The Darwinian Returns to Scale

David Rezza Baqaee
UCLA

Emmanuel Farhi*
Harvard

May 11, 2020

Abstract

How does an increase in the size of the market due to fertility, immigration, or trade integration, affect welfare and GDP? We study this question using a model with heterogeneous firms, fixed costs, and monopolistic competition. We show how the shape of residual demand curves endogenously determines the variation of markups with firm size and the inefficiency of the decentralized equilibrium. We decompose changes in welfare from increased scale into changes in technical efficiency and changes in allocative efficiency due to reallocation. We non-parametrically identify residual demand curves with firm-level data and, using these estimates, quantify our theoretical results. We find that somewhere between 70% to 90% of the aggregate returns to scale are due to changes in allocative efficiency. In bigger markets, competition endogenously toughens and triggers Darwinian reallocations: big firms expand, small firms shrink and exit, and new firms enter. However, important as they are, the improvements in allocative efficiency are not primarily driven by oft-emphasized reductions in markups or deaths of unproductive firms. Instead, they are mostly caused by a composition effect that reallocates resources from low-markup to high-markup firms. Our analysis implies that the aggregate return to scale is an endogenous outcome shaped by frictions and market structure and likely varies with time, place, and policy. Furthermore, even mild increasing returns to scale at the micro level can give rise to large increasing returns to scale at the macro level.

Keywords: Increasing Returns to Scale, Monopolistic Competition, Heterogenous Firms, Endogenous Markups, Incomplete Pass-Throughs, Misallocation.

*Emails: baqaee@econ.ucla.edu, efarhi@harvard.edu. 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, and Marc Melitz, for useful comments. Emmanuel Farhi acknowledges research financial support from the Ferrante fund at Harvard University.

1 Introduction

Aggregate increasing returns to scale underlie some of the most fundamental problems in economics, ranging from the mechanics of growth, to the gains from trade, to the benefits from industrial and competition policy. Despite their importance, the origins and magnitudes of these forces remain mysterious and controversial.

We argue that to a large extent increasing returns to scale at the aggregate level reflect changes in allocative rather than technical efficiency. This implies that the aggregate return to scale is not an exogenous technological primitive. Instead, it is an endogenous outcome, affected by frictions and market structure and likely varying with time, place, and policy. Furthermore, even mild increasing returns at the micro level can catalyze large increasing returns at the macro level. This papers fleshes out the theory behind this hypothesis and provides empirical evidence supporting it.

We consider economies with the sort of realistic ingredients that give rise to aggregate increasing returns to scale: fixed costs, entry and exit, and monopolistic competition between heterogeneous firms. We then study how welfare and real GDP change in response to an increase in the population which could be due to immigration, fertility, or trade integration. Because of technological economies of scale, increases in population raise individual welfare even when the allocation of resources is held constant, since the fixed costs are now spread over a larger population. However, the change in population also triggers reallocations of resources, and these reallocations could further increase welfare. It is these endogenous changes in allocative efficiency that we focus on in this paper.[1]

When the initial equilibrium is efficient, consumer welfare behaves like the solution to a planning problem. The logic of the envelope theorem then implies that the resulting changes in welfare come only from changes in technical efficiency and not from changes in allocative efficiency. Precisely because the marginal benefit of any input is equated across competing uses, reallocations of resources have no first-order effect on welfare.

For example, classic disaggregated models of increasing returns which emphasize the importance of reallocations, such as Melitz (2003), actually have efficient equilibria because they assume Constant-Elasticity-of-Substitution (CES) demand systems. In these models, reallocations may take place, but they are irrelevant for welfare.[2] Therefore, in such settings, we can understand how a change in population affects welfare by studying only the technological aspects of the problem which are the only ones relevant to a social planner.

[1] We define changes in allocative efficiency in response to a shock as the changes in welfare or real GDP that can be attributed to the equilibrium reallocation of resources caused by the shock. See footnote 15 for more details and see Baqaee and Farhi (2019) for a formal discussion.

[2] A similar remark applies to Hopenhayn (1992), who assumes perfect competition and decreasing returns at the margin at the producer level. This model also has efficient equilibria and, as a result, equilibrium reallocations are irrelevant for welfare.

However, the CES demand system is unique in delivering efficient equilibria under monopolistic competition, and unfortunately, its theoretical elegance is firmly rejected by empirical evidence. In particular, CES demand imposes constant and uniform markups with complete pass-through of marginal costs into prices. The data, on the other hand, features substantial hetereogeneity in both markups and pass-throughs. Therefore, matching the empirical heterogeneity of markups and pass-throughs requires deviating from the CES benchmark. This in turn necessitates confronting inefficiencies and changes in allocative efficiency due to reallocations. Doing so typically leads to the conclusion that increases in market size intensify competition amongst firms. These Darwinian forces trigger reallocations along several margins, and because the economy is inefficient to begin with, these reallocations impact welfare. In other words, increasing returns to aggregate scale fundamentally entangle changes in technical and allocative efficiency.

In this paper, we analyze aggregate returns to scale in a model with monopolistic competition, fixed entry and overhead costs, entry and exit, and heterogeneous productivities. We use the flexible non-parametric Kimball (1995) demand system, which can generate any downward-sloping residual demand curve while remaining tractable. This flexibility is important for matching the empirical patterns of markups and pass-throughs.

We propose a new approach for decomposing and interpreting comparative statics in this type of model. We characterize changes in technical and allocative efficiency separately. We decompose changes in allocative efficiency into the different reallocations driven by adjustments along different margins of firm behavior. We quantify this decomposition by non-parametrically estimating the model using firm-level data.

We find that changes in allocative efficiency are more important than changes in technical efficiency in determining aggregate increasing returns to scale since they account for between 70% and 90% of the overall effect. Furthermore, we show the typical mechanisms for reallocation emphasized in the previous literature, such as the toughening of selection and reductions in markups, are not the key drivers of the improvements in allocative efficiency. On the contrary, we observe that these particular mechanisms can easily be harmful. Instead, we identify a different reallocation channel, from firms with low markups to firms with high markups, that unambiguously improves welfare and that turns out to be the main source of welfare gains.[3]

To gain some intuition, consider the residual *per-capita* demand curve for a product

$$\frac{p}{P} = \Upsilon'(\frac{y}{Y}),$$

[3]This reallocation from low to high markups firms improves efficiency for the same reasons as in Baqaee and Farhi (2019). That paper showed that these reallocations can explain a significant fraction of aggregate TFP growth. This paper raises the possibility that increases in scale, perhaps driven by globalization, could be responsible for these reallocations.

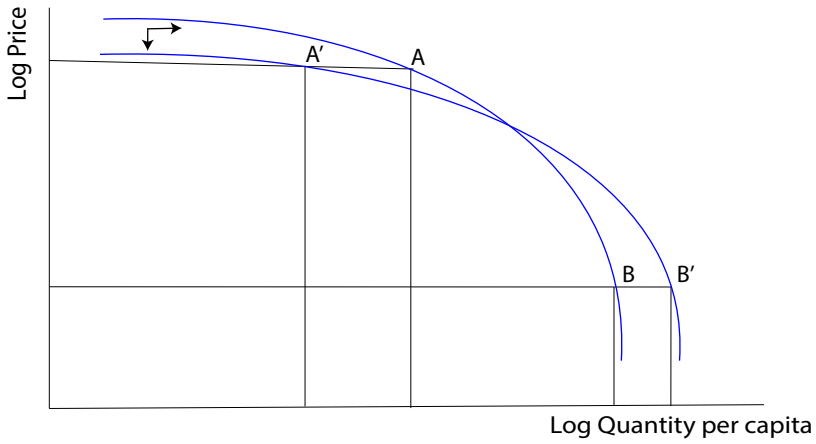


Figure 1: Illustration of the reallocation effect due to increased entry (holding fixed markups and the selection cutoff).

where p and y are the price and quantity per capita of some product variety, P is an aggregate price index, Y is an aggregate per-capita utility index, and Υ' is a decreasing function. For concreteness, suppose that larger firms face more inelastic demand curves and hence charge higher markups than smaller firms. This assumption, which is empirically justified, implies that the demand curve is log-concave (concave in log-log space), as is illustrated in Figure 1.

To understand how a change in population affects welfare, we decompose the equilibrium response using a succession of restricted equilibria. To start with, imagine how welfare changes if we hold the distribution of resources constant. That is, hold constant the share of resources dedicated to fixed versus variable costs, as well as the distribution of resources across firm types. Then an increase in population results in more varieties, and more varieties increase welfare. This is what we refer to as the improvement in technical efficiency. When the equilibrium is efficient, this effect coincides with the average micro technological return to scale, and is all we need to consider.

However, in an inefficient equilibrium, there are also changes in efficiency due to reallocations. To understand these reallocation effects, in a first step, we let firms adjust their entry behavior, but hold exit behavior and markups constant. Next in the second step, we consider the restricted equilibrium where firms can adjust their entry and exit behavior, but cannot adjust their markups. In the third and final step, we allow firms to also adjust their markups, which coincides with the full equilibrium outcome.

Consider first the equilibrium response allowing only for adjustment in entry behavior but holding exit behavior and markups constant. The increase in population shifts the demand curve down, because the aggregate price index P falls, and to the right, because the utility index rises. This means that demand per capita falls by more for the relatively elastic firms (A

to A') as compared to the inelastic ones (B to B'). In fact, the demand for the highly inelastic firms can even increase if the rightward shift is large enough (this is the case illustrated in the figure). So, resources are reallocated away from the elastic firms with low markups, towards the inelastic firms with high markups. Because the low-markup firms were too large and the high-markup ones were too small to begin with, these reallocations unambiguously increase welfare.[4] The positive changes in allocative efficiency associated with this composition effect turn out to be the largest positive force in the model when we take it to the data.

The fact that this composition effect improves welfare does not rely on log-concavity of the demand curve. For example, if the demand curve were log-convex, so that smaller firms had lower elasticities and higher markups, they would be too small to begin with, and they would be the ones to grow in response to entry.

On the other hand, the effect does rely critically on the fact that the firm size distribution is non-degenerate and that the residual demand curve is not iso-elastic. The effect would disappear If the firm-size distribution were degenerate (homogeneous firms), or if Υ' were iso-elastic and the demand curve log-linear (CES demand system).

Of course, in equilibrium, exit behavior and markups adjust as well, leading to further changes in allocative efficiency. Consistent with Asplund and Nocke (2006) and Melitz and Ottaviano (2008), we find that with log-concave demand curves, increases in population drive up the marginal/cutoff productivity level, reduce the probability of survival, and drive down markups. However, unlike the effect outlined in Figure 1, the toughening of selection and the reduction in markups have theoretically ambiguous effects on welfare. Empirically, we find they actually tend to reduce allocative efficiency.

The increase in the minimum productivity cutoff, also known as the selection effect, can easily reduce welfare. This is because the marginal firm may create more infra-marginal value for the consumer than the average firm. Hence, driving these firms out of business can hurt the consumer. Intuitively, when a firm enters, its value to the consumer is given by the average area under the demand curve. Similarly, the exit of a firm takes away value from the consumer according to the area under the demand curve at the productivity cutoff. If the latter is larger than the former, then the toughening of selection reduces allocative efficiency because there was too much selection to begin with. Empirically, we find these changes in allocative efficiency to be small and negative.

The reduction in markups, also known as the pro-competitive effect, can also turn out to be deleterious for welfare. All firms cut their markups in response to the decrease in the price index. This reduces profits and entry, which is beneficial if there is too much entry to begin with. But this is not the end of the story. In empirically relevant scenarios, high-markup firms cut their markups by more than the low-markup firms. However, low-markup firms

[4]The positive effect on welfare is mediated through entry. Basically, because they improve efficiency, these reallocations free up some labor which ends up being used in entry. Since product variety is desirable, this always improves welfare, irrespective of whether there was too much or too little entry to begin with.

face much more elastic demand curves than high-markup firms. Therefore, whether there is a reallocation away from or towards the high markup firms depends on a race between the reduction in markup and the elasticity of demand. In our empirical application, we find changes in allocative efficiency due to the change in markups to be small and negative.

Our theoretical results are non-parametric and allow the residual demand curve facing individual producers to have any downward-sloping shape. This is useful since there is no consensus on a parametric functional form for representing preferences, and our approach ensures that parametric restrictions are not inadvertently shutting down important channels. We use cross-sectional firm-level information from Belgium on pass-throughs (from Amiti et al., 2019) to non-parametrically identify household preferences, and given these estimates, we quantify how welfare changes in response to shocks. Non-parametric identification allows us the freedom to match the data in terms of both pass-throughs and sales shares, whereas typical parametric specifications of preferences, for instance CES, quadratic, or Klenow and Willis (2016) have counterfactual properties.[5]

Finally, by separately characterizing the behavior of welfare and real GDP, we also clarify some potentially confusing issues. 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. In fact, we show that changes in real GDP are entirely driven by reductions in markups, and do not depend at all on the reallocation effects that are so crucial for welfare. Therefore, intuitions that apply to real GDP cannot naively be used to understand the behavior of welfare and vice versa.

Relatedly, in the literature, aggregate TFP is sometimes conflated with an ad-hoc productivity index obtained as an average of firm productivities (for example, Melitz and Ottaviano, 2008). However, the growth in this index is not aggregate TFP growth as defined by or measured using national income statistics. We formally show that changes in ad-hoc productivity indices and changes in aggregate TFP behave in different ways. In particular, increases in ad-hoc average productivity indices due to a right-shift in the productivity cutoff do not, by themselves, lead to an increase aggregate TFP.

Many of the ideas that we develop regarding the response of the economy to changes in population apply to changes in other parameters. 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.

Finally, in addition to our comparative static results, in the appendix, we analytically characterize (to a second-order) the distance of the economy from the efficient frontier as well as the optimal combination of industrial and competition policy that restores efficiency.

[5]With monopolistic competiton, CES preferences imply that the pass-through of marginal cost into prices must be constant and equal to one. Klenow and Willis (2016) preferences imply that pass-throughs go to zero too quickly (markups increase too rapidly) for very productive firms, so that without demand shifters, it becomes impossible to match the fat right-tail of the firm size distribution (see Amiti et al., 2019, for a discussion of these issues). Quadratic preferences used by Melitz and Ottaviano (2008) force pass-throughs to start at 0.5 and decline to zero.

We also compute the distance to the frontier in our empirical application. The losses from misallocation are between 2.5% and 6.3% in Belgium.

Related Literature. This paper builds on the vast literature on entry and monopolistic competition, with its origin in the work of Chamberlin (1933) and Robinson (1933). We base our analysis on the foundation of a representative consumer with a taste for variety, following Spence (1976) and Dixit and Stiglitz (1977).

Initially, the theoretical analysis of monopolistic competition was undertaken under the assumption that firms are homogeneous, for example Krugman (1979), Mankiw and Whinston (1986), Vives (1999), or Venables (1985). The heterogeneous-firm case has been studied by Melitz (2003) when efficient, and by Asplund and Nocke (2006), Melitz and Ottaviano (2008), Zhelobodko et al. (2012), Dhingra and Morrow (2019), Mrázová and Neary (2017), and Mrázová and Neary (2019) when inefficient. For efficient models, like Melitz (2003), the envelope theorem implies that reallocations, for example the movement in the cutoff, have no direct effect on welfare to a first-order. Dhingra and Morrow (2019) is the closest paper to ours, since they also study inefficient models, but their focus is primarily on comparing the decentralized equilibrium to the first-best, and providing qualitative conditions under which the effect of market expansion can be signed.

Relative to these papers, our paper makes several contributions. First, we provide analytical non-parametric comparative statics in the second best. Second, we decompose the overall effect into technical and allocative efficiency, and further decompose changes in allocative efficiency into adjustments along different margins. This helps highlight a new mechanism, distinct from movements in the cutoff or changes in markups, through which market size affects efficiency. Third, we provide an empirical strategy for backing out demand curves from the data, allowing us to quantify our comparative statics. Fourth, we analyze real GDP as well as welfare, clarifying the similarities and differences between the object we typically measure (real GDP) and the one we care about (welfare). Finally, we provide analytical formulas for optimal policy and for the economy's distance from the efficient frontier.

Our results on optimal policy and the associated efficiency gains allow us to link up with the vast literature on cross-sectional misallocation (e.g. Restuccia and Rogerson, 2008, Hsieh and Klenow, 2009, and Baqaee and Farhi, 2019) and the welfare costs of markups (e.g. Edmond et al., 2018, Bilbiie et al., 2019, or Behrens et al., 2018). In particular, we generalize results in Baqaee and Farhi (2019) to economies with entry, exit, and endogenous markups.

The reallocation of resources towards firms with high markups, which drives most of the increasing returns to aggregate scale in the present paper, offers another point of contact with our previous work in Baqaee and Farhi (2019). There, we showed that the well-documented increase in average markups in the U.S. since 1980 is in large part driven by a composition effect whereby firms with high markups are getting larger. We also showed that these reallocations generated improvements in allocative efficiency which explained close to 50% of aggregate

TFP growth over the past twenty years. An interesting possibility, suggested by this paper, is that these reallocations are in part caused by an increase in scale, perhaps due to globalization.

Finally, although our model is static, it is related to models of endogenous growth in which increasing returns to scale can lead to sustained growth (e.g. Romer, 1986, Romer, 1990, Grossman and Helpman, 1991, Aghion and Howitt, 1992, Jones, 1995), and to the analysis of Perla et al. (2020) who find large gains from trade from increased technology adoption in a growth model with inefficiencies stemming from a technology-adoption externality.

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 describes the solution strategy and discusses efficiency. Section 4 analyzes the case where firms are homogeneous. Section 5 considers the case with heterogeneous firms. Section 6 backs out demand curves from the data and quantifies our results. Section 7 summarizes a number of extensions developed in the appendix: additional comparative static results characterizing the responses of the economy to changes in fixed costs or productivities and decomposing them into changes in technical and allocative efficiency; and an analytical characterization of optimal policy and a second-order approximation of the distance to the efficient frontier, as well as a quantification in the context of our empirical application. Section 8 concludes. The appendix contains all the proofs.

2 Model

In this section, we specify the model and describe the equilibrium.

2.1 Set Up

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 $\omega \in \mathbb{R}^+$. Consumers have homothetic Kimball (1995) preferences, with per-capita utility Y defined implicitly in units of consumption by

$$\int_0^\infty \Upsilon(\frac{y_\omega}{Y})d\omega = 1, \tag{1}$$

where y_ω is the per-capita consumption of variety ω and Υ is an increasing and concave function in units of utils with $\Upsilon(0) = 0$.[6] CES preferences are a special case of equation (1)

[6]Kimball preferences are a natural benchmark. They are flexible enough to generate individual demand curves of any shape. They are also homothetic with linear Engel curves for varieties of different sizes within a sector. We view this property as desirable given the absence of precise empirical guidance on possible deviations from it. An implication of homotheticity is that scale effects do not appear in the absence of fixed costs and entry. We have also derived similar versions of our results (available upon request) using non-homothetic separable preferences which do generate scale effects in the absence of fixed cost and entry (as in Krugman, 1979 or Dhingra and Morrow, 2019).

when Υ is a power function.

