The Darwinian Returns to Scale

David Rezza Baqaee
UCLA

Emmanuel Farhi
Harvard

Kunal Sangani*
Harvard

March 24, 2023

Abstract

How does an increase in market size, say due to globalization, affect welfare? We study this question using a model with monopolistic competition, heterogeneous markups, and fixed costs. We characterize changes in welfare and decompose changes in allocative efficiency into three different effects: (1) reallocations across firms with heterogeneous price elasticities due to intensifying competition, (2) reallocations due to the exit of marginally profitable firms, and (3) reallocations due to changes in firms' markups. Whereas the second and third effects have ambiguous implications for welfare, the first effect, which we call the Darwinian effect, always increases welfare regardless of the shape of demand curves. We non-parametrically calibrate demand curves with data from Belgian manufacturing firms and quantify our results. We find that mild increasing returns at the micro level can catalyze large increasing returns at the macro level. Between 70–90% of increasing returns to scale come from improvements in how a larger market allocates resources. The lion's share of these gains are due to the Darwinian effect, which increases the aggregate markup and concentrates sales and employment in high-markup firms. This has implications for policy: an entry subsidy, which harnesses Darwinian reallocations, can improve welfare even when there is more entry than in the first-best.

First version: November, 2019. Emmanuel Farhi tragically passed away in July, 2020. Emmanuel was a one-in-a-lifetime collaborator and friend. We thank Cédric Duprez and Oleg Itskhoki for sharing their data. We thank Maria Voronina and Sihwan Yang for outstanding research assistance. We thank Pol Antras, Andrew Atkeson, Ariel Burstein, Elhanan Helpman, Chad Jones, Kiminori Matsuyama, Marc Melitz, and Simon Mongey for helpful comments. We acknowledge research financial support from the Ferrante fund at Harvard University and NSF grant #1947611.

1 Introduction

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

In this paper, we propose a framework for decomposing these effects theoretically and quantitatively. We consider an economy with fixed entry and overhead costs, entry and exit, monopolistic competition, and heterogeneous markups. We argue that, to a large extent, increasing returns to scale at the aggregate level may reflect changes in allocative rather than technical efficiency. That is, a large share of the gains from an increase in market size—say due to immigration, fertility, or globalization—arise from how intensified competition reallocates resources across firms. Furthermore, we show that even mild increasing returns at the micro level (measured by the average ratio of marginal to average cost) can catalyze large increasing returns at the macro level.[1] Our findings hinge on the fact that we relax the popular constant-elasticity-of-substitution (CES) assumption.[2]

Models of monopolistic competition and entry commonly feature CES demand due to its tractability. The classic reference is Melitz (2003), which is a workhorse model of reallocation. However, since the equilibrium in this model is efficient, reallocations have no first-order effect on welfare. This is because efficiency ensures that the marginal social benefit of any input is equated across competing uses. Hence, reshuffling resources across uses cannot raise welfare. Moreover, efficiency also implies that micro- and macro-level returns to scale must be the same since, on the margin, allocating all incremental inputs to a single firm must yield the same aggregate return as the equilibrium allocation.

This simple elegance of CES demand comes at the expense of realism. CES demand imposes constant markups in both the cross-section and the time-series with complete pass-through of marginal costs into prices. In contrast, the data feature substantial heterogeneity in both markups and pass-throughs. Matching the empirical heterogeneity of markups and pass-throughs requires deviating from CES. This, in turn, introduces distortions in the equilibrium

[1] See Basu and Fernald (1997) on how aggregation can amplify micro returns to scale in distorted economies.

[2] We are not the first to consider deviations from CES in models of free entry and monopolistic competition. We discuss how our approach and findings differ from other papers that relax CES below.

and opens the door for reallocations triggered by shocks to primitives to affect welfare.

We relax CES by using a generalized homothetic demand system introduced by Matsuyama and Ushchev (2017).[3] This allows us to depart from Melitz (2003) in two ways. First, we allow each firm's price elasticity to vary with its position on its demand curve. Second, and in contrast to most existing studies (e.g. Zhelobodko et al. 2012 and Dhingra and Morrow 2019), we allow firms to face differently shaped residual demand curves and to have different overhead costs. This added flexibility is useful for matching data, but, more importantly, it allows us to cleanly isolate different channels of reallocation using special cases.

We characterize how welfare changes in response to an increase in market size. The response of welfare consists of a change in technical efficiency (i.e., an increase in welfare holding the allocation of resources across uses constant) and a change in allocative efficiency due to endogenous reallocations. We decompose these reallocations into three distinct channels that we call (1) the Darwinian effect, (2) the selection effect, and (3) the pro/anti-competitive effect. We briefly discuss these effects.

The Darwinian effect (1) captures how firms with different price-elasticities are differentially affected by changes in the aggregate price index holding fixed their markups. To understand this effect, consider the loglinearized per-capita demand curve for variety θ:

$$d \log y_\theta = -\sigma_\theta \left[d \log p_\theta - d \log P \right] - d \log P,$$

where y_θ is quantity, p_θ is the price, σ_θ is the price elasticity, P is a market-level price index, and per-capita spending is the numeraire. When the market expands and new firms enter, the market-level price index P falls and intensifies competition for all firms. However, not all varieties are exposed in the same way. Varieties with more inelastic demand are relatively insulated from changes in the price index.

Holding markups constant, firms with relatively inelastic demand thus expand relative to firms with more elastic demand. Since the markup of each firm is inversely related to its demand elasticity, this means that high-markup firms expand relative to low-markup firms. From a social perspective, high-markup firms are too small relative to low-markup firms in the initial equilibrium. Hence, this effect always improves efficiency regardless of the shape of demand curves. We call this the *Darwinian* effect because a more competitive environment automatically selects and expands the "fittest" firms (those with the higher markups).

