," *The Review of Economic Studies*, 2019, *86* (1), 46–80.

Asplund, Marcus and Volker Nocke, "Firm turnover in imperfectly competitive markets," *The Review of Economic Studies*, 2006, *73* (2), 295–327.

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," *The Quarterly journal of economics*, 2020, *135* (2), 645–709.

Baqaee, David Rezza and Emmanuel Farhi, "Productivity and Misallocation in General Equilibrium.," Technical Report, National Bureau of Economic Research 2019.

Basu, Susanto and John G. Fernald, "Returns to scale in US Production: Estimates and Implications," *Journal of Political Economy*, 1997, *105* (2), 249–283.

Bilbiie, Florin O, Fabio Ghironi, and Marc J Melitz, "Endogenous entry, product variety, and business cycles," *Journal of Political Economy*, 2012, *120* (2), 304–345.

_ , _ , **and** _ , "Monopoly power and endogenous product variety: Distortions and remedies," *American Economic Journal: Macroeconomics*, 2019, *11* (4), 140–74.

Corcos, Gregory, Massimo Del Gatto, Giordano Mion, and Gianmarco I. Ottaviano, "Productivity and Firm Selection: Quantifying the "New" Gains from Trade," *The Economic Journal*, 2012, *122* (561), 754–798.

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

Dixit, Avinash K and Joseph E Stiglitz, "Monopolistic competition and optimum product diversity," *The American economic review*, 1977, *67* (3), 297–308.

Edmond, Chris, Virgiliu Midrigan, and Daniel Yi Xu, "How costly are markups?," Technical Report, National Bureau of Economic Research 2018.

Epifani, Paolo and Gino Gancia, "Trade, markup heterogeneity and misallocations," *Journal of International Economics*, 2011, *83* (1), 1–13.

Feenstra, Robert C., "Restoring the product variety and pro-competitive gains from trade with heterogeneous firms and bounded productivity," *Journal of International Economics*, 2018, *110*, 16–27.

_ **and David E. Weinstein**, "Globalization, Markups, and US Welfare," *Journal of Political Economy*, 2017, *125* (4), 1040–1074.

Helpman, Elhanan and Paul R Krugman, *Market structure and foreign trade: increasing returns, imperfect competition, and the international economy*, MIT Press, 1985.

Hulten, Charles R, "Growth Accounting with Intermediate Inputs," *The Review of Economic Studies*,

1978, pp. 511–518.

Jaimovich, Nir and Max Floetotto, "Firm dynamics, markup variations, and the business cycle," *Journal of monetary Economics*, 2008, *55* (7), 1238–1252.

Kimball, Miles, "The Quantitative Analytics of the Basic Neomonetarist Model," *Journal of Money, Credit and Banking*, 1995, *27* (4), 1241–77.

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

Krugman, Paul R, "Increasing returns, monopolistic competition, and international trade," *Journal of international Economics*, 1979, *9* (4), 469–479.

Lipsey, Richard G. and Kelvin Lancaster, "The general theory of second best," *The Review of Economic Studies*, 1956, *24* (1), 11–32.

Loecker, Jan De, Jan Eeckhout, and Gabriel Unger, "The rise of market power and the macroeconomic implications," *The Quarterly journal of economics*, 2020, *135* (2), 561–644.

_ , Pinelopi K. Goldberg, Amit K. Khandelwal, and Nina Pavcnik, "Prices, markups, and trade reform," *Econometrica*, 2016, *84* (2), 445–510.

Mankiw, N. Gregory and Michael D. Whinston, "Free Entry and Social Inefficiency," *RAND Journal of Economics*, Spring 1986, *17* (1), 48–58.

Matsuyama, Kiminori and Philip Ushchev, "Beyond CES: Three Alternative Classes of Flexible Homothetic Demand Systems," 2017. Working paper.

_ and _ , "Constant Pass-Through," 2020.

_ and _ , "When Does Procompetitive Entry Imply Excessive Entry?," 2020.

Mayer, Thierry, Marc J. Melitz, and Gianmarco I. Ottaviano, "Market size, competition, and the product mix of exporters," *American Economic Review*, 2014, *104* (2), 495–536.

Melitz, Marc J., "The Impact of Trade on Intra-Industry Reallocations and Aggregate Industry Productivity," *Econometrica*, November 2003, *71* (6), 1695–1725.

Melitz, Marc J, "Competitive effects of trade: theory and measurement," *Review of World Economics*, 2018, *154* (1), 1–13.

_ and Gianmarco IP Ottaviano, "Market size, trade, and productivity," *The review of economic studies*, 2008, *75* (1), 295–316.

Melitz, Marc J. and Stephen J. Redding, "New trade models, new welfare implications," *American Economic Review*, 2015, *105* (3), 1105–46.

Mrázová, Monika and J Peter Neary, "Not so demanding: Demand structure and firm behavior," *American Economic Review*, 2017, *107* (12), 3835–74.

_ and _ , "IO For Export(s)," 2019.

Olley, G Steven and Ariel Pakes, "The Dynamics of Productivity in the Telecommunications Equipment Industry," *Econometrica*, 1996, *64* (6), 1263–1297.

Pavcnik, Nina, "Trade liberalization, exit, and productivity improvements: Evidence from Chilean plants," *The Review of Economic Studies*, 2002, *69* (1), 245–276.

Pugsley, Benjamin W, Petr Sedlacek, and Vincent Sterk, "The nature of firm growth," 2018.

Spence, Michael, "Product selection, fixed costs, and monopolistic competition," *The Review of economic studies*, 1976, *43* (2), 217–235.

Trefler, Daniel, "The long and short of the Canada-US Free Trade Agreement," *American Economic Review*, 2004, *94* (4), 870–895.

Venables, Anthony J, "Trade and trade policy with imperfect competition: The case of identical products and free entry," *Journal of International Economics*, 1985, *19* (1-2), 1–19.

Vives, Xavier, *Oligopoly pricing: old ideas and new tools*, MIT press, 1999.

Zhelobodko, Evgeny, Sergey Kokovin, Mathieu Parenti, and Jacques-François Thisse, "Monopolistic competition: Beyond the constant elasticity of substitution," *Econometrica*, 2012, *80* (6), 2765–2784.

Online Appendix

Appendix A Details of Empirical Implementation

Amiti et al. (2019) provide estimates of the average sales-weighted pass-through (denoted by α) for Belgian manufacturing firms conditional on the firms being smaller than a certain size as measured by their numbers of employees. These estimates are based on information from Prodcom, which is a subsample of Belgian manufacturing firms. Inclusion in Prodcom requires that firms have turn-overs above 1 million euros, which means that the sample is not representative of all manufacturers. The estimates are in Table 5.

No of employees	Share of observations	Share of employment	Share of sales	α
100	0.76313963	0.14761668	0.23096292	0.9719
200	0.85435725	0.22086396	0.3389753	0.8689
300	0.88848094	0.28832632	0.4083223	0.9295
400	0.92032149	0.33549505	0.48074553	0.8303
500	0.93746047	0.38345889	0.54008827	0.6091
600	0.94523549	0.41987701	0.58209142	0.6612
1000	0.96365488	0.52280162	0.66820585	0.6229
8000	0.99996915	0.99999999	0.99999174	0.6497

Table 5: Estimates from Amiti et al. (2019).

Our objective is to infer the pass-through ρ as a function of firm size. With some abuse of notation, let $\theta \in [0,1]$ be the fraction of observations in Prodcom up to some sales value. Let $\lambda(\theta)$ be the sales share density of Prodcom firms of type θ. Then the variable "Share of sales" is defined as

$$\Lambda(\theta) = \int_0^\theta \lambda(x)dx.$$

We fit a smooth curve to $\Lambda(\theta)$, then the pdf of sales shares $\lambda(\theta)$ is given by

$$\lambda(\theta) = \frac{d\Lambda}{d\theta}.$$

The curve we fit has the form $\exp(c_0 + c_1\theta + c_2\theta^{c_3})$, where c_0, c_1, c_2, c_3 are chosen to minimize the mean squared error.

Next, the variable $\alpha(\theta)$ satisfies

$$\alpha(\theta) = \frac{\int_0^\theta \lambda(x)\rho(x)dx}{\int_0^\theta \lambda(x)dx},$$

$$= \frac{\int_0^\theta \lambda(x)\rho(x)dx}{\Lambda(\theta)},$$

where $\lambda(x)$ is the sales-share of firms of type x. Next we fit a flexible spline function to $\alpha(\theta)$. The fitted curve is shown in Figure 2a.

To recover the pass-throughs $\rho(\theta)$, we write

$$\frac{d\alpha}{d\theta} = \frac{\lambda(\theta)\rho(\theta)}{\int_0^\theta \lambda(x)dx} - \frac{\lambda(\theta)}{\int_0^\theta \lambda(x)dx}\alpha(\theta).$$

In other words, we can recover the pass-through function via

$$\begin{aligned}\rho(\theta) &= \frac{\left(\int_0^\theta \lambda(x)dx\right)}{\lambda(\theta)}\frac{d\alpha}{d\theta} + \alpha(\theta), \\ &= \frac{\Lambda(\theta)}{\lambda(\theta)}\frac{d\alpha}{d\theta} + \alpha(\theta).\end{aligned}$$

This gives us pass-throughs as a function of the number of employees.

Next, we use information from VAT declaration in Belgium for the year 2014 to recover the sales distribution of Belgian manufacturers (overcoming the sample selection issues in Prodcom). Table 6 displays the underlying data.

Number of employees	Share of sales	Share of Observations
1	0.004559	0.16668
2	0.00826	0.284539
3	0.014786	0.375336
5	0.022269	0.489659
10	0.043011	0.652879
20	0.076444	0.779734
30	0.111713	0.843161
50	0.163492	0.906204
75	0.198242	0.932729
100	0.231815	0.947413
200	0.325376	0.974629
300	0.386449	0.983547
400	0.449491	0.989237
500	0.486108	0.991927
600	0.655522	0.994311
1000	0.740656	0.997386
8000	0.970654	0.999923

Table 6: Firm size distribution for manufacturing firms from VAT declarations in Belgium for 2014.

As before, we let $\theta \in [0,1]$ index the fraction of observations up to some size. Then the variable "Share of sales" is defined as

$$\Lambda(\theta) = \int_0^\theta \lambda(x)dx,$$

where (abusing notation) λ is the sales share density of all manufacturing firms (rather than just the ones in Prodcom). We fit a smooth curve to $\Lambda(\theta)$, then the pdf of sales shares $\lambda(\theta)$ is given by

$$\lambda(\theta) = \frac{d\Lambda}{d\theta}.$$

The curve we fit has the form $\exp(c_0 + c_1\theta + c_2\theta^{c_3})$, displayed in Figure 2b. Finally, we merge our pass-through information from Prodcom with the sales density from VAT declarations by assuming that the pass-through ρ of a firm with a given number of employees in Prodcom is the same as it is in the bigger dataset. We then fit a smooth spline to this pass-through data from $[0,1]$ assuming that the pass-through for the smallest firm is 1 and declines monotonically from the smallest firm to the first observation (which is a pass-through of 0.97 for firms with 100 employees). Given a smooth curve for both λ_θ and ρ_θ we follow the procedure outlined in Section 6.1, solving the differential equations numerically using the Runge-Kutta algorithm on a large grid.

Appendix B Product-Level Data

In the body of the paper, we assume that different products produced by a single firm are perfect substitutes from the perspective of the consumer, and so we use overall sales of a firm as the sales of each variety. An alternative approach is to instead to treat each product as a single variety instead. In Table 7 we display the average number of products each firm in Prodcom sells, for each firm-size bin.

To map each product to a variety, we take the sales density for firms and divide the density for firms of a given size by the average number of products (renormalizing the density so that it still integrates to one). Mapping the model to the data in this way results in less dispersion in sales, a left tail which is slightly less thick, and as a result, less dispersed estimates of productivities and markups. The comparative statics for this version of the model are in Table 8. The basic qualitative message of our previous results in Table 1 is unchanged, and the Darwinian effects are still overwhelmingly the dominant force in the model.

No of Employees	No of Products	No of firms
5	1.3636364	22
10	2.0550459	109
20	2.200495	404
30	2.4203297	728
50	2.4203895	873
75	2.3727506	389
100	3.294686	207
200	3.225	400
300	3.3308824	136
400	3.6511628	86
500	5.2162162	37
600	4.1724138	29
1000	8.3095238	42
8000	8.8780488	41

Table 7: Number of products on average from Prodcom sample in 2014.

	$\bar{\mu} = 1.045$		$\bar{\mu} = 1.090$	
	$\bar{\delta} = \delta_{\theta^*}$	$\bar{\delta} = \bar{\mu}$	$\bar{\delta} = \delta_{\theta^*}$	$\bar{\delta} = \bar{\mu}$
Welfare: $d \log Y$	0.080	0.133	0.176	0.294
Technical efficiency: $d \log Y^{tech}$	0.020	0.045	0.042	0.090
Allocative efficiency: $d \log Y^{alloc}$	0.060	0.088	0.134	0.204
Adj. of Entry: $d \log Y^{\epsilon} - d \log Y^{tech}$	0.056	0.136	0.126	0.327
Adj. of Exit: $d \log Y^{\epsilon,\theta^*} - d \log Y^{\epsilon}$	0.000	-0.037	0.000	-0.094
Adj. of Markups: $d \log Y^{\epsilon,\theta^*,\mu} - d \log Y^{\epsilon,\theta^*}$	0.004	-0.012	0.008	-0.029
Real GDP per capita	0.015	0.016	0.032	0.035

Table 8: The elasticity of welfare and real GDP per capita to population following Propositions 1 and 1 for heterogeneous firms case using product-level data.

Appendix C Robustness to Boundary Conditions

Table 9 provides the elasticity of welfare and changes in allocative efficiency, following Proposition 1 for different boundary conditions. Although the magnitude of $d \log Y / d \log L$ changes as we change the boundary conditions, the contribution of allocative efficiency to the overall total is at least 50% of the overall effect. Table 10 breaks down the overall effect on allocative efficiency into the different margins of adjustment (Darwinian, selection, and pro-competitive). The Darwinian effect is always responsible for the bulk of the positive effect. As mentioned, for a given $\bar{\mu}$, the selection effect is strongest when δ_{θ^*} is lowest, but even for $\delta_{\theta^*} = 1$, the selection effect is negligible.