Consumers maximize their utility Y subject to the following budget constraint

$$\int_0^\infty p_\omega y_\omega d\omega = 1, \tag{2}$$

where p_ω is the price of variety ω. Normalize the nominal wage to one so that the income of each consumer is equal to one. This expression for the budget anticipates the fact that in equilibrium, there are no profits because of free entry.

The per-capita inverse-demand curve for an individual variety ω is

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

where the aggregate *price index* P is defined as

$$P = \frac{\bar{\delta}}{Y} \tag{4}$$

and the *demand index* $\bar{\delta}$ is defined as

$$\frac{1}{\bar{\delta}} = \int_0^\infty \Upsilon'(\frac{y_\omega}{Y})\frac{y_\omega}{Y} d\omega.$$

Equation (3) demonstrates the appeal of Kimball preferences — by choosing Υ, we can generate demand curves of any desired (downward-sloping) shape. Equation (3) can be thought of as a relative demand curve for y_ω/Y as a function of the relative price p_ω/P. This is the sense in which P acts like an aggregate price index for substitution.

However, we warn the reader that P does *not* coincide with the *ideal price index* P^Y for the representative consumer: it cannot be used to obtain welfare by deflating income.[7] In fact, $d\log P = d\log\bar{\delta} + d\log P^Y$. Since in general, $\bar{\delta}$ is not a constant, $d\log P \neq d\log P^Y$. It is only in the CES case, when $\Upsilon(x) = x^{(\sigma-1)/(\sigma)}$, that these two price indices coincide since then $\bar{\delta} = \sigma/(\sigma-1)$ is constant.[8] Going forward, we refer to P as "the" price index without further qualification, despite the fact that it is not the ideal price index.

This inverse demand curve can be inverted into a per-capita demand curve for an individual variety. We denote the price elasticity of this demand curve for an individual variety by

$$\sigma(\frac{y}{Y}) = \frac{\Upsilon'(\frac{y}{Y})}{-\frac{y}{Y}\Upsilon''(\frac{y}{Y})}. \tag{5}$$

[7] Let $e(p_\omega, Y)$ be the expenditure function of a household as a function of the price of all varieties p_ω and welfare Y, where the price of unavailable varieties is set to infinity. Since preferences are homothetic, we can write $e(\{p_\omega\}, Y) = P^Y(\{p_\omega\})Y$, where $P^Y(\{p_\omega\})$ is the ideal price index.

[8] Changes in the price index $d\log P$ also do not correspond to the way the deflators, such as the GDP deflator $\int_0^\infty p_\omega y_\omega d\log p_\omega d\omega$, are measured in the data.

Monopolistic competition models with Kimball demand are parsimonious in the sense that competition across varieties is mediated by the price index. They are also flexible since they allow different varieties to face different intensities of competition as captured by the different elasticities of the demand curve at different points.

Firms. Each variety is supplied by a single firm seeking to maximize profits under monopolistic competition. Firms can enter to supply new varieties by incurring a fixed entry cost of f_e units of labor. Upon entry, firms draw a type $\theta \in \mathbb{R}^+$ from a distribution with density $g(\theta)$ and cumulative distribution function $G(\theta)$. Each firm's productivity A_θ is an increasing function of its type θ. Having drawn its type, the firm then decides whether to produce or to exit. Production requires paying an overhead cost of f_o units of labor, with a constant marginal cost of $1/A_\theta$ units of labor per unit of the good sold. Finally, the firm decides what price to set, taking as given their residual demand curve.

The profit-maximizing price p_θ of a producing firm of type θ is a markup μ_θ over its marginal cost $1/A_\theta$. Its per-capita quantity y_θ is the demand at that price. The price, markup, and per-capita quantity are determined implicitly by

$$p_\theta = \frac{\mu_\theta}{A_\theta}, \quad y_\theta = y(p_\theta), \quad \text{and} \quad \mu_\theta = \mu(\frac{y_\theta}{Y}), \tag{6}$$

where the markup function is given by the usual Lerner formula

$$\mu(\frac{y}{Y}) = \frac{1}{1 - \frac{1}{\sigma(\frac{y}{Y})}}. \tag{7}$$

A firm of type θ chooses to produce if, and only if,

$$Lp_\theta y_\theta \left(1 - \frac{1}{\mu_\theta}\right) \geq f_o. \tag{8}$$

Hence, there is an endogenous cutoff θ^* such that firms of type $\theta \geq \theta^*$ decide to produce, and firms of type $\theta < \theta^*$ exit. The zero-profit condition, associated with entry, is

$$\frac{1}{\Delta} \int_{\theta^*}^{\infty} \left[Lp_\theta y_\theta \left(1 - \frac{1}{\mu_\theta}\right) - f_o w \right] g(\theta) d\theta \geq f_e. \tag{9}$$

The parameter Δ is introduced to allow the equations to represent a repeated version of the static model with an infinite number of periods $0, 1, \cdots, \infty$, where each producing firm has an exogenous probability Δ of being forced to exit in every period $t = 0, 1, \cdots, \infty$. In the absence of discounting, the expected net present value of profits is then given by the left-hand side of the entry equation.

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.

2.2 Summary of the Equilibrium Conditions

Since data is usually recorded in terms of sales rather than physical quantities, we restate the model's conditions in terms of sales. Define the sales share density

$$\lambda_\theta = (1 - G(\theta^*))Mp_\theta y_\theta, \tag{10}$$

where M is the mass of entrants and θ^* is the selection cutoff so that $(1 - G(\theta^*))M$ is the mass of surviving firms. The sales share density is such that the aggregate sales share (as a fraction of income) of firms with type in $(\theta, \theta + d\theta)$ is $\lambda_\theta g(\theta)/(1 - G(\theta^*))d\theta$. Quantities per capita y_θ and prices p_θ can all be recovered from the sales share density λ_θ and markups μ_θ:

$$y_\theta = \frac{\lambda_\theta A_\theta}{\mu_\theta(1 - G(\theta^*))M} \quad \text{and} \quad p_\theta = \frac{\mu_\theta}{A_\theta}. \tag{11}$$

It follows that all the equilibrium conditions can also be written entirely in terms of the endogenous equilibrium variables M, Y, λ_θ, μ_θ, and exogenous parameters L, f, $f_e\Delta$, and A_θ.

Using the expectation conditional on survival with measure $[g(\theta)/(1 - G(\theta^*))]1_{\{\theta \geq \theta^*\}}d\theta$, consumer welfare per capita is given implicitly by

$$1 = (1 - G(\theta^*))M\mathbb{E}\left[\Upsilon(\frac{\lambda_\theta A_\theta}{\mu_\theta(1 - G(\theta^*))MY})\right], \tag{12}$$

restating the definition in equation (1). Similarly, the free entry condition is

$$\frac{Mf_e\Delta}{L} = \mathbb{E}\left[\lambda_\theta\left(1 - \frac{1}{\mu_\theta}\right) - \frac{(1 - G(\theta^*))Mf_o}{L}\right], \tag{13}$$

restating that entry costs exactly offset the aggregate variable profit share net of the overhead costs. The selection condition is

$$\frac{(1 - G(\theta^*))Mf_o}{L} = \lambda_{\theta^*}\left(1 - \frac{1}{\mu_{\theta^*}}\right), \tag{14}$$

restating that the overhead costs exactly offset the variable profit share for the marginal producer type θ^*. The individual markup equations is

$$\mu_\theta = \mu(\frac{\lambda_\theta A_\theta}{\mu_\theta(1 - G(\theta^*))MY}). \tag{15}$$

restating the Lerner condition. Individual demand is

$$\frac{\mu_\theta}{A_\theta} = P\Upsilon'\left(\frac{\lambda_\theta A_\theta}{\mu_\theta(1-G(\theta^*))MY}\right). \tag{16}$$

Finally, the price and demand index equations are

$$P = \frac{\bar{\delta}}{Y} \tag{17}$$

and

$$\frac{1}{\bar{\delta}} = (1-G(\theta^*))M\mathbb{E}\left[\frac{\lambda_\theta A_\theta}{\mu_\theta(1-G(\theta^*))MY}\Upsilon'\left(\frac{\lambda_\theta A_\theta}{\mu_\theta(1-G(\theta^*))MY}\right)\right]. \tag{18}$$

To streamline the exposition, we have made use of the following convention. For two variable $x_\theta > 0$ and z_θ, define

$$\mathbb{E}_x[z_\theta] = \frac{\int_{\theta^*}^{\infty} x_\theta z_\theta \frac{g(\theta)}{1-G(\theta^*)} d\theta}{\int_{\theta^*}^{\infty} x_\theta \frac{g(\theta)}{1-G(\theta^*)} d\theta}. \tag{19}$$

We write $\mathbb{E}$ to denote $\mathbb{E}_x$ when $x_\theta = 1$ for all θ. The operator $\mathbb{E}_x$ operates a change of measure by putting more weight on types θ with higher values of x_θ. In the rest of the paper, we will often encounter $\mathbb{E}$, $\mathbb{E}_\lambda$, and $\mathbb{E}_{\lambda(1-1/\mu)}$, which respectively correspond to integrals with respect to the physical density, the sales share density, and the profit share density.

3 Central Concepts and Solution Strategy

In this section, we introduce some central concepts and describe the solution strategy.

3.1 Markups, Pass-Throughs, and Infra-Marginal Surplus

In order to characterize changes in consumer welfare and in real GDP, we will need the following definitions for markups, pass-throughs, and infra-marginal surplus ratios.

Markups, pass-throughs, and Marshall's second laws of demand. It is important to define a number of notions related to markups and their behavior. We have already defined the markup function $\mu(y/Y)$ and described its relation to the elasticity of the individual demand function

$$\mu\left(\frac{y}{Y}\right) = \frac{1}{1-\frac{1}{\sigma\left(\frac{y}{Y}\right)}}. \tag{20}$$

We define the individual pass-through of a variety as the elasticity of its price to its productivity $\rho = -d\log p/d\log A = 1 - d\log\mu/d\log A$. This elasticity is necessarily positive, and can be

computed by differentiating the individual demand equation $\mu(\frac{y}{Y})/A = P\Upsilon'(\frac{y}{Y})$ with respect to A, holding Y and P constant (since each firm is infinitesimal):

$$\rho(\frac{y}{Y}) = \frac{1}{1 + \frac{\frac{y}{Y}\mu'(\frac{y}{Y})}{\mu(\frac{y}{Y})}\sigma(\frac{y}{Y})}. \tag{21}$$

The markup and pass-through of a variety of type θ are denoted by $\mu_\theta = \mu(y_\theta/Y)$ and $\rho_\theta = \rho(y_\theta/Y)$.

A natural assumption is that markups are increasing with productivity. This condition is called Marshall's (weak) second law of demand and is well known to be equivalent to the requirement that the individual demand curve be log-concave (see e.g. Melitz, 2018). Given that productivity is increasing in the type of a firm, it is also equivalent to the requirement that μ_θ be increasing in θ, or equivalently,

$$\mu'(\frac{y}{Y}) \geq 0. \tag{22}$$

The illustration in Figure 1 satisfies Marshall's weak second law of demand.

Marshall's strong second law of demand is the requirement that pass-throughs be less than one and decreasing with productivity. The strong law implies the weak law, and is equivalent to the requirement that individual marginal revenue curve be log-concave. Given that productivity is increasing in the type of a firm, it is also equivalent the requirement that $\rho_\theta \leq 1$ be decreasing in θ, or equivalently

$$\rho(\frac{y}{Y}) \leq 1 \quad \text{and} \quad \rho'(\frac{y}{Y}) \leq 0. \tag{23}$$

Both the weak and strong version have empirical support (see e.g. Amiti et al., 2019) and turn out to be useful benchmarks for understanding the model. They are also verified in our empirical application. However, our theoretical analysis, while consistent with these empirical regularities, does not require them and so we do not impose them.

Infra-marginal surplus ratio and aligned preferences. We also define the infra-marginal surplus ratio of a variety δ as the amount of infra-marginal surplus per unit sales. More precisely, δ is the ratio of the consumption equivalent utility $\bar{\delta}\Upsilon(y/Y)$ from a marginal variety to its sales share $py = \bar{\delta}\Upsilon'(y/Y)y/Y$. It is given by the infra-marginal surplus ratio function

$$\delta(\frac{y}{Y}) = \frac{\Upsilon(\frac{y}{Y})}{\frac{y}{Y}\Upsilon'(\frac{y}{Y})} \geq 1. \tag{24}$$

Figure 2 gives a visual intuition for δ as the ratio of consumer surplus $A + B$ to revenues A.[9]

[9]When goods enter and exit at a choke price we naturally have $B = 0$ for these firms. In these cases, the entry-exit margin can be said to be "neoclassical" in the sense that revenues reflect consumer surplus. As can be seen from

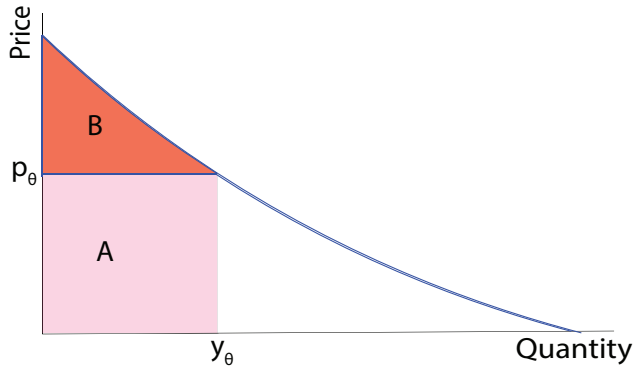


Figure 2: Graphical illustration of δ_θ.

We denote the equilibrium infra-marginal surplus ratio of a variety of type θ by $\delta_\theta = \delta(y_\theta/Y)$. There is an important relationship between the demand index $\bar{\delta}$ and the infra-marginal surplus ratios δ_θ: the former is the sales-weighted average of the latter[10]

$$\mathbb{E}_\lambda[\delta_\theta] = \bar{\delta}. \tag{25}$$

We say that (social and private) preferences are aligned if the infra-marginal consumer surplus ratio varies with productivity in the same direction as markups, or in other words if $\delta'(y/Y)$ and $\mu'(y/Y)$ have the same sign.[11] For example, if Marshall's weak second law of demand holds, so that $\mu'(y/Y) \geq 0$, preferences are aligned if, and only if, δ also increases with productivity. It is easy to see that a necessary and sufficient condition in terms of demand primitives is

$$\delta'(\frac{y}{Y}) = \frac{1 - [\sigma(\frac{y}{Y}) - 1][\delta(\frac{y}{Y}) - 1]}{\frac{y}{Y}\sigma(\frac{y}{Y})} \geq 0. \tag{26}$$

Whereas Marshall's second law of demand has empirical support, whether or not preferences are aligned has not been empirically documented one way or another.

3.2 Consumer Welfare and Real GDP Per Capita

We are interested in changes in consumer welfare per capita and real GDP per capita in response to changes in the exogenous parameters. For brevity, we simply write consumer welfare instead of consumer welfare per capita. The change in consumer welfare, measured using the equivalent variation, is

$$d \log Y. \tag{27}$$

equation (29), in such models, the equivalence between real GDP and welfare is restored.

[10]This follows from the definitions $\bar{\delta} = 1/\mathbb{E}[(1 - G(\theta^*))M\Upsilon'(y_\theta/Y)y_\theta/Y]$ and $1 = (1 - G(\theta^*))M\mathbb{E}(\Upsilon(y_\theta/Y))$.

[11]This terminology, due to Dhingra and Morrow (2019), captures the idea that when preferences are "aligned," then private gains (which are increasing in markups) are aligned with social preferences (which are increasing in the infra-marginal consumer surplus).

Changes in real GDP per capita are defined using idealized versions of the procedures used by statistical agencies. That is, using Divisia indices for continuing varieties present before and after the change.[12] Hence, the change in real GDP per capita is defined to be nominal income deflated by the GDP deflator. In other words,

$$d\log Q = -\mathbb{E}_\lambda[d\log p_\theta]. \tag{28}$$

Since the supply of the primary factor (labor) is exogenous in our model, changes in real GDP per capita $d\log Q$ are equal to changes in aggregate TFP. An important theme of this paper is that changes in welfare and changes in aggregate TFP are different objects with different determinants.

If the model did not allow for the creation and destruction of varieties, then changes in consumer welfare $d\log Y$, and changes in real GDP per capita/aggregate TFP $d\log Q$ would coincide.[13] Because the model does allow for entry and exit, changes in consumer welfare and real GDP/aggregate TFP do not coincide.[14]

More concretely,

$$d\log Y = \left(\mathbb{E}_\lambda[\delta_\theta] - 1\right) d\log M + \left(\mathbb{E}_\lambda[\delta_\theta] - \delta_{\theta^*}\right)\lambda_{\theta^*}\frac{g(\theta^*)}{1-G(\theta^*)}d\theta^* + \mathbb{E}_\lambda\left[d\log(\frac{A_\theta}{\mu_\theta})\right], \tag{29}$$

and

$$d\log Q = \mathbb{E}_\lambda\left[d\log(\frac{A_\theta}{\mu_\theta})\right]. \tag{30}$$

Intuitively, consumer welfare changes $d\log Y$ incorporate the infra-marginal 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 the expression. In contrast, changes in real GDP per capita/aggregate TFP $d\log Q$ do not, and instead only take into account changes in the intensive margin of prices $d\log p_\theta = d\log(\mu_\theta/A_\theta)$.

[12] In principle, changes in real GDP can either be defined using a quantity index or a price index. In the body of the paper, we use the price index definition. We include a discussion of the quantity index in Appendix G.

[13] As mentioned before, this is also true in models of entry devoid of non-convexities featuring no fixed costs and demand curves with choke prices, where prices and quantities at the variety level change smoothly. If new goods enter smoothly, then $\delta_\theta = 1$ for all entrants, and real GDP and aggregate TFP in equation (30) coincide with welfare in equation (1). This applies, for example, to Arkolakis et al. (2018). They consider an open economy with an export margin where export quantities vary smoothly from zero (at the choke price) because there are no fixed costs in exporting. They analyze the effect of changes in iceberg trade costs (somewhat analogous to productivity shocks) as opposed to scale shocks. In their model, entry is not affected by iceberg costs and so plays no role in their analysis. For these reasons, our results are not comparable to theirs.

[14] A related point is that aggregate TFP is not equal to an average of productivity levels across firms. For example, the sales weighted average productivity level $\mathbb{E}_\lambda[A_\theta]$ is *not* a sensible measure of aggregate TFP. Such an average is ill-defined because it is unit-dependent, and does not measure economic efficiency. To see the latter point, note that even in an efficient model, where welfare and real GDP are maximized, $\mathbb{E}_\lambda[A_\theta]$ is not being maximized and can be driven to be arbitrarily high by changing units or by introducing distorting taxes that reduce welfare and real GDP but nonetheless increase $\mathbb{E}_\lambda[A_\theta]$. See Baqaee and Farhi (2019) for more details.