In contrast, the selection and pro/anti-competitive effect, which have been studied in detail in previous work, have theoretically ambiguous effects on welfare. The selection effect (2) results from the fact that, as the market expands, the minimum level of profitability a firm

[3]The preferences we use, which Matsuyama and Ushchev (2017) call homothetic with a single aggregator (HSA), nest CES, separable translog, and linear expenditure shares as special cases. We also derive our results using generalized Kimball (1995) preferences. The results are similar both qualitatively and quantitatively. We discuss this extension in Section 8.

must have to survive can change. This mechanism is important in models with overhead costs and is emphasized by Asplund and Nocke (2006), Melitz and Ottaviano (2008), Corcos et al. (2012), and Melitz and Redding (2015), among others.[4] As pointed out by Dhingra and Morrow (2019), whether or not the selection effect increases or reduces welfare is ambiguous. A toughening of the selection cutoff improves welfare only if the consumer surplus generated by the marginal variety relative to its sales is less than the average.[5]

Lastly, the pro/anti-competitive effect (3) results from the fact that firms' markups may change as the market expands. Of the three channels, the pro/anti-competitive effect is the sole change in allocative efficiency arising in homogeneous firm models such as Krugman (1979). If firms have incomplete pass-through, as is the case considered by Krugman (1979), then as the price index falls due to an increase in market size, firms cut their markups (pro-competitive effect). Recent studies exploring the pro/anti-competitive effect include Edmond et al. (2015), De Loecker et al. (2016), Feenstra and Weinstein (2017), Feenstra (2018), Arkolakis et al. (2019), and Matsuyama and Ushchev (2020b). We show that whether these changes in markups raise or lower welfare is also ambiguous.

Together, these three channels describe how an increase in market size affects allocative efficiency. To assess the importance of these channels, we develop a strategy for taking the model to data. Using cross-sectional firm-level information from Belgium on pass-throughs (from Amiti et al., 2019), we non-parametrically solve for the shape of the residual demand curve that can exactly rationalize the distributions of firm sales and pass-throughs. We then use our calibrated model to quantify the role reallocations play in aggregate returns to scale.

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

[4] In the absence of overhead costs, an increase in market size may still lead to a change in the selection cutoff if there is a choke price. However, this change in the cutoff will have no first-order effect on welfare because the consumer surplus from marginal varieties is zero at the cutoff.

[5] As we discuss in detail in the body of the paper, the selection and Darwinian effect are different. When a variety exits or enters, due to a change in the selection cutoff, consumers lose or gain all the inframarginal surplus that variety generates. However, when a variety shrinks or expands, due to the Darwinian effect, consumers lose or gain only on the margin. We show that the welfare effect of the former depends on the area under the demand curve whereas the latter depends on the elasticity of the demand curve.

of a market expansion to increases in concentration.[6]

These reallocative forces also have implications for policy. In particular, we show that a marginal entry subsidy may improve welfare even when entry is above the first-best. This is a consequence of the general theory of the second best (Lipsey and Lancaster, 1956)— since all optimality conditions cannot be satisfied, the second-best involves changing the amount of entry away from its first-best value. In our calibration, we find that subsidizing entry above the first-best level is desirable since entry triggers Darwinian reallocations that alleviate cross-sectional misallocation.

Many of the ideas that we develop regarding the response of the economy to changes in market size apply to changes in other parameters and to other demand systems. In the appendix, we provide analytical results for how welfare responds to changes in entry and overhead costs. We also show how the results change, qualitatively and quantitatively, if we use a generalization of Kimball (1995) preferences instead.

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

The first analyses of how market size affect welfare assume that firms are homogeneous, such as Krugman (1979), Mankiw and Whinston (1986), Vives (2001), or Venables (1985). For example, Krugman (1979) shows that, in an economy with homogeneous firms, an increase in market size affects welfare through two channels: the entry of new varieties, and the decrease in markups as the relative share of each variety in total consumption falls. Chaney and Ossa (2013) enrich this result to show that improvements in within-firm productivity (as measured by average cost) can additionally arise from a greater division of labor. This line of research has also been extended by Bilbiie et al. (2012) and Bilbiie et al. (2019) to a dynamic context, and by Matsuyama and Ushchev (2020b) for more general classes of homothetic preferences.

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

Dhingra and Morrow (2019) compare the gains from an increase in market size in an economy with heterogeneous firms compared to an economy with homogeneous firms un-

[6] Baqaee and Farhi (2019) show that this type of reallocation—a reallocation from low-markup firms to high-markup firms—can explain a significant fraction of aggregate TFP growth in the US over the last two decades. De Loecker et al. (2020), Kehrig and Vincent (2021), and Autor et al. (2020) document a similar reallocation of market share to high-markup and high-revenue-productivity firms over time. Our paper raises the possibility that increases in scale, perhaps driven by globalization, could be responsible for these reallocations.

der (non-homothetic) directly additive preferences. They show that certain restrictions on demand are sufficient for gains in a heterogeneous firm economy to be greater.[7] We instead decompose the change in welfare into different margins of adjustment (entry, exit, and changes in markups). This allows us to isolate the Darwinian effect, which can be signed without restrictions on the shape of demand curves. In addition, we use a homothetic demand system and allow for multiple sources of exogenous heterogeneity besides physical productivity.

Mrázová and Neary (2019) show that when markups are increasing in quantity, an increase in scale increases the profits of large firms — an effect they call the "Matthew Effect." While their focus on firm profits is different from our focus on consumer welfare, we show that the Darwinian effect leads to a reallocation of employment and market share to high-markup firms. In our quantitative application, markups and firm size are positively related and increases in market size raise market concentration consistent with Mrázová and Neary (2019).[8]

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

Finally, compared to previous work, we provide a new strategy for calibrating our non-parametric model. Using this strategy, we quantify the importance of the Darwinian, selection, and pro-competitive channels. Our approach offers significant advantages compared to calibrating an off-the-shelf functional form, since common parametric specifications are unable to match important features of the data and this matters for counterfactuals.[9] Our non-parametric

[7] The condition is that the markup is monotonically increasing and the elasticity of utility is monotonically decreasing in quantity. Alternatively, gains in a heterogeneous-firm economy are also greater than in a homogeneous-firm economy if instead the markup is decreasing and elasticity of utility is increasing with quantity, if the product of price elasticities and pass-throughs is also increasing in quantity.

[8] We provide more discussion in Footnote 30 after we present our formal results.

[9] For example, two common alternatives to CES are symmetric translog (Feenstra and Weinstein 2017) and Klenow and Willis (2016). Symmetric translog preferences impose that pass-throughs start at 0.5 for the smallest firms and increase with firm size, which is at odds with the data (see, e.g., Figure 2a). Klenow and Willis (2016) preferences cannot simultaneously match the distributions of pass-throughs and markups (see Appendix N). The failure of these popular functional forms to match the data on sales, markups, and pass-throughs implies that comparative statics with respect to market size calculated under these functional forms are not correct.

demand system, which can simultaneously match a realistic sales, markup, and pass-through distribution can be used for other quantitative applications, and we provide standalone code for evaluating this demand system on our websites.