Table 9: Change in log welfare and allocative efficiency for different boundary conditions

		δ_{θ^*}									
		1	2	3	4	5	6	7	8	9	10
$\bar{\mu}$	1.05	[0.144,0.126]	[0.150,0.122]	[0.155,0.117]	[0.161,0.112]	[0.166,0.107]	[0.171,0.102]	[0.177,0.097]	[0.182,0.092]	[0.187,0.086]	[0.192,0.082]
	1.06	[0.180,0.158]	[0.186,0.153]	[0.191,0.148]	[0.196,0.144]	[0.202,0.139]	[0.207,0.134]	[0.212,0.128]	[0.218,0.123]	[0.223,0.118]	[0.228,0.113]
	1.07	[0.213,0.187]	[0.218,0.183]	[0.224,0.178]	[0.229,0.173]	[0.235,0.168]	[0.240,0.163]	[0.245,0.158]	[0.251,0.153]	[0.256,0.148]	[0.261,0.143]
	1.08	[0.255,0.225]	[0.260,0.220]	[0.266,0.215]	[0.271,0.211]	[0.276,0.206]	[0.282,0.201]	[0.287,0.196]	[0.292,0.190]	[0.297,0.185]	[0.302,0.180]
	1.09	[0.293,0.260]	[0.299,0.255]	[0.304,0.251]	[0.310,0.246]	[0.315,0.241]	[0.321,0.236]	[0.326,0.231]	[0.331,0.226]	[0.336,0.220]	[0.341,0.215]
	1.10	[0.336,0.299]	[0.341,0.294]	[0.347,0.289]	[0.352,0.284]	[0.357,0.279]	[0.363,0.274]	[0.368,0.269]	[0.373,0.264]	[0.378,0.259]	[0.383,0.254]
	1.11	[0.382,0.341]	[0.387,0.336]	[0.393,0.331]	[0.398,0.326]	[0.403,0.321]	[0.409,0.316]	[0.414,0.311]	[0.419,0.306]	[0.424,0.301]	[0.429,0.296]
	1.12	[0.433,0.388]	[0.438,0.383]	[0.443,0.378]	[0.449,0.373]	[0.454,0.368]	[0.459,0.363]	[0.464,0.357]	[0.469,0.352]	[0.474,0.347]	[0.479,0.342]
	1.13	[0.489,0.439]	[0.494,0.435]	[0.499,0.430]	[0.505,0.424]	[0.510,0.419]	[0.515,0.414]	[0.520,0.409]	[0.525,0.404]	[0.530,0.398]	[0.535,0.393]
	1.14	[0.551,0.498]	[0.557,0.493]	[0.562,0.487]	[0.567,0.482]	[0.572,0.477]	[0.577,0.472]	[0.582,0.466]	[0.587,0.461]	[0.592,0.455]	[0.596,0.450]
	1.15	[0.622,0.563]	[0.627,0.558]	[0.632,0.553]	[0.637,0.548]	[0.642,0.542]	[0.647,0.537]	[0.651,0.531]	[0.656,0.526]	[0.661,0.520]	[0.665,0.515]

Each cell reports $[d \log Y / d \log L, d \log Y^{alloc} / d \log L]$ for different boundary conditions. Each column is a different value for the boundary condition δ_{θ^*} and each row is a different aggregate markup $\bar{\mu}$. Cells that approximately correspond to efficient selection are colored in blue and cells that approximately correspond to efficient entry are colored in yellow. The bulk of the changes in welfare are due to reallocation effects.

Table 10: Change in allocative efficiency for different boundary conditions

		δ_{θ^*}				
		1	3	5	7	9
$\bar{\mu}$	1.05	[0.129,0.001,-0.004]	[0.329,-0.169,-0.0430]	[0.619,-0.428,-0.0850]	[1.068,-0.842,-0.130]	[1.84,-1.576,-0.177]
	1.06	[0.162,0.002,-0.005]	[0.376,-0.180,-0.0470]	[0.689,-0.459,-0.0910]	[1.182,-0.915,-0.139]	[2.05,-1.743,-0.189]
	1.07	[0.192,0.002,-0.007]	[0.420,-0.191,-0.0510]	[0.755,-0.490,-0.0970]	[1.292,-0.986,-0.147]	[2.263,-1.914,-0.201]
	1.08	[0.232,0.003,-0.009]	[0.476,-0.205,-0.0560]	[0.843,-0.531,-0.106]	[1.439,-1.085,-0.159]	[2.559,-2.157,-0.216]
	1.09	[0.269,0.003,-0.012]	[0.530,-0.219,-0.0610]	[0.927,-0.572,-0.114]	[1.586,-1.184,-0.171]	[2.867,-2.415,-0.231]
	1.10	[0.310,0.004,-0.015]	[0.591,-0.234,-0.0670]	[1.023,-0.620,-0.124]	[1.756,-1.303,-0.184]	[3.244,-2.736,-0.249]
	1.11	[0.355,0.004,-0.019]	[0.658,-0.252,-0.0750]	[1.131,-0.675,-0.135]	[1.956,-1.445,-0.200]	[3.715,-3.145,-0.269]
	1.12	[0.406,0.005,-0.023]	[0.735,-0.273,-0.0840]	[1.257,-0.741,-0.149]	[2.195,-1.619,-0.218]	[4.322,-3.682,-0.293]
	1.13	[0.463,0.006,-0.029]	[0.822,-0.298,-0.0940]	[1.403,-0.819,-0.164]	[2.485,-1.836,-0.240]	[5.134,-4.415,-0.321]
	1.14	[0.527,0.007,-0.036]	[0.922,-0.327,-0.107]	[1.576,-0.916,-0.184]	[2.847,-2.114,-0.266]	[6.280,-5.470,-0.355]
	1.15	[0.601,0.008,-0.046]	[1.039,-0.362,-0.123]	[1.785,-1.036,-0.207]	[3.31,-2.481,-0.298]	[8.023,-7.107,-0.396]

Each cell reports $[d\log Y^{\epsilon} - d\log Y^{tech}, d\log Y^{\epsilon,\theta^*} - d\log Y^{\epsilon}, d\log Y^{\epsilon,\theta^*,\mu} - d\log Y^{\epsilon,\theta^*}]$ for different boundary conditions. Each column is a different value for the boundary condition δ_{θ^*} and each row is a different aggregate markup $\bar{\mu}$. The bulk of the positive changes in allocative are due to the Darwinian effect. The pro-competitive and selection effects are either unimportant or harmful.

Appendix D Propagation and Aggregation Equations

In this section, we summarize the propagation and aggregation equations for the model with heterogeneous firms. We expand the equilibrium equations presented in Section 2 to the first order in the shocks. Changes in all the equilibrium variables are expressed via propagation equations as functions of changes in consumer welfare. Changes in consumer welfare are then expressed as as functions of the changes in the equilibrium variable via an aggregation equation. Putting propagation and aggregation together yields a fixed point in changes in consumer welfare.

Welfare. Differentiating the implicit definition of the welfare Y, we find

$$\bar{\delta} d\log M - \lambda_{\theta^*}\delta_{\theta^*}\frac{g(\theta^*)}{1-G(\theta^*)}d\theta^* + \mathbb{E}_{\lambda}\left[d\log(\frac{y_\theta}{Y})\right] = 0.$$

Aggregate price index. Differentiating the definition of the price index, we find

$$-d\log P = d\log Y + d\log M - \lambda_{\theta^*}\frac{g(\theta^*)}{1-G(\theta^*)}d\theta^* + \mathbb{E}_{\lambda}\left[\left(1-\frac{1}{\sigma_\theta}\right)d\log(\frac{y_\theta}{Y})\right].$$

Prices. Differentiating the inverse-demand curve facing each variety, we get

$$d\log p_\theta - d\log P = -\frac{1}{\sigma_\theta}d\log(\frac{y_\theta}{Y}).$$

Markups. Differentiating the markup equation, we get

$$d\log\mu_\theta = \frac{1}{\sigma_\theta}\frac{1-\rho_\theta}{\rho_\theta}d\log(\frac{y_\theta}{Y}).$$

Quantities. Differentiating the individual demand function, we find

$$d\log(\frac{y_\theta}{Y}) = \sigma_\theta\left(d\log(\frac{A_\theta}{\mu_\theta}) + d\log P\right).$$

Combining with the equation for markups, we get

$$d\log(\frac{y_\theta}{Y}) = \rho_\theta\sigma_\theta\left(d\log A_\theta + d\log P\right).$$

Sales shares. Differentiating the sales shares equation, we find

$$d\log\lambda_\theta = d\log p_\theta + d\log(\frac{y_\theta}{Y}) + \frac{-g(\theta^*)}{1-G(\theta^*)}d\theta^* + d\log M + d\log Y.$$

Ratio of variable profits to overhead costs. Differentiating our definition of X_θ, we get

$$d\log X_\theta = \left(\frac{1}{\mu_\theta - 1}\right)d\log\mu_\theta + d\log\lambda_\theta - d\log f_{0,\theta}.$$

Selection. Differentiating the selection condition, we get

$$d\log X_{\theta^*} + \left[\frac{\partial\log X_\theta}{\partial\theta}\bigg|_{\theta^*}\right]d\theta^* = \frac{-g(\theta^*)}{1-G(\theta^*)}d\theta^* + d\log M - d\log L.$$

We define

$$\frac{1}{\gamma_{\theta^*}} = \frac{1-G(\theta^*)}{g(\theta^*)}\left[\frac{\partial\log X_\theta}{\partial\theta}\bigg|_{\theta^*}\right] = \frac{1-G(\theta^*)}{g(\theta^*)}\left[\frac{-\sigma_\theta}{\rho_\theta}\frac{\partial\log\mu_\theta}{\partial\theta} + \left(\frac{\sigma_\theta}{\rho_\theta}-1\right)\frac{\partial\log A_\theta}{\partial\theta} - \frac{\partial\log f_{o,\theta}}{\partial\theta}\bigg|_{\theta^*}\right],$$

which allows us to write the selection condition more simply as

$$d\log X_{\theta^*} + \frac{1}{\gamma_{\theta^*}}\frac{g(\theta^*)}{1-G(\theta^*)}d\theta^* = \frac{-g(\theta^*)}{1-G(\theta^*)}d\theta^* + d\log M - d\log L.$$

Entry. Differentiating the free-entry condition yields

$$d\log L + \left(1 - \left[\mathbb{E}_\lambda\left[\frac{1}{\sigma_\theta}\right]\right]^{-1}\frac{\lambda_{\theta^*}}{\sigma_{\theta^*}}\right)\frac{g(\theta^*)}{1-G(\theta^*)}d\theta^* - d\log M + \mathbb{E}_{\lambda\left(1-\frac{1}{\mu}\right)}\left[d\log f_{o,\theta} + d\log X_\theta\right]$$
$$= \frac{f_e d\log(f_e) - f_{o,\theta^*}g(\theta^*)d\theta^* + (1-G(\theta^*))\,\mathbb{E}\left[f_{o,\theta}\right]\mathbb{E}_{f_o}\left[d\log f_{o,\theta}\right]}{f_e + (1-G(\theta^*))\,\mathbb{E}\left[f_{o,\theta}\right]}.$$

System of equations for a change in market size. To solve for the change in welfare following a change in market size, $d \log L$, we take the system of log-linearized equations above and set $d \log A_\theta = d \log f_{o,\theta} = 0$. We get the following system of eight equations:

$$
\begin{aligned}
0 &= \bar{\delta} d \log M - \lambda_{\theta^*} \delta_{\theta^*} \frac{g(\theta^*)}{1 - G(\theta^*)} d\theta^* + \mathbb{E}_\lambda \left[d \log(\frac{y_\theta}{Y}) \right]. \\
-d \log P &= d \log Y + d \log M - \lambda_{\theta^*} \frac{g(\theta^*)}{1 - G(\theta^*)} d\theta^* + \mathbb{E}_\lambda \left[\left(1 - \frac{1}{\sigma_\theta} \right) d \log(\frac{y_\theta}{Y}) \right]. \\
d \log p_\theta - d \log P &= -\frac{1}{\sigma_\theta} d \log(\frac{y_\theta}{Y}). \\
d \log \mu_\theta &= \frac{1}{\sigma_\theta} \frac{1 - \rho_\theta}{\rho_\theta} d \log(\frac{y_\theta}{Y}). \\
d \log X_\theta &= \left(\frac{1}{\mu_\theta - 1} \right) d \log \mu_\theta + d \log \lambda_\theta. \\
d \log \lambda_\theta &= d \log p_\theta + d \log(\frac{y_\theta}{Y}) + \frac{-g(\theta^*)}{1 - G(\theta^*)} d\theta^* + d \log M + d \log Y. \\
d \log X_{\theta^*} + \frac{1}{\gamma_{\theta^*}} \frac{g(\theta^*)}{1 - G(\theta^*)} d\theta^* &= \frac{-g(\theta^*)}{1 - G(\theta^*)} d\theta^* + d \log M - d \log L. \\
0 &= d \log L + \frac{g(\theta^*)}{1 - G(\theta^*)} d\theta^* - d \log M + \mathbb{E}_{\lambda\left(1 - \frac{1}{\mu}\right)} \left[d \log X_\theta \right].
\end{aligned}
$$

We will now solve for the fixed point of this system. To start, we eliminate all firm-level terms, $d \log \mu_\theta, d \log p_\theta, d \log y_\theta / Y, d \log X_\theta$, and $d \log \lambda_\theta$. We are left with a system of four equations that together pin down the change in welfare, the mass of entrants, the selection cutoff, and the aggregate price index following a change in market size.

$$
\begin{aligned}
0 &= \bar{\delta} d \log M - \lambda_{\theta^*} \delta_{\theta^*} \frac{g(\theta^*)}{1 - G(\theta^*)} d\theta^* + \mathbb{E}_\lambda \left[\rho_\theta \sigma_\theta \right] d \log P. \\
-d \log P &= d \log Y + d \log M - \lambda_{\theta^*} \frac{g(\theta^*)}{1 - G(\theta^*)} d\theta^* + \mathbb{E}_\lambda \left[(\sigma_\theta - 1) \rho_\theta \right] d \log P. \\
\frac{g(\theta^*)}{1 - G(\theta^*)} d\theta^* &= -\gamma_{\theta^*} \sigma_{\theta^*} d \log P - \gamma_{\theta^*} \left(d \log L + d \log Y \right). \\
d \log P &= -\mathbb{E}_\lambda \left[\frac{1}{\sigma_\theta} \right] \left(d \log L + d \log Y \right).
\end{aligned}
$$

The last equation gives intuition for how the aggregate price index moves as the market size increases. An increase in market size lowers the price index due to new entry. This decrease in the price index then increases welfare due to beneficial reallocations, and the increase in welfare further reduces the price index. The result is that the technical efficiency and allocative efficiency effects are amplified via a multiplier due to adjustments in the price index and welfare.

With some manipulation, we can express the change in welfare as a function of the change in

market size and the aggregate price index:

$$d\log Y = (\bar{\delta} - 1)\, d\log L - \left(\xi^{\epsilon} + \xi^{\theta^*} + \xi^{\mu}\right) \mathbb{E}_{\lambda}\left[\sigma_{\theta}^{-1}\right]^{-1} d\log P. \tag{41}$$

The first term captures the change in welfare due to technical efficiency while the second term captured the change in welfare due to allocative efficiency, which is entirely mediated by the aggregate price index.

By plugging in the equation for the price index above and solving the fixed point for $d\log Y$, we get the result in Theorem 1.