3.3 Changes in Technical and Allocative Efficiency

To understand changes in consumer welfare and real GDP in response to changes in exogenous parameters, it will be useful to decompose them into changes in technical and allocative efficiency. 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 equilibrium reallocation of resources triggered by the shock. Because the economy is inefficient to begin with, these reallocations typically have nonzero first-order effects.

To make this precise, following Baqaee and Farhi (2019), we define the allocation vector $\mathcal{X} = (l_e, l_o, \{l_\theta\})$. It describes the fractions of labor allocated to the following activities: entry, overhead, and variable production of varieties of type θ. Together with the technology vector $\mathcal{A} = (L, f_e\Delta, f_o, \{A_\theta\})$, it entirely describes any feasible allocation. Let $\mathcal{Y}(\mathcal{A}, \mathcal{X})$ be the associated level of consumer welfare.

We decompose changes in consumer welfare into changes in technical and changes in allocative efficiency as

$$d\log Y = \underbrace{\frac{\partial \log \mathcal{Y}}{\partial \log \mathcal{A}} d\log \mathcal{A}}_{\text{technical efficiency}} + \underbrace{\frac{\partial \log \mathcal{Y}}{\partial \mathcal{X}} d\mathcal{X}}_{\text{allocative efficiency}}, \tag{31}$$

Changes in technical efficiency are the changes in welfare that are directly due to changes in technology $d\log\mathcal{A}$, holding the allocation of resources $\mathcal{X}$ constant. Changes in allocative efficiency are the changes in welfare that are due to the equilibrium reallocation of resources $d\mathcal{X} = (d\mathcal{X}/d\log\mathcal{A})d\log\mathcal{A}$ triggered by the shocks, holding technology $\mathcal{A}$ constant.[15]

In efficient economies, the envelope theorem implies there are only changes in technical efficiency and no changes in allocative efficiency. In inefficient economies, there are both changes in technical efficiency and changes in allocative efficiency.[16]

[15]There are different notions of changes in allocative efficiency. Each is obtained by comparing changes in welfare in equilibrium against changes in welfare under a different benchmark allocation rule. Our notion captures changes in allocative efficiency due to reallocation: the benchmark allocation rule keeps the allocation of resources constant. A different notion is the change in the distance to the efficient frontier: the benchmark allocation rule is the first-best allocation. The first notion identifies changes in allocative efficiency as changes in welfare due reallocation, holding technology constant. Since technology is held constant, welfare can only increase if the efficiency of the allocation improves. The second notion identifies changes in allocative efficiency as changes in the distance to the efficient frontier due to technology and reallocation. Both notions are used in the literature but they are different. See Baqaee and Farhi (2019) for a discussion.

[16]The breakdown between changes in technical and allocative efficiency depends on how the allocation matrix is defined. In the paper, holding the allocation matrix constant means sending the same share of labor to entry, overhead, and variable production by type. Under this definition, and recalling that $\bar{\delta} = \mathbb{E}[\delta_\theta]$, changes in technical efficiency are $(\bar{\delta} - 1)d\log L$ and reflect the entry of new varieties. An alternative notion could be to send all the new labor to variable production and hence increasing each existing producer's variable production by the same proportional amount. Under this different definition, and denoting by $\bar{\mu} = 1/\mathbb{E}[1/\mu_\theta]$ the harmonic average of markups, changes in technical efficiency would be $(\bar{\mu} - 1)d\log L$ instead of $(\bar{\delta} - 1)d\log L$ and would reflect the average across producers of the micro technological return to scale. They would be larger or smaller than under our

3.4 Solution Strategy

In the following sections, we provide analytical characterizations of first-order comparative statics of the model with respect to changes in parameters. In the main text of the paper, we focus exclusively on shocks to population. We treat shocks to other parameters, like productivity or fixed costs, in Appendix E.

Steps for comparative statics. The representation of the equilibrium in Section 2.2 allows such characterizations to be broken down into the following steps.

First, characterize changes in entry $d \log M$, markups $d \log \mu_\theta$, and the selection cutoff $d\theta^*$, as a function of changes in consumer welfare $d \log Y$, using the free-entry condition, the selection condition, the individual markup equation, the individual demand equation, and the demand index equation. Second, aggregate these changes into changes in consumer welfare $d \log Y$ using the formulas in Section 3.2. Solve the resulting fixed point. Third aggregate these changes into changes in aggregate GDP per capita $d \log Q$ using the formula in Section 3.2. Fourth, decompose these changes into changes in technical and allocative efficiency along the lines of Section 3.3.

Non-parametric sufficient statistics. These characterizations will avoid putting any additional parametric structure on the model. For example, they will not impose any specific functional form on the Kimball aggregator or the productivity distribution. Instead, they will be expressed in terms of ex-ante measurable non-parametric sufficient statistics introduced in Section 3.1: sales shares λ_θ, markups μ_θ, pass-throughs ρ_θ, and relative infra-marginal consumer surplus ratios δ_θ. They will also depend on the hazard rate $\gamma_{\theta^*} = g_a(\log A_{\theta^*})/[1 - G_a(\log A_{\theta^*})]$ of the log-productivity distribution at the selection cutoff, where $g_a(\log A_\theta) = g(\theta)/(\partial \log A_\theta/\partial\theta)$. Will make contact with the data through these sufficient statistics.

4 Homogeneous Firms

To build intuition, we start by analyzing the case where firms are homogeneous. This case is obtained by assuming that all types have the same productivity $A_\theta = A$. We denote the common markup, pass-through, and individual demand elasticity by μ, ρ, and σ.

We proceed as follows. We start by discussing the sources inefficiency. We then study shocks to population. We end by analyzing the special case of CES preferences.

To aid with the intuition, we state our results using both markups μ and elasticities σ, but the two are of course connected via $\mu = 1/(1 - 1/\sigma)$ or equivalently $\sigma = 1/(1 - 1/\mu)$.

definition depending on whether there is too much or too little entry to begin with. Furthermore, we would then have to add $(\bar{\delta} - \bar{\mu})d \log L$ to the expression for changes allocative efficiency effect, which would be smaller or larger depending on whether there is too much or too little entry to begin with. Since $\bar{\delta} - \bar{\mu}$ is small in our quantitative application, this change would be relatively inconsequential and would not alter our main message.

4.1 Sources of Inefficiency

With homogeneous firms, the only margin that can be distorted is the allocation of labor to entry (and overhead) vs. variable production. As a result, social efficiency boils down to entry efficiency. This will no longer be true with heterogeneous firms, where distortions will arise on several other margins.

The allocation matrix gives us an intuitive way to think about entry efficiency. Starting at the initial equilibrium, change the allocation of resources by increasing the fraction of labor allocated to entry and overhead and decreasing the fraction of labor allocated to variable production. Compute the resulting change in consumer welfare. We say that there is too much entry if the change in consumer welfare is negative and that there is too little entry if it is positive.[17] There is too much (too little) entry if, and only if, the following condition is verified (violated)

$$\delta < \mu. \tag{32}$$

Rearranging (26) shows this condition is automatically verified under weak second Marshall law of demand and aligned preferences.

To understand this result, we apply the following formula, which can easily be obtained by simple differentiation of the consumer welfare definition

$$d \log Y = \delta d \log M + d \log y. \tag{33}$$

The intuition for this formula is straightforward: new varieties capture a total sales share equal to $d \log M$, with an effect $\delta d \log M$ on consumer welfare; the per-capita quantity of each existing variety changes by $d \log y$, with an effect $d \log y$ on consumer welfare.

Note that the initial allocation of labor allocates a fraction $l = 1/\mu$ to variable production and $l_e + l_o = 1 - 1/\mu$ to entry and overhead. Consider a reduction in the fraction of labor allocated to variable production $d \log l < 0$ and a complementary increase in the fraction of labor allocated to entry and overhead $d \log l_e = d \log l_o = -[1/(\mu - 1)] d \log l > 0$. The change in consumer welfare is $d \log Y = [1 - (\delta - 1)/(\mu - 1)] d \log l$, which is negative (too much entry) if and only if $\delta < \mu$.

Indeed, in this experiment, we have $d \log y = d \log l - d \log M < 0$, since the amount of labor per capita available for each variety is reduced by the reallocation of labor away from variable production and by the labor required to produce the new varieties. We also have $d \log M = -1/(\mu - 1) d \log l > 0$, because of the reallocation of labor to entry and overhead. The result follows.

[17] The comparative static underlying this definition is a feasible allocation, but not an equilibrium allocation. It can only be supported as an equilibrium allocation by introducing a subsidy on entry. The defining question can then be reformulated as whether this subsidy on entry decreases or increases equilibrium consumer welfare.

Another way to understand the result draws on the intuition in Mankiw and Whinston (1986). Whether or not there is too much or too little entry depends on the relative strength of two offsetting effects. First, there is the non-appropriability effect. It pushes in the direction of too little entry because entering firms do not internalize the infra-marginal consumer surplus that they create. Firm revenues do not reflect total consumer surplus. The non-appropriability effect is commensurate with the relative gap $\delta - 1$. It is stronger, the higher is the infra-marginal surplus ratio δ. Second, there is the business stealing effect. It pushes in the direction of too much entry. Entering firms steal revenues from incumbent firms. They do not internalize the corresponding loss of profits. The business stealing effect is commensurate with $\mu - 1$. There is too much entry if non-appropriability effect is weaker than the business stealing effect: $\delta - 1 < \mu - 1$. Conversely, there is too little entry if the non-appropriability effect is stronger than the business stealing effect: $\delta - 1 > \mu - 1$.

4.2 Shocks to Population

Increases in population can be interpreted literally as immigration or increased fertility, or in a more stylized way as the trade integration of more and more countries which are otherwise operating under autarky. Changes in consumer welfare and real GDP per capita in response to increases in population can therefore be interpreted as either gains from scale or as gains from trade.

Proposition 1. *Suppose that firms have the same productivity $A_\theta = A$. In response to changes in population $d \log L$, changes in consumer welfare are given by*

$$d \log Y = \underbrace{(\delta - 1) d \log L}_{\textit{technical efficiency}} + \underbrace{\delta \frac{\xi}{1 - \xi} d \log L}_{\textit{allocative efficiency}}, \tag{34}$$

where

$$\xi = \left(1 - \rho\right)\left(1 - \frac{\delta - 1}{\mu - 1}\right)\frac{1}{\sigma} = \left(1 - \rho\right)\left(1 - \frac{\delta}{\mu}\right). \tag{35}$$

We assume throughout that $\xi < 1$, which guarantees that $0 < d \log Y < \infty$. The first expression for ξ is more complex but we list it here because it will be useful to understand the intuition and to compare with the case with heterogeneous firms below.

In response to a positive population shock $d \log L > 0$, there are in general both changes in technical efficiency and changes in allocative efficiency. Changes in technical efficiency are given by $(\delta - 1) d \log L > 0$. Recall that the technical efficiency effect holds fixed the fraction of labor allocated to entry, overhead, and variable production. Because we hold the fraction of labor allocated to entry (and overhead) constant, the increase in population implies a proportional increase $d \log L > 0$ in entry. The sales share captured by these new varieties is

also $d \log L$. Therefore the increase in the number of varieties increases consumer welfare by $\delta d \log L > 0$. On the other hand, the increase in the number of varieties reduces the amount of labor per capita allocated to the production of each variety, and hence the per-capita quantity of each variety, by $d \log L$. This implies a reduction $-d \log L < 0$ in consumer welfare. The overall effect balances out these two offsetting effects.

There are also changes in allocative efficiency. Changes in allocative efficiency are given by $\delta[\xi/(1-\xi)]d \log L$. The shock reduces the price index by $d \log P = -(1/\sigma)(d \log Y + d \log L) < 0$ (since $\xi < 1$ implies that $d \log Y > 0$). Assuming that the weak second Marshall law of demand holds ($\rho < 1$), this triggers a reduction in markups by $d \log \mu = (1-\rho)d \log P < 0$ and reduces the variable profit share and hence entry by $[1/(\mu-1)](1-\rho)d \log P < 0$. This in turn changes consumer welfare by $[(\delta-1)/(\mu-1)-1](1-\rho)d \log P$. These changes in consumer welfare are positive if and only if there is too much entry to begin with ($\delta < \mu$). The result in the proposition is obtained by replacing $d \log P$ for its expression as a function of $d \log Y$ and solving for the fixed point.

The fact that markups decrease with market size is called the *pro-competitive* effect of scale. In the homogeneous-firm case, these pro-competitive effects are the exclusive source of changes in allocative efficiency. In the next section, we will see that in the presence of heterogeneity, there are other drivers of allocative efficiency that are not related to the pro-competitive effect. In fact, quantitatively, we will find that these much-talked about pro-competitive effects are quantitatively much less significant than the other sources of changes in allocative efficiency.

Proposition 1 allows us to easily determine the sign of changes in allocative efficiency, and hence whether changes in allocative efficiency amplify or mitigate the effects of the shocks.

Corollary 1. *Suppose that firms have the same productivity $A_\theta = A$. Increased population increases allocative efficiency if and only if $\xi > 0$. As long as pass-through is incomplete ($\rho < 1$), this is equivalent to there being too much entry $\delta < \mu$. There is always too much entry under weak second Marshall law of demand and aligned preferences.*

Now, consider the effects of the population shock on real GDP per capita.

Proposition 2. *Suppose that firms have the same productivity $A_\theta = A$. In response to changes in population $d \log L$, changes in real GDP per capita are given by*

$$d \log Q = \frac{1-\rho}{\sigma}(d \log Y + d \log L), \tag{36}$$

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

An increase in population leads to a reduction in markups, and this this pro-competitive effect increases increase real GDP per capita $d \log Q = -d \log p$.

4.3 CES Example

It is interesting to study the CES benchmark, obtained by setting $\Upsilon(x) = x^{(\sigma-1(/\sigma}$, where with some abuse of notation, $\sigma > 1$ is some scalar. In this case, the elasticity of substitution is constant and equal to σ, markups are constant $\mu = 1/(1 - 1/\sigma)$, pass-throughs are equal to one $\rho = 1$, and the infra-marginal surplus ratio is constant $\delta = \sigma/(\sigma - 1)$. Moreover, entry is efficient since $\delta = \mu$.

Changes in consumer welfare are given by

$$d \log Y = \underbrace{(\delta - 1) d \log L}_{\text{technical efficiency}} + \underbrace{0}_{\text{allocative efficiency}} . \tag{37}$$

Unlike in the general case, shocks to population only lead to changes in technical efficiency and do not lead to any change in allocative efficiency. The reason, which is twofold, is straightforward: first, there is no change in markups ($\rho = 1$) and hence no change in the fraction of resources allocated to entry; second entry is actually efficient to begin with ($\delta = \mu$), so even if resources were reallocated, it would not matter for welfare.

The response of real GDP per capita is

$$d \log Q = 0, \tag{38}$$

since markups and productivity shifters do not change.

4.4 Discussion

Assuming that the weak second Marshall law of demand holds, an increase in population, which as discussed above can also be interpreted as an increase in trade integration, impacts consumer welfare, through different channels.

The positive changes in technical efficiency arise from increasing economies of scale. When the allocation of resources is kept constant, the total quantity sold by each firm remains constant, the per-capita quantity declines, and new firms enter.

There are also changes in allocative efficiency which come into play as the allocation of resources adjusts to the shock. Because existing firms reduce their markups, the total quantities sold by these firms expands at the expense of entry. If there was too much entry to begin with, this reallocation of resources is beneficial.

The beneficial nature of these reallocation effects hinges entirely on second-best principles. It depends on whether there was too much or too little entry from a social perspective to begin with, and on whether reallocations decrease or increase entry.

An increase in population is always pro-competitive in the sense that it always reduces markups and expands the total (not per capita) quantities produced by existing firms. How-

ever, these pro-competitive effects do not necessarily increase consumer welfare. They do so only if there was too much entry to begin with, because the reduction in markups has the effect of reducing profit shares, and hence entry. By contrast, the effect on real GDP per capita is unambiguous: the reduction in markup always increases real GDP per capita because it is associated with decreases in prices of existing firms.

5 Heterogeneous Firms

In this section, we turn to the case of heterogeneous firms. We start by discussing social inefficiency. We then study shocks to population. We end by analyzing the special case of CES preferences.

Our results are expressed in terms of the following sufficient statistics: sales shares λ_θ, markups μ_θ, pass-throughs ρ_θ, the infra-marginal surplus ratio δ_θ, and the hazard rate of log productivity at the selection cutoff $\gamma_{\theta^*} = g_a(\log A_{\theta^*})/[1 - G_a(\log A_{\theta^*})]$, where $g_a(\log A_\theta) = g(\theta)/(\partial \log A_\theta/\partial\theta)$.

5.1 Sources of Inefficiency

As before, understanding the comparative statics relies on second-best principles. Therefore, it helps to first consider the different margins of adjustment, and what efficiency looks like along each margin. With homogeneous firms, the only margin that could be distorted was the allocation of labor to entry (and overhead) vs. production. With heterogeneous firms, more margins can be distorted: the allocation of labor to entry (and overhead) vs. variable production, but also the selection cutoff determining which varieties are allocated labor for variable production at all, and the allocation of labor for variable production across active varieties.

The allocation matrix continues to give us an intuitive way to think about efficiency along these different margins. The following expression for changes in consumer welfare, derived from the definition of consumer welfare, will be useful for this purpose

$$d\log Y = \mathbb{E}_\lambda[\delta_\theta] d\log M - \delta_{\theta^*}\lambda_{\theta^*}\frac{g(\theta^*)}{1-G(\theta^*)}d\theta^* + \mathbb{E}_\lambda[d\log y_\theta]. \tag{39}$$

Entry efficiency. We say that there is too much (too little) entry if consumer welfare increases (decreases) when labor is reallocated from variable production to entry and overhead, but keeping the selection cutoff and the relative allocation of labor across non-exiting varieties constant. There is too much (too little) entry if and only if the following condition is verified (violated)

$$\mathbb{E}_\lambda[\delta_\theta] < \mathbb{E}_\lambda\left[\mu_\theta^{-1}\right]^{-1}, \tag{40}$$

which is a comparison of the sales-weighted average of the infra-marginal surplus ratios and the harmonic average of markups.

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 $l_e + l_o = 1 - \mathbb{E}_\lambda[1/\mu]$ to entry and overhead. Consider a reduction in the fraction of labor allocated to variable production $d\log l_\theta = d\log l < 0$ and a complementary increase in the fraction of labor allocated to entry and overhead $d\log l_e = d\log l_o = -[\mathbb{E}_\lambda[1/\mu_\theta]/(1 - \mathbb{E}_\lambda[1/\mu_\theta])]d\log l > 0$. The reduction in the per-capita quantity of each variety $d\log y_\theta = d\log l - d\log M < 0$ reduces consumer welfare by $\mathbb{E}_\lambda[d\log y_\theta] < 0$. We also have an increase in entry $d\log M = -[\mathbb{E}_\lambda[1/\mu_\theta]/(1 - \mathbb{E}_\lambda[1/\mu_\theta])]d\log l > 0$, which increases consumer welfare by $\mathbb{E}_\lambda[\delta_\theta]d\log M > 0$. Finally, we have no change in selection $d\theta^* = 0$. The overall effect on consumer welfare is given by $d\log Y = [1 - (\mathbb{E}_\lambda[\delta_\theta] - 1)\mathbb{E}_\lambda[1/\mu_\theta]/(1 - \mathbb{E}_\lambda[1/\mu_\theta])]d\log l$, which is negative (too much entry) if and only if $1/\mathbb{E}_\lambda[1/\mu_\theta] < \mathbb{E}_\lambda[\delta_\theta]$. In the homogeneous-firm case, this condition collapses to the simple $\delta < \mu$.