Structure of the paper. The structure of the rest of the paper is as follows. Section 2 sets up the model and defines the equilibrium. Section 3 decomposes changes in welfare into changes in technical and allocative efficiency and introduces sufficient statistics that we use to state our results. Section 4 shows how welfare responds to an increase in market size and isolates the role of certain reallocations using special cases. Section 5 draws out the implications of these reallocations for how welfare responds to a tax or subsidy on entry. Section 6 introduces a calibration strategy allowing us to take the model to the data non-parametrically. Section 7 is a quantitative application. Section 8 summarizes extensions, and Section 9 concludes. The appendix contains all the proofs.

2 Model Setup

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

Households. There is a population of L identical consumers. Each consumer supplies one unit of labor and has homothetic preferences over varieties of final goods indexed by a type θ. The expenditure share of each variety of type θ is

$$\frac{p_\theta y_\theta}{I} = s_\theta(\frac{p_\theta}{P}), \tag{1}$$

where y_θ is the per-capita consumption of the variety, p_θ is its price, I is per-capita income, P is a *price aggregator*, and $s_\theta(\cdot)$ is a decreasing function. The price aggregator P is defined implicitly by the requirement that expenditure shares sum to one. That is,

$$\int_\Theta s_\theta(\frac{p_\theta}{P})dF(\theta) = 1, \tag{2}$$

where the set Θ contains all potential types, and $dF(\theta)$ is a measure of varieties of type θ.[10] We return to the definitions of Θ and $dF(\theta)$ with more precision when we discuss the firm side of the economy below.

Consumers maximize money-metric per-person utility Y subject to their budget constraint. Define P^Y to be the ideal price index and let per-capita income be the numeraire so that

[10] We assume that $s_\theta(x)$ is strictly decreasing when $s_\theta(x) > 0$. We also assume that $\lim_{x\to 0} s_\theta(x) = \infty$ and $\lim_{x\to\infty} s_\theta(x) = 0$. These conditions guarantee that demand curves for each variety are downward sloping and that the demand system described can be rationalized by a monotone, convex, continuous, and homothetic rational preference relation (see Matsuyama and Ushchev 2017).

$P^Y Y = I = 1$.[11] CES preferences are a special case of equation (1) when $s_\theta(x) = s(x) = x^{1-\sigma}$. These preferences also nest separable translog and linear expenditure shares as special cases.[12] The appeal of these preferences is that, by choosing s_θ, we can match residual expenditure functions of any desired (downward-sloping) shape. Furthermore, since s_θ can vary by θ, different varieties can face different residual demand curves.

Equation (1) also makes clear that the demand for a variety is determined by the ratio of its price, p_θ, to the price aggregator, P. Hence, the price aggregator P mediates competition between each variety and all other available goods. Outside of the CES special case, the price aggregator P is distinct from the ideal price index P^Y.[13] Whereas P is the price aggregator that disciplines expenditure switching, P^Y is the price aggregator that matters for welfare.

Firms. Each firm supplies a single variety and seeks to maximize profits under monopolistic competition similar to the production structure in Melitz (2003).[14] To enter, firms incur a fixed entry cost of f_e units of labor. Upon entry, firms draw their type $\theta \in [0, 1]$ from a distribution with density $g(\theta)$ and cumulative distribution function $G(\theta)$. Having drawn its type, each firm then decides whether to produce or to exit. Production requires paying an overhead cost of $f_{o,\theta}$ units of labor and a constant marginal cost of $1/A_\theta$ units of labor per unit of the good produced. Finally, the firm decides what price to set, taking as given its residual demand curve. We allow the firm's residual demand curve (controlled by s_θ), overhead cost $f_{o,\theta}$, and productivity A_θ to vary with the firm's type θ.

From (1), the price-elasticity of demand facing a variety of type θ, denoted σ_θ, is given by

$$\sigma_\theta(\frac{p}{P}) = -\frac{\partial \log y_\theta}{\partial \log p_\theta} = 1 - \frac{\frac{p}{P} s'_\theta(\frac{p}{P})}{s_\theta(\frac{p}{P})}. \tag{3}$$

Conditional on operating, a firm of type θ will set its price equal to a markup μ_θ times its marginal cost $1/A_\theta$. The profit-maximizing markup is given by the usual Lerner formula,[15]

$$\mu_\theta(\frac{p}{P}) = \frac{1}{1 - \frac{1}{\sigma_\theta(\frac{p}{P})}}. \tag{4}$$

To ensure that each firm's profit-maximizing price is unique, we assume restrictions on s_θ such

[11] Matsuyama and Ushchev (2017) show that under (1) and (2), the ideal price index P^Y is related to the price aggregator P by $\log P^Y = \log P - \int_\Theta \left[\int_{p_\theta/P}^\infty (s_\theta(\xi)/\xi) d\xi \right] dF(\theta)$.

[12] Kimball (1995) preferences are an alternative way to generalize CES preferences while maintaining homotheticity. We discuss how our results change if we use these preferences instead in Section 8.

[13] See Matsuyama and Ushchev (2017) for a proof.

[14] For an extension with oligopolistic competition, see Appendix K.

[15] In our model, firms set markups to maximize static profits. A rich literature describes why consumption habits, financial frictions, customer acquisition costs, or other factors may lead firms to set markups that differ from their static profit-maximizing markups (see e.g., Ravn et al. 2006, Gilchrist et al. 2017, Johnson and Myatt 2006). Since our objective is to compare long-run steady states, we abstract from these considerations.

that marginal revenue curves are strictly downward sloping.[16] When preferences are CES, firms have constant and symmetric price-elasticities of demand $\sigma_\theta = \sigma$, and hence markups $\mu_\theta = \sigma/(\sigma - 1)$ are constant in the cross-section and time-series. The generalized preferences we consider instead allow firms' markups to vary with type θ and relative prices p_θ/P.