Proof of Lemma 1. To derive (18), note that the initial allocation of labor allocates a fraction $l = \mathbb{E}[l_\theta] = \mathbb{E}_\lambda[1/\mu_\theta]$ to variable production, and the remainder to entry and overhead. Suppose we take reduce the fraction of labor allocated to variable production (while preserving the proportions of variable production labor allocated across firms) by $d\log l_\theta = d\log l$. Reallocating that labor to entry and overhead costs allows us to increase consumer welfare by

$$\mathbb{E}_\lambda[\delta_\theta]\; d\log M = \mathbb{E}_\lambda[\delta_\theta]\; d\log l_e = \mathbb{E}_\lambda[\delta_\theta] \frac{\mathbb{E}_\lambda[1/\mu_\theta]}{1 - \mathbb{E}_\lambda[1/\mu_\theta]} (-d\log l) > 0,$$

where $d\log l_e$ is the increase in labor allocated to entry. This gain is consumer welfare is offset by a reduction in the per-capita quantity consumed of each variety, equal to $\mathbb{E}_\lambda[d\log y_\theta] = d\log l - d\log M$. Rearranging, we find that the net change in welfare from reducing the fraction of labor allocated to variable production and increasing the allocation to entry is positive if and only if the average consumer surplus ratio exceeds the harmonic average of markups, yielding the condition in (18) above. ■

Proof of Lemma 2. To derive this condition, suppose that we increase the selection cut-off by $d\theta^* > 0$, and reallocate the labor previously allocated to the variable production and overhead of varieties with type in $[\theta^*, \theta^* + d\theta^*)$ proportionately to entry, overhead, and variable production. The exiting varieties reduce consumer welfare by $-\delta_{\theta^*}\lambda_{\theta^*}[g(\theta^*)/(1 - G(\theta^*))]d\theta^*$. The new varieties $d\log M = \lambda_{\theta^*}[g(\theta^*)/(1 - G(\theta^*))]d\theta^*$ increases consumer welfare by $\mathbb{E}_\lambda[\delta_\theta]d\log M$. There is no change in the production of existing varieties $d\log y_\theta = 0$. Plugging these perturbations into (16), the overall effect on welfare is $(\mathbb{E}_\lambda[\delta_\theta] - \delta_{\theta^*})\lambda_{\theta^*}[g(\theta^*)/(1 - G(\theta^*))]d\theta^*$, which is positive (too little selection) if and only if $\delta_{\theta^*} < \mathbb{E}_\lambda[\delta_\theta]$. ■

Proof of Lemma 3. The intuition is the following. Consider a reduction $d\log l_{\theta'} < 0$ in the fraction of labor allocated to the supply of varieties in $(\theta', \theta' + d\theta')$ and a complementary increase $d\log l_\theta = -(g(\theta')/g(\theta))(l_{\theta'}/l_\theta)d\log l_{\theta'} > 0$ in the fraction of labor allocated to the supply of varieties in $(\theta, \theta + d\theta')$, which, using the fact that $l_{\theta'}/l_\theta = (\lambda_{\theta'}/\mu_{\theta'})/(\lambda_\theta/\mu_\theta)$, can be rewritten as $d\log l_\theta = -(g(\theta')/g(\theta))(\lambda_{\theta'}/\mu_{\theta'})/(\lambda_\theta/\mu_\theta)d\log l_{\theta'} > 0$. This leads to a decrease $d\log y_{\theta'} = d\log l_{\theta'} < 0$ in the quantity of the former varieties and an increase $d\log y_\theta = -(g(\theta')/g(\theta))(\lambda_{\theta'}/\mu_{\theta'})/(\lambda_\theta/\mu_\theta)d\log l_{\theta'} >$

0 in the quantity of the latter varieties. The net effect on welfare is $g(\theta')\lambda_{\theta'}d\log y_{\theta'}d\theta' + g(\theta)\lambda_\theta d\log y_\theta d\theta' = -(\mu_\theta/\mu_{\theta'} - 1)\lambda_{\theta'}g(\theta')d\theta' d\log l_{\theta'}$, which is positive if and only $\mu_\theta > \mu_{\theta'}$.

■

D.1 Conditions for a Locally Unique Equilibrium

In this subsection, we develop conditions under which the model equilibrium exists and is locally unique. We first begin with a definition of a feasible set of statistics (sales densities, consumer surplus ratios, markups, pass-throughs, variable profit to overhead cost ratios, and selection cutoff), then show that a condition on these statistics is sufficient to prove that the equilibrium exists and is locally unique (Proposition 3). Finally, we provide a set of simpler (but stricter) sufficient conditions that guarantee existence and local uniqueness (Corollary 7).

Definition 1. A collection of sales densities, consumer surplus ratios, markups, pass-throughs, variable profit to overhead cost ratios, and selection cutoff $\left\{\{\lambda_\theta, \delta_\theta, \mu_\theta, \rho_\theta, X_\theta\}_{\theta\in\Theta}, \theta^*\right\}$ is *feasible* if

1. $\int_{\theta\in\Theta}\lambda_\theta d\theta = 1$ and $\lambda_\theta \geq 0$ for all θ,
2. $\delta_\theta, \mu_\theta \geq 1$ for all θ,
3. $\rho_\theta \geq 0$ for all θ,
4. $X_\theta \geq 0$ and $\frac{\partial \log X_\theta}{\partial\theta} > 0$ for all θ, and
5. $X_{\theta^*} = 0$.

Proposition 3 (Existence and Local Uniqueness)**.** *For any feasible* $\left\{\{\lambda_\theta, \delta_\theta, \mu_\theta, \rho_\theta, X_\theta\}_{\theta\in\Theta}, \theta^*\right\}$*, the equilibrium exists and is locally unique if*

$$0 \leq \xi^\epsilon + \xi^{\theta^*} + \xi^\mu < 1,$$

where ξ^ϵ*,* ξ^{θ^*}*, and* ξ^μ *are functions of* $\left\{\{\lambda_\theta, \delta_\theta, \mu_\theta, \rho_\theta, X_\theta\}_{\theta\in\Theta}, \theta^*\right\}$ *as defined in Theorem 1.*

Proof. We first show that a collection of feasible $\left\{\{\lambda_\theta, \delta_\theta, \mu_\theta, \rho_\theta, X_\theta\}_{\theta\in\Theta}, \theta^*\right\}$ can be rationalized via some collection of primitives $\{\Upsilon_\theta, A_\theta, f_{o,\theta}\}$. Then, by the inverse function theorem, the equilibrium is locally unique if the Jacobian determinant is non-zero at the equilibrium point.

First, note that the collection $\{\lambda_\theta, \delta_\theta, \mu_\theta, \rho_\theta, X_\theta\}_{\theta\in\Theta}$ can be expressed in terms of some underlying $\{\Upsilon_\theta, A_\theta, f_{o,\theta}\}$:

$$\begin{aligned}
\lambda_\theta &= \bar{\delta}\frac{y_\theta}{Y}\Upsilon'_\theta(\frac{y_\theta}{Y})M(1-G(\theta^*)),\\
\delta_\theta &= \frac{\Upsilon_\theta(\frac{y_\theta}{Y})}{\frac{y_\theta}{Y}\Upsilon'_\theta(\frac{y_\theta}{Y})},\\
\mu_\theta &= \frac{1}{1-\frac{-\frac{y_\theta}{Y}\Upsilon''_\theta(\frac{y_\theta}{Y})}{\Upsilon'_\theta(\frac{y_\theta}{Y})}},
\end{aligned}$$

$$\rho_\theta = \frac{1}{\mu_\theta\left[\frac{\frac{y_\theta}{Y}\Upsilon'''_\theta(\frac{y_\theta}{Y})}{-\Upsilon''_\theta(\frac{y_\theta}{Y})} - 1\right]},$$
$$X_\theta = \frac{\lambda_\theta}{f_{o,\theta}}\left(1 - \frac{1}{\mu_\theta}\right).$$

To rationalize the observed statistics, first choose $\Upsilon'_\theta(\frac{y_\theta}{Y})$ to match the sales densities λ_θ. Then, choose $\left\{\Upsilon_\theta(\frac{y_\theta}{Y}), \Upsilon''_\theta(\frac{y_\theta}{Y}), \Upsilon'''_\theta(\frac{y_\theta}{Y})\right\}$ to match $\{\delta_\theta, \mu_\theta, \rho_\theta\}$. Finally, given λ_θ and μ_θ, choose $\{f_{o,\theta}\}$ to match $\{X_\theta\}$.

By the inverse function theorem, the equilibrium defined by $\left\{\{\lambda_\theta, \delta_\theta, \mu_\theta, \rho_\theta, X_\theta\}_{\theta\in\Theta}, \theta^*\right\}$ and the set $\{\Upsilon_\theta, A_\theta, f_{o,\theta}\}$ is locally unique if the Jacobian determinant is well-defined and non-zero at the equilibrium point. Following Theorem 1, this is the case as long as

$$\xi^\epsilon + \xi^{\theta^*} + \xi^\mu < 1.$$

and

$$\xi^\epsilon + \xi^{\theta^*} + \xi^\mu \neq 1 - \mathbb{E}_\lambda[\delta_\theta].$$

The requirement $0 \leq \xi^\epsilon + \xi^{\theta^*} + \xi^\mu < 1$ ensures both conditions are met. ■

Corollary 7 lists three stricter conditions that are sufficient (but not necessary) to ensure that the condition on $\left\{\{\lambda_\theta, \delta_\theta, \mu_\theta, \rho_\theta, X_\theta\}_{\theta\in\Theta}, \theta^*\right\}$ from Proposition 3 is met.

Corollary 7 (Sufficient Conditions for Existence and Local Uniqueness)**.** *Sufficient conditions for the equilibrium to exist and be locally unique are:*

1. *Firm pass-throughs are $\rho_\theta \leq 1$ for all θ.*
2. *There is a maximum price-elasticity of demand faced by a firm, $\sigma^{\max}$, which satisfies $(\sigma^{\max} - 1)(\mathbb{E}_\lambda[\delta_\theta] - 1) \leq 4$.*
3. *At the cutoff, the price-elasticity of demand and consumer surplus ratio are both weakly greater than average. ($\delta_{\theta^*} \geq \mathbb{E}_\lambda[\delta_\theta]$ and $\sigma_{\theta^*} \geq \mathbb{E}_\lambda[\sigma_\theta]$).*

Proof. Rearranging terms from Theorem 1, the condition that $\xi^\epsilon + \xi^{\theta^*} + \xi^\mu < 1$ is equivalent to:

$$\mathbb{E}_\lambda\left[\rho_\theta\left(\sigma_\theta\left(1 - \frac{1}{\mathbb{E}_\lambda[\delta_\theta]}\right) - 1\right)\right] + 1 + \left(\frac{\mathbb{E}_\lambda[\delta_\theta] - \delta_{\theta^*}}{\mathbb{E}_\lambda[\delta_\theta]}\right)\lambda_{\theta^*}\gamma_{\theta^*}\left(\sigma_{\theta^*} - \mathbb{E}_\lambda\left[\sigma_\theta^{-1}\right]^{-1}\right) < \mathbb{E}_\lambda\left[\sigma_\theta^{-1}\right]^{-1}. \quad (42)$$

We can bound the left-hand side:

$$\begin{aligned}\text{LHS} &= \mathbb{E}_\lambda\left[\rho_\theta\left(\sigma_\theta\left(1 - \frac{1}{\mathbb{E}_\lambda[\delta_\theta]}\right) - 1\right)\right] + 1 + \left(\frac{\mathbb{E}_\lambda[\delta_\theta] - \delta_{\theta^*}}{\mathbb{E}_\lambda[\delta_\theta]}\right)\lambda_{\theta^*}\gamma_{\theta^*}\left(\sigma_{\theta^*} - \mathbb{E}_\lambda\left[\sigma_\theta^{-1}\right]^{-1}\right) \\ &\leq \mathbb{E}_\lambda\left[\rho_\theta\left(\sigma_\theta\left(1 - \frac{1}{\mathbb{E}_\lambda[\delta_\theta]}\right) - 1\right)\right] + 1\end{aligned}$$

$$\leq \left(1 - \frac{1}{\mathbb{E}_\lambda [\delta_\theta]}\right) \mathbb{E}_\lambda [\sigma_\theta],$$

where the second line uses assumption (3) and the third line uses assumption (1). We can thus restate our condition as:

$$\mathbb{E}_\lambda [\sigma_\theta] \mathbb{E}_\lambda \left[\sigma_\theta^{-1}\right] - 1 < \frac{1}{\mathbb{E}_\lambda [\delta_\theta] - 1}. \tag{43}$$

Again, we can bound the left-hand side:

$$\begin{aligned}
\mathbb{E}_\lambda [\sigma_\theta] \mathbb{E}_\lambda \left[\sigma_\theta^{-1}\right] - 1 &= -Cov_\lambda \left[\sigma_\theta, \sigma_\theta^{-1}\right] \\
&\leq \left(Var_\lambda [\sigma_\theta] \, Var_\lambda \left[\sigma_\theta^{-1}\right]\right)^{1/2} \\
&\leq \frac{1}{4} \left(\frac{\sigma^{\max} - 1}{\sigma^{\max}}\right) (\sigma^{\max} - 1) \\
&< \frac{1}{4} (\sigma^{\max} - 1),
\end{aligned}$$

where the second line applies the Cauchy-Schwarz inequality and the third line applies Popoviciu's inequality.[32] Hence, we have $\xi^\epsilon + \xi^{\theta^*} + \xi^\mu < 1$ if

$$\frac{1}{4} (\sigma^{\max} - 1) \leq \frac{1}{\mathbb{E}_\lambda [\delta_\theta] - 1}, \tag{44}$$

which is satisfied under assumption (2). For context, under our baseline calibration where $\mathbb{E}_\lambda [\delta_\theta] = 1.045$, assumption (2) implies that the price-elasticity of demand is at most $\sigma^{\max} = 89$ (i.e., the lowest desired markup of any firm is 1.011).

■

Appendix E Welfare Response to an Entry Tax

This appendix presents the proof of Proposition 2, which characterizes the response of welfare to a marginal tax on entry.

We modify our setup to allow for an entry tax. As in the main text, welfare is defined implicitly by

$$\int_\Theta \Upsilon_\theta(\frac{y_\theta}{Y}) dF(\theta) = 1.$$

Now, however, the household's budget constraint includes both labor earnings and distributed revenues from the entry tax, which we assume is returned to households in a lump-sum transfer. We will use g to denote the per-capita rebate of tax revenue and Λ_L to denote the share of household

[32] These bounds are quite loose; we could further relax assumption (2) by considering tighter bounds on both inequalities.

income coming from labor earnings,

$$\int_{\Theta} p_\theta y_\theta dF(\theta) = w + g, \qquad \text{and} \qquad \Lambda_L = \frac{w}{w+g}. \tag{45}$$

We continue to use the wage as the numeraire, normalizing $w = 1$ throughout. The household's inverse-demand curve for each variety remains

$$\frac{p_\theta}{P} = \Upsilon'_\theta(\frac{y_\theta}{Y}),$$

but with the price aggregator P now taking into account the labor share,

$$P = \frac{1}{\Lambda_L Y} \frac{1}{\int_{\Theta} \Upsilon'_\theta(\frac{y_\theta}{Y}) \frac{y_\theta}{Y} dF(\theta)}. \tag{46}$$

On the production side, firms' profit-maximizing prices and markups are unchanged, and the selection condition remains unchanged. The entry condition now incorporates a tax on entry, which we denote τ:

$$\int_{\theta^*}^{\infty} \left[\left(1 - \frac{1}{\mu_\theta}\right) p_\theta y_\theta wL - f_{o,\theta} \right] g(\theta) d\theta = (1+\tau) f_e. \tag{47}$$