Selection efficiency. We say that there is too little (too much) selection if consumer welfare increases (decreases) when the selection cutoff is increased and the labor previously allocated to variable production and overhead of the newly exiting varieties is reallocated proportionately to entry, overhead, and to variable production. There is too little (too much) selection if and only if the following condition is verified (violated)

$$\delta_{\theta^*} < \mathbb{E}_\lambda[\delta_\theta]. \tag{41}$$

Suppose that we increase the selection cutoff 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 $d\log y_\theta = 0$. The overall effect on consumer 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 $\delta_{\theta^*} < \mathbb{E}_\lambda[\delta_\theta]$.

Relative production efficiency. Finally, we say that the supply of a variety is too large (too small) compared to another if consumer welfare increases (decreases) when labor is reallocated from the former to the latter. The supply of variety θ' is too large (too small) compared to that of variety θ if, and only if, the following condition is verified (violated)

$$\mu_{\theta'} < \mu_\theta. \tag{42}$$

Following Baqaee and Farhi (2019), 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. This effect on consumer 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 $\mu_\theta > \mu_{\theta'}$.

5.2 Shocks to Population

As before, increases in population can be interpreted literally as immigration or increased fertility, or in a more stylized way as the trade integration of more and more countries which are otherwise operating under autarky. Changes in consumer welfare and real GDP per capita in response to increases in population can therefore be interpreted as either gains from scale or as gains from trade.

Proposition 3. *In response to changes in population $d\log L$, changes in consumer welfare are given by*

$$d\log Y = \underbrace{\left(\mathbb{E}_\lambda[\delta_\theta] - 1\right) d\log L}_{\textit{technical efficiency}} + \underbrace{\frac{\xi^\epsilon + \xi^\mu + \xi^{\theta^*}}{1 - \xi^\epsilon - \xi^\mu - \xi^{\theta^*}} \left(\mathbb{E}_\lambda[\delta_\theta]\right) d\log L}_{\textit{allocative efficiency}}, \tag{43}$$

where

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

We assume throughout that $\xi^\epsilon + \xi^\mu + \xi^{\theta^*} < 1$, which guarantees that $0 < d\log Y < \infty$. We examine this expression term by term and explain its underlying intuition.

The intuition for the changes in technical efficiency is the same as in the case of homogeneous firms covered in Section 4. The only difference is that the infra-marginal surplus ratios δ_θ are heterogeneous and matter through their average $\mathbb{E}_\lambda[\delta_\theta]$. We now dissect the allocative efficiency term.

Each of ξ^ϵ, ξ^μ, and ξ^{θ^*} relates to adjustments along a specific margin. Starting at the initial equilibrium an increase in population $d\log L > 0$ leads each firm to adjust its behavior on three margins: entry, exit, and markup. We decompose the general equilibrium response by analyzing three successive equilibrium allocations which allow firms to adjust along more and more margins. All three equilibrium allocations feature the same technical efficiency effect,

but different changes in allocative efficiency, driven by different changes in the allocation of resources.

Entry only margin. First, consider the equilibrium where firms can only adjust their entry behavior (free entry) but can neither adjust exit behavior nor their markups. The resulting change in consumer welfare is given by Proposition 3 but with $\xi^{\mu} = \xi^{\theta^*} = 0$.

Changes in allocative efficiency are strictly positive ($\xi^{\epsilon} > 0$) as long as there is non-trivial heterogeneity. To see this, note that

$$\mathbb{E}_{\lambda}[\sigma_\theta] - \mathbb{E}_{\lambda(1-1/\mu)}[\sigma_\theta] = \frac{Cov_{\lambda}[\sigma_\theta, \frac{1}{\mu_\theta}]}{\mathbb{E}_{\lambda}[1 - \frac{1}{\mu_\theta}]},$$

where $Cov_{\lambda}[\sigma_\theta, 1/\mu_\theta]$ denotes the covariance of σ_θ and $1/\mu_\theta$ using under the sales share probability measure $\lambda_\theta g(\theta)/[1 - G(\theta^*)]$. Since the price elasticity of demand σ_θ and the reciprocal of the markup $1/\mu_\theta$ covary positively, ξ^{ε} is positive as long as there is any heterogeneity.

Intuitively, the reduction in the price index triggers bigger reductions in per-capita quantities and sales for firms with higher elasticities and lower markups than for firms with lower elasticities and higher markups. Since the former were too large and the latter too small to begin with, this reallocation improves welfare. Basically, because it improves efficiency, this reallocation frees up labor, which ends up being used in entry, and this always improves welfare, irrespective of whether there was too much or too little entry to begin with. This is the composition effect that is graphically illustrated in Figure 1 in the introduction. This effect, which is unambiguously positive, will turn out to be the dominant quantitative force in the model.

When the weak second Marshall law of demand holds, then it is the bigger firms that have have higher markups, and the beneficial reallocations are away from small firms and towards big firms. However, the composition effect is always positive, irrespective of whether the weak second Marhsall law of demand holds or not.[18]

The detailed intuition for the changes in allocative efficiency captured by ξ^{ϵ} is as follows. The shock increases consumer welfare $d \log Y > 0$. It leads to a reduction in the price index by $d \log P = -\mathbb{E}_{\lambda}[1/\sigma_\theta](d \log Y + d \log L) < 0$. As a result, sales shares change by $(\sigma_\theta - \mathbb{E}_{\lambda}[\sigma_\theta])d \log P$. This reallocates resources towards varieties with lower elasticities σ_θ, which also have higher markups μ_θ, and increases the aggregate variable profit share and entry by $-(\mathbb{E}_{\lambda}[\sigma_\theta] - \mathbb{E}_{\lambda(1-1/\mu)}[\sigma_\theta])d \log P > 0$. This then increases consumer welfare by $-(\mathbb{E}_{\lambda}[\delta_\theta] -$

[18] This composition effect is related to the "Matthew Effect" discussed by Mrázová and Neary (2019), who show that when Marshall's weak second law holds, an increase in scale increases the profits of large firms. The difference here is that we relate this composition effect to welfare, and that the welfare relevant reallocation is in terms of employment as opposed to profits. Furthermore, inspection of ξ^{ϵ} shows that this effect is always positive regardless of whether or not the weak second Marshall law of demand holds. In particular, if the demand curve is log-convex, even though it still increases welfare, the composition effect becomes an "anti"-Matthew effect.

$1)(\mathbb{E}_\lambda[\sigma_\theta] - \mathbb{E}_{\lambda(1-1/\mu)}[\sigma_\theta])d\log P > 0$. The result in the proposition is obtained by replacing $d\log P$ by its expression as a function of $d\log Y$ and solving for the fixed point.

Entry and selection only margins. Second, consider the equilibrium where firms can adjust their entry behavior and also their exit behavior (that is, firms can choose whether or not to exit after drawing their type), but cannot adjust their markups. The resulting changes in consumer welfare are given by Proposition 3 but with $\xi^\mu = 0$.

There is a new source of changes in allocative efficiency captured by $\xi^{\theta^*} \neq 0$. The intuition is as follows. Suppose that the weak second Marshall law of demand holds, so that markups are increasing with productivity. The demand elasticity is then higher at the selection cutoff than above it. This implies that the reduction in the price index disproportionately hurts firms at the selection cutoff and hence that the selection cutoff increases $d\theta^* > 0$. The sales shares of the exiting varieties with $\theta \in [\theta^*, \theta^* + d\theta^*)$ is $\lambda_{\theta^*}(g(\theta^*)/[1 - G(\theta^*)])d\theta^*$. It is equal to $\lambda_{\theta^*}\gamma_{\theta^*}d\log A_{\theta^*}$, where $d\log A_{\theta^*}$ is the change in productivity associated with a change in type from θ^* to $\theta^* + d\theta^*$. This reallocates sales from exiting firms to the average surviving firm and changes consumer welfare by $(\mathbb{E}[\delta_\theta] - \delta_{\theta^*})\lambda_{\theta^*}(g(\theta^*)/[1-G(\theta^*)])d\theta^*$. These changes in allocative efficiency are positive ($\xi^{\theta^*} > 0$) if there is too little selection to begin with ($\mathbb{E}[\delta_\theta] > \delta_{\theta^*}$).

It is important to note that the fact that θ^* increases is not, on its own, evidence of an improvement in allocative efficiency. In other words, increases in the cutoff θ^*, due to intensifying competition, are only socially desirable if the marginal firm provides households with less infra-marginal surplus than the average surviving firm. In fact, in our empirical Section 6 we find that increases in selection cutoff reduce welfare.

To complete the intuition, we now discuss the change in the selection cutoff $d\theta^*$. It can be shown that the change in variable profits at the selection cutoff is given by $(\sigma_{\theta^*} - \mathbb{E}_{\lambda(1-1/\mu)}[\sigma_\theta])d\log P$, where the change in the price index is given by $d\log P = -\mathbb{E}_\lambda[1/\sigma_\theta](d\log Y + d\log L) < 0$. Variable profits at θ^* decrease as long as the marginal firm is more price elastic than the average firm $\sigma_{\theta^*} > \mathbb{E}_{\lambda(1-1/\mu)}[\sigma_\theta]$, which is guaranteed under Marshall's weak second law of demand. Under this assumption, the reduction in variable profits at the cutoff requires an offsetting increase in the selection cutoff $d\theta^* > 0$ so that productivity increases by $d\log A_{\theta^*} = -[1/(\sigma_{\theta^*} - 1)](\sigma_{\theta^*} - \mathbb{E}_{\lambda(1-1/\mu)}[\sigma_\theta])d\log P > 0$. The result in the proposition is obtained by replacing $d\log P$ by its expression as a function of $d\log Y$ and solving for the fixed point.

Entry, exit, and pricing/markup margins. Last, consider the equilibrium where firms can not only adjust their entry and exit behavior, but also their pricing/markup behavior. The resulting changes in consumer welfare are the full-blown changes in equilibrium welfare given by Proposition 3.

There is a new source of changes in allocative efficiency captured by $\xi^\mu \neq 0$. Adjustments along this margin (markups) are the source of pro-competitive effects of scale.

The intuition for the additional changes in allocative efficiency captured by ξ^{μ} is very similar to that presented in the case of homogeneous firms in Section 4. The only difference is that the terms in ξ^{μ} are appropriately averaged versions of the now heterogeneous underlying sufficient statistics. In the homogeneous-firm case, the ξ^{ϵ} and ξ^{θ^*} were equal to zero because there was no heterogeneity, so we only had to contend with the markup margin. In that case, the reduction in markups increased welfare if, and only if, there was too much entry to begin with. In the heterogeneous-firm case, these changes in allocative efficiency are positive ($\xi^{\mu} > 0$) if and only if $\mathbb{E}_{\lambda(1-\rho)}\left[1 - (\mathbb{E}_{\lambda}[\delta_{\theta}] - 1)/(\mu_{\theta} - 1)\right] > 0$.

It is not clear in general whether this condition is weaker or stronger than the condition that there is too much entry. The reason is subtle. It easiest to understand when the strong second Marshall law of demand holds, which we shall assume here. There is a general reduction in markups, which reduces entry, and increases consumer welfare if, and only if, there was too much entry to begin with. But there is also a bigger reduction in markups for firms with low pass-throughs, which under the strong second Marshall law of demand, also have lower elasticities and higher markups. That they have lower pass-throughs pushes for a reallocation of resources towards them, but that they have lower price-elasticities pushes in the other direction. Whether or not resources are reallocated towards these high-markup firms, and hence whether or not the associated reallocation effects increase or decrease consumer welfare, depends on whether or not the pass-through effect dominates the elasticity effect. In the former case excessive entry is a sufficient condition for $\xi^{\mu} > 0$ but not in the latter case.

When reallocations all three margins are beneficial so that $\xi^{\epsilon} \geq 0$, $\xi^{\theta^*} \geq 0$, and $\xi^{\mu} \geq 0$, we can unambiguously sign the changes in allocative efficiency.

Corollary 2. *Sufficient conditions for positive changes in allocative efficiency in response to increases in population are: (1) that the strong second Marshall law of demand hold; (2) that the condition* $\mathbb{E}_{\lambda}\left[(1-\rho_{\theta})\left[1 - (\mathbb{E}_{\lambda}[\delta_{\theta}] - 1)/(\mu_{\theta} - 1)\right]\right] > 0$*, which can be stronger or weaker than the condition for excessive entry, hold; and (3) that there be too little selection* $\mathbb{E}_{\lambda}[\delta_{\theta}] > \delta_{\theta^*}$*, which is automatically verified if (1) holds and if, in addition, preferences are aligned. More precisely, we always have* $\xi^{\epsilon} \geq 0$*, with a strict inequality as long as there is non trivial heterogeneity; (1) and (3) imply* $\xi^{\theta^*} > 0$*; and (2) implies that* $\xi^{\mu} > 0$*.*

Appealing as they may be, these conditions will not always be verified in our empirical application, and in fact, we will sometimes find $\xi^{\theta^*} \leq 0$ or $\xi^{\mu} \leq 0$. Nonetheless, we will find that $\xi^{\epsilon} > 0$ will always be large enough to dominate the other terms, resulting in large positive changes in allocative efficiency from increases in scale.

Finally, we can also characterize the change in real GDP per capita as an appropriately-averaged version of that presented in the case of homogeneous firms in Section 4.

Proposition 4. *In response to changes in population $d \log L$, changes in real GDP per capita are*

$$d \log Q = \left(\mathbb{E}_\lambda \left[(1 - \rho_\theta) \right] \right) \left(\mathbb{E}_\lambda \left[\frac{1}{\sigma_\theta} \right] \right) \left(d \log Y + d \log L \right), \tag{44}$$

where $d \log Y$ is given by Proposition 3.

5.3 CES Example

Once again, we consider the CES benchmark, $\Upsilon(x) = x^{(\sigma-1)/\sigma}$, where $\sigma > 1$ is some scalar. With heterogeneous firms, this example is a closed-economy version of Melitz (2003). The markups are constant $\mu = 1/(1 - 1/\sigma)$, pass-throughs are equal to one $\rho = 1$, and the infra-marginal surplus ratio is constant $\delta = \sigma/(\sigma - 1)$. Moreover, entry is efficient since $\bar{\delta} = \bar{\mu}$. Furthermore, since $\mathbb{E}_\lambda(\delta_\theta) = \delta_{\theta^*}$, the exit/selection margin is also efficient. Finally, since markups are constant, there is no adjustment of markups. This means that changes in consumer welfare are simply given by technical efficiency effects

$$d \log Y = \underbrace{(\delta - 1)\, d \log L}_{\text{technical efficiency}} + \underbrace{0}_{\text{allocative efficiency}}. \tag{45}$$

The fact that changes in allocative efficiency are zero is not surprising: the CES model is efficient, so this result is a consequence of the envelope theorem.

Next, consider changes in real GDP per capita, which are

$$d \log Q = 0, \tag{46}$$

since markups and productivity shifters do not change in response to changes in population, despite the fact that welfare increases.

5.4 Discussion

We can now take stock and dispel some common misconceptions. Although we conduct this discussion for shocks to population, which as discussed above, can also be interpreted as shocks to trade integration, the spirit of our remarks applies more broadly to other shocks.

There are positive changes in technical efficiency arising from economies of scale because fixed entry and overhead costs can be spread across a larger population. When the allocation of resources is kept constant, the total quantity sold by each firm remains constant, the per-capita quantity declines, and new firms enter.

There are also changes in allocative efficiency which come into play as the allocation of resources adjusts to the shock. For the purposes of discussion, assume that Marshall's strong second law of demand holds.

First, holding exit and pricing/markup behavior constant, there is a reallocation away from small firms with higher elasticities and lower markups, towards big firms with lower elasticities and higher markups. Since the former were too large and the latter too small to begin with, this reallocation is beneficial. It frees up labor, which ends up being used in entry, and this is always beneficial, irrespective of whether there was too much or too little entry to begin with.

Second, holding pricing/markup behavior constant, the smallest firms exit and make room for new firms. If there was too little selection to begin with, this reallocation of resources is beneficial.

Third, because existing firms reduce their markups, the total quantities sold by these firms expands at the expense of entry. This last effect is the only that operates in models with homogeneous firms. This reallocation of resources is beneficial If there was too much entry to begin with. However, this beneficial expansion of existing firms at the expense of entry is complicated by ambiguous reallocation effects across existing firms: on the one hand, large firms reduce their markups more than small firms which tends to reallocate resources towards larger firms; on the other hand, large firms are less elastic so that a given markup reduction induces less reallocation towards them than for small firms. The former effect is beneficial but the latter is detrimental.

The beneficial nature of these three reallocation effects hinges entirely on second-best principles: in which direction the underlying margin is distorted, and in which direction the reallocation effect is moving this margin. It has nothing to do with the productivities of expanding and shrinking or disappearing firms per se. Such misleading intuitions are routinely invoked in economic writings, for example when discussing the popular model of Melitz (2003). In particular, that model has CES preferences and is therefore efficient. As a result, reallocations of resources, such as movements in the selection cutoff, have no impact on consumer welfare to the first order.[19]

The relative expansion of large firms at the expense of small firms underlying the reallocation effects ξ^{ϵ} increases consumer welfare only because large firms were too small to begin with from a social perspective since they charge higher markups, not because they have "higher productivities".[20] Similarly, the exit of the smallest firms underlying the second reallocation effect ξ^{θ^*} increases consumer welfare only if selection was too weak to begin with, which amounts to assuming that the marginal firm has lower infra-marginal consumer surplus ratio than average. Finally, the relative expansion of existing firms at the expense of entry underlying part of the third reallocation effect increases consumer welfare if and only if there was to much entry to begin with, which is the case when the average infra-marginal

[19] With CES preferences, reallocations also have no effect on real GDP or aggregate TFP either.

[20] In fact, since firms are producing differentiated varieties, it makes little sense to try to compare the "level" of their productivity; their output is not measured in comparable units.

consumer surplus ratio is less than the average markup.[21]

Similarly, these reallocations affect consumer welfare and real GDP or aggregate TFP through very different channels. In other words, the nature of the consumer welfare gains is distinct from that of aggregate TFP gains, and intuitions supporting the latter should not be used to shed light on the former. Indeed, changes in aggregate TFP only arise from changes in markups. For example, holding pricing/markups constant, aggregate TFP is completely unaffected. In particular, the oft-emphasized increase in the minimal productivity cutoff has no effect on aggregate TFP.