Since y_θ is the per-capita output of the firm, the firm's total output is Ly_θ. A firm of type θ chooses to produce if, and only if, its total variable profits exceed its overhead cost of production, i.e.,

$$Lp_\theta y_\theta \left(1 - \frac{1}{\mu_\theta}\right) \geq f_{o,\theta}. \tag{5}$$

Denote the ratio of variable profits to overhead costs by

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

and assume that firm types are ordered so that profitability X_θ is strictly increasing and continuously differentiable in $\theta \in [0,1]$.[17] Define θ^* to be the infimum of the set $\{\theta \in [0,1] : X_\theta \geq 1\}$. Firms with types $\theta \geq \theta^*$ decide to produce, since variable profits for these firms exceed overhead costs, and firms of type $\theta < \theta^*$ do not produce and exit.

Following Melitz (2003), we assume no discounting and suppose that each firm faces an exogenous probability Δ of being forced to exit each period. Free entry implies that firms enter until expected lifetime variable profits minus overhead costs are equal to the entry cost:

$$\frac{1}{\Delta} \int_{\theta^*}^{1} \left[Lp_\theta y_\theta \left(1 - \frac{1}{\mu_\theta}\right) - f_{o,\theta} \right] g(\theta) d\theta \geq f_e. \tag{6}$$

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

Equilibrium. Consumers maximize utility taking prices as given, firms maximize profits taking other prices as given, and markets clear. The equilibrium is determined by equations (1), (2), (4), (5), and (6).

[16] In terms of primitives, we assume that $xs''_\theta(x) < \left[\frac{xs'_\theta(x)}{s_\theta(x)} - 1\right] s'_\theta(x)$ for all x and all θ.

[17] We require firm types to be one-dimensional so that there is a one-to-one mapping from type θ to profitability X_θ and thus a single cutoff type θ^*. In terms of primitives, firms are ordered such that $\frac{-\sigma_\theta}{\rho_\theta} \frac{\partial \log \mu_\theta}{\partial \theta} + \left(\frac{\sigma_\theta}{\rho_\theta} - 1\right) \frac{\partial \log A_\theta}{\partial \theta} - \frac{\partial \log f_{o,\theta}}{\partial \theta} > 0$, where ρ_θ is the pass-through function defined in terms of primitives by (7). In the absence of overhead costs, we do not need to order types by profitability, and hence firm types could instead be multi-dimensional.

Notation. Denote the sales share density by

$$\lambda_\theta = (1 - G(\theta^*))Mp_\theta y_\theta.$$

This is a density because it is always non-negative and integrates to one.[18] For two variables $x_\theta \geq 0$ and z_θ, denote the x-weighted average of z_θ by

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

Denote the x-weighted covariance of any two variables w_θ and z_θ by

$$Cov_x[w_\theta, z_\theta] = \mathbb{E}_x[w_\theta z_\theta] - \mathbb{E}_x[w_\theta]\mathbb{E}_x[z_\theta].$$

Finally, denote the aggregate markup—the ratio of total sales to total variable costs—by $\bar{\mu}$. The aggregate markup is equal to the sales-weighted harmonic average of firm markups,

$$\bar{\mu} = \mathbb{E}_\lambda \left[\mu_\theta^{-1}\right]^{-1}.$$

3 Central Concepts

In this section, we introduce some central concepts that will guide our analysis. First, we introduce statistics related to the shape of the demand curve that help characterize welfare changes. Second, we discuss how welfare is determined in terms of some intuitive, but endogenous, variables. Third, we describe the distortions in the decentralized equilibrium and show how reallocations affect welfare. We build on the definitions in this section to prove our main results in Sections 4 and 5.

3.1 Pass-Throughs and Consumer Surplus Ratios

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

$$\rho_\theta(\frac{p}{P}) = \frac{\partial \log p_\theta}{\partial \log mc_\theta} = 1 + \frac{\partial \log \mu_\theta}{\partial \log mc_\theta} = \frac{1}{1 - \frac{\frac{p}{P}\mu_\theta'(\frac{p}{P})}{\mu_\theta(\frac{p}{P})}}, \tag{7}$$

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

where the markup function is given by (4). Under CES preferences, firms' markups are constant, and hence firms exhibit "complete pass-through" ($\rho_\theta = 1$). In general, however, a firm's desired markup may vary with its position on the demand curve. For example, if a firm's desired markup is decreasing in its price, the firm exhibits "incomplete pass-through" ($\mu'_\theta(\frac{p}{P}) < 0$ and thus $\rho_\theta < 1$). This is sometimes referred to as *Marshall's second law* of demand.

Denote the ratio of the area under the demand curve to sales for each variety by δ_θ. That is,

$$\delta_\theta = \frac{\int_0^{y_\theta} p_\theta(y)dy}{p_\theta y_\theta} = 1 + \frac{\int_{p_\theta/P}^{\infty} \frac{s_\theta(\xi)}{\xi} d\xi}{s_\theta(\frac{p}{P})}, \tag{8}$$

where $p_\theta(y)$ is the inverse residual demand curve for variety θ. Figure 1 illustrates that $\delta_\theta = (A + B)/A$, where B is consumer surplus and A is revenues for variety θ. We call δ_θ the *consumer surplus ratio*. Naturally, the consumer surplus ratio $\delta_\theta \geq 1$ for all θ. In a CES model, δ_θ measures the "love-of-variety" effect and is equal to $\sigma/(\sigma - 1)$.[19] In general, δ_θ is a function of both the variety's type θ and its location on its demand curve (determined by p_θ/P).

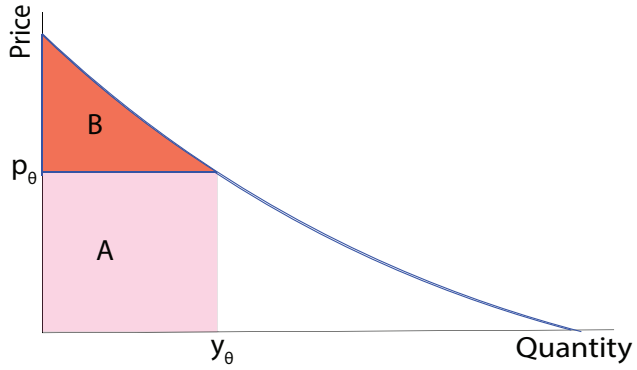


Figure 1: Graphical illustration of δ_θ as the area under the residual demand curve divided by revenues. That is $\delta_\theta = (A + B)/A \geq 1$.