To ensure that sales densities λ_θ still integrate to one, we adjust the definition of the sales density to

$$\lambda_\theta = \Lambda_L p_\theta y_\theta (1 - G(\theta^*)) M.$$

Finally, we add a government budget constraint, which sets the amount rebated to households equal to the amount collected in taxes,

$$\tau f_e M = gL. \tag{48}$$

We combine this equation with (45) to solve for the labor share in terms of the entry tax,

$$\Lambda_L = \frac{1}{1 + \tau f_e \frac{M}{L}}. \tag{49}$$

By differentiating the above conditions, we find that the response of welfare to a change in the entry tax is the fixed point of the following system of equations:

$$0 = d\log M - \Lambda_L \lambda_{\theta^*} \frac{\delta_{\theta^*}}{\bar{\delta}} \frac{g(\theta^*)}{1 - G(\theta^*)} d\theta^* + \frac{1}{\bar{\delta}} \Lambda_L \mathbb{E}_\lambda \left[d\log(\frac{y_\theta}{Y}) \right].$$

$$d\log p_\theta - d\log P = -\frac{1}{\sigma_\theta} d\log \frac{y_\theta}{Y}$$

$$-d\log P = d\log \Lambda_L + d\log Y + d\log M - \lambda_{\theta^*} \frac{g(\theta^*)}{1 - G(\theta^*)} d\theta^* + \mathbb{E}_\lambda \left[\left(1 - \frac{1}{\sigma_\theta}\right) d\log(\frac{y_\theta}{Y}) \right].$$

$$d\log \mu_\theta = \frac{1}{\sigma_\theta} \frac{1 - \rho_\theta}{\rho_\theta} d\log(\frac{y_\theta}{Y}).$$

$$d\log X_\theta = \frac{1}{\mu_\theta - 1} d\log \mu_\theta + d\log \Lambda_L + d\log p_\theta + d\log \frac{y_\theta}{Y} + d\log Y + \frac{-g(\theta^*)}{1-G(\theta^*)} d\theta^* + d\log M.$$

$$\frac{1}{\gamma_{\theta^*}} \frac{g(\theta^*)}{1-G(\theta^*)} d\theta^* = -d\log X_{\theta^*} + \frac{-g(\theta^*)}{1-G(\theta^*)} d\theta^* + d\log M + d\log \Lambda_L.$$

$$0 = \frac{g(\theta^*)}{1-G(\theta^*)} d\theta^* - d\log M - d\log \Lambda_L + \mathbb{E}_{\lambda\left(1-\frac{1}{\mu}\right)}\left[d\log X_\theta\right] - \frac{(1+\tau) f_e}{(1+\tau) f_e + (1-G(\theta^*))\,\mathbb{E}\left[f_{o,\theta}\right]} d\log(1+\tau).$$

$$d\log \Lambda_L = -\frac{(1+\tau) f_e \frac{M}{L}}{1+\tau f_e \frac{M}{L}} d\log(1+\tau) - \frac{\tau f_e \frac{M}{L}}{1+\tau f_e \frac{M}{L}} d\log M.$$

We evaluate this system at the point where the tax is zero, and hence $\tau = 0, \Lambda_L = 1$. With some manipulation, we can express the change in welfare, the mass of entrants, the price aggregator, the labor share, and the selection cutoff in terms of the marginal tax $d\tau$.

$$0 = \bar{\delta} d\log M - \lambda_{\theta^*}\delta_{\theta^*} \frac{g(\theta^*)}{1-G(\theta^*)} d\theta^* + \mathbb{E}_\lambda\left[\sigma_\theta \rho_\theta\right] d\log P.$$

$$0 = d\log \Lambda_L + d\log Y + d\log M - \lambda_{\theta^*} \frac{g(\theta^*)}{1-G(\theta^*)} d\theta^* + \mathbb{E}_\lambda\left[\sigma_\theta - (\sigma_\theta - 1)(1-\rho_\theta)\right] d\log P.$$

$$0 = \left[\sigma_{\theta^*} d\log P + d\log Y\right] + \frac{1}{\gamma_{\theta^*}} \frac{g(\theta^*)}{1-G(\theta^*)} d\theta^*.$$

$$\frac{f_e M}{L} d\tau = d\log P + \mathbb{E}_\lambda\left[\frac{1}{\sigma_\theta}\right] d\log Y.$$

$$d\log \Lambda_L = -\frac{f_e M}{L} d\tau.$$

Solving the fixed point yields,

$$d\log Y = \frac{1 - \bar{\delta}\mathbb{E}_\lambda\left[\frac{1}{\mu_\theta}\right] - (\bar{\delta} - \delta_{\theta^*})\lambda_{\theta^*}\gamma_{\theta^*}\mathbb{E}_\lambda\left[\frac{\sigma_{\theta^*}}{\sigma_\theta}\right] - \mathbb{E}_\lambda\left[\sigma_\theta(1-\rho_\theta)\left[1-\frac{\bar{\delta}}{\mu_\theta}\right]\right]\mathbb{E}_\lambda\left[\frac{1}{\sigma_\theta}\right] - (\bar{\delta}-1)\,Cov_\lambda\left[\sigma_\theta, \frac{1}{\mu_\theta}\right]}{1 - (\bar{\delta} - \delta_{\theta^*})\lambda_{\theta^*}\gamma_{\theta^*}\left(\mathbb{E}_\lambda\left[\frac{\sigma_{\theta^*}}{\sigma_\theta}\right] - 1\right) - \mathbb{E}_\lambda\left[\sigma_\theta(1-\rho_\theta)\left[1-\frac{\bar{\delta}}{\mu_\theta}\right]\right]\mathbb{E}_\lambda\left[\frac{1}{\sigma_\theta}\right] - (\bar{\delta}-1)\,Cov_\lambda\left[\sigma_\theta, \frac{1}{\mu_\theta}\right]} \times \frac{f_e}{f_e + (1-G(\theta^*))\,\mathbb{E}\left[f_{o,\theta}\right]} d\tau.$$

We use the definitions of ξ^ϵ, ξ^{θ^*}, and ξ^μ in the main text to simplify this expression to the result in Proposition 2.

Appendix F Distance to Efficient Frontier

In this appendix, we focus on the distance to the efficient frontier, that is the amount of misallocation in the decentralized equilibrium compared to the first-best allocation.

In Appendix F.1, we provide an analytical second-order approximation which neatly decom-

poses the contributions of the different margins of inefficiency to the overall amount of misallocation. The proof of the main proposition can be found in Appendix F.3. In Appendix F.2, we compute the distance to the frontier in our empirical application.

F.1 Analytical Second-Order Approximation

In this section, we calculate the social costs of the distortions caused by monopolistic competition around the efficient CES benchmark. We index the Kimball aggregator Υ_t by some parameter t, where $t = 0$ gives an iso-elastic form for Υ (CES), and moving from $t = 0$ perturbs the Kimball aggregator away from iso-elasticity in a smooth fashion. The proposition below provides a second-order approximation in t of the distance to the efficient frontier, providing a link between our framework and the literature on the social costs of misallocation with entry (for example, Epifani and Gancia, 2011).

Proposition 4 (Distance to Frontier)**.** *The difference between welfare at the first-best allocation and the decentralized equilibrium can be approximated around $t = 0$ by*

$$\log \frac{\Upsilon^{opt}}{\Upsilon} \approx \frac{1}{2}\mathbb{E}_\lambda\left[\sigma_\theta\left(\frac{\mu_\theta}{\mathbb{E}_\lambda\left[\delta_\theta\right]} - \frac{\mathbb{E}_\lambda\left[\mu_\theta\right]}{\mathbb{E}_\lambda\left[\delta_\theta\right]}\right)^2\right] + \frac{1}{2}\mathbb{E}_\lambda\left[\sigma_\theta\right]\left(\frac{\mathbb{E}_\lambda\left[\mu_\theta\right]}{\mathbb{E}_\lambda\left[\delta_\theta\right]} - 1\right)^2 + \frac{1}{2}\lambda_{\theta^*}\gamma_{\theta^*}\left(\mathbb{E}_\lambda\left[\delta_\theta\right] - \delta_{\theta^*}\right)^2,$$

where the remainder term is order t^3.

The first term, familiar from the misallocation literature, captures distortions in the relative sizes of existing firms. It scales with the dispersion of the ratios of markups to the average consumer surplus ratio $\mu_\theta/\mathbb{E}_\lambda[\delta_\theta]$. It also scales with the elasticities of substitution σ_θ.[33]

The second term captures the distortions due to inefficient entry. It scales with the squared distance to unity of the ratio of the average markup to the average consumer surplus ratio $\mathbb{E}_\lambda[\mu_\theta]/\mathbb{E}_\lambda[\delta_\theta]$. It also scales with the elasticities of substitution σ_θ.

The third and final term captures the distortions due to inefficient selection. It scales with the squared difference between the consumer surplus ratio of the marginal firm δ_{θ^*} and that of the average $\mathbb{E}_\lambda(\delta_\theta)$. It also scales with the hazard rate of the log productivity distribution for the marginal firm γ^*_θ (rather than the price elasticity of demand), which captures the relevant elasticity of the selection margin.[34]

In the CES case, markups are constant across varieties $\mu_\theta = \mathbb{E}_\lambda[\mu_\theta]$, the average markup is equal to the average consumer surplus ratio $\mathbb{E}_\lambda[\mu_\theta] = \mathbb{E}_\lambda[\delta_\theta]$, and consumer surplus ratios are constant across varieties $\delta_{\theta^*} = \mathbb{E}_\lambda[\delta_\theta]$. As a result, all three terms are zero.

[33]The first term is a particular case of the formulas in Baqaee and Farhi (2019) applied to the relevant distortions $\mu_\theta/\mathbb{E}_\lambda[\delta_\theta]$ in the presence of entry (rather than to μ_θ when there is no entry).

[34]If there are many firms at the cut-off (high λ_{θ^*}) or the cut-off moves very quickly (high γ_{θ^*}) in response to distortions, then the losses from selection inefficiency $\delta_{\theta^*} \neq \mathbb{E}_\lambda(\delta_\theta)$ are amplified.

F.2 Quantitative Results

In this appendix, we compute the distance to the efficient frontier in our empirical application.

	$\bar{\mu} = 1.045$		$\bar{\mu} = 1.090$	
	$\bar{\delta} = \delta_{\theta^*}$	$\bar{\delta} = \bar{\mu}$	$\bar{\delta} = \delta_{\theta^*}$	$\bar{\delta} = \bar{\mu}$
Heterogeneous firms	0.024	0.027	0.057	0.065
Homogeneous firms	0.021	0.000	0.041	0.000

Table 11: Distance to the efficient frontier $\log(Y^{opt}/Y)$.

We finish by computing the distance to the efficient frontier. The results are reported in Table 11 both for the case with heterogeneous firms and for the case with homogeneous firms.

With heterogeneous firms, and with average markups $\bar{\mu} = 1.045$ the distance to the frontier is around 2.5%. The distance to the frontier is higher with higher average markups $\bar{\mu} = 1.09$ at around 6%. In both cases, the numbers are similar for efficient entry and efficient selection.

While these numbers are sizable, one might think that they are not large enough. Indeed, in Section 7, we saw in the decentralized equilibrium, cumulated changes in allocative efficiency are large relative to cumulated changes in technical efficiency even for large increases in population. If the distance to the frontier is sizable but not very large, doesn't that mean that the economy should quickly approach the frontier as we increase population? And then shouldn't this source of welfare gains grounded in misallocation quickly peter out? The answer to these questions is no and the reason is the following. At the first-best allocation, increases in population only increase welfare by improving technical efficiency. But changes in technical efficiency for the first-best allocation (at the frontier) turn out to be much larger than changes in technical efficiency for the decentralized equilibrium (inside the frontier). And so the distance to the efficient frontier remains sizable even for large increases in population.[35]

With homogeneous firms, the distance to the frontier is zero when $\bar{\delta} = \bar{\mu}$ since then entry, which is the only margin that can be distorted, is efficient. Otherwise the distance to the frontier is smaller than with heterogeneous firms, but not considerably so. Again, and for the same reasons as those explained above, this does not contradict the earlier observation that changes in allocative efficiency are small at the decentralized equilibrium with homogeneous firms.

F.3 Proof of Proposition 4

To do this, imagine a social planner who can implement the efficient allocation by regulating markups and imposing sales taxes. A sufficient condition is to set markups according to the

[35]This discussion goes back to our definition of changes in allocative efficiency as the changes in welfare that arise from the reallocation of resources as opposed to the change in the distance to the efficient frontier discussed in footnote 12 and Appendix F.

consumer surplus each firm generates $\mu_\theta^{opt} = \delta_\theta$ and sales taxes to be the reciprocal of markups $\tau_\theta^{opt} = 1/\mu_\theta$. The markups provide socially optimal incentives along the extensive margin and the output taxes undo the inefficiencies brought about by dispersed markups. See Edmond et al. (2018) for an alternative implementation of the optimal allocation using taxes.[36] This section contributes to the literature by providing an analytical approximation for distance to the efficient frontier.

At the decentralized monopolistically competitive equilibrium, we instead have $\mu_\theta = (1 - 1/\sigma_\theta)^{-1}$ and $\tau_\theta = 1$. The equilibrium equations are

$$(1 - G(\theta^*))M \int_{\theta^*}^{\infty} \Upsilon(\frac{y_\theta}{Y}) \frac{g(\theta)}{1 - G(\theta^*)} d\theta = 1,$$

$$\Lambda_L = \int_{\theta^*}^{\infty} \frac{\lambda_\theta}{\tau_\theta} \frac{g(\theta)}{1 - G(\theta^*)} d\theta,$$

$$\frac{M\Lambda_L f_e}{L} = \int_{\theta^*}^{\infty} \left(\lambda_\theta \frac{1}{\tau_\theta} \left(1 - \frac{1}{\mu_\theta} \right) - \frac{(1 - G(\theta^*))M\Lambda_L f_o}{L} \right) \frac{g(\theta)}{1 - G(\theta^*)} d\theta,$$

$$\lambda_{\theta^*} \frac{1}{\tau_{\theta^*}} \left(1 - \frac{1}{\mu_{\theta^*}} \right) = \frac{(1 - G(\theta^*))M\Lambda_L f_o}{L},$$

$$\lambda_\theta = (1 - G(\theta^*))M \frac{\tau_\theta \mu_\theta \Lambda_L y_\theta}{A_\theta},$$

$$\frac{\tau_\theta \mu_\theta \Lambda_L}{A_\theta} = P\Upsilon'(\frac{y_\theta}{Y}),$$

$$P = \frac{\bar{\delta}}{Y},$$

$$\frac{1}{\bar{\delta}} = (1 - G(\theta^*))M \int_{\theta^*}^{\infty} \frac{y_\theta}{Y} \Upsilon'(\frac{y_\theta}{Y}) \frac{g(\theta)}{1 - G(\theta^*)} d\theta.$$

Efficiency requires

$$\mu_\theta = \frac{1}{\tau_\theta} = \frac{\Upsilon_\theta}{\frac{y_\theta}{Y} \Upsilon'_\theta}.$$

In step 1, we log-differentiate the equilibrium equations (at an arbitrary point). In step 2, we specialize these equations to the monopolistically competitive equilibrium with changes in markups and taxes towards the efficient point. We use the resulting formulas to compute the distance to the efficient frontier by dividing the first order effect (of moving towards the efficient point) by 1/2. This is because we know that the derivative once we reach the efficient point is zero, and the average of two first-order approximations yields a second-order approximation.

[36] Bilbiie et al. (2019) also consider related issues in a dynamic context.

Step 1:

In the first step, we generalize the propagation equations to allow for policy.