This last observation may be of independent interest. In the literature, aggregate TFP is sometimes conflated with some ad-hoc productivity index obtained as an average of firm productivities (for example, Melitz and Ottaviano, 2008). However, the growth in this index is not aggregate TFP growth defined as real GDP growth net of the effects of factor growth. While the increase in the minimal productivity cutoff directly increases these ad-hoc productivity indices, it has no independent impact on aggregate TFP. Of course, in equilibrium, aggregate TFP does increase, but only because of the reduction in markups.

6 Empirical Application

In this section, we take the theory to the data. We first describe our non-parametric estimation procedure. We then implement it using Belgian data and decompose returns to aggregate scale into changes in technical and allocative efficiency following Proposition 3.

6.1 Non-Parametric Model Estimation Approach

To derive a non-parametric estimate of the Kimball aggregator Υ, we use a restriction imposed by the demand system: the time-series pass-through capturing how a firm changes its markup in response to a productivity shock is equal to the cross-sectional pass-through capturing how markups increase as we move up the productivity distribution across firms.

We will take two objects as data: (1) the density of sales shares λ_θ, and (2) the distribution of pass-throughs ρ_θ. Sales are readily available in our dataset, and we use estimates of pass-throughs by firm size from Amiti et al. (2019). Since pass-throughs are third derivatives of the Kimball aggregator, we can recover Υ by solving a series of differential equations. For boundary conditions, we need to take a stand on the average levels of first and second derivatives, i.e. on the average markup $\bar{\mu}$ and on the average infra-marginal consumption

[21] Although it is not directly relevant for welfare, it is possible to characterize how the absolute size of firms changes in response to an increase in population. Assuming that the strong second Marshall law of demand holds, big firms always expand and small firms shrink in absolute terms as long as their pass-throughs are close enough to one since $d\log(l_\theta L) = (1 - \rho_\theta\sigma_\theta/\mathbb{E}_{\lambda(1-1/\mu)}[\sigma_\theta])(d\log Y + d\log L)$. Similarly, the size of the smallest firms (at the cutoff) increases since $d\log(l_{\theta^*}L) = [1/(\sigma_{\theta^*} - 1)][\sigma_{\theta^*}/\mathbb{E}_{\lambda(1-1/\mu)}[\sigma_\theta] - 1 + \sigma_{\theta^*}(1 - \rho_{\theta^*})(1 - 1/\mathbb{E}_{\lambda(1-1/\mu)}[\sigma_\theta])](d\log Y + d\log L)$.

surplus ratio $\bar{\delta}$ (these will be constants of integration). We will present our estimates for different values of these variables.

Observation of sales λ_θ and pass-throughs ρ_θ will allow us to back out productivities A_θ up to a normalizing constant and markups μ_θ up to the average markup $\bar{\mu}$. Using $\sigma_\theta = 1/(1-1/\mu_\theta)$ to recover elasticities will then allow us to back out infra-marginal surplus ratios δ_θ and the whole Kimball aggregator up to the average infra-marginal surplus ratio $\bar{\delta}$.

Basically, cross-sectional observations on pass-throughs allows us to trace the individual demand curve and hence to back out the Kimball aggregator Υ up to some constants $\bar{\delta}$ and $\bar{\mu}$. Through this procedure, we will therefore be able to recover the whole nonlinear structure of the model.[22]

In our empirical application, we use a uniform type distribution with $g(\theta) = 1$ and $G(\theta) = 1 - \theta$ by ranking firms by increasing size and associating their type to the fraction of firms with smaller sizes. The type distribution itself is irrelevant: the only thing that matters is the relation of the measure over types to the measure over sales.

In principle, one could alternatively use markups μ_θ (second derivatives of the Kimball aggregator) or infra-marginal surplus ratios δ_θ (first derivatives of the Kimball aggregator) in conjunction with sales λ_θ to recover Υ. However, these objects are much harder to estimate, requiring either strong structural assumptions in the case of μ_θ, or in the case of δ_θ experimental data tracing out individual demand curves. In comparison, estimating pass-throughs is less daunting. The downside is that it will require outside information or informed guesses to pin down $\bar{\mu}$ and $\bar{\delta}$.

Productivities, quantities, elasticities, and infra-marginal surplus ratios. Productivities A_θ and markups μ_θ must simultaneously solve the two differential equations

$$\frac{d\log\lambda_\theta}{d\theta} = \frac{\rho_\theta}{\mu_\theta - 1}\frac{d\log A_\theta}{d\theta}, \tag{47}$$

$$\frac{d\log\mu_\theta}{d\theta} = (1-\rho_\theta)\frac{d\log A_\theta}{d\theta}. \tag{48}$$

The intuition for the first differential equation for sales shares is the following: compared to a firm of type θ, a firm with type $\theta+d\theta$ has higher productivity $\log A_{\theta+d\theta} - \log A_\theta = d\log A_\theta/d\theta$, lower price $\log p_{\theta+d\theta} - \log p_\theta = \rho_\theta d\log A_\theta/d\theta$, and higher sales $\log\lambda_{\theta+d\theta} - \log\lambda_\theta = (\sigma_\theta - 1)\rho_\theta d\log A_\theta/d\theta$ with $\sigma_\theta - 1 = 1/(\mu_\theta - 1)$. The intuition for the second differential equation for markups is the following: compared to a firm of type θ, a firm with type $\theta + d\theta$ has higher

[22] We refer the reader to Appendix F for the discussion of a model with taste shocks and heterogeneous overhead costs in which cross-section and times-series are entirely disconnected. Non-parametric estimation of this richer model would require additional data on markups μ_θ as well as taking a stand on the distribution of infra-marginal surplus ratios δ_θ and overhead costs $f_{o,\theta}$. And even then, only local non-parametric estimation could be achieved, which would be just enough to allow the computation of local counterfactuals along the lines of the formulas presented in the paper.

markup $\log \mu_{\theta+d\theta} - \log \mu_\theta = (1-\rho_\theta) d \log A_\theta / d\theta$.

Combining the two differential equations yields

$$\frac{d \log \mu_\theta}{d\theta} = (\mu_\theta - 1) \frac{1-\rho_\theta}{\rho_\theta} \frac{d \log \lambda_\theta}{d\theta}. \tag{49}$$

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^*}. The constant $\mu_{\theta^*} \geq 1$ can be chosen to match a given value of the (harmonic) sales-weighted average markup $\bar{\mu} \geq 1$.

Either of the two differential equations for sales shares or markups then allows us to recover productivities up to a constant A_{θ^*}. This constant A_{θ^*} can be normalized to 1. For example, using the differential equation for sales shares, we get

$$\frac{d \log A_\theta}{d\theta} = \frac{\mu_\theta - 1}{\rho_\theta} \frac{d \log \lambda_\theta}{d\theta},$$

with initial condition $A_{\theta^*} = 1$.

Next we can then recover quantities using

$$y_\theta = \frac{\lambda_\theta A_\theta}{(1 - G(\theta^*)) M \mu_\theta}. \tag{50}$$

Quantities are increasing in θ since $d \log y_\theta / d \log \theta = \mu_\theta d \log \lambda_\theta / d \log \theta$. We denote by $\theta(y)$ the reverse mapping giving a firm type as a function of the quantity that it produces.

Finally, we can recover infra-marginal consumption surplus ratios using the differential equation

$$\frac{d \log \delta_\theta}{d\theta} = \frac{\mu_\theta - \delta_\theta}{\delta_\theta} \frac{d \log \lambda_\theta}{d\theta}, \tag{51}$$

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

Kimball aggregator. We can then recover the Kimball aggregator by combining the definition of δ_θ with the residual demand curve

$$\Upsilon(\frac{y}{Y}) = \frac{\lambda_{\theta(y)}}{(1 - G(\theta^*)) M} \frac{\delta_{\theta(y)}}{\bar{\delta}}.$$

Fixed costs and Selection Cutoff. The information so far does not reveal the cutoff 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. Hence, in their first year, the unconditional probability of exit is higher than the exogenous death rate. We then fit a quasi-hyperbolic process to firm exit

probability by age as reported by Pugsley et al. (2018). The difference between the probability of exit in the first period versus later periods identifies θ^*. Conditional on θ^*, we can back out the fixed costs using the free-entry condition

$$\frac{f_e\Delta}{L} + (1 - G(\theta^*))\frac{f_o}{L} = \frac{1}{M}\mathbb{E}\left[\lambda_\theta\left(1 - \frac{1}{\mu_\theta}\right)\right], \tag{52}$$

and the selection condition

$$(1 - G(\theta^*))\frac{f_o}{L} = \frac{1}{M}\lambda_{\theta^*}\left(1 - \frac{1}{\mu_{\theta^*}}\right), \tag{53}$$

where total population L, mass of firms M, and Δ can be normalized to 1.

6.2 Empirical Implementation

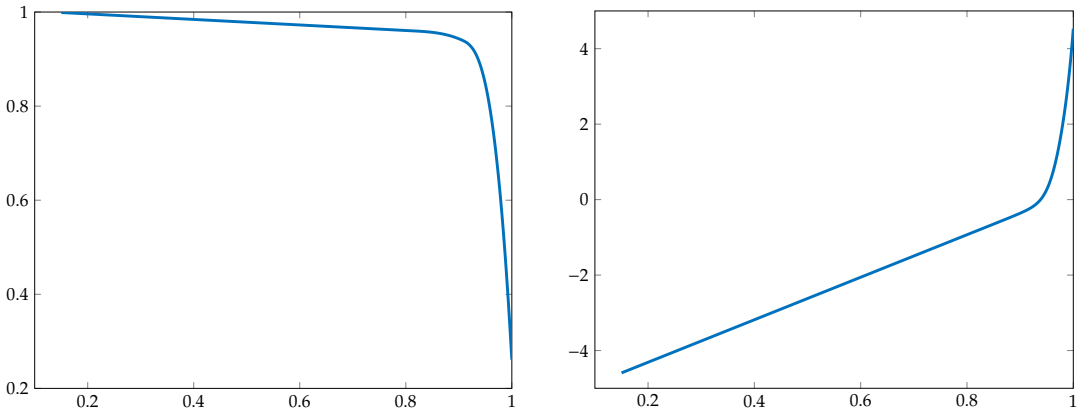
In this section, we implement the non-parametric estimation procedure described above using estimates of the firm-level pass-throughs and the distribution of firms sales. We then use the estimated model in conjunction with our theoretical formulas to investigate quantitatively the effects of population shocks and the origins of increasing aggregate returns to scale.

Data sources. We only give a brief account of our data sources and procedures.The full details can be found in Appendix A. We rely on Amiti et al. (2019) who report estimates of pass-throughs by firm size for manufacturing firms in Belgium. They use annual administrative firm-product level data (Prodcom) from 1995-2007, which contains information on prices and sales, collected by Statistics Belgium. They merge this with Customs data, and using exchange rate shocks as instruments for changes in marginal cost, they show that they can identify the partial equilibrium pass-through by firm size (under assumptions that are consistent with our model).

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 (as a function of size) with the sales distribution λ for the universe of Belgian manufacturing firms (from VAT declarations). 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. This is consistent with the estimates of Amiti et al. (2019) who find that the average pass-through for the smallest 75% of firms in Prodcom is already 0.97.[23]

Boundary Conditions. Our results require taking a stand on the average infra-marginal surplus ratio $\bar{\delta} = \mathbb{E}_\lambda[\delta_\theta]$ and on the average markup $\bar{\mu} = 1/[\mathbb{E}_\lambda[1/\mu_\theta]]$. To set $\bar{\delta}$, we consider

[23]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. Although this affects numbers change, the overall message does not change.



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

Figure 3: Pass-throughs and sales share density.

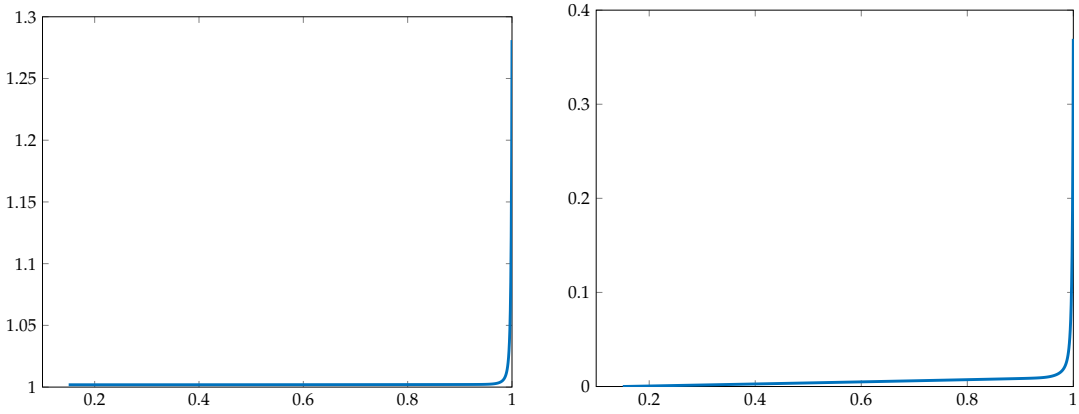
two benchmark calibrations: (1) efficient entry $\bar{\delta} = \bar{\mu}$, and (2) efficient selection $\bar{\delta} = \delta_{\theta^*}$. We consider two different values for the average markup $\bar{\mu} = 1.045$ and $\bar{\mu} = 1.090$, which are chosen so that $d \log Y / d \log L \approx 0.13$ under the first assumption, and $d \log Y / d \log L \approx 0.30$ under the second assumption. The number 0.13 is broadly in line with the empirical estimates from Bartelme et al. (2019) (which we take to represent the trade literature) and the number 0.30 is how Jones (2019) calibrates aggregate returns to scale (which we take to represent the growth literature).[24]

Estimation Results. Figures 3a and 3b display pass-throughs ρ_θ and log sales $\log \lambda_\theta$ as a function of firm type θ. Sales are initially increasing exponentially (linear in logs), but become super-exponential towards the end reflecting a high degree of concentration in the tail. Pass-throughs decrease from 1 for the smallest firms to about 0.3 for the largest firms. Marhsall's strong second law of demand therefore holds.

For brevity, we only show graphs of the estimates for $\bar{\mu} = 1.045$ but the patterns are similar for the other case (though obviously, markups are higher when we set $\bar{\mu} = 1.09$). Figure 4a shows that markups μ_θ are increasing and convex in θ. The net markup ranges from close to zero for the smallest firms to around 30% for the very largest firms when $\bar{\mu}$ is calibrated to equal 1.045. The heterogeneity in markups is a consequence of the vast dispersion in the firm size distribution and estimated pass-throughs. Figure 4b shows the log productivity/quality distribution. As with the sales density, the productivity density is also initially exponential, and becomes super exponential in the tail, however, the pattern is more pronounced. The main reason is that elasticities are decreasing in θ.

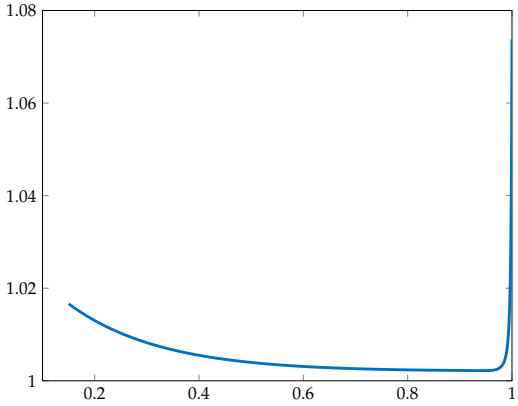
Figures 4d and 4c show the infra-marginal surplus ratio δ for the efficient-selection case

[24] In the Dixit-Stiglitz model, the former would correspond to setting an elasticity of substitution around 8 whilst the latter is an elasticity of substitution around 4.

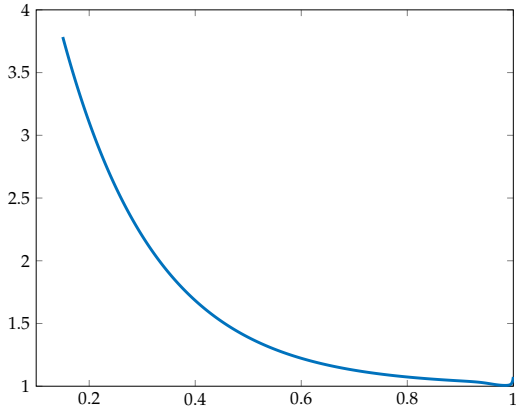


(a) Markup μ_θ

(b) Log productivity A_θ



(c) Infra-marginal surplus ratio δ_θ (efficient selection)



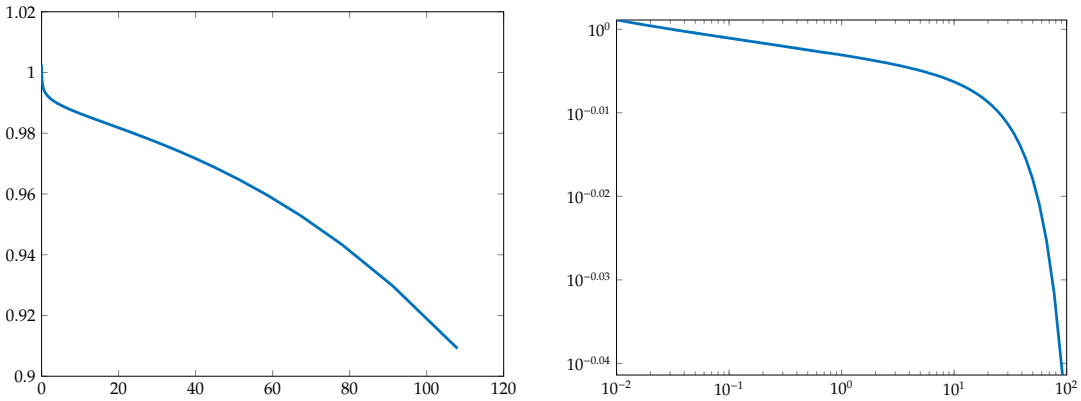
(d) Infra-marginal surplus ratio δ_θ (efficient entry)

Figure 4: Markups and infra-marginal surplus ratios with $\bar{\mu} = 1.045$.

($\delta_{\theta^*} = \bar{\delta}$) and the efficient-entry case ($\bar{\mu} = \bar{\delta}$). In both cases, δ_θ is U-shaped, although in the latter case it starts at a much higher level than the former case, meaning that preferences are not aligned.

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-iso-elastic shape, indicating substantial departures from CES. The Lerner formula ties the markup to the price elasticity of demand σ_θ, which means that σ_θ is greater than 100 for the very smallest firms, and around 4 for the very largest firms. The log-log plot also shows that the residual demand curve is log-concave, which confirms that Marshall's weak second of law of demand holds, and that markups are increasing in size.

Implications. Markups are increasing in firm size (since Marshall's strong second law holds, the weak second law holds a fortiori). However, since the infra-marginal surplus ratios are U-shaped, preferences are not globally aligned.



(a) Inverse residual demand curve (b) Log-log inverse residual demand curve

Figure 5: Residual demand curve (price against quantity) for the efficient-selection case with $\bar{\mu} = 1.045$. The results for the efficient-entry case are similar.