3.2 Welfare

We are interested in how per-capita welfare responds to changes in market size. To a first-order, this is

$$d\log Y = \underbrace{(\mathbb{E}_\lambda[\delta_\theta] - 1)\, d\log M}_{\substack{\text{Consumer surplus} \\ \text{from entry of new varieties}}} - \underbrace{(\delta_{\theta^*} - 1)\, \lambda_{\theta^*} \frac{g(\theta^*)}{1 - G(\theta^*)} d\theta^*}_{\substack{\text{Consumer surplus loss} \\ \text{from exit of varieties } d\theta^*}} - \underbrace{\mathbb{E}_\lambda\left[d\log p_\theta\right]}_{\substack{\text{Marginal surplus} \\ \text{from price changes}}}. \tag{9}$$

[19] As noted by Spence (1976) and Mankiw and Whinston (1986), firms may not appropriate the entire surplus they generate for consumers. In our model, δ_θ also measures the degree of "non-appropriability": as δ_θ increases, the firm captures a smaller portion of the surplus it generates for consumers in revenues. This concept is important because firms' willingness to pay the entry cost depends on the fraction of surplus they can appropriate. The "degree of preference for variety" defined by Vives (2001) in his model of directly additive preferences is proportional to $1 - 1/\delta_\theta$. The elasticity of utility defined by Dhingra and Morrow (2019) is proportional to $1/\delta_\theta$.

Intuitively, welfare changes $d \log Y$ incorporate the consumer surplus brought about by the entry of new varieties $d \log M$ or destroyed by the exit of varieties $d\theta^*$ via the first two terms on the right-hand side of (9). The final term is Shephard's lemma and captures how changes in prices of continuing varieties affect the consumer. If the model did not allow creation and destruction of varieties, then the first two terms of (9) would be zero and changes in welfare would simply be the sales-weighted average change in prices.

One can also interpret Y as a measure of productivity (aggregate output per worker). This welfare-relevant notion of productivity, which we study and decompose, is different to another notion of "productivity" studied, for example, by Baily et al. (1992), Olley and Pakes (1996), Foster et al. (2001) and Melitz and Polanec (2015). In that literature, changes in aggregate productivity are proxied using changes in an index defined as a weighted average of firm productivity levels, e.g., $\bar{A} = \mathbb{E}_\lambda[A_\theta]$. Changes in this index are given by

$$d \log \bar{A} = \lambda_{\theta^*} \left(1 - \frac{A_{\theta^*}}{\bar{A}}\right) \frac{g(\theta^*)}{1 - G(\theta^*)} d\theta^* + Cov_\lambda \left[\frac{A_\theta}{\bar{A}}, d \log \lambda_\theta\right] + \mathbb{E}_\lambda[d \log A_\theta]. \tag{10}$$

Comparing (10) to (9) reveals important differences. Increases in $\bar{A}$ cannot be interpreted as improvements in efficiency. For example, even starting from an optimal point, a reallocation that moves sales from low-A_θ to high-A_θ firms raises $\bar{A}$, contradicting the optimality of the initial point. Furthermore, as pointed out by Petrin and Levinsohn (2012) and Baqaee and Farhi (2019), these statistical decompositions can detect "improvements" in $\bar{A}$ even in cases where reallocations actually reduce welfare and aggregate output.

3.3 Sources of Inefficiency

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

In what follows, we define *local* efficiency for each margin. That is, whether a marginal reallocation along some dimension improves or decreases welfare. This is distinct from global efficiency which compares the allocation to the first-best allocation. These local notions of efficiency are the ones that are relevant for understanding how reallocations affect welfare on the margin in the decentralized equilibrium.

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

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

$$\bar{\mu} < \mathbb{E}_\lambda[\delta_\theta]. \tag{11}$$

If this inequality is reversed, entry is excessive.

In words, there is too little entry if the aggregate markup is less than the sales-weighted average consumer surplus ratio. Intuitively, raising entry by one percent raises welfare according to $\mathbb{E}_\lambda[\delta_\theta]$, but reduces variable production per variety (and hence welfare) by $\bar{\mu}$ percent.[20] In a CES model, (11) holds as an equality and so the CES model has efficient entry.

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

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

$$\delta_{\theta^*} < \mathbb{E}_\lambda[\delta_\theta]. \tag{12}$$

If this inequality is reversed, selection is too tough.

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

If the inequality in (12) is reversed, then an increase in the selection cutoff $d\theta^* > 0$ reduces efficiency and welfare. Therefore, tougher selection and the exit of marginally profitable firms is not, ipso facto, evidence that efficiency is rising. In a CES model, (12) holds as an equality and so the CES model has efficient exit.

[20] Equivalently, Lemma 1 can be understood through the lens of the non-appropriability and business stealing externalities discussed by Mankiw and Whinston (1986). A marginal entrant generates consumer surplus over and above the revenues it captures, on average by $\mathbb{E}_\lambda[\delta_\theta] - 1$, but causes all existing firms to contract output, resulting in an aggregate loss of profits equal to $\bar{\mu} - 1$. If $\bar{\mu} - 1 < \mathbb{E}_\lambda[\delta_\theta] - 1$, the additional consumer surplus generated by the marginal entrant dominates the business stealing externality, and entry is insufficient.

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

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

$$\mu_{\theta'} < \mu_{\theta}. \tag{13}$$

Intuitively, firms with higher markups are inefficiently small in the cross-section compared to firms with lower markups. Hence, reallocating labor from a low-markup firm to a high-markup firm increases allocative efficiency.[21] Crucially, it is a comparison of markups μ_θ, and not productivities A_θ, that determines whether or not one firm should be larger than another from a social perspective. If markups happen to be positively associated with productivity, then an expansion of more productive firms increases welfare, but this is only because "high productivity" proxies for "high markup."[22] In a CES model, (13) holds as an equality and so the CES model has an efficient cross-sectional allocation of resources.

Note that correcting relative size inefficiencies is distinct from choosing whether marginally profitable firms should operate. For example, suppose the marginally profitable firm is a mom-and-pop store with markup μ_{θ^*} and consumer surplus ratio δ_{θ^*}. If μ_{θ^*} is less than average ($\mu_{\theta^*} < \bar{\mu}$), then a planner can raise welfare by moving variable production labor from θ^* to the rest of the economy. However, this does not mean that shutting down the mom-and-pop store is beneficial. In fact, if $\delta_{\theta^*} > \mathbb{E}_\lambda[\delta_\theta]$, then shutting down the mom-and-pop store results in a greater loss in welfare than the gain from using those resources for new entry.