Aggregate price index:

$$-d\log P = \frac{d\log M + d\log Y - \lambda_{\theta^*}\frac{g(\theta^*)}{1-G(\theta^*)}d\theta^*}{1+\int_{\theta^*}^{\infty}\lambda_\theta\left(\frac{\Upsilon'_\theta}{-\frac{y_\theta}{Y}\Upsilon''_\theta}-1\right)\frac{g(\theta)}{1-G(\theta^*)}d\theta} - \frac{\int_{\theta^*}^{\infty}\lambda_\theta\left(\frac{\Upsilon'_\theta}{-\frac{y_\theta}{Y}\Upsilon''_\theta}-1\right)(d\log\mu_\theta + d\log\tau_\theta + d\log\Lambda_L)\frac{g(\theta)}{1-G(\theta^*)}d\theta}{1+\int_{\theta^*}^{\infty}\lambda_\theta\left(\frac{\Upsilon'_\theta}{-\frac{y_\theta}{Y}\Upsilon''_\theta}-1\right)\frac{g(\theta)}{1-G(\theta^*)}d\theta}.$$

Sales shares:

$$d\log\lambda_\theta = d\log M - \frac{g(\theta^*)}{1-G(\theta^*)}d\theta^* + d\log Y - \left(\frac{\Upsilon'_\theta}{-\frac{y_\theta}{Y}\Upsilon''_\theta}-1\right)(d\log\mu_\theta + d\log\tau_\theta + d\log\Lambda_L) + \frac{\Upsilon'_\theta}{-\frac{y_\theta}{Y}\Upsilon''_\theta}d\log P.$$

Variable profits:

$$d\log\left(\frac{\lambda_\theta}{\Lambda_L}\frac{1}{\tau_\theta}\left(1-\frac{1}{\mu_\theta}\right)\right) = d\log M - \frac{g(\theta^*)}{1-G(\theta^*)}d\theta^* + d\log Y - \frac{\Upsilon'_\theta}{-\frac{y_\theta}{Y}\Upsilon''_\theta}(d\log\tau_\theta + d\log\Lambda_L) + \left(\frac{1}{\mu_\theta - 1} - \left(\frac{\Upsilon'_\theta}{-\frac{y_\theta}{Y}\Upsilon''_\theta}-1\right)\right)d\log\mu_\theta + \frac{\Upsilon'_\theta}{-\frac{y_\theta}{Y}\Upsilon''_\theta}d\log P.$$

Quantities:

$$d\log\left(\frac{y_\theta}{Y}\right) = -\frac{\Upsilon'_\theta}{-\frac{y_\theta}{Y}\Upsilon''_\theta}(d\log\mu_\theta + d\log\tau_\theta + d\log\Lambda_L - d\log P).$$

Labor share:

$$d\log\Lambda_L = \frac{g(\theta^*)}{1-G(\theta^*)}d\theta^* + \frac{\int_{\theta^*}^{\infty}\frac{\lambda_\theta}{\tau_\theta}(d\log\lambda_\theta - d\log\tau_\theta)\frac{g(\theta)}{1-G(\theta^*)}d\theta}{\int_{\theta^*}^{\infty}\frac{\lambda_\theta}{\tau_\theta}\frac{g(\theta)}{1-G(\theta^*)}d\theta,} - \frac{\frac{\lambda_{\theta^*}}{\tau_{\theta^*}}\frac{g(\theta^*)}{1-G(\theta^*)}d\theta^*}{\int_{\theta^*}^{\infty}\frac{\lambda_\theta}{\tau_\theta}\frac{g(\theta)}{1-G(\theta^*)}d\theta,}.$$

Entry:

$$d\log M = \frac{g(\theta)}{1-G(\theta^*)}d\theta^* + \frac{\int_{\theta^*}^{\infty}\left(\frac{\lambda_\theta}{\Lambda_L}\frac{1}{\tau_\theta}\left(1-\frac{1}{\mu_\theta}\right)\right)\left[d\log\left(\lambda_\theta\frac{1}{\tau_\theta}\left(1-\frac{1}{\mu_\theta}\right)\right) - d\log\Lambda_L\right]\frac{g(\theta)}{1-G(\theta^*)}d\theta}{\int_{\theta^*}^{\infty}\frac{\lambda_\theta}{\Lambda_L}\frac{1}{\tau_\theta}\left(1-\frac{1}{\mu_\theta}\right)\frac{g(\theta)}{1-G(\theta^*)}d\theta}.$$

Replacing to get aggregate price index:

$$-d\log P + d\log\Lambda_L = \frac{d\log Y}{\int_{\theta^*}^{\infty}\frac{\lambda_\theta}{\Lambda_L}\frac{1}{\tau_\theta}\left(1-\frac{1}{\mu_\theta}\right)\frac{\Upsilon'_\theta}{-\frac{y_\theta}{Y}\Upsilon''_\theta}\frac{g(\theta)}{1-G(\theta^*)}d\theta} - \frac{\int_{\theta^*}^{\infty}\frac{\lambda_\theta}{\Lambda_L}\frac{1}{\tau_\theta}\left(1-\frac{1}{\mu_\theta}\right)\frac{\Upsilon'_\theta}{-\frac{y_\theta}{Y}\Upsilon''_\theta}d\log\tau_\theta\frac{g(\theta)}{1-G(\theta^*)}d\theta}{\int_{\theta^*}^{\infty}\frac{\lambda_\theta}{\Lambda_L}\frac{1}{\tau_\theta}\left(1-\frac{1}{\mu_\theta}\right)\frac{\Upsilon'_\theta}{-\frac{y_\theta}{Y}\Upsilon''_\theta}\frac{g(\theta)}{1-G(\theta^*)}d\theta}$$
$$+\frac{\int_{\theta^*}^{\infty}\frac{\lambda_\theta}{\Lambda_L}\frac{1}{\tau_\theta}\left(1-\frac{1}{\mu_\theta}\right)\left(\frac{1}{\mu_\theta-1}-\left(\frac{\Upsilon'_\theta}{-\frac{y_\theta}{Y}\Upsilon''_\theta}-1\right)\right)d\log\mu_\theta\frac{g(\theta)}{1-G(\theta^*)}d\theta}{\int_{\theta^*}^{\infty}\frac{\lambda_\theta}{\Lambda_L}\frac{1}{\tau_\theta}\left(1-\frac{1}{\mu_\theta}\right)\frac{\Upsilon'_\theta}{-\frac{y_\theta}{Y}\Upsilon''_\theta}\frac{g(\theta)}{1-G(\theta^*)}d\theta}+\frac{\left(\frac{M(1-G(\theta^*))f_o}{L}-\lambda_\theta\frac{1}{\tau_\theta}\left(1-\frac{1}{\mu_\theta}\right)\right)\frac{g(\theta^*)}{1-G(\theta^*)}d\theta^*}{\int_{\theta^*}^{\infty}\frac{\lambda_\theta}{\Lambda_L}\frac{1}{\tau_\theta}\left(1-\frac{1}{\mu_\theta}\right)\frac{\Upsilon'_\theta}{-\frac{y_\theta}{Y}\Upsilon''_\theta}\frac{g(\theta)}{1-G(\theta^*)}d\theta}.$$

Replacing to get entry:

$$\begin{aligned}d\log M = -d\log Y + \lambda_{\theta^*}\frac{g(\theta^*)}{1-G(\theta^*)}d\theta^* \\ -\left(1+\int_{\theta^*}^{\infty}\lambda_\theta\left(\frac{\Upsilon'_\theta}{-\frac{y_\theta}{Y}\Upsilon''_\theta}-1\right)\frac{g(\theta)}{1-G(\theta^*)}\right)d\log P \\ +\int_{\theta^*}^{\infty}\lambda_\theta\left(\frac{\Upsilon'_\theta}{-\frac{y_\theta}{Y}\Upsilon''_\theta}-1\right)\left(d\log\mu_\theta+d\log\tau_\theta+d\log\Lambda_L\right)\frac{g(\theta)}{1-G(\theta^*)}d\theta.\end{aligned}$$

Selection cut-off:

$$\left(\frac{\Upsilon'_{\theta^*}}{-\frac{y_{\theta^*}}{Y}\Upsilon''_{\theta^*}}-1\right)\left.\frac{\partial\log A_\theta}{\partial\theta}\right|_{\theta=\theta^*}d\theta^* = -d\log\left(\frac{\lambda_{\theta^*}}{\Lambda_L}\frac{1}{\tau_{\theta^*}}\left(1-\frac{1}{\mu_{\theta^*}}\right)\right)+d\log M-\frac{g(\theta^*)}{1-G(\theta^*)}d\theta^*.$$

Welfare:

$$\begin{aligned}d\log Y = d\log M(\bar\delta-1) - \int_{\theta^*}^{\infty}\lambda_\theta\left(d\log\mu_\theta+d\log\tau_\theta+d\log\Lambda_L\right)\frac{g(\theta)}{1-G(\theta^*)}d\theta \\ -\left(\frac{\Upsilon_{\theta^*}}{\frac{y_{\theta^*}}{Y}\Upsilon'_{\theta^*}}-1\right)\lambda_{\theta^*}\frac{g(\theta^*)}{1-G(\theta^*)}d\theta^*,\end{aligned}$$

or

$$\begin{aligned}\bar\delta d\log Y = -(\bar\delta-1)\left(1+\int_{\theta^*}^{\infty}\lambda_\theta\left(\frac{\Upsilon'_\theta}{-\frac{y_\theta}{Y}\Upsilon''_\theta}-1\right)\frac{g(\theta)}{1-G(\theta^*)}\right)d\log P \\ -\int_{\theta^*}^{\infty}\lambda_\theta\left[1-(\bar\delta-1)\left(\frac{\Upsilon'_\theta}{-\frac{y_\theta}{Y}\Upsilon''_\theta}-1\right)\right]\left(d\log\mu_\theta+d\log\tau_\theta+d\log\Lambda_L\right)\frac{g(\theta)}{1-G(\theta^*)}d\theta \\ +\left(\bar\delta-\frac{\Upsilon_{\theta^*}}{\frac{y_{\theta^*}}{Y}\Upsilon'_{\theta^*}}\right)\lambda_{\theta^*}\frac{g(\theta^*)}{1-G(\theta^*)}d\theta^*.\end{aligned}$$

Step 2

We proceed in two steps.

Applying the formula at the monopolistic competitive equilibrium. We start at the monopolistic competitive equilibrium. We can simplify the equations to get

$$d\log\Lambda_L = -\int_{\theta^*}^{\infty}\lambda_\theta d\log\tau_\theta\frac{g(\theta)}{1-G(\theta^*)}d\theta,$$

$$-d\log P + d\log\Lambda_L = d\log Y - \int_{\theta^*}^{\infty}\lambda_\theta d\log\tau_\theta\frac{g(\theta)}{1-G(\theta^*)}d\theta,$$

$$\begin{aligned}d\log M = -d\log Y + \lambda_{\theta^*}\frac{g(\theta^*)}{1-G(\theta^*)}d\theta^* \\ -\left(1+\int_{\theta^*}^{\infty}\lambda_\theta\left(\frac{\Upsilon'_\theta}{-\frac{y_\theta}{Y}\Upsilon''_\theta}-1\right)\frac{g(\theta)}{1-G(\theta^*)}\right)d\log P \\ +\int_{\theta^*}^{\infty}\lambda_\theta\left(\frac{\Upsilon'_\theta}{-\frac{y_\theta}{Y}\Upsilon''_\theta}-1\right)\left(d\log\mu_\theta+d\log\tau_\theta+d\log\Lambda_L\right)\frac{g(\theta)}{1-G(\theta^*)}d\theta,\end{aligned}$$

$$\left(\frac{\Upsilon'_{\theta^*}}{-\frac{y_{\theta^*}}{Y}\Upsilon''_{\theta^*}}-1\right)\left.\frac{\partial\log A_\theta}{\partial\theta}\right|_{\theta=\theta^*}d\theta^* = -d\log Y + \frac{\Upsilon'_{\theta^*}}{-\frac{y_{\theta^*}}{Y}\Upsilon''_{\theta^*}}\left(d\log\tau_{\theta^*}-d\log P+d\log\Lambda_L\right).$$

The solution (apart from $d\log M$ which we do not need for what follows) is

$$d\log\Lambda_L = -\int_{\theta^*}^{\infty}\lambda_\theta d\log\tau_\theta\frac{g(\theta)}{1-G(\theta^*)}d\theta,$$

$$-d\log P = d\log Y,$$

$$\left.\frac{\partial\log A_\theta}{\partial\theta}\right|_{\theta=\theta^*}d\theta^* = d\log Y + \frac{\frac{\Upsilon'_{\theta^*}}{-\frac{y_{\theta^*}}{Y}\Upsilon''_{\theta^*}}}{\frac{\Upsilon'_{\theta^*}}{-\frac{y_{\theta^*}}{Y}\Upsilon''_{\theta^*}}-1}\left(d\log\tau_{\theta^*}-\int_{\theta^*}^{\infty}\lambda_\theta d\log\tau_\theta\frac{g(\theta)}{1-G(\theta^*)}d\theta\right).$$

Plugging into welfare, we get

$$\begin{aligned}\left[1-\int_{\theta^*}^{\infty}\lambda_\theta(\bar\delta-1)\left(\frac{\Upsilon'_\theta}{-\frac{y_\theta}{Y}\Upsilon''_\theta}-1\right)\frac{g(\theta)}{1-G(\theta^*)}d\theta-\left(\bar\delta-\frac{\Upsilon_{\theta^*}}{\frac{y_{\theta^*}}{Y}\Upsilon'_{\theta^*}}\right)\lambda_{\theta^*}\frac{g(\theta^*)}{1-G(\theta^*)}d\theta^*\right]d\log Y = \\ -\int_{\theta^*}^{\infty}\lambda_\theta\left[1-(\bar\delta-1)\left(\frac{\Upsilon'_\theta}{-\frac{y_\theta}{Y}\Upsilon''_\theta}-1\right)\right]\left(d\log\mu_\theta+d\log\tau_\theta\right)\frac{g(\theta)}{1-G(\theta^*)}d\theta \\ +\int_{\theta^*}^{\infty}\lambda_\theta\left[1-(\bar\delta-1)\left(\frac{\Upsilon'_\theta}{-\frac{y_\theta}{Y}\Upsilon''_\theta}-1\right)\right]\left(\int_{\theta^*}^{\infty}\lambda_\theta d\log\tau_\theta\frac{g(\theta)}{1-G(\theta^*)}d\theta\right)\frac{g(\theta)}{1-G(\theta^*)}d\theta \\ +\lambda_{\theta^*}\gamma_{\theta^*}\left(\bar\delta-\frac{\Upsilon_{\theta^*}}{\frac{y_{\theta^*}}{Y}\Upsilon'_{\theta^*}}\right)\frac{\frac{\Upsilon'_{\theta^*}}{-\frac{y_{\theta^*}}{Y}\Upsilon''_{\theta^*}}}{\frac{\Upsilon'_{\theta^*}}{-\frac{y_{\theta^*}}{Y}\Upsilon''_{\theta^*}}-1}\left(d\log\tau_{\theta^*}-\int_{\theta^*}^{\infty}\lambda_\theta d\log\tau_\theta\frac{g(\theta)}{1-G(\theta^*)}d\theta\right).\end{aligned}$$

Applying to changes in markups and taxes towards the efficient point. Efficiency requires markups $\mu_\theta = \frac{\Upsilon_\theta}{\frac{y_\theta}{Y}\Upsilon'_\theta}$ and taxes on production $\tau_\theta = 1/\frac{\Upsilon_\theta}{\frac{y_\theta}{Y}\Upsilon'_\theta}$. Hence we use the forcing variables (the endogenous response of $\frac{\Upsilon_\theta}{\frac{y_\theta}{Y}\Upsilon'_\theta}$ is second order)

$$d\log\mu_\theta \approx -\log\left(\frac{\mu_\theta}{\frac{\Upsilon_\theta}{\frac{y_\theta}{Y}\Upsilon'_\theta}}\right),$$

$$d\log\tau_\theta \approx -\log\left(\frac{\Upsilon_\theta}{\frac{y_\theta}{Y}\Upsilon'_\theta}\right).$$