In the efficient entry case, we have $\delta_{\theta^*} > \bar{\delta}$, implying that selection is too tough to begin with. Welfare could be improved by allowing more small firms to operate, and an increase in the selection cutoff worsens welfare. Nevertheless, this does not imply that we would want small firms to become larger — since small firms have lower markups, efficiency would be improved by having them produce less than they already do. The fact that an increase in the cutoff θ^* reduces welfare may be counterintuitive. In fact, the proposition that increases in the selection cutoff ipso facto increase efficiency is sometimes taken for granted, but as our results show, this argument is flawed. In the efficient selection case, $\bar{\delta} < \bar{\mu}$, meaning that there is excessive entry. In both cases, markups μ_θ are increasing in θ indicating that small firms are too large and large firms are too small along the intensive margin.

6.3 Shocks to Population

In this section, we report the elasticities of consumer welfare and real GDP per capita to changes in population. As stressed before, increases in population can also be interpreted as certain forms of trade integration.

Baseline. Table 1 implements Proposition 3. It reports the elasticity of consumer welfare to population and its decomposition into changes in technical efficiency and changes in allocative efficiency

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

The table also breaks down the allocative efficiency effect by considering the different margins of adjustment. Welfare under the entry-only allocation $\mathrm{d}\log Y^{\epsilon}$ (holding fixed θ^* and markups μ_θ); welfare allowing entry and selection to adjust $\mathrm{d}\log Y^{\epsilon,\theta^*}$ (holding fixed markups μ_θ); and welfare when all three margins can adjust $\mathrm{d}\log Y^{\epsilon,\theta^*,\mu} = \mathrm{d}\log Y$. Table 1 presents the

contributions of endogenous entry $\mathrm{d}\log Y^{\epsilon} - \mathrm{d}\log Y^{tech}$, exit $\mathrm{d}\log Y^{\epsilon,\theta^*} - \mathrm{d}\log Y^{\epsilon}$, and markups $\mathrm{d}\log Y^{\epsilon,\theta^*,\mu} - \mathrm{d}\log Y^{\epsilon,\theta^*}$ to changes in allocative efficiency. In other words, the sum of the three adjustment rows gives the overall change in allocative efficiency in equilibrium.

	$\bar{\mu} = 1.045$		$\bar{\mu} = 1.090$	
	$\bar{\delta} = \delta_{\theta^*}$	$\bar{\delta} = \bar{\mu}$	$\bar{\delta} = \delta_{\theta^*}$	$\bar{\delta} = \bar{\mu}$
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
Adj. of Entry: $\mathrm{d}\log Y^{\epsilon} - \mathrm{d}\log Y^{tech}$	0.117	0.408	0.272	1.396
Adj. of Exit: $\mathrm{d}\log Y^{\epsilon,\theta^*} - \mathrm{d}\log Y^{\epsilon}$	0.000	-0.251	0.000	-1.006
Adj. of Markups: $\mathrm{d}\log Y^{\epsilon,\theta^*,\mu} - \mathrm{d}\log Y^{\epsilon,\theta^*}$	-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 Propositions 3 and 4 for heterogeneous firms case.

In our discussions, 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 entry-efficiency first. 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 population therefore brings about considerable improvements in allocative efficiency, and these improvements are about nine times larger than the technical efficiency effects arising directly from technological increasing returns.

The changes in allocative efficiency from endogenous entry are large and positive at 0.117. Increases in population lead to a reduction in the aggregate price index. This causes a larger reduction in quantity per capita for small firms with high elasticities than for large firms with low elasticities. This reallocation towards large firms, which were too small to begin with, and away from small firms, which were too small to begin with, improves allocative efficiency. In other words, the composition effect highlighted in Figure 1 of the introduction is large.

The changes in allocative efficiency from endogenous exit is zero by construction despite the fact that the cutoff increases in response to the increase in population.

The changes in allocative efficiency from changes in markups are slightly negative -0.004. There are several effects to consider. First, increases in in population lead to a reduction in the price index. This triggers a pro-competitive effect by causing an overall reduction in markups, and a reduction in entry, which is beneficial since there was too much entry to begin with (the average markup is higher than the average infra-marginal surplus ratio). However, the changes in markups are not uniform, and larger firms cut their markups by more

than smaller firms since they have lower pass-through. As before, large firms also face less elastic demand curves than small firms, so that the overall reallocation across large and small firms is, in principle, ambiguous and depends on whether the pass-through effect dominates the elasticity effect. Here, the detrimental reallocation effect dominates the beneficial pro-competitive effect, and the overall effect of changes in markups is negative.

The elasticity of real GDP per capita is small at 0.024. This is much smaller than the elasticity of consumer welfare. This is a consequence of the well-known result that the welfare benefits of new goods are not reflected in changes in real GDP.[25] Indeed, the positive changes in real GDP can be entirely attributed to the reduction in markups of existing firms. In particular, it also worth noting as before, that the movement in the minimal productivity cutoff on its own plays no role in determining real GDP or aggregate TFP, even though the model is not efficient. Hence, an increase in the cutoff does not translate into an increase in aggregate TFP.

Next, consider the efficient-entry case. 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 composition effect from endogenous entry is now much larger at 0.408. The intuition is the same as before: increases in population lead to a reduction in the aggregate price index; this shifts resources away from small firms facing elastic demand and charging low markups towards large firms facing less elastic demand and charging higher markups; this also increases the profit share and triggers entry. The main reason the effects are so much larger than they were in the efficient-selection case 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. Furthermore, since the new entry moves the price index, and the changes in the price index cause large firms to expand relative to small firms, there is a feedback loop from new entry, to changes in the price index, to changes in aggregate profits, back to entry, amplifying the effect.

The selection effect from the adjustment of the exit cutoff is now non-zero and negative at -0.27. The reason for this can be seen from inspecting Figure 4d, which shows that the infra-marginal surplus at the cutoff is much higher than average, hence, as the cutoff 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 markup effect is now more negative than it was before is because in the efficient-selection case there was too much entry, so the overall reductions

[25] 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 infra-marginal 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.

in markups had a positive effect (over and above the detrimental reallocation across existing firms). Since we are now imposing entry efficiency, this latter effect no longer operates, and the overall contribution of changing markups is more negative than before.

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 cutoff) as a driver of 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 the infra-marginal firms is perfectly horizontal. In other words, the marginal firms produce no infra-marginal 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 technical efficiency effect of 0.016 and an allocative efficiency effect of 0.114. The adjustment of entry still accounts for the bulk of changes in allocative efficiency at 0.116. The contribution of the adjustment of selection to welfare is now positive, but it is still close to insignificant at 0.001. And the pro-competitive effects through the endogenous adjustment of markups are 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 adjustment of entry (holding fixed markups and selection) at 0.269, with a negligible contribution of the adjustment of selection at 0.003 and a small and negative pro-competitive effect through the endogenous adjustment of markups at -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. In fact, given our choice of $\bar{\mu}$, it is impossible to choose initial conditions that give selection a bigger role to play than the numbers in the paragraph above.

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. The results can be found in Table 2.

The most striking difference is that the elasticity of consumer welfare to population is much smaller, because changes in allocative efficiency become negligible. With efficient entry $\bar{\delta} = \bar{\mu} = 1.045$ and $\bar{\delta} = \bar{\mu} = 1.090$, there are no changes in allocative efficiency since the model

	$\bar{\mu} = 1.045$		$\bar{\mu} = 1.090$	
	$\bar{\delta} = \delta_{\theta^*}$	$\bar{\delta} = \bar{\mu}$	$\bar{\delta} = \delta_{\theta^*}$	$\bar{\delta} = \bar{\mu}$
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 population following Propositions 1 and 2 for homogeneous firms.

is efficient. With efficient selection, the changes in allocative efficiency are positive but fairly small. They are small because without heterogeneity, there are no longer changes in allocative efficiency from the entry margin (holding fixed selection and markups) or from the selection margin (holding fixed markups). Instead, all the changes in allocative efficiency come from the pro-competitive effect through the reduction of markups. The changes in allocative efficiency are positive because $\bar{\delta} < \bar{\mu}$, and so the reduction in markups, which reduces entry, improves welfare because there was too much entry to begin with.

This exercise emphasizes once again that the much discussed pro-competitive effects are small in comparison the changes in allocative efficiency arising from reallocations across heterogeneous firms that we discussed above.

The effects of population increases on GDP per capita are also smaller than with heterogeneous firms. This is simply because, just like the change in welfare, the change in the price index which triggers the reduction in markups is smaller.

Are there larger increasing returns at the macro vs. micro levels? We also compare 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 $1/\mathbb{E}[1/(ac_\theta/mc_\theta - 1)] = \bar{\mu} - 1$.[26]

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.[27]

[26] 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 $l_e + l_o = 1 - 1/\bar{\mu}$.

[27] The technical efficiency effect $(\bar{\delta} - 1)\,\mathrm{d} \log L$ in the decomposition of the aggregate increasing returns to scale does not exactly coincide with the average microeconomic technological return to scale $(\bar{\mu} - 1)\,\mathrm{d} \log L$ because of our choice for the allocation matrix. However, the difference is small. The reason for this small difference is that under our chosen definition, keeping the allocation matrix constant results in the creation of new varieties. Had we

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. This result would neither hold if the economy were efficient nor if there were no heterogeneity.

	$\bar{\mu} = 1.045$		$\bar{\mu} = 1.090$	
	$\bar{\delta} = \delta_{\theta^*}$	$\bar{\delta} = \bar{\mu}$	$\bar{\delta} = \delta_{\theta^*}$	$\bar{\delta} = \bar{\mu}$
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
Adj. of Entry: $\Delta \log Y^{\epsilon} - \Delta \log Y^{tech}$	0.066	0.107	0.145	0.272
Adj. of Exit: $\Delta \log Y^{\epsilon,\theta^*} - \Delta \log Y^{\epsilon}$	0.000	-0.065	0.000	-0.176
Adj. of Markups: $\Delta \log Y^{\epsilon,\theta^*,\mu} - \Delta \log Y^{\epsilon,\theta^*}$	0.008	0.008	0.017	0.021
Real GDP per capita	0.025	0.024	0.054	0.051

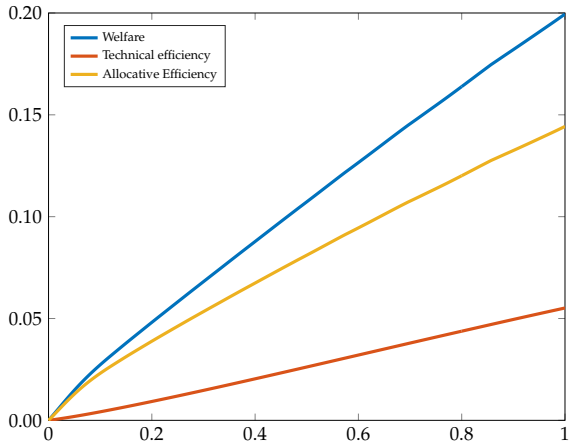
Table 3: The average elasticity of welfare and real GDP per capita to population for a large shock $\Delta \log L = 0.5$.

Are there big nonlinearities? Table 1 shows that changes in allocative efficiency account for the bulk of the marginal increase in welfare from a marginal increase in population. More precisely, most of the effect comes from the reallocations from small firms with low markups to big firms with high markup that take place when only the entry margin is allowed to adjust, with a more limited role for the reallocations coming from the adjustment of markups and exit. One might worry that these composition effects could peter out quickly if we kept increasing the size of the population. Fortunately, since the model is identified globally, we can not only solve the model for small shocks but also for big shocks and thereby analyze potential nonlinearities.[28] The results in Table 3 and Figure 6 below show that the forces identified for small shocks by Proposition 3 continue to apply for large shocks.

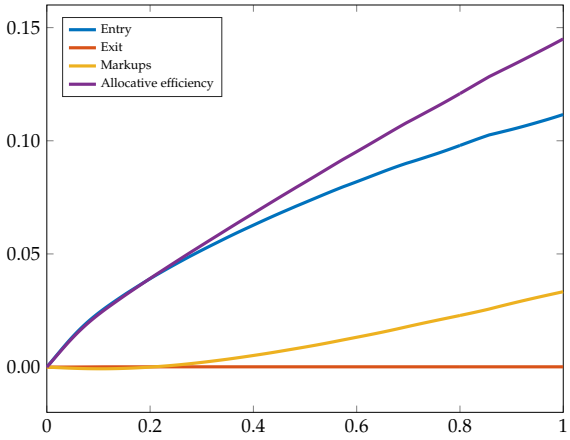
Table 3 reports the average (rather than the marginal) elasticity of welfare to a 0.5 log point increase in population (a roughly 68% increase). The magnitude of and the decomposition of the average effects are similar to those for the marginal effects reported in Table 1. Although the model is far from being log-linear, the conclusion that changes in allocative efficiency account for the bulk of aggregate increasing returns to scale, and in particular those changes driven the reallocation from small firms with low markups to big firms with high markups

adopted the alternative definition described in footnote 16, keeping the allocation matrix constant would not result in the creation of new firms and instead would only scale up existing producers. The resulting technical efficiency effect would then exactly coincide with the average microeconomic technological return to scale.

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



(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 3, 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.

that take place when only the entry margin is allowed to adjust, with a more limited role for the reallocations coming from adjustment of markups and exit. Interestingly, for the calibrations assuming efficient selection with $\bar{\delta} = \delta_{\theta^*}$ (columns 1 and 3), adjustments in exit continue to play a negligible role even though the selection cutoff increases with the size of the shock.

Figure 6 shows cumulated changes in welfare, their breakdown into in technical and allocative efficiency, and the breakdown of allocative efficiency into the reallocation effects arising from the different margins of adjustment of firm behavior (entry, exit, and markups), focusing on the calibration assuming efficient selection with $\bar{\delta} = \delta_{\theta^*}$ and average markups $\bar{\mu} = 1.09$ (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 reallocations from the adjustment of markups start to account for a non-trivial part of overall changes in allocative efficiency. This happens because as we increase population, average markups actually increase through the composition effect reallocating resources away from low-markup firms and towards high-markup firms, and despite the fact that all individual markups decrease. This means that entry becomes more excessive, and hence that reallocations triggered by individual markup reductions improve allocative efficiency more.

7 Extensions

Before concluding, we describe some extensions which are developed in the appendix.

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

Optimal policy and distance to the efficient frontier. In the main text of the paper, we have focused exclusively on comparative statics. In Appendix D.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 neatly decomposes the contributions of the different margins of inefficiency (entry, selection, and relative production) to the overall amount of misallocation. In Appendix D.2 we compute the distance to the efficient frontier in our empirical application. There, we quantify the extent of misallocation in the decentralized equilibrium compared to the first-best allocation, or in other words, the gains from optimal policy. We find the number to be somewhere between 2.5% and 6.8% in Belgium. We also explain why this number is consistent with our finding of large cumulated changes in allocative efficiency even for large increases in population.

8 Conclusion

We analyze how changes in market size affect welfare in a model with monopolistic competition, heterogeneous firms, variable markups and pass-throughs, entry, and exit. We decompose the overall change into changes in technical and allocative efficiency. We use firm-level information to non-parametrically recover demand curves and quantify this decomposition.

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. The most important reallocation is a composition effect that shifts resources from small firms with low markups towards big firms with high markups. The toughening of selection and the pro-competitive reduction in markups play only minor roles in comparison.

Two important implications are: that the aggregate return to scale is an endogenous equilibrium outcome shaped by frictions and market structure, and not an exogenous technological primitive; and that even mild increasing returns to scale at the micro level can give rise to large increasing returns to scale at the macro level.

Natural extensions of our analysis would be to consider alternative market structures and entry processes, to add explicit dynamics, to model rich input-output structures, and to explicitly take trade into account. We are pursuing these generalizations in ongoing work.

References

Aghion, Philippe and Peter Howitt, "A Model of Growth through Creative Destruction," *Econometrica*, March 1992, *60* (2), 323–351.

Amiti, Mary, Oleg Itskhoki, and Jozef Konings, "International Shocks, Variable Markups, and Domestic Prices," *The Review of Economic Studies*, 02 2019.

Arkolakis, Costas, Arnaud Costinot, Dave Donaldson, and Andrés Rodríguez-Clare, "The elusive pro-competitive effects of trade," *The Review of Economic Studies*, 2018, *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.

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

Bartelme, Dominick G., Arnaud Costinot, Dave Donaldson, and Andres Rodriguez-Clare, "The Textbook Case for Industrial Policy: Theory Meets Data," NBER Working Papers 26193, National Bureau of Economic Research, Inc August 2019.

Behrens, Kristian, Giordano Mion, Yasusada Murata, and Jens Suedekum, "Quantifying the gap between equilibrium and optimum under monopolistic competition," *Higher School of Economics Research Paper No. WP BRP*, 2018, *185.*

Bilbiie, Florin O, Fabio Ghironi, and Marc J Melitz, "Monopoly power and endogenous product variety: Distortions and remedies," *American Economic Journal: Macroeconomics*, 2019, *11* (4), 140–74.

Chamberlin, Edward Hastings, *Theory of monopolistic competition: A re-orientation of the theory of value*, Oxford University Press, London, 1933.

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.

Grossman, Gene M. and Elhanan Helpman, "Quality Ladders in the Theory of Growth," *Review of Economic Studies*, 1991, *58* (1), 43–61.

Hopenhayn, Hugo A, "Entry, exit, and firm dynamics in long run equilibrium," *Econometrica: Journal of the Econometric Society*, 1992, pp. 1127–1150.

Hsieh, Chang-Tai and Peter J Klenow, "Misallocation and manufacturing TFP in China and India," *The Quarterly journal of economics*, 2009, *124* (4), 1403–1448.

Hulten, Charles R, "Growth Accounting with Intermediate Inputs," *The Review of Economic Studies*, 1978, pp. 511–518.

Jones, Charles I, "R&D-Based Models of Economic Growth," *Journal of Political Economy*,

August 1995, *103* (4), 759–784.

_ , "The End of Economic Growth? Unintended Consequences of a Declining Population," Technical Report, Working Paper, Stanford University 2019.

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.

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

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.

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.

Perla, Jesse, Christopher Tonetti, and Michael E. Waugh, "Equilibrium Technology Diffusion, Trade, and Growth," Technical Report January 2020.

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

Restuccia, Diego and Richard Rogerson, "Policy distortions and aggregate productivity with heterogeneous establishments," *Review of Economic dynamics*, 2008, *11* (4), 707–720.

Robinson, Joan, *The economics of imperfect competition*, Springer, 1933.

Romer, Paul M, "Increasing Returns and Long-run Growth," *Journal of Political Economy*, October 1986, *94* (5), 1002–1037.

_ , "Endogenous Technological Change," *Journal of Political Economy*, October 1990, *98* (5), 71–102.

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

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 4.

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 4: 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 7.