That is, Lemma 3 shows that the welfare effect of marginally expanding and shrinking firms involves a comparison of their markups, whereas Lemma 1 shows that the welfare effect of shutting down and starting firms depends on a comparison of their consumer surplus ratios.

[21] In reality, there may be other distortions that make it sub-optimal to reallocate resources to high-markup firms. For example, suppose firms that charge high markups also receive subsidies on inputs (e.g., by lobbying public officials). If these subsidies are large enough, then on net these high-markup firms are too large relative to other firms, and reallocating more resources to them is harmful for welfare. In our model, this is not the case because all firms buy inputs at the same price and sell directly to households, and markups are the only distortionary wedges in the economy that vary across firms. Even in more complex models with input-output linkages, Baqaee and Farhi (2019) show that reallocating resources to more distorted parts of the economy, taking distortions along the entire supply chain into account, improves efficiency.

[22] In general, the level of A_θ is irrelevant for whether a reallocation improves or worsens efficiency. This contrasts with statistical decompositions, like the one in (10), which consider a reallocation towards firms with higher *levels* of productivity A_θ as improving efficiency. See Section 3.2.

4 Changes in Market Size

In this section, we characterize how an increase in market size, L, affects welfare. We also consider how statistics like the aggregate markup and real GDP respond to an increase in market size.[23] As in Krugman (1979), one can think of an increase in L as capturing the effect of trade integration of symmetric economies. Suppose we have N countries with identical tastes and technologies, with populations $L_1, L_2, ..., L_N$. The market equilibrium if these N countries trade freely is the same as the market equilibrium in a single, closed economy with size $L_1 + L_2 + ... + L_N$; hence, comparative statics of the equilibrium with respect to L can be interpreted as the effect of opening to trade with symmetric foreign markets.

4.1 Decomposition into Technical and Allocative Efficiency

As noted by Helpman and Krugman (1985), reallocations associated with increased competition can mitigate or exacerbate pre-existing distortions. To understand these reallocations, we decompose welfare changes into changes due to *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 shock resulting from endogenous reallocations triggered by the shock.[24]

Following Baqaee and Farhi (2019), let the *allocation vector* $\mathcal{X}$ capture the share of labor allocated to entry, overhead, and variable production of each variety. For any L, every feasible allocation is described by some $\mathcal{X}$. Let $\mathcal{Y}(L, \mathcal{X})$ be the associated level of consumer welfare. Our analysis decomposes changes in welfare into changes in technical and allocative efficiency as

$$d\log Y = \underbrace{\frac{\partial \log \mathcal{Y}}{\partial \log L} d\log L}_{\substack{\text{technical efficiency} \\ \text{(i.e., holding } \mathcal{X} \text{ fixed)}}} + \underbrace{\frac{\partial \log \mathcal{Y}}{\partial \mathcal{X}} \frac{d\mathcal{X}}{d\log L} d\log L}_{\substack{\text{allocative efficiency} \\ \text{(i.e., due to reallocations)}}}. \tag{14}$$

At the efficient allocation, the envelope theorem implies that changes in allocative efficiency are zero to a first-order. Inefficiencies in the initial allocation open the door for reallocations to have first-order effects on welfare. Hence, in the general case, our model will feature changes in both technical and allocative efficiency following an increase in market size.[25]

[23] Although we focus on changes in market size in the body of the paper, in Appendix H we show that similar results can be derived for changes in overhead and entry costs.

[24] Our notion of allocative efficiency compares changes in welfare due to reallocations against a benchmark where the allocation of resources is held constant. A different notion of allocative efficiency measures changes in the distance to the efficient frontier. Changes in that measure of allocative efficiency depend on an extra term, which is how fast the efficient frontier moves when market size changes. See Appendix F.2 for a derivation.

[25] Appendix F explicitly characterizes $\partial \log \mathcal{Y}/\partial \mathcal{X}$ and $d\mathcal{X}/d\log L$ separately. Our Theorem 1, below, follows from combining these formulas as in Equation (14).

4.2 Welfare and Changes in Market Size

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

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

$$d \log Y = \underbrace{\left(\mathbb{E}_\lambda[\delta_\theta] - 1\right) d \log L}_{\textit{technical efficiency}} + \underbrace{\left(\xi^\epsilon + \xi^{\theta^*} + \xi^\mu\right) \bar{\mu}\, d \log L}_{\textit{allocative efficiency}}, \tag{15}$$

where

$$\begin{aligned}
(\textit{Darwinian Effect}) \quad & \xi^\epsilon = (\mathbb{E}_\lambda[\delta_\theta] - 1)\, Cov_\lambda\left[\sigma_\theta, \frac{1}{\mu_\theta}\right] \geq 0, \\
(\textit{Selection Effect}) \quad & \xi^{\theta^*} = (\mathbb{E}_\lambda[\delta_\theta] - \delta_{\theta^*})\, \lambda_{\theta^*} \gamma_{\theta^*} \left(\mathbb{E}_\lambda\left[\frac{\sigma_{\theta^*}}{\sigma_\theta}\right] - 1\right) \gtreqless 0, \\
(\textit{Pro/Anti-competitive Effect}) \quad & \xi^\mu = \mathbb{E}_\lambda\left[(1 - \rho_\theta)\, \sigma_\theta \left(1 - \frac{\mathbb{E}_\lambda[\delta_\theta]}{\mu_\theta}\right)\right] \mathbb{E}_\lambda\left[\frac{1}{\sigma_\theta}\right] \gtreqless 0,
\end{aligned}$$

and $\gamma_{\theta^*} > 0$ *is the hazard rate of the profitability distribution* X_θ *at the selection cutoff. That is,* $\gamma_{\theta^*} = g(\theta^*)/(1 - G(\theta^*))(\partial \log \theta / \partial \log X)$.[26]

Equation (15) decomposes the change in welfare into a technical and allocative efficiency effect according to the definition in (14). We start by discussing the technical efficiency term before discussing the allocative efficiency term.