Plugging into welfare, we get

$$\begin{aligned}
&\left[1-\int_{\theta^*}^{\infty}\lambda_\theta\left(\bar\delta-1\right)\left(\frac{\Upsilon'_\theta}{-\frac{y_\theta}{Y}\Upsilon''_\theta}-1\right)\frac{g(\theta)}{1-G(\theta^*)}-\left(\bar\delta-\frac{\Upsilon_{\theta^*}}{\frac{y_{\theta^*}}{Y}\Upsilon'_{\theta^*}}\right)\lambda_{\theta^*}\frac{g(\theta^*)}{1-G(\theta^*)}\theta^*\right]d\log Y\approx\\
&\quad-\int_{\theta^*}^{\infty}\lambda_\theta\left[1-\left(\bar\delta-1\right)\left(\frac{\Upsilon'_\theta}{-\frac{y_\theta}{Y}\Upsilon''_\theta}-1\right)\right]\left(-\log\left(\frac{\mu_\theta}{\frac{\Upsilon_\theta}{\frac{y_\theta}{Y}\Upsilon'_\theta}}\right)-\log\left(\frac{\Upsilon_\theta}{\frac{y_\theta}{Y}\Upsilon'_\theta}\right)\right)\frac{g(\theta)}{1-G(\theta^*)}d\theta\\
&\quad-\int_{\theta^*}^{\infty}\lambda_\theta\left[1-\left(\bar\delta-1\right)\left(\frac{\Upsilon'_\theta}{-\frac{y_\theta}{Y}\Upsilon''_\theta}-1\right)\right]\left(\int_{\theta^*}^{\infty}\lambda_\theta\log\left(\frac{\Upsilon_\theta}{\frac{y_\theta}{Y}\Upsilon'_\theta}\right)\frac{g(\theta)}{1-G(\theta^*)}d\theta\right)\frac{g(\theta)}{1-G(\theta^*)}d\theta\\
&\quad+\lambda_{\theta^*}\gamma_{\theta^*}\left(\bar\delta-\frac{\Upsilon_{\theta^*}}{\frac{y_{\theta^*}}{Y}\Upsilon'_{\theta^*}}\right)\frac{\frac{\Upsilon'_{\theta^*}}{-\frac{y_{\theta^*}}{Y}\Upsilon''_{\theta^*}}}{\frac{\Upsilon'_{\theta^*}}{-\frac{y_{\theta^*}}{Y}\Upsilon''_{\theta^*}}-1}\left(-\log\left(\frac{\Upsilon_{\theta^*}}{\frac{y_{\theta^*}}{Y}\Upsilon'_{\theta^*}}\right)+\int_{\theta^*}^{\infty}\lambda_\theta\log\left(\frac{\Upsilon_\theta}{\frac{y_\theta}{Y}\Upsilon'_\theta}\right)\frac{g(\theta)}{1-G(\theta^*)}d\theta\right).
\end{aligned}$$

And the loss function encapsulating the distance to the efficient frontier is

$$\mathcal{L}\approx\frac{1}{2}d\log Y.$$

Using the notation in the paper, we therefore get

$$\mathcal{L}\approx-\frac{1}{2}\mathbb{E}_\lambda\left[\left(1-\frac{\mathbb{E}_\lambda\left[\delta_\theta\right]-1}{\mu_\theta-1}\right)\log\left(\frac{\mathbb{E}_\lambda\left[\delta_\theta\right]}{\mu_\theta}\right)\right]+\frac{1}{2}\lambda_{\theta^*}\gamma_{\theta^*}\left(\mathbb{E}_\lambda\left[\delta_\theta\right]-\delta_{\theta^*}\right)\mu^*_\theta\log\left(\frac{\mathbb{E}_\lambda\left[\delta_\theta\right]}{\delta_{\theta^*}}\right),$$

or

$$\mathcal{L}\approx\frac{1}{2}\mathbb{E}_\lambda\left[\frac{\left(\frac{\mu_\theta}{\mathbb{E}_\lambda[\delta_\theta]}-1\right)^2}{\mu_\theta-1}\mathbb{E}_\lambda\left[\delta_\theta\right]\frac{\mathbb{E}_\lambda\left[\delta_\theta\right]}{\mu_\theta}\right]+\frac{1}{2}\lambda_{\theta^*}\frac{\mu_{\theta^*}}{\delta_{\theta^*}}\gamma_{\theta^*}\left(\mathbb{E}_\lambda\left[\delta_\theta\right]-\delta_{\theta^*}\right)^2,$$

or

$$\mathcal{L}\approx\frac{1}{2}\mathbb{E}_\lambda\left[\frac{\mu_\theta}{\mu_\theta-1}\left(\frac{\mu_\theta}{\mathbb{E}_\lambda\left[\delta_\theta\right]}-1\right)^2\right]+\frac{1}{2}\lambda_{\theta^*}\gamma_{\theta^*}\left(\mathbb{E}_\lambda\left[\delta_\theta\right]-\delta_{\theta^*}\right)^2,$$

or

$$\mathcal{L}\approx\frac{1}{2}\mathbb{E}_\lambda\left[\sigma_\theta\left(\frac{\mu_\theta}{\mathbb{E}_\lambda\left[\delta_\theta\right]}-1\right)^2\right]+\frac{1}{2}\lambda_{\theta^*}\gamma_{\theta^*}\left(\mathbb{E}_\lambda\left[\delta_\theta\right]-\delta_{\theta^*}\right)^2,$$

or

$$\mathcal{L} \approx \frac{1}{2}\mathbb{E}_\lambda\left[\sigma_\theta\left[\left(\frac{\mu_\theta}{\mathbb{E}_\lambda\left[\delta_\theta\right]} - \frac{\mathbb{E}_\lambda\left[\mu_\theta\right]}{\mathbb{E}_\lambda\left[\delta_\theta\right]}\right)^2 + \left(\frac{\mathbb{E}_\lambda\left[\mu_\theta\right]}{\mathbb{E}_\lambda\left[\delta_\theta\right]} - 1\right)^2\right]\right] + \frac{1}{2}\lambda_{\theta^*}\gamma_{\theta^*}\left(\mathbb{E}_\lambda\left[\delta_\theta\right] - \delta_{\theta^*}\right)^2,$$

Appendix G Additional Comparative Statics

In this section, we characterize comparative statics with respect to shocks to the fixed costs and shocks to the productivity distribution. We start with fixed cost shocks, and then examine productivity shocks.

G.1 Shocks to Fixed Costs

For simplicity, we consider the case where overhead costs are identical across firms, $f_{o,\theta} = f_o$. Proposition 5 characterizes the response of welfare to a change in fixed costs of entry and overhead costs.

Proposition 5. *In response to changes in fixed costs of entry $d\log f_e$ and fixed overhead costs $d\log f_o$, changes in consumer welfare are given by*

$$d\log Y = \underbrace{-\left(\mathbb{E}_\lambda[\delta_\theta] - 1\right)\frac{f_e d\log f_e + f_o d\log f_o}{f_e + (1-G(\theta^*))f_o}}_{\textit{technical efficiency}}$$
$$\underbrace{-\frac{\xi^\epsilon + \xi^\mu + \xi^{\theta^*}}{1-\xi^\epsilon-\xi^\mu-\xi^{\theta^*}}\left(\mathbb{E}_\lambda[\delta_\theta]\right)\frac{f_e d\log f_e + (1-G(\theta^*))f_o d\log f_o}{f_e + (1-G(\theta^*))f_o}}_{\textit{allocative efficiency}}$$
$$\underbrace{-\frac{\zeta^{\theta^*}}{1-\xi^\epsilon-\xi^\mu-\xi^{\theta^*}}\frac{f_e[d\log f_e - d\log f]}{f_e + (1-G(\theta^*))f_o}}_{\textit{allocative efficiency}},$$

where ξ^ϵ, ξ^{θ^*}, and ξ^μ are given in Theorem 1 and

$$\zeta^{\theta^*} = \left(\mathbb{E}_\lambda[\delta_\theta] - \delta_{\theta^*}\right)\left(\lambda_{\theta^*}\gamma_{\theta^*}\frac{1}{\sigma_{\theta^*}-1}\right). \tag{50}$$

To understand these results, it is useful to observe that the model is homogeneous of degree zero in fixed costs and population f_e, f_o, and L. This is because they only matter through fixed costs per capita f_e/L and f_o/L. This means that joint proportional reductions in fixed costs of entry and fixed overhead costs $d\log f_e = d\log f_o < 0$ have exactly the same effects on consumer welfare as equivalent increases in population $d\log L = -d\log f_e = -d\log f_o > 0$.

Consider first a reduction in the fixed cost of entry $d\log f_e < 0$. This reduces the total (entry

and overhead) fixed cost per entering variety in proportion to the share of the fixed cost of entry in the total fixed cost $[(f_e)/[f_e + (1 - G(\theta^*))f_o]]d \log f_e < 0$. This reduction in fixed cost acts like an equivalent increase in population coupled with an equivalent increase in the fixed overhead cost. The effect of the former was analyzed in Theorem 1. The effect of the latter is to further increase the sales shares of exiting varieties by $-[\lambda_{\theta^*}\gamma_{\theta^*}/(\sigma_{\theta^*}-1)][(f_e)/[f_e+(1-G(\theta^*))f_o]]d \log f_e > 0$. This in turn increases consumer welfare by $-[(\mathbb{E}[\delta_\theta] - \delta_{\theta^*})\lambda_{\theta^*}\gamma_{\theta^*}/(\sigma_{\theta^*} - 1)][(f_e)/[f_e + (1 - G(\theta^*))f_o]]d \log f_e > 0$ as long as there is too little selection ($\mathbb{E}_\lambda[\delta_\theta] > \delta_{\theta^*}$). The result in the proposition is obtained by solving the fixed point in $d \log Y$.

Consider now a reduction in the fixed overhead cost $d \log f < 0$. The effect on the selection cut-off is reversed compared to the case of a reduction in the fixed cost of entry: compared to an increase in population by $-[(1 - G(\theta^*))f_o/[f_e + (1 - G(\theta^*))f_o]]d \log(f_o) > 0$, the increase in the fixed overhead cost reduces the selection cut-off, which typically overcomes the increase in selection associated with the equivalent increase in population. If this is the case, the overall change in consumer welfare from the change in selection is positive if and only if there is too much selection ($\mathbb{E}_\lambda[\delta_\theta] < \delta_{\theta^*}$).

In both cases, and exactly as for population shocks, we can decompose the general equilibrium response by analyzing three successive equilibrium allocations which allow firms to adjust along more and more margins: entry, entry and exit, and entry, exit and markups. All three equilibrium allocations feature the same changes in technical efficiency, but different changes in allocative efficiency, driven by different changes in the allocation of resources. The corresponding changes in consumer welfare are respectively given by Proposition 5, but with $\xi^\mu = \xi^{\theta^*} = 0$ and $\zeta^{\theta^*} = 0$, $\xi^\mu = 0$, and without any modification.

We can also perform the same decomposition for changes in real GDP per capita.

Proposition 6. *In response to changes in fixed costs of entry $d \log f_e$ and fixed overhead costs $d \log f$, changes in real GDP per capita are given by*

$$d \log Q = \left(\mathbb{E}_\lambda\left[(1 - \rho_\theta)\right]\right)\left(\mathbb{E}_\lambda\left[\frac{1}{\sigma_\theta}\right]\right)\left(d \log Y + \frac{f_e d \log f_e + (1 - G(\theta^*))f_o d \log f_o}{f_e + (1 - G(\theta^*))f_o}\right), \tag{51}$$

where $d \log Y$ is given by Proposition 5.

G.2 Shocks to Productivity

Now, we consider shocks to the distribution of productivity shifters.

Proposition 7. *In response to changes in productivity $d \log A_\theta$, changes in consumer welfare are given by*

$$d \log Y = \underbrace{\mathbb{E}_\lambda\left[d \log A_\theta\right]}_{\textit{technical efficiency}} + \underbrace{\frac{\nu^\epsilon\left[d \log A_\theta\right] + \nu^{\theta^*}\left[d \log A_\theta\right] + \nu^\mu\left[d \log A_\theta\right]}{1 - \xi^\epsilon - \xi^\mu - \xi^{\theta^*}}}_{\textit{allocative efficiency}}$$

$$+\underbrace{\frac{\xi^{\epsilon}+\xi^{\mu}+\xi^{\theta^*}}{1-\xi^{\epsilon}-\xi^{\mu}-\xi^{\theta^*}}\left(\mathbb{E}_{\lambda(1-1/\mu)}\left[(\sigma_\theta-1)d\log A_\theta\right]+\mathbb{E}_{\lambda}\left[d\log A_\theta\right]\right)}_{\textit{allocative efficiency}},$$

where ξ^{ϵ}, ξ^{θ^*}, and ξ^{μ} are given in Proposition 1 and

$$\begin{aligned}
\nu^{\epsilon}\left[d\log A_\theta\right] &= \left(\mathbb{E}_{\lambda}\left[\delta_\theta\right]-1\right)\left(\mathbb{E}_{\lambda(1-1/\mu)}\left[(\sigma_\theta-1)d\log A_\theta\right]-\mathbb{E}_{\lambda}\left[(\sigma_\theta-1)d\log A_\theta\right]\right),\\
\nu^{\theta^*}\left[d\log A_\theta\right] &= -\left(\mathbb{E}_{\lambda}\left[\delta_\theta\right]-\delta_{\theta^*}\right)\lambda_{\theta^*}\gamma_{\theta^*}\left(\sigma_{\theta^*}d\log A_{\theta^*}-\mathbb{E}_{\lambda(1-1/\mu)}\left[\sigma_\theta d\log A_\theta\right]\right),\\
\nu^{\mu}\left[d\log A_\theta\right] &= -\left(\mathbb{E}_{\lambda}\left[(1-\rho_\theta)\left[1-\frac{\mathbb{E}_{\lambda}\left[\delta_\theta\right]-1}{\mu_\theta-1}\right]d\log A_\theta\right]\right).
\end{aligned}$$

Exactly as for shocks to population and to fixed costs, we can decompose the general equilibrium response by analyzing three successive equilibrium allocations which allow firms to adjust along more and more margins: entry, entry and exit, and entry, exit and markups. All three equilibrium allocations feature the same changes in technical efficiency given by the sales-weighted changes in productivities, exactly as in Hulten's theorem (Hulten, 1978). These three equilibrium allocations feature different changes in allocative efficiency, driven by different changes in the allocation of resources. The corresponding changes in consumer welfare are respectively given by Proposition 7, but with $\xi^{\mu}=\xi^{\theta^*}=0$ and $\nu^{\mu}[d\log A_\theta]=\nu^{\theta^*}[d\log A_\theta]=0$, $\xi^{\mu}=0$ and $\nu^{\mu}[d\log A_\theta]=0$, and without any modification.

Changes in allocative efficiency are given by the sum of two sets of terms. The first set of terms $\nu^{\epsilon}\left[d\log A_\theta\right]$, $\nu^{\theta^*}\left[d\log A_\theta\right]$, and $\nu^{\mu}\left[d\log A_\theta\right]$ captures the effects of changes in productivities $d\log A_\theta$ holding the aggregate price index $\bar{\delta}/Y$ constant. The second set of terms capture the effects of changes in the aggregate price index $d\log P=(\mathbb{E}_{\lambda(1-1/\mu)}[(\sigma_\theta-1)d\log A_\theta]+d\log Y)\mathbb{E}_{\lambda}[1/\sigma_\theta]$.