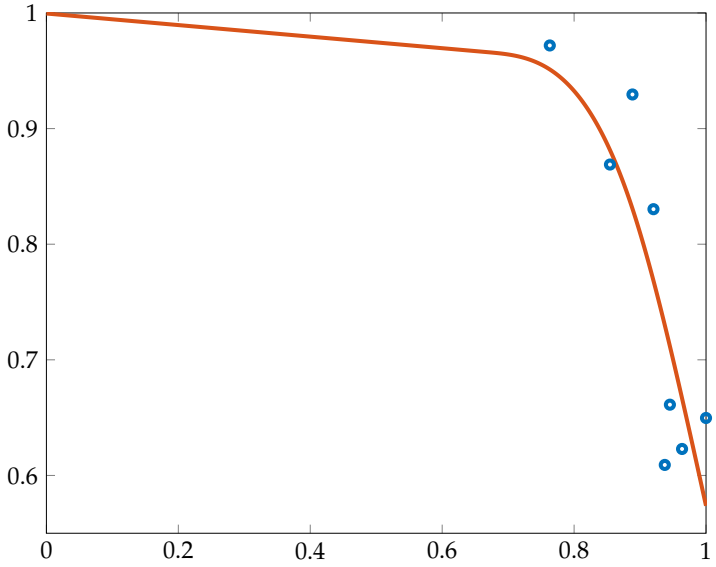


Figure 7: Average pass-through for firms up to a certain size α from Prodcom.

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 5 displays the underlying data.

As before, we let $\theta \in [0, 1]$ index the fraction of observations up to some size. Then the

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

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 8. 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.

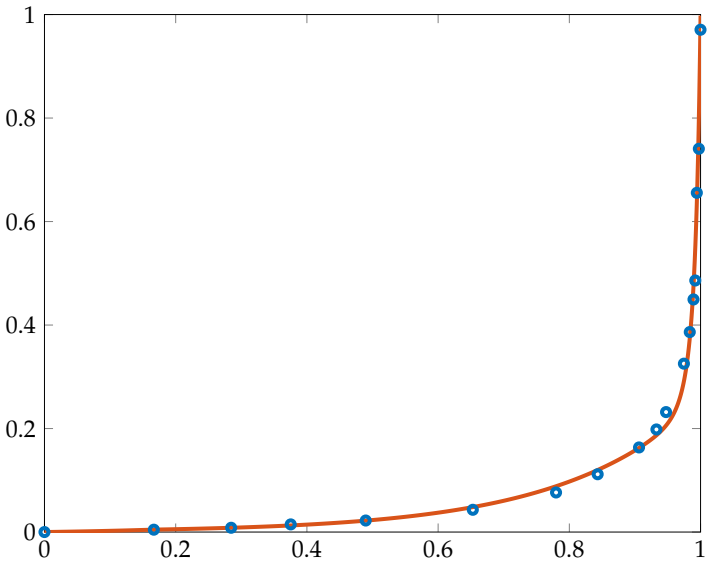


Figure 8: Cumulative share distribution Λ_θ from VAT declarations.

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 6 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 7. The basic qualitative message of our previous results in Table 1 is unchanged, and the composition effects from the adjustment of the entry margin (holding fixed markups and selection) are still overwhelmingly the dominant force in the model.

Appendix C 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 2.2 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

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

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.

Aggregate price index. Differentiating the definition of the price and demand indices, we find

$$-d\log P = -\lambda_{\theta^*}\frac{g(\theta^*)}{1-G(\theta^*)}d\theta^* + d\log M + d\log Y + \mathbb{E}_\lambda\left[\left(1-\frac{1}{\sigma_\theta}\right)d\log(\frac{y_\theta}{Y})\right]. \tag{54}$$

Combining this equation with the equation for quantities and markups, we get

$$-d\log P = -\lambda_{\theta^*}\frac{g(\theta^*)}{1-G(\theta^*)}d\theta^* + d\log M + d\log Y + \mathbb{E}_\lambda\left[\rho_\theta(\sigma_\theta-1)\left(d\log A_\theta + d\log P\right)\right]. \tag{55}$$

Finally, combining with the second equation for entry derived below, we find

$$d\log P = \frac{-d\log Y - \mathbb{E}_{\lambda(1-1/\mu)}\left[(\sigma_\theta-1)d\log A_\theta\right] + \frac{\frac{f_e\Delta}{L}d\log(\frac{f_e\Delta}{L})+[1-G(\theta^*)]\frac{f}{L}d\log(\frac{f}{L})}{\frac{f_e\Delta}{L}+[1-G(\theta^*)]\frac{f}{L}}}{\mathbb{E}_{\lambda(1-1/\mu)}\left[\sigma_\theta\right]}. \tag{56}$$

This equation for the aggregate price index can be replaced in all the equations below.

	$\bar{\mu} = 1.045$		$\bar{\mu} = 1.090$	
	$\bar{\delta} = \delta_{\theta^*}$	$\bar{\delta} = \bar{\mu}$	$\bar{\delta} = \delta_{\theta^*}$	$\bar{\delta} = \bar{\mu}$
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: $d \log Y^{\epsilon} - \Delta \log Y^{tech}$	0.056	0.136	0.126	0.327
Adj. of Exit: $d \log Y^{\epsilon,\theta^*} - \Delta \log Y^{\epsilon}$	0.000	-0.037	0.000	-0.094
Adj. of Markups: $d \log Y^{\epsilon,\theta^*,\mu} - \Delta \log Y^{\epsilon,\theta^*}$	0.004	-0.012	0.008	-0.029
Real GDP per capita	0.024	0.016	0.051	0.052

Table 7: The elasticity of welfare and real GDP per capita to population following Propositions 3 and 4 for heterogeneous firms case using product-level data.

Entry. We derive two equations for free entry. The first equation is obtained as follows. Differentiating the free-entry condition, we find

$$\frac{\frac{f_e\Delta}{L} d\log(\frac{f_e\Delta}{L}) + [1-G(\theta^*)]\frac{f}{L} d\log(\frac{f}{L})}{\frac{f_e\Delta}{L} + [1-G(\theta^*)]\frac{f}{L}} + d\log M - \frac{g(\theta^*)}{1-G(\theta^*)} d\theta^* = \mathbb{E}_{\lambda(1-1/\mu)}\left[d\log\left(\lambda_\theta\left(1-\frac{1}{\mu_\theta}\right)\right)\right]. \tag{57}$$

Combining with the equation for variable profit shares, we get

$$\begin{aligned} d\log M = &-\frac{\frac{f_e\Delta}{L} d\log(\frac{f_e\Delta}{L}) + [1-G(\theta^*)]\frac{f}{L} d\log(\frac{f}{L})}{\frac{f_e\Delta}{L} + [1-G(\theta^*)]\frac{f}{L}} + \lambda_{\theta^*}\frac{g(\theta^*)}{1-G(\theta^*)} d\theta^* \\ &+ \mathbb{E}_{\lambda(1-1/\mu)}\left[(\sigma_\theta - 1)\left(d\log A_\theta + d\log P\right)\right] - \mathbb{E}_{\lambda}\left[\rho_\theta(\sigma_\theta - 1)\left(d\log A_\theta + d\log P\right)\right]. \end{aligned}$$

We can also use the equation for the demand index to get

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

Sales shares. Differentiating the sales shares equation, we find

$$d\log\lambda_\theta = d\log M - \frac{g(\theta^*)}{1-G(\theta^*)} + d\log Y + (\sigma_\theta - 1) d\log(\frac{A_\theta}{\mu_\theta}) + \sigma_\theta d\log P. \tag{58}$$

Combining with the third equation for entry, we get

$$d\log\lambda_\theta = (\lambda_{\theta^*} - 1)\frac{g(\theta^*)}{1-G(\theta^*)} d\theta^* + \rho_\theta(\sigma_\theta - 1)\left(d\log A_\theta + d\log P\right)$$

$$- \mathbb{E}_\lambda \left[\rho_\theta (\sigma_\theta - 1) \left(d \log A_\theta + d \log P \right) \right].$$

Markups. Differentiating the markup equation, we get

$$d \log \mu_\theta = (1 - \rho_\theta) \left(d \log A_\theta + d \log P \right). \tag{59}$$

Variable profit shares. Combining the equations for sales shares and for markups, we get

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

Quantities. Differentiating the individual demand function, we find

$$d \log(\frac{y_\theta}{Y}) = \sigma_\theta \left(d \log(\frac{A_\theta}{\mu_\theta}) + d \log P \right). \tag{60}$$

Combining with the equation for markups, we get

$$d \log(\frac{y_\theta}{Y}) = \rho_\theta \sigma_\theta \left(d \log A_\theta + d \log P \right). \tag{61}$$

Selection. Differentiating the selection condition, we get

$$(\sigma_{\theta^*} - 1) \left(\frac{\partial \log A_\theta}{\partial \theta} |_{\theta = \theta^*} \right) d\theta^* = -d \log \left(\lambda_{\theta^*} \left(1 - \frac{1}{\mu_{\theta^*}} \right) \right) + d \log M - \frac{g(\theta^*)}{1 - G(\theta^*)} d\theta^* + d \log(\frac{f_o}{L}). \tag{62}$$

Combining with the equations for variable profits shares and entry, we get

$$(\sigma_{\theta^*} - 1) \left(\frac{\partial \log A_\theta}{\partial \theta} |_{\theta = \theta^*} \right) d\theta^* = -(\sigma_{\theta^*} - 1) \left(d \log A_{\theta^*} + d \log P \right) - d \log P - d \log Y + d \log(\frac{f_o}{L}), \tag{63}$$

where we note that

$$\frac{\partial \log A_\theta}{\partial \theta} |_{\theta = \theta^*} = \frac{g(\theta^*)}{g_a(\log A_{\theta^*})}. \tag{64}$$

Welfare. Differentiating the consumer welfare equation, we get

$$d \log Y = -(\delta_{\theta^*} - 1) \lambda_{\theta^*} \frac{g(\theta^*)}{1 - G(\theta^*)} d\theta^* + (\bar{\delta} - 1) \, d \log M + \mathbb{E}_\lambda \left[d \log(\frac{A_\theta}{\mu_\theta}) \right]. \tag{65}$$

Combining with the equation for markups, we get

$$d \log Y = -(\delta_{\theta^*} - 1) \lambda_{\theta^*} \frac{g(\theta^*)}{1 - G(\theta^*)} d\theta^* + (\bar{\delta} - 1) d \log M + \mathbb{E}_\lambda \left[\rho_\theta d \log A_\theta - (1 - \rho_\theta) d \log P \right]. \tag{66}$$

Combining with the equations for the aggregate price index and entry leads to a fixed point in $d \log Y$.

Appendix D 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 D.1, we provide an analytical second-order approximation which neatly decomposes 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 D.3. In Appendix D.1, we compute the distance to the frontier in our empirical application.

D.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 5. *The difference between welfare at the first-best allocation and the decentralized equilibrium can be approximated around $t = 0$ by*

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

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 infra-marginal surplus ratio $\mu_\theta/\mathbb{E}_\lambda[\delta_\theta]$. It also scales with the elasticities of substitution σ_θ.[29]

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 infra-marginal 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 infra-marginal surplus of the marginal firm δ_{θ^*} and that of

[29] 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).

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.[30]

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

D.2 Empirical Application

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 8: 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 8 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 6.3, 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.[31]

[30] If there are many firms at the cutoff (high λ_{θ^*}) or the cutoff 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.

[31] 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 already discussed in footnotes 1 and 15.

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.

D.3 Proof of Proposition 5

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 infra-marginal 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.[32] 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

$$(1 - G(\theta^*))M \int_{\theta^*}^{\infty} \Upsilon(\frac{y_\theta}{Y}) \frac{g(\theta)}{1 - G(\theta^*)} d\theta = 1, \tag{67}$$

$$\Lambda_L = \int_{\theta^*}^{\infty} \frac{\lambda_\theta}{\tau_\theta} \frac{g(\theta)}{1 - G(\theta^*)} d\theta, \tag{68}$$

$$\frac{M\Lambda_L f_e \Delta}{L} = \int_{\theta^*}^{\infty} \left(\lambda_\theta \frac{1}{\tau_\theta} \left(1 - \frac{1}{\mu_\theta} \right) - \frac{(1 - G(\theta^*))M\Lambda_L f_o}{L} \right) \frac{g(\theta)}{1 - G(\theta^*)} d\theta, \tag{69}$$

$$\lambda_{\theta^*} \frac{1}{\tau_{\theta^*}} \left(1 - \frac{1}{\mu_{\theta^*}} \right) = \frac{(1 - G(\theta^*))M\Lambda_L f_o}{L}, \tag{70}$$

$$\lambda_\theta = (1 - G(\theta^*))M \frac{\tau_\theta \mu_\theta \Lambda_L y_\theta}{A_\theta}, \tag{71}$$

$$\frac{\tau_\theta \mu_\theta \Lambda_L}{A_\theta} = P\Upsilon'(\frac{y_\theta}{Y}), \tag{72}$$

$$P = \frac{\bar{\delta}}{Y}, \tag{73}$$

$$\frac{1}{\bar{\delta}} = (1 - G(\theta^*))M \int_{\theta^*}^{\infty} \frac{y_\theta}{Y} \Upsilon'(\frac{y_\theta}{Y}) \frac{g(\theta)}{1 - G(\theta^*)} d\theta. \tag{74}$$

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

Efficiency requires

$$\mu_\theta = \frac{1}{\tau_\theta} = \frac{\Upsilon_\theta}{\frac{y_\theta}{Y}\Upsilon'_\theta}. \tag{75}$$

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.

Step 1:

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

Aggregate price index:

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

Sales shares:

$$d\log\lambda_\theta = d\log M - \frac{g(\theta^*)}{1-G(\theta^*)}d\theta^* + d\log Y - \left(\frac{\Upsilon'_\theta}{-\frac{y_\theta}{Y}\Upsilon''_\theta}-1\right)(d\log\mu_\theta + d\log\tau_\theta + d\log\Lambda_L) + \frac{\Upsilon'_\theta}{-\frac{y_\theta}{Y}\Upsilon''_\theta}d\log P. \tag{76}$$

Variable profits:

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

Quantities:

$$d\log\left(\frac{y_\theta}{Y}\right) = -\frac{\Upsilon'_\theta}{-\frac{y_\theta}{Y}\Upsilon''_\theta}\left(d\log\mu_\theta + d\log\tau_\theta + d\log\Lambda_L - d\log P\right). \tag{77}$$

Labor share:

$$d\log\Lambda_L = \frac{g(\theta^*)}{1-G(\theta^*)}d\theta^* + \frac{\int_{\theta^*}^{\infty}\frac{\lambda_\theta}{\tau_\theta}\left(d\log\lambda_\theta - d\log\tau_\theta\right)\frac{g(\theta)}{1-G(\theta^*)}d\theta}{\int_{\theta^*}^{\infty}\frac{\lambda_\theta}{\tau_\theta}\frac{g(\theta)}{1-G(\theta^*)}d\theta,} - \frac{\frac{\lambda_{\theta^*}}{\tau_{\theta^*}}\frac{g(\theta^*)}{1-G(\theta^*)}d\theta^*}{\int_{\theta^*}^{\infty}\frac{\lambda_\theta}{\tau_\theta}\frac{g(\theta)}{1-G(\theta^*)}d\theta,}. \tag{78}$$

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:

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

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 cutoff:

$$\left(\frac{\Upsilon'_{\theta^*}}{-\frac{y_{\theta^*}}{Y}\Upsilon''_{\theta^*}}-1\right)\frac{\partial\log A_\theta}{\partial\theta}\Big|_{\theta=\theta^*}d\theta^* = -d\log\left(\frac{\lambda_{\theta^*}}{\Lambda_L}\frac{1}{\tau_{\theta^*}}\left(1-\frac{1}{\mu_{\theta^*}}\right)\right) + d\log M - \frac{g(\theta^*)}{1-G(\theta^*)}d\theta^*. \tag{79}$$

Welfare:

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

or

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

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

$$d\log\Lambda_L = -\int_{\theta^*}^{\infty}\lambda_\theta d\log\tau_\theta\frac{g(\theta)}{1-G(\theta^*)}d\theta, \tag{80}$$

$$-d\log P + d\log\Lambda_L = d\log Y - \int_{\theta^*}^{\infty}\lambda_\theta d\log\tau_\theta\frac{g(\theta)}{1-G(\theta^*)}d\theta, \tag{81}$$

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

$$\left(\frac{\Upsilon'_{\theta^*}}{-\frac{y_{\theta^*}}{Y}\Upsilon''_{\theta^*}}-1\right)\left.\frac{\partial\log A_\theta}{\partial\theta}\right|_{\theta=\theta^*}d\theta^* = -d\log Y + \frac{\Upsilon'_{\theta^*}}{-\frac{y_{\theta^*}}{Y}\Upsilon''_{\theta^*}}\left(d\log\tau_{\theta^*}-d\log P+d\log\Lambda_L\right). \tag{82}$$

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

$$d\log\Lambda_L = -\int_{\theta^*}^{\infty}\lambda_\theta d\log\tau_\theta\frac{g(\theta)}{1-G(\theta^*)}d\theta, \tag{83}$$

$$-d\log P = d\log Y, \tag{84}$$

$$\left.\frac{\partial\log A_\theta}{\partial\theta}\right|_{\theta=\theta^*}d\theta^* = d\log Y + \frac{\frac{\Upsilon'_{\theta^*}}{-\frac{y_{\theta^*}}{Y}\Upsilon''_{\theta^*}}}{\frac{\Upsilon'_{\theta^*}}{-\frac{y_{\theta^*}}{Y}\Upsilon''_{\theta^*}}-1}\left(d\log\tau_{\theta^*}-\int_{\theta^*}^{\infty}\lambda_\theta d\log\tau_\theta\frac{g(\theta)}{1-G(\theta^*)}d\theta\right). \tag{85}$$

Plugging into welfare, we get

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

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)

$$d\log\mu_\theta\approx-\log\left(\frac{\mu_\theta}{\frac{\Upsilon_\theta}{\frac{y_\theta}{Y}\Upsilon'_\theta}}\right),\tag{86}$$

$$d\log\tau_\theta\approx-\log\left(\frac{\Upsilon_\theta}{\frac{y_\theta}{Y}\Upsilon'_\theta}\right).\tag{87}$$

Plugging into welfare, we get

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

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

$$\mathcal{L}\approx\frac{1}{2}d\log Y.\tag{88}$$

Using the notation in the paper, we therefore get

$$\mathcal{L} \approx -\frac{1}{2}\mathbb{E}_\lambda\left[\left(1-\frac{\mathbb{E}_\lambda\left[\delta_\theta\right]-1}{\mu_\theta-1}\right)\log\left(\frac{\mathbb{E}_\lambda\left[\delta_\theta\right]}{\mu_\theta}\right)\right]+\frac{1}{2}\lambda_{\theta^*}\gamma_{\theta^*}\left(\mathbb{E}_\lambda\left[\delta_\theta\right]-\delta_{\theta^*}\right)\mu_\theta^*\log\left(\frac{\mathbb{E}_\lambda\left[\delta_\theta\right]}{\delta_{\theta^*}}\right), \quad (89)$$

or

$$\mathcal{L} \approx \frac{1}{2}\mathbb{E}_\lambda\left[\frac{\left(\frac{\mu_\theta}{\mathbb{E}_\lambda[\delta_\theta]}-1\right)^2}{\mu_\theta-1}\mathbb{E}_\lambda\left[\delta_\theta\right]\frac{\mathbb{E}_\lambda\left[\delta_\theta\right]}{\mu_\theta}\right]+\frac{1}{2}\lambda_{\theta^*}\frac{\mu_{\theta^*}}{\delta_{\theta^*}}\gamma_{\theta^*}\left(\mathbb{E}_\lambda\left[\delta_\theta\right]-\delta_{\theta^*}\right)^2, \quad (90)$$

or

$$\mathcal{L} \approx \frac{1}{2}\mathbb{E}_\lambda\left[\frac{\mu_\theta}{\mu_\theta-1}\left(\frac{\mu_\theta}{\mathbb{E}_\lambda\left[\delta_\theta\right]}-1\right)^2\right]+\frac{1}{2}\lambda_{\theta^*}\gamma_{\theta^*}\left(\mathbb{E}_\lambda\left[\delta_\theta\right]-\delta_{\theta^*}\right)^2, \quad (91)$$

or

$$\mathcal{L} \approx \frac{1}{2}\mathbb{E}_\lambda\left[\sigma_\theta\left(\frac{\mu_\theta}{\mathbb{E}_\lambda\left[\delta_\theta\right]}-1\right)^2\right]+\frac{1}{2}\lambda_{\theta^*}\gamma_{\theta^*}\left(\mathbb{E}_\lambda\left[\delta_\theta\right]-\delta_{\theta^*}\right)^2, \quad (92)$$

or

$$\mathcal{L} \approx \frac{1}{2}\mathbb{E}_\lambda\left[\sigma_\theta\left[\left(\frac{\mu_\theta}{\mathbb{E}_\lambda\left[\delta_\theta\right]}-\frac{\mathbb{E}_\lambda\left[\mu_\theta\right]}{\mathbb{E}_\lambda\left[\delta_\theta\right]}\right)^2+\left(\frac{\mathbb{E}_\lambda\left[\mu_\theta\right]}{\mathbb{E}_\lambda\left[\delta_\theta\right]}-1\right)^2\right]\right]+\frac{1}{2}\lambda_{\theta^*}\gamma_{\theta^*}\left(\mathbb{E}_\lambda\left[\delta_\theta\right]-\delta_{\theta^*}\right)^2, \quad (93)$$

Appendix E 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.