The first term in Equation (15) captures changes in technical efficiency: the welfare gains from an increase in market size holding the proportional allocation of resources across uses (entry, overhead, and variable production) fixed. Because the fraction of labor allocated to entry is held fixed, the increase in population implies a proportional increase in entry. This has two offsetting effects. On the one hand, the new varieties increase consumer welfare by $\mathbb{E}_\lambda[\delta_\theta] d \log L$, since the consumer's surplus associated with the new varieties will average to $\mathbb{E}_\lambda[\delta_\theta]$. On the other hand, the increase in the number of varieties reduces the per-capita consumption of existing varieties by $d \log L$. The net effect balances these two offsetting effects. Since $\delta_\theta \geq 1$, the technical efficiency term is always positive. In a CES model, this is the only effect.

The second term in (15) captures changes in allocative efficiency: the welfare gains due to changes in the allocation of resources. Each of ξ^ϵ, ξ^{θ^*}, and ξ^μ relates to a particular type

[26] In terms of primitives, this is

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

of reallocation. In fact, the general equilibrium response can be analyzed as a series of three successive allocations, each of which allows firms to adjust along a greater number of margins.[27] In the first restricted allocation, we allow free entry, but hold markups and the selection cutoff constant (i.e., μ_θ and θ^* are fixed using implicit taxes). The change in welfare in this allocation is the same as in Theorem 1, but setting $\xi^{\theta^*} = \xi^\mu = 0$. In the second allocation, firms can also change their decision to operate but still cannot alter their markups. The change in welfare in this allocation is equal to Theorem 1, but setting $\xi^\mu = 0$. Finally, the third allocation allows firms to adjust on all three margins: entry, exit, and choice of markup.

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

Darwinian Effect. To isolate the role of the Darwinian effect, consider an economy in which there are no overhead costs ($f_{o,\theta} = 0$) so that $\theta^* = 0$. Furthermore, assume that preferences are given by[28]

$$s_\theta(\frac{p_\theta}{P}) = \left(\frac{p_\theta}{P}\right)^{1-\sigma_\theta}. \tag{16}$$

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

Corollary 1 (Darwinian Effect)**.** *When preferences are given by* (16) *and overhead costs are zero, the change in welfare from an increase in market size is given by*

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

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

$$\xi^\epsilon = (\mathbb{E}_\lambda[\delta_\theta] - 1) Cov_\lambda\left[\sigma_\theta, \frac{1}{\mu_\theta}\right] = -(\mathbb{E}_\lambda[\delta_\theta] - 1) Cov_\lambda\left[\sigma_\theta, \frac{1}{\sigma_\theta}\right] \geq 0. \tag{17}$$

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

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

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

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

$$d \log y_\theta = -\sigma_\theta d \log p_\theta + (\sigma_\theta - 1) d \log P = (\sigma_\theta - 1) d \log P.$$

The second equality follows from the fact that in this example $d \log p_\theta = d \log \mu_\theta = 0$. Consider how an increase in market size affects demand for this firm. As explained in the introduction, an increase in market size and the entry of new firms causes the price aggregator to fall $d \log P < 0$. The reduction in the price aggregator triggers bigger reductions in per-capita quantities for firms that face more elastic demand. The result is that low-markup firms (who have high price-elasticities of demand) shrink more than high-markup firms (who have low price-elasticities). By Lemma 3, high-markup firms were initially too small relative to low-markup firms, so this reallocation reduces relative productive inefficiencies and improves welfare. We call this a *Darwinian* effect because a more competitive environment, from a reduction in the price index, shifts resources towards the "fittest" firms (those with higher markups and more inelastic demand).[29] The $(\mathbb{E}_\lambda[\delta_\theta] - 1)$ in (17) appears because the reallocations caused by the Darwinian effect save on labor, and these extra resources are funneled into additional entry.[30]

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

Corollary 2 (Darwinian and Selection Effect)**.** *When preferences are given by* (16) *and overhead costs are nonzero, the change in welfare from an increase in market size is given by*

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

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

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

[29] Appendix M discusses conditions under which the Darwinian effect persists when we depart from the specific assumptions of our model.

[30] Mrázová and Neary (2019) show that when Marshall's second law holds (markups are increasing in size or, equivalently, demand curves are log-concave), an increase in scale increases the profits of large firms (which they term the "Matthew Effect"). The Darwinian effect we isolate concerns the reallocation of employment, not profits, which is the welfare-relevant reallocation. Furthermore, we show that this reallocation is welfare-increasing regardless of whether Marshall's second law holds. For example, the demand curves generated by (16) are log-linear. In fact, if the demand curve is log-convex, even though it still increases welfare, the Darwinian effect becomes an "anti"-Matthew effect because it reallocates labor to small, rather than large, firms.

This happens, for example, if consumer surplus ratio at the cutoff δ_{θ^*} is lower than average $\mathbb{E}_\lambda[\delta_\theta]$, and the price elasticity σ_{θ^*} is higher than average $\mathbb{E}_\lambda[\sigma_\theta]$. The second condition ensures that the selection cutoff increases in response to an increase in market size since marginal firms are more exposed to competition than the average firm, and the first condition ensures that the exit of marginal firms is beneficial since selection was too weak to begin with (following Lemma 2).

As discussed above, an increase in the selection cutoff, $d\theta^* > 0$, is not, on its own, evidence of an improvement in allocative efficiency, unless the marginal firm provides households with less consumer surplus than reallocating that labor to entry and other surviving firms. Indeed, in our quantitative application in Section 7, we find that increases in the selection cutoff are welfare-reducing.

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

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

$$d\log Y = \underbrace{\big(\delta - 1\big)\, d\log L}_{\textit{technical efficiency}} + \underbrace{\xi^\mu \mu\, d\log L}_{\textit{allocative efficiency}},$$

Homogeneity of firms implies that $\xi^\epsilon = \xi^{\theta^*} = 0$ and that ξ^μ simplifies to

$$\xi^\mu = (1-\rho)\left(1 - \frac{\delta}{\mu}\right). \tag{18}$$

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

The literature typically refers to the idea that markups may fall with market size as the *pro-competitive effect* of scale. In this example, the pro-competitive effect is captured entirely by $\rho < 1$: markups decrease since each firm's price rises relative to the aggregate price index. As (18) makes clear, the welfare impact of these pro-competitive effects then depends on the

initial efficiency of entry.[31]

4.3 Response of Other Variables

We finish this section by characterizing how a change in market size affects two other quantities of interest—the aggregate markup and real GDP.