We have already discussed the effects of changes in the aggregate price index, for example in Section 4.2. We therefore focus our discussion on the effects of changes in productivities holding the aggregate price index constant. We quickly discuss the intuition for the terms $\nu^{\epsilon}\left[d\log A_\theta\right]$, $\nu^{\theta^*}\left[d\log A_\theta\right]$, and $\nu^{\mu}\left[d\log A_\theta\right]$. These terms are then amplified by a multiplier $1/[1-(\xi^{\epsilon}+\xi^{\mu}+\xi^{\theta^*})]$ arising from solving the fixed point in $d\log Y$.

The intuition for the term $\nu^{\epsilon}\left[d\log A_\theta\right]$ is the following. Productivity shocks change prices for given markups, exit behavior, and aggregate price index. The sales shares of varieties with high markups tend to increase if they experience sufficiently higher relative productivity shocks to offset their relatively lower elasticities. If they do, the variable profit share increases, which increases entry by $\mathbb{E}_{\lambda(1-1/\mu)}[(\sigma_\theta-1)d\log A_\theta]-\mathbb{E}_{\lambda}[(\sigma_\theta-1)d\log A_\theta]$ and welfare by $(\mathbb{E}_{\lambda}[\delta_\theta]-1)(\mathbb{E}_{\lambda(1-1/\mu)}[(\sigma_\theta-1)d\log A_\theta]-\mathbb{E}_{\lambda}[(\sigma_\theta-1)d\log A_\theta])$.

The intuition for the term $\nu^{\theta^*}\left[d\log A_\theta\right]$ is the following. Productivity shocks change exit behavior for given markups and aggregate price index. The selection cut-off tends to decrease if the productivity increases relatively more and if the elasticity of substitution is rela-

tively higher at the cut-off. If they do does, the sales share of exiting varieties decreases by $(\sigma_{\theta^*} d \log A_{\theta^*} - \mathbb{E}_{\lambda(1-1/\mu)}[\sigma_\theta d \log A_\theta])/(\sigma_{\theta^*} - 1)$, which changes welfare by $-(\mathbb{E}_\lambda[\delta_\theta] - \delta_{\theta^*})(\sigma_{\theta^*} d \log A_{\theta^*} - \mathbb{E}_{\lambda(1-1/\mu)}[\sigma_\theta d \log A_\theta])/(\sigma_{\theta^*} - 1)$.

The intuition for the term $\nu^\mu [d \log A_\theta]$ is the following. Productivity shocks lead to changes in markups for a given aggregate price index. Increases in productivity lead to increases in markups, which increases the variable profit share. This in turn increases entry and changes welfare by $-\mathbb{E}_\lambda[(1-\rho_\theta)[1 - [(\mathbb{E}_\lambda[\delta_\theta] - 1)/(\mu_\theta - 1)] d \log A_\theta]]$.

Signing the overall changes in allocative efficiency is difficult because of offsetting effects. For example if all productivity shocks are identical $d \log A_\theta = d \log A$, then there are no changes in allocative efficiency, since just like in the case with homogeneous firms, the model is homothetic with respect to such shocks. In this special case, the terms capturing the effects of changes in productivities given the aggregate price index exactly offset (term by term) the terms capturing the effects of changes in the aggregate price index given productivities: the terms in $\nu^\epsilon [d \log A_\theta]$ exactly offset the terms in ξ^ϵ, the terms in $\nu^{\theta^*} [d \log A_\theta]$ exactly offset the terms in ξ^{θ^*}, and the terms in $\nu^\mu [d \log A_\theta]$ exactly offset the terms in ξ^μ. This shows that changes in allocative efficiency from productivity shocks depend finely on the distribution of these shocks across types.

The response of real GDP to productivity shocks is given in Proposition 8.

Proposition 8. *In response to changes in productivities $d \log A_\theta$, changes in real GDP per capita are given by*

$$d \log Q = \mathbb{E}_\lambda \left[\rho_\theta d \log A_\theta \right] + \left(\mathbb{E}_\lambda \left[(1 - \rho_\theta) \right] \right) \left(\mathbb{E}_\lambda \left[\frac{1}{\sigma_\theta} \right] \right) \left(d \log Y + \mathbb{E}_{\lambda(1-1/\mu)} \left[(\sigma_\theta - 1) d \log A_\theta \right] \right), \quad (52)$$

where $d \log Y$ is given by Proposition 7.

Appendix H Homothetic with a Single Aggregator (HSA) Preferences

In this appendix, we develop a version of our results using an alternative demand system to the generalized Kimball preferences we use in the main text. We use homothetic demand with a single aggregator (HSA) preferences, as defined by Matsuyama and Ushchev (2017). These preferences nest separable translog preferences and linear expenditure shares as special cases. The CES demand system is the only point of union between HSA preferences and the generalized Kimball preferences used in the main text. Nevertheless, our theoretical and quantitative results are quite similar when we use HSA preferences instead.

This appendix is organized as follows. In Section H.1, we set up the consumer and firm problems and describe firm elasticities, markups, pass-throughs, and consumer surplus ratios in terms of primitives. In Section H.2, we present theoretical results analogous to Theorem 1

and Proposition 1 in the main text. Finally, we show that the system of differential equations used to calibrate the model remain valid under HSA preferences and provide quantitative results analogous to Table 1 and Table 2. The results are qualitatively and quantitatively similar to those in the main text.

H.1 Setup

Under HSA preferences, the per-capita quantity y_θ consumed of a variety θ is:

$$y_\theta = \frac{w}{p_\theta} s_\theta\left(\frac{p_\theta}{P}\right), \tag{53}$$

where p_θ is the price of the variety, $s_\theta(\cdot)$ are the expenditures on variety θ as a fraction of the consumer's budget, and P is the price aggregator. As in the main text, we have anticipated the fact that free entry will force firm profits to zero in equilibrium, and we normalize the wage $w = 1$.

The price aggregator P is implicitly defined so that expenditure shares add to one:

$$\int_\Theta s_\theta(\frac{p_\theta}{P}) dF(\theta) = 1. \tag{54}$$

We assume there exists some choke constant $(p/P)^{\max}$, such that for any $p_\theta/P \geq (p/P)^{\max}$, $s_\theta(\frac{p_\theta}{P}) = 0$. The relationship between the ideal price index, P^Y, and the price aggregator P, is

$$\log P^Y = \log P - \int_\Theta \left[\int_{p_\theta/P}^{(p/P)_{\max}} \frac{s_\theta(\xi)}{\xi} d\xi \right] dF(\theta). \tag{55}$$

Again, consumers maximize welfare Y under the budget constraint,

$$\int_{\theta \in \Theta} p_\theta y_\theta dF(\theta) = P^Y Y = 1. \tag{56}$$

The firm side of the economy remains exactly the same as in the main text: upon entry, firms draw a type θ from a distribution with density $g(\theta)$ and cumulative density function $G(\theta)$. Each firm then decides whether to operate, and if so, what price to charge. The firm's maximization problem is

$$\max_{\text{operate}, p_\theta} \begin{cases} \left(p_\theta - \frac{1}{A_\theta}\right) L y_\theta - f_{o,\theta} & \text{if the firm operates} \\ 0 & \text{if the firm does not operate} \end{cases} \tag{57}$$

subject to the household per-capita demand curve in (53).

For firms that operate, the price that maximizes firm profits can be written as a markup μ_θ times the firms marginal cost, where the markup is given by the Lerner formula,

$$\mu_\theta(\frac{p}{P}) = \frac{1}{1 - \frac{1}{\sigma_\theta(\frac{p}{P})}}, \tag{58}$$

and the price-elasticity of demand is given by,

$$\sigma_\theta(\frac{p}{P}) = \frac{-d\log y_\theta}{d\log p_\theta} = 1 - \frac{\frac{p_\theta}{P}s'_\theta\left(\frac{p_\theta}{P}\right)}{s_\theta\left(\frac{p_\theta}{P}\right)}. \tag{59}$$

Firms are ordered by the ratio X_θ of variable profits to overhead costs, so there is an endogenous cutoff type θ^* such that

$$\left(p_{\theta^*} - \frac{1}{A_{\theta^*}}\right)Ly_{\theta^*} = f_{o,\theta^*}, \tag{60}$$

firms with types $\theta \geq \theta^*$ operate, and firms with types $\theta < \theta^*$ exit the market. Free entry leads expected profits to be equal to entry costs in equilibrium,

$$\int_{\theta^*}^{\infty}\left[\left(1 - \frac{1}{\mu_\theta}\right)p_\theta y_\theta wL - f_{o,\theta}\right]g(\theta)d\theta = f_e. \tag{61}$$

We use the set Θ to denote types that operate in equilibrium: $\Theta = \{\theta|\theta \geq \theta^*\}$. We use M to denote the mass of entrants, so that the mass of surviving firms is $(1 - G(\theta^*))M$. Accordingly, the density of varieties available to the consumer $dF(\theta) = Mg(\theta)d\theta$.

We will use the same definitions of pass-throughs and consumer surplus ratios as in the main text. In terms of primitives, the pass-through and the consumer surplus ratio are now

$$\rho_\theta(\frac{p}{P}) = \frac{1}{1 - \frac{\frac{p}{P}\mu'_\theta(\frac{p}{P})}{\mu_\theta(\frac{p}{P})}}, \qquad \text{and} \qquad \delta_\theta = 1 + \frac{1}{s_\theta(\frac{p}{P})}\int_{p/P}^{(p/P)_{\max}}\frac{s_\theta(\xi)}{\xi}d\xi. \tag{62}$$

The sales density is defined as $\lambda_\theta = s_\theta(\frac{p_\theta}{P})M(1 - G(\theta^*))$. We denote the sales-weighted average consumer surplus ratio $\bar{\delta} = \mathbb{E}_\lambda[\delta_\theta]$ and the harmonic (sales-weighted) average of markups $\bar{\mu} = \mathbb{E}_\lambda[\mu_\theta^{-1}]^{-1}$.

In equilibrium, consumers maximize utility, firms maximize profits, and resource constraints are satisfied. The equilibrium is defined by the consumer's demand for each variety (53), the implicit definition of the price aggregator (54), the relationship of the price aggregator to the ideal price index (55), firms' profit-maximizing markups (58), the selection cutoff (60), and the free entry condition (61).

H.2 Response to Change in Market Size

Theorem 2 characterizes the change in welfare following an exogenous change in market size under HSA preferences.

Theorem 2. *In response to changes in population $d\log L$, changes in consumer welfare are given by*

$$d\log Y = \underbrace{(\bar{\delta}-1)d\log L}_{\text{technical efficiency}} + \underbrace{\bar{\mu}\left(\xi^{\epsilon}+\xi^{\theta^*}+\xi^{\mu}\right)d\log L}_{\text{allocative efficiency}}, \tag{63}$$

where

$$\begin{aligned}
\xi^{\epsilon} &= (\bar{\delta}-1)\,Cov_{\lambda}\left[\sigma_{\theta},\frac{1}{\mu_{\theta}}\right],\\
\xi^{\theta^*} &= (\bar{\delta}-\delta_{\theta^*})\,\lambda_{\theta^*}\gamma_{\theta^*}\left(\mathbb{E}_{\lambda}\left[\frac{\sigma_{\theta^*}}{\sigma_{\theta}}\right]-1\right),\\
\xi^{\mu} &= \mathbb{E}_{\lambda}\left[(1-\rho_{\theta})\,\sigma_{\theta}\left(1-\frac{\bar{\delta}}{\mu_{\theta}}\right)\right]\mathbb{E}_{\lambda}\left[\frac{1}{\sigma_{\theta}}\right].
\end{aligned}$$

Compared to the results under generalized Kimball preferences in the main text, the change in technical efficiency following a change in market size is the same, but the change in allocative efficiency is somewhat different. Note, however, that the change in allocative efficiency depends on the same three margins of adjustment: the Darwinian margin (ξ^{ϵ}), the selection margin (ξ^{θ^*}), and pro/anti-competitive (ξ^{μ}). The terms ξ^{ϵ}, ξ^{θ^*}, and ξ^{μ}, are exactly as defined in the main text. For a given collection of $\xi^{\varepsilon},\xi^{\theta^*},\xi^{\mu}$, the generalized Kimball model will generate stronger reallocation effects as long as $\xi^{\varepsilon}+\xi^{\theta^*}+\xi^{\mu}\in[0,1]$. Intuitively, this is because Kimball preferences feature a feedback loop from increases in Y driving reductions in P and reductions in P driving increases in Y. HSA preferences lack this feedback loop. Quantitatively however, we find very similar results when we calibrate the HSA version of the model.

Proposition 9 describes the response of real GDP to a change in market size.

Proposition 9. *In response to changes in population $d\log L$, changes in real GDP per capita are given by*

$$d\log Q = \mathbb{E}_{\lambda}\left[1-\rho_{\theta}\right]\mathbb{E}_{\lambda}\left[\frac{1}{\sigma_{\theta}}\right]\bar{\mu}d\log L. \tag{64}$$

Proof. In response to an exogenous change in market size $d\log L$, the following system of log-linearized equations describe the movements of all endogenous variables.

$$\mathbb{E}_{\lambda}\left[(1-\sigma_{\theta})\right]d\log P = d\log M - \lambda_{\theta^*}\frac{g(\theta^*)}{1-G(\theta^*)}d\theta^* + \mathbb{E}_{\lambda}\left[(1-\sigma_{\theta})\,d\log p_{\theta}\right]$$

$$d\log y_{\theta} = -\sigma_{\theta}d\log\frac{p_{\theta}}{P} - d\log P.$$

$$d\log Y = (\bar{\delta}-1)\,d\log M - \lambda_{\theta^*}\left(\delta_{\theta^*}-1\right)\frac{g(\theta^*)}{1-G(\theta^*)}d\theta^* - \mathbb{E}_{\lambda}\left[d\log p_{\theta}\right].$$

$$d\log\mu_\theta = \frac{\rho_\theta - 1}{\rho_\theta} d\log\left(\frac{p_\theta}{P}\right).$$

$$d\log X_\theta = (\sigma_\theta - 1)\, d\log p_\theta + d\log\lambda_\theta.$$

$$d\log\lambda_\theta = d\log p_\theta + d\log y_\theta + \frac{-g(\theta^*)}{1 - G(\theta^*)} d\theta^* + d\log M.$$

$$d\log X_{\theta^*} + \frac{1}{\gamma_{\theta^*}} \frac{g(\theta^*)}{1 - G(\theta^*)} d\theta^* = \frac{-g(\theta^*)}{1 - G(\theta^*)} d\theta^* + d\log M - d\log L.$$

$$d\log L + \frac{g(\theta^*)}{1 - G(\theta^*)} d\theta^* - d\log M + \mathbb{E}_{\lambda\left(1-\frac{1}{\mu}\right)}\left[d\log X_\theta\right] = 0.$$

The first two equations, which describe the change in the price aggregator and the change in the consumption of individual varieties, are different from the analogous equations under generalized Kimball preferences, since the consumer demand curve and the price aggregator are now different. The remaining equations are unchanged from the derivation under generalized Kimball preferences.

Solving the fixed point of this system yields Theorem 2 and Proposition 9.

H.3 Calibration

For calibration, we impose the restriction that the expenditure function is identical across types, $s_\theta(\cdot) = s(\cdot)$. We also assume that overhead costs are homogenous across firms, $f_{o,\theta} = f_o$, so that the sole source of exogenous variation across firm types is due to differing productivities A_θ. Under this restriction, we can use the cross-sectional variation in pass-throughs and sales shares to solve for markups and consumer surplus ratios, up to boundary conditions.