E.1 Shocks to Fixed Costs

As with population shocks, we begin by focusing on the homogeneous-firm case.

E.1.1 Homogeneous Firms

Proposition 6. *Suppose that firms have the same productivity $A_\theta = A$. In response to changes in population $d\log L$, changes in consumer welfare are given by*

$$d\log Y = \underbrace{(\delta-1)d\log L}_{\textit{technical efficiency}} + \underbrace{\delta\frac{\xi}{1-\xi}d\log L}_{\textit{allocative efficiency}}, \quad (94)$$

where

$$\xi = \left(1-\rho\right)\left(1-\frac{\delta-1}{\mu-1}\right)\frac{1}{\sigma} = \left(1-\rho\right)\left(1-\frac{\delta}{\mu}\right). \quad (95)$$

Changes in entry costs $d\log(f_e\Delta)$ *and in overhead costs* $d\log f_0$ *respectively have the same effects on consumer welfare as change in population shocks* $d\log L = -[f_e\Delta/(f_e\Delta + f_0)]d\log(f_e\Delta)$ *and* $d\log L = -[f_o/(f_e\Delta + f_0)]d\log(f_o)$.

Proposition 7. *Suppose that firms have the same productivity* $A_\theta = A$. *In response to changes in population* $d\log L$, *changes in real GDP per capita are given by*

$$d\log Q = \frac{1-\rho}{\sigma}(d\log Y + d\log L), \tag{96}$$

where $d\log Y$ *is given by Proposition 1. Changes in entry costs* $d\log(f_e\Delta)$ *and in overhead costs* $d\log f_0$ *respectively have the same effects on these variables as change in population shocks* $d\log L = -[f_e\Delta/(f_e\Delta + f_o)]d\log(f_e\Delta)$ *and* $d\log L = -[f_o/(f_e\Delta + f_o)]d\log(f_o)$.

E.1.2 Heterogeneous Firms

Now, we consider shocks to fixed costs when firms are heterogeneous.

Proposition 8. *In response to changes in fixed costs of entry* $d\log(f_e\Delta)$ *and fixed overhead costs* $d\log f_o$, *changes in consumer welfare are given by*

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

where ξ^ϵ, ξ^{θ^*}, *and* ξ^μ *are given in Proposition 3 and*

$$\zeta^{\theta^*} = \left(\mathbb{E}_\lambda[\delta_\theta]-\delta_{\theta^*}\right)\left(\lambda_{\theta^*}\gamma_{\theta^*}\frac{1}{\sigma_{\theta^*}-1}\right). \tag{97}$$

As with population shocks, we can provide sufficient conditions under which changes in allocative efficiency amplify or mitigate the effects of the shocks.

Corollary 3. *Sufficient conditions for positive changes in allocative efficiency in response to decreases in the fixed cost of entry are the same as in Corollary 2. Indeed, (1), (2), and (3) imply* $\xi^\epsilon > 0$, $\xi^{\theta^*} > 0$, *and* $\xi^\mu > 0$. *Furthermore, (1) and (3) imply* $\zeta^{\theta^*} > 0$. *Sufficient conditions for positive changes in*

allocative efficiency in response to decreases in the fixed overhead cost if selection decreases ($d\theta^* < 0$) are that (1) and (2) hold but that (3) fail (too much selection).

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

With homogeneous firms, shocks to fixed costs act like scaled population shocks even in isolation. The equivalent shock to population is inversely related to the shock to the total fixed cost $-[f_e\Delta d\log(f_e\Delta) + (1-G(\theta^*))fd\log f]/[f_e\Delta + (1-G(\theta^*))f]$. This is no longer true with heterogeneous firms because the two fixed costs impact selection in different ways.

Consider first a reduction in the fixed cost of entry $d\log(f_e\Delta) < 0$. 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 $[(f_e\Delta)/[f_e\Delta + (1-G(\theta^*))f]]d\log(f_e\Delta) < 0$. 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 Proposition 3 and Corollary 2. The effect of the latter is to further increase the sales shares of exiting varieties by $-[\lambda_{\theta^*}\gamma_{\theta^*}/(\sigma_{\theta^*}-1)][(f_e\Delta)/[f_e\Delta + (1-G(\theta^*))f]]d\log(f_e\Delta) > 0$. This in turn increases consumer welfare by $-[(\mathbb{E}[\delta_\theta] - \delta_{\theta^*})\lambda_{\theta^*}\gamma_{\theta^*}/(\sigma_{\theta^*}-1)][(f_e\Delta)/[f_e\Delta + (1-G(\theta^*))f]]d\log(f_e\Delta) > 0$ 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 cutoff is reversed compared to the case of a reduction in the fixed cost of entry: compared to an increase in population by $-[(1-G(\theta^*))f/[f_e\Delta + (1-G(\theta^*))f]]d\log(f) > 0$, the increase in the fixed overhead cost reduces the selection cutoff, 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 8, 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 9. *In response to changes in fixed costs of entry $d\log(f_e\Delta)$ and fixed overhead costs*

$d \log f$, changes in real GDP per capita are given by

$$d \log Q = \left(\mathbb{E}_\lambda \left[(1-\rho_\theta)\right]\right)\left(\mathbb{E}_\lambda \left[\frac{1}{\sigma_\theta}\right]\right)\left(d \log Y + \frac{f_e \Delta d \log(f_e \Delta) + (1 - G(\theta^*)) f d \log f}{f_e \Delta + (1 - G(\theta^*)) f}\right), \tag{98}$$

where $d \log Y$ is given by Proposition 8.

Proposition 9 can be used to decompose real GDP per capita along the same lines as the decomposition of welfare in Proposition 8. Setting $\xi^\mu = \xi^{\theta^*} = 0$, $\zeta^{\theta^*} = 0$ and $\rho_\theta = 1$ holds fixed markups and selection but allows entry, setting $\xi^\mu = 0$ and $\rho_\theta = 1$ holds fixed markups but allows entry and selection to adjust, and finally apply Proposition 9 without any modification allows all margins to adjust.

E.2 CES Example

The CES case is once again very simple. We have $\sigma_\theta = \sigma$, $\mu_\theta = \mu = 1/(1 - 1/\sigma)$, $\rho = 1$, and $\delta = \sigma/(\sigma - 1)$. This implies that $\xi^\epsilon = \xi^{\theta^*} = \xi^\mu = 0$. The simplicity of this expression is a consequence of the fact that the equilibrium is efficient.

E.3 Shocks to Productivity

Now, we consider shocks to the distribution of productivity shifters, starting with the homogeneous-firm case before moving onto the heterogeneous case.

E.3.1 Homogeneous Firms

Whereas the model is not homothetic in population and fixed costs L, f_e, and f, it is homothetic in productivity A.

Proposition 10. *Suppose that firms have the same productivity $A_\theta = A$. In response to changes in productivity $d \log A$, changes in consumer welfare are given by*

$$d \log Y = \underbrace{d \log A}_{\text{technical efficiency}} + \underbrace{0}_{\text{allocative efficiency}}. \tag{99}$$

In response to a positive productivity shock $d \log A > 0$, individual quantities and consumer welfare all increase proportionately with the shock $d \log y = d \log Y = d \log A$. As a result, there is no change in markups $d \log \mu = 0$, and hence individual prices decrease proportionately with the shock $d \log p = -d \log A$. Entry remains unchanged $d \log M = 0$. More generally the allocation of resources actually stays unchanged, that is, the fractions of labor allocated to entry, overhead, and variable production remain unchanged. The absence of

reallocations in turn implies that there are no changes in allocative efficiency. There are only changes in technical efficiency.

Proposition 11. *Suppose that firms have the same productivity $A_\theta = A$. In response to changes in productivity $d\log A$, changes in real GDP per capita are given by*

$$d\log Q = d\log A. \tag{100}$$

Basically, the price of each variety is reduced by the amount of the productivity shock, with no change in markups.

E.3.2 Heterogeneous Firms

Finally, we consider shocks to productivities when the firm-size distribution is heterogeneous.

Proposition 12. *In response to changes in productivity $d\log A_\theta$, changes in consumer welfare are given by*

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

where ξ^ϵ, ξ^{θ^*}, and ξ^μ are given in Proposition 3 and

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

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 12, 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 $\nu^{\epsilon}[d \log A_\theta]$, $\nu^{\theta^*}[d \log A_\theta]$, and $\nu^{\mu}[d \log A_\theta]$ 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 5.2. 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 $\nu^{\epsilon}[d \log A_\theta]$, $\nu^{\theta^*}[d \log A_\theta]$, and $\nu^{\mu}[d \log A_\theta]$. 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 $\nu^{\epsilon}[d \log A_\theta]$ 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 $\nu^{\theta^*}[d \log A_\theta]$ is the following. Productivity shocks change exit behavior for given markups and aggregate price index. The selection cutoff tends to decrease if the productivity increases relatively more and if the elasticity of substitution is relatively higher at the cutoff. 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.

It turns out to be easier to determine if changes in consumer welfare are greater than sales- and pass-through-weighted changes in productivity.

Corollary 4. *Sufficient conditions for changes in consumer welfare to be greater than sales- and pass-through-weighted changes in productivity*

$$d\log Y > \mathbb{E}_\lambda\left[\rho_\theta d\log A_\theta\right] \tag{101}$$

in response to positive changes in productivity are the conditions (1), (2), and (3) of Corollary 2, together with two conditions ensuring that productivity shocks are sufficiently skewed towards large firms

$$\mathbb{E}_{\lambda(1-1/\mu)}\left[(\sigma_\theta-1)d\log A_\theta\right]-\mathbb{E}_\lambda\left[\rho_\theta(\sigma_\theta-1)d\log A_\theta\right]>0, \tag{102}$$

and

$$\mathbb{E}_{\lambda(1-1/\mu)}\left[(\sigma_\theta-1)d\log A_\theta\right]-(\sigma_{\theta^*}-1)d\log A_{\theta^*}>0. \tag{103}$$

Finally, 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 markups. The corresponding changes in real GDP per capita are respectively given by Proposition 13 below, but with $\xi^\mu=\xi^{\theta^*}=0$ and $\nu^\mu[d\log A_\theta]=\nu^{\theta^*}[d\log A_\theta]=0$ and $\rho_\theta=1$, $\xi^\mu=0$ and $\nu^\mu[d\log A_\theta]=0$ and $\rho_\theta=1$, and without any modification.

Proposition 13. *In response to changes in productivities $d\log A_\theta$, changes in real GDP per capita are given by*

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

where $d\log Y$ is given by Proposition 12.

Appendix F Differences in Tastes and Overhead Costs

In this section, we extend the model to allow for differences in tastes and overhead costs, by allowing the Kimball aggregator $\Upsilon(\frac{\theta}{Y};\theta)$ and the overhead cost $f_o(\theta)$ to depend on the type θ of the variety. Instead of ranking types by productivity, we rank them in increasing order of variable profits to overhead cost ratio so that $X_\theta=\lambda_\theta(1-1/\mu_\theta)/f_{o,\theta}$ is increasing in θ. The formulas in the paper continue to apply, with one exception: changes in selection are now given by

$$\left(\frac{\partial\log X_\theta}{\partial\theta}|_{\theta=\theta^*}\right)d\theta^*=-(\sigma_{\theta^*}-1)\left(d\log A_\theta-d\log P\right)-d\log P-d\log Y+d\log(\frac{f_{o,\theta^*}}{L}). \tag{104}$$

This implies that in all the formulas, we must now use $\gamma^*_\theta = [g_x(\log X_{\theta^*})/[1-G_x(\log X_{\theta^*})]]/(\sigma_{\theta^*}-1)$ where $g_x(\log X_\theta) = g(\theta)/(\partial \log X_\theta/\partial\theta)$.

Empirical implementation requires more data than the strategy described in Section 6.1. This is because the model is richer. To simplify the discussion, assume that overhead costs are homogeneous so that $f_{o,\theta} = f_o$.

The model without taste shocks required data on sales λ_θ and pass throughs ρ_θ as well as taking a stand on the average markup $\bar{\mu} = 1/[\mathbb{E}_\lambda[1/\mu_\theta]]$ and the average infra-marginal surplus ratio $\bar{\delta} = \mathbb{E}_\lambda[\delta_\theta]$. The nonlinear model could then be perfectly identified, allowing us to perform local and global counterfactuals.

Identification of the model with taste shocks requires additional data: we need data on markups μ_θ and we need to take a stand on the whole distribution of infra-marginal consumption surplus ratios δ_θ. Even with this data, we only have a local identification of the model, allowing us only to perform local first-order counterfactuals.

The reason is that in the model without taste shocks, a bigger firm is a smaller firm which received a positive productivity shock. Cross-sectional observations then allow us to trace the whole individual demand curve and hence to back out the Kimball aggregator up to some constants. This simplification disappears in the model with taste shocks.

Appendix G 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

$$d\log Q^q = \mathbb{E}_\lambda[d\log y_\theta]. \tag{105}$$

This is equal to

$$d\log Q^q = -d\log M + \lambda_{\theta^*}\frac{g(\theta^*)}{1-G(\theta^*)}d\theta^* + \mathbb{E}_\lambda\left[d\log(\frac{A_\theta}{\mu_\theta})\right], \tag{106}$$

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 $\boldsymbol{Q}^q(\mathcal{A},\mathcal{X})$:[33]

$$d\log Q^q = \underbrace{\frac{\partial \log \boldsymbol{Q}^q}{\partial \log \mathcal{A}} d\log \mathcal{A}}_{\text{technical efficiency}} + \underbrace{\frac{\partial \log \boldsymbol{Q}^q}{\partial \mathcal{X}} d\mathcal{X}}_{\text{allocative efficiency}} . \tag{107}$$

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.

Homogeneous Firms

Proposition 14. *Suppose that firms have the same productivity $A_\theta = A$. In response to changes in population $d\log L$, changes in real GDP per capita are given by*

$$d\log Q^q = \underbrace{-d\log L}_{\textit{technical efficiency}} + \underbrace{(1-\rho)(d\log Y + d\log L)}_{\textit{allocative efficiency}}, \tag{108}$$

where $d\log Y$ is given by Proposition 1. Changes in entry costs $d\log(f_e\Delta)$ and in overhead costs $d\log f_0$ respectively have the same effects on these variables as change in population shocks $d\log L = -[f_e\Delta/(f_e\Delta + f_o)]d\log(f_e\Delta)$ and $d\log L = -[f_o/(f_e\Delta + f_o)]d\log(f_o)$.

Changes in real GDP per capita measured with quantities are given by $d\log Q^q = d\log y$ so that $d\log Q^q = d\log Y + d\log(y/Y) = d\log Y - \rho(d\log Y + d\log L)$. They can be decomposed into changes in technical efficiency $-d\log L$ and changes in allocative efficiency $(1-\rho)d\log Y + (1-\rho)d\log L$.

Holding the allocation of resources constant, an increase in population $d\log L > 0$ leads to a proportional reduction $-d\log L < 0$ in the per-capita quantity of each variety because the number of varieties increases by $d\log L > 0$. The new varieties do not contribute at all

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

to changes in real GDP measured with quantities. This explains, in this case, the negative changes in technical efficiency $-d \log L < 0$.

Turning to changes in allocative efficiency, the pro-competitive reduction in markups reduces entry and increases the per-capita quantity of each variety. This explains, in this case, the positive changes in allocative efficiency $(1-\rho)(d \log Y + d \log L) > 0$.

CES Example Changes in real GDP per capita are given by

$$d \log Q^q = \underbrace{-d \log L}_{\text{technical efficiency}} + \underbrace{0}_{\text{allocative efficiency}}. \tag{109}$$

Even though the CES model is efficient, and there are no changes in allocative efficiency, increases in population reduce real GDP measured using the quantity index. Intuitively, the production of new goods means that fewer units of existing goods are produced per capita. Since the quantity index only measures changes in the quantity of existing goods per capita, it falls in response to the shock.

Heterogeneous Firms

Proposition 15. *In response to changes in population $d \log L$, changes in real GDP per capita are*

$$d \log Q^q = \underbrace{-d \log L}_{\textit{technical efficiency}} + \underbrace{\left(1 - \mathbb{E}_\lambda \left[\rho_\theta \sigma_\theta\right] \mathbb{E}_\lambda \left[\frac{1}{\sigma_\theta}\right]\right)\left(d \log Y + d \log L\right)}_{\textit{allocative efficiency}}, \tag{110}$$

where $d \log Y$ is given by Proposition 3.

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 4, but setting $\xi^\mu = \xi^{\theta^*} = 0$ and $\rho_\theta = 1$ (which holds fixed markups and the cutoffs), $\xi^\mu = 0$ and $\rho_\theta = 1$ (which holds fixed markups but allows the cutoff 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 4, 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.