The same differential equations used to solve for markups and consumer surplus ratios in the Kimball case apply under HSA preferences. To see why, note that the markups and sales-shares vary with productivity according to:

$$\frac{d\log\mu_\theta}{d\theta} = (1 - \rho_\theta) \frac{d\log A_\theta}{d\theta}, \tag{65}$$

$$\frac{d\log\lambda_\theta}{d\theta} = \frac{\rho_\theta}{\mu_\theta - 1} \frac{d\log A_\theta}{d\theta}. \tag{66}$$

Rearranging yields the differential equation,

$$\frac{d\log\mu_\theta}{d\theta} = (\mu_\theta - 1) \frac{1 - \rho_\theta}{\rho_\theta} \frac{d\log\lambda_\theta}{d\theta}, \tag{67}$$

from which we solve for markups up to a boundary condition using pass-throughs and sales shares.

For consumer surplus ratios, recall that we can write

$$\int_{p_\theta/P}^{(p/P)_{\max}} \frac{s(\xi)}{\xi} d\xi = s(\frac{p_\theta}{P}) \left[\delta(\frac{p_\theta}{P}) - 1 \right]. \tag{68}$$

Differentiating both sides and rearranging, we find a differential equation relating consumer surplus ratios to markups,

$$\frac{d \log \delta_\theta}{d\theta} = \frac{\mu_\theta - \delta_\theta}{\delta_\theta} \frac{d \log \lambda_\theta}{d\theta}, \tag{69}$$

which we use to solve for consumer surplus ratios up to a boundary condition. Since both differential equations are identical to those derived under Kimball preferences in the main text, the estimates of sufficient statistics are unchanged.

Table 12 shows the elasticity of welfare and real GDP per capita to market size. The elasticity of welfare to market size is further decomposed into changes in technical and allocative efficiency, including the three margins of adjustment (entry, exist, and markups) discussed in the main text. The results are quantitatively similar to those in the main text. In particular, the majority of gains from an increase in market size are due allocative efficiency effects arising from entry; the selection and pro-competitive channels have zero or mildly deleterious effects on welfare.

	$\bar{\mu} = 1.045$		$\bar{\mu} = 1.090$	
	$\bar{\delta} = \delta_{\theta^*}$	$\bar{\delta} = \bar{\mu}$	$\bar{\delta} = \delta_{\theta^*}$	$\bar{\delta} = \bar{\mu}$
Welfare: $\mathrm{d} \log Y$	0.122	0.137	0.253	0.283
Technical efficiency: $\mathrm{d} \log Y^{tech}$	0.017	0.045	0.034	0.090
Allocative efficiency: $\mathrm{d} \log Y^{alloc}$	0.105	0.091	0.219	0.192
Darwinian effect: $\mathrm{d} \log Y^{\epsilon} - \mathrm{d} \log Y^{tech}$	0.108	0.294	0.228	0.613
Selection effect: $\mathrm{d} \log Y^{\epsilon,\theta^*} - \mathrm{d} \log Y^{\epsilon}$	0.000	-0.157	0.000	-0.325
Pro-competitive effect: $\mathrm{d} \log Y^{\epsilon,\theta^*,\mu} - \mathrm{d} \log Y^{\epsilon,\theta^*}$	-0.003	-0.046	-0.008	-0.095
Real GDP per capita	0.022	0.022	0.043	0.043

Table 12: The elasticity of welfare and real GDP per capita to population following Theorem 2.

Table 13 replicates the analysis in a setting with homogeneous firms. Again, firm heterogeneity appears to play a significant role. Without heterogeneity, we find that the elasticity of welfare to changes in market size are much smaller than in the calibration with heterogeneous firms.

	$\bar{\mu} = 1.045$		$\bar{\mu} = 1.090$	
	$\bar{\delta} = \delta_{\theta^*}$	$\bar{\delta} = \bar{\mu}$	$\bar{\delta} = \delta_{\theta^*}$	$\bar{\delta} = \bar{\mu}$
Welfare: $\mathrm{d} \log Y$	0.030	0.045	0.060	0.090
Technical efficiency: $\mathrm{d} \log Y^{tech}$	0.017	0.045	0.034	0.090
Allocative efficiency: $\mathrm{d} \log Y^{alloc}$	0.014	0.000	0.027	0.000
Real GDP per capita	0.022	0.022	0.043	0.043

Table 13: The elasticity of welfare and real GDP per capita to population for homogeneous firms.

Appendix I Klenow-Willis Calibration

In the main text, we caution that using an off-the-shelf functional form may mute important features of the data. As an illustration, we present the results of our model using Klenow and Willis (2016) preferences, a parametric form for the Kimball aggregator that is used often in the literature. We show that Klenow and Willis (2016) preferences are unable to match the empirical data. When calibrated using standard parameters from the literature, these preferences overstate the importance of technical efficiency changes and understate the importance of allocative efficiency changes.

Under Klenow and Willis (2016) preferences, the markup and pass-through functions are

$$\mu_\theta = \mu(\frac{y_\theta}{Y}) = \frac{1}{1 - \frac{1}{\sigma}(\frac{y_\theta}{Y})^{\frac{\epsilon}{\sigma}}}, \tag{70}$$

$$\rho_\theta = \rho(\frac{y_\theta}{Y}) = \frac{1}{1 + \frac{\epsilon}{\sigma - (\frac{y_\theta}{Y})^{\frac{\epsilon}{\sigma}}}} = \frac{1}{1 + \frac{\epsilon}{\sigma}\mu_\theta}. \tag{71}$$

where the parameters σ and ϵ are the elasticity and superelasticity (i.e., the rate of change in the elasticity) that firms would face in a symmetric equilibrium. This functional form imposes a maximum output of $(y_\theta/Y)^{\max} = \sigma^{\frac{\sigma}{\epsilon}}$, at which markups approach infinity.

These preferences are unable to match the empirical distribution of firm pass-throughs without counterfactually large markups. To see why, note that the pass-through function $\rho(\cdot)$ is strictly decreasing, and that the maximum pass-through admissible (for a firm with $y_\theta/Y = 0$) is

$$\rho^{\max} = \frac{1}{1 + \epsilon/\sigma}. \tag{72}$$

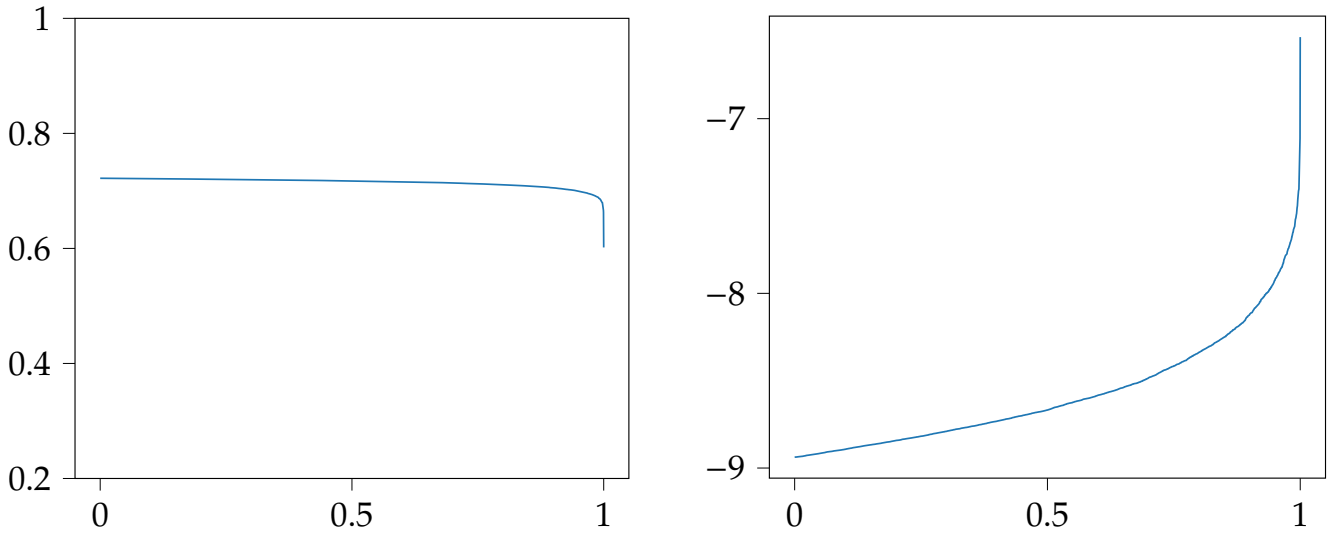
Amiti et al. (2019) estimate the average pass-through for the smallest 75% of firms in ProdCom is 0.97. In order to match the nearly complete pass-through for small firms, we must choose ϵ/σ to be around $0.01 - 0.03$.

This makes it difficult, however, to match the incomplete pass-throughs estimated for the

largest firms. To match a pass-through of $\rho_\theta = 0.3$ with $\epsilon/\sigma \in [0.01, 0.03]$, for example, we need a markup of $\mu_\theta \in [78, 233]$ for the largest firms. In contrast, our non-parametric procedure matches the pass-through distribution with realistic markups of around 2 for the largest firms (shown in the main text, Figure 4a). This roughly accords with estimates of markups by De Loecker et al. (2020).

Rather than attempting to match the empirical pass-through distribution, suppose we used a set of parameters from the literature. We adopt the calibration from Appendix D of Amiti et al. (2019): $\sigma = 5$, $\epsilon = 1.6$, and firm productivities are drawn from a Pareto distribution with shape parameter equal to 8.[37] The simulated distributions of firm pass-throughs and sales shares are shown in Figure 8. Over the range of drawn productivities, we see little variation in pass-through.

Figure 8: Pass-through ρ_θ and sales share density $\log \lambda_\theta$ under Klenow and Willis (2016) preferences.



(a) Pass-through ρ_θ

(b) Log sales share $\log \lambda_\theta$

Table 14 shows the response of welfare and real GDP per capita to an increase in market size for Klenow and Willis (2016) preferences, with the results from the main text for comparison. We find that the calibration of Klenow and Willis (2016) preferences attributes nearly all gains to technical efficiency gains, rather than allocative efficiency gains. In particular, the parametric preferences dramatically understate the importance of the Darwinian channel.

[37] We calibrate the model by drawing 10,000 firms and finding a fixed point in output. Since the Pareto distribution is unbounded, we could theoretically draw firms with zero pass-throughs and infinite sales shares; the simulated distributions are bounded away from these extremes.

	Benchmark $\bar{\mu} = 1.090$		Klenow-Willis
	$\bar{\delta} = \delta_{\theta^*}$	$\bar{\delta} = \bar{\mu}$	
Welfare: $\mathrm{d}\log Y$	0.293	0.323	0.276
Technical efficiency: $\mathrm{d}\log Y^{tech}$	0.034	0.090	0.271
Allocative efficiency: $\mathrm{d}\log Y^{alloc}$	0.260	0.233	0.004
Darwinian effect: $\mathrm{d}\log Y^{\epsilon} - \mathrm{d}\log Y^{tech}$	0.272	1.396	0.019
Selection effect: $\mathrm{d}\log Y^{\epsilon,\theta^*} - \mathrm{d}\log Y^{\epsilon}$	0.000	-1.006	-0.004
Pro-competitive effect: $\mathrm{d}\log Y^{\epsilon,\theta^*,\mu} - \mathrm{d}\log Y^{\epsilon,\theta^*}$	-0.012	-0.157	-0.011
Real GDP per capita	0.051	0.052	0.073

Table 14: Comparison of the elasticity of welfare and real GDP per capita to population in the benchmark and Klenow and Willis (2016) calibrations.

Appendix J Real GDP via a Quantity Index

In a neoclassical setting (without non-convexities), real GDP can in principle be measured in two equivalent ways, either using a Divisia quantity index or a Divisia price index. In this model, since new goods enter with finite sales, this breaks the equivalence between the two indices. The price index is the definition we adopt in the body of the paper, however, for completeness, we also discuss the quantity index. The quantity index measures the change in individual quantities at constant prices

$$d\log Q^q = \mathbb{E}_\lambda[d\log y_\theta]. \tag{73}$$

This is equal to

$$d\log Q^q = -d\log M + \lambda_{\theta^*}\frac{g(\theta^*)}{1-G(\theta^*)}d\theta^* + \mathbb{E}_\lambda\left[d\log(\frac{A_\theta}{\mu_\theta})\right], \tag{74}$$

The two notions of changes in real GDP per capita differ. For the rest of this section, denote the price-index notion (that we use in the body of the paper) using $d\log Q^p$: this is the change in real GDP per capita measured at constant quantities (more precisely, the price index is measured at constant quantities, and then changes in real GDP are defined to be changes in nominal GDP deflated by the price index). Changes in real GDP per capita measured with quantities $d\log Q^p$ depend only on changes in prices $d\log(p_\theta/w) = d\log(\mu_\theta/A_\theta)$. For given prices $p_\theta/w = \mu_\theta/A_\theta$, they do not depend on the allocation of spending between new, existing, and disappearing varieties. By contrast, changes in real GDP measured with quantities do depend on the allocation of spending for given prices. In fact, $d\log Q^q$ penalizes new product creation since the quantity of new products produced is not included in the measure, but the reduction in the quantity of existing products is included. The reduction in the quantity of existing products comes about from the fact that, in

order to produce new products, less of the old products must be produced.

Since real GDP measured at constant prices has a physical interpretation, we can write real GDP per capita measured with quantities $Q^q(\mathcal{A}, \mathcal{X})$:[38]

$$d \log Q^q = \underbrace{\frac{\partial \log Q^q}{\partial \log \mathcal{A}} d \log \mathcal{A}}_{\text{technical efficiency}} + \underbrace{\frac{\partial \log Q^q}{\partial \mathcal{X}} d\mathcal{X}}_{\text{allocative efficiency}} . \tag{75}$$

Note that changes in allocative efficiency are different for consumer welfare $d \log Y$ and for changes in real GDP per capita at constant prices $d \log Q^q$. Changes in allocative efficiency are changes in the object of interest originating in reallocation effects. It is therefore natural that they depend on the object of interest.

Proposition 10. *In response to changes in population $d \log L$, changes in real GDP per capita are*

$$d \log Q^q = \underbrace{-d \log L}_{\textit{technical efficiency}} + \underbrace{\left(1 - \mathbb{E}_\lambda \left[\rho_\theta \sigma_\theta\right] \mathbb{E}_\lambda \left[\frac{1}{\sigma_\theta}\right]\right)\left(d \log Y + d \log L\right)}_{\textit{allocative efficiency}}, \tag{76}$$

where $d \log Y$ is given by Theorem 1.

We can apply the same decomposition as above into three different equilibrium allocations incorporating more and more margins of adjustment: entry, entry and exit, and entry, exit and pricing/markups. The corresponding changes in real GDP per capita are respectively given by Proposition 1, but setting $\xi^\mu = \xi^{\theta^*} = 0$ and $\rho_\theta = 1$ (which holds fixed markups and the cut-offs), $\xi^\mu = 0$ and $\rho_\theta = 1$ (which holds fixed markups but allows the cut-off to adjust), and without any modification (allowing all margins to adjust).

For changes in real GDP per capita, it is actually even more interesting to study this decomposition in reverse order, because of the more central role played by pricing/markups in the evolution of these variables. This means incorporating more and more margins of adjustment as follows: pricing/markups, pricing/markups and exit, and pricing/markups, entry and exit. The corresponding changes in real GDP per capita are respectively given by Proposition 1, but with $\xi^\epsilon = \xi^{\theta^*} = 0$, $\xi^\epsilon = 0$, and without any modification. For example, under assumptions (1), (2), and (3), changes in real GDP per capita measured with prices increase as more and more margins of adjustment are incorporated.

[38] However, no such representation is available for real GDP measured with prices Q^p.