Aggregate Markup and Income Shares. An increase in market size changes the aggregate markup for both within-firm and between-firm reasons. In this model, the share of income earned by production labor is inversely related to the aggregate markup $1/\bar{\mu}$. The remainder of income, $1 - 1/\bar{\mu}$, is variable profits dissipated by the costs of entry (i.e. quasi-rents). Proposition 1 characterizes the change in the aggregate markup, and hence the share of income going to variable profits, following a change in market size.

Proposition 1 (Aggregate Markup Effect of Change in Market Size)**.** *In response to changes in population* $d \log L$*, changes in the aggregate markup are*

$$d \log \bar{\mu} = \left(\zeta^{\epsilon} + \zeta^{\theta^*} + \zeta^{\mu}\right) \bar{\mu}\, d \log L,$$

where

$$\begin{aligned}
(\textit{Darwinian Effect}) \quad & \zeta^{\epsilon} = (\bar{\mu} - 1)\, Cov_{\lambda}\left[\sigma_{\theta}, \frac{1}{\mu_{\theta}}\right] \geq 0, \\
(\textit{Selection Effect}) \quad & \zeta^{\theta^*} = \lambda_{\theta^*} \gamma_{\theta^*} \left(\frac{\bar{\mu}}{\mu_{\theta^*}} - 1\right)\left(\mathbb{E}_{\lambda}\left[\frac{\sigma_{\theta^*}}{\sigma_{\theta}}\right] - 1\right) \geq 0, \\
(\textit{Pro-competitive Effect}) \quad & \zeta^{\mu} = -\mathbb{E}_{\lambda}\left[\frac{\bar{\mu} - 1}{\sigma_{\theta}}\right] \mathbb{E}_{\lambda}\left[(\sigma_{\theta} - 1)\,(1 - \rho_{\theta})\right] \leq 0, \quad \textit{if } \rho_{\theta} \leq 1.
\end{aligned}$$

The change in the aggregate markup is composed of three distinct effects that are familiar from our discussion of changes in allocative efficiency above. First, increased entry causes a reallocation toward high-markup firms (the Darwinian effect), which always increases the aggregate markup. Second, changes in market size affect the exit cutoff (the selection effect). The selection effect also increases the aggregate markup because either the cutoff firm's elasticity is higher than average and markup is lower than average—which means an increase in market size toughens selection and causes the exit of low-markup firms—or the cutoff firm's elasticity is lower than average and markup is higher than average—in which case an increase in market size weakens selection and leads the market to retain more high-markup

[31] This discussion is closely related to the contemporaneous findings from Matsuyama and Ushchev (2020b), who show that if entry is globally pro-competitive, then entry is excessive in models with homogeneous firms. When there is cross-sectional heterogeneity, the effect of the pro-competitive effect is complicated by cross-sectional misallocation. We discuss this in more detail in Footnote 40.

firms. The Darwinian and selection effects are mitigated by the third effect, which captures firms' markup adjustments (the pro-competitive effect). The pro-competitive effect always decreases the aggregate markup when pass-through is incomplete ($\rho_\theta < 1$), since incomplete pass-through leads firms to adjust their markups downward as the aggregate price index falls.

Whether an increase in market size leads to an increase in the aggregate markup on net depends on whether the Darwinian and selection effects outweigh the pro-competitive effect. If firms are homogeneous, then only the pro-competitive effect remains, and an increase in market size will lead to a decrease in the aggregate markup. In our calibrated model, the Darwinian and selection effects dominate and the aggregate markup increases when the market becomes larger.

Real GDP. Statistical agencies calculate real GDP using the change in prices for varieties present before and after a change. This means that product entry and exit are ignored in the computation of real GDP (see e.g., Aghion et al. 2019).

Proposition 2 (Real GDP Effect of Change in Market Size)**.** *In response to changes in population* $d \log L$*, changes in real GDP per capita are*

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

The first equation shows that changes in real GDP are equal to the last term in (9). Hence, changes in real GDP and welfare coincide only if there is no consumer surplus from entry and exit.[32] The second equation shows that when pass-throughs are incomplete ($\rho_\theta < 1$), an increase in market size leads to a reduction in markups and hence an increase in measured real GDP per capita. On the other hand, if pass-throughs are complete (as in Corollaries 1 and 2), real GDP per capita is invariant to market size, even though welfare increases as the market expands. Hence, measured real GDP may provide a poor description of how welfare changes with market size.

5 Policy Interventions

In this section, we consider the implications of our results for policy. Section 3.3 discussed the three margins along which the decentralized allocation can be distorted—entry inefficiency, selection inefficiency, and relative production inefficiencies. The policy that obtains

[32]For example, welfare and real GDP coincide in the absence of fixed and entry costs, where goods enter and exit according to a choke price. If goods enter and exit smoothly from a choke price (as in Arkolakis et al. 2019, for example), then $\delta_\theta = 1$ for all entrants and exiters, so the first two terms in (9) are zero. Our decomposition of efficiency in (9) does not impose these restrictions. This also separates the decomposition in (9) from decompositions that do not capture how entering/exiting varieties affect welfare, such as Petrin and Levinsohn (2012) and Baqaee and Farhi (2019).

the first-best allocation eliminates all three types of distortion. However, achieving the first-best requires at least as many policy instruments as there are firm types, since the first-best eliminates variation in markups across firm types. Moreover, the planner also needs to regulate selection by comparing consumer surplus at the cutoff against the average. Whereas such extensive interventions in the market are impracticable, subsidizing entry is, in comparison, straightforward.[33]

In this section, we consider how a marginal entry tax affects welfare, and show that an entry tax can trigger similar reallocative forces to those in Theorem 1. The tax on entry, τ, modifies the free entry condition given in (6), so that each entering firm now pays $(1+\tau)f_e$ units of labor upon entry:

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

Revenues from the tax are rebated lump-sum to households.

For brevity, we include details of how these changes affect the system of equilibrium conditions in Appendix E and continue now to the welfare result. Proposition 3 characterizes the response of welfare to a tax on entry, starting from the point where entry is untaxed.

Proposition 3 (Welfare Effect of an Entry Tax)**.** *Suppose entry is initially untaxed (unsubsidized). The response of welfare to a marginal tax on entry is given by*

$$d\log Y = \left(1-\frac{\mathbb{E}_\lambda\left[\delta_\theta\right]}{\bar{\mu}} - \left[\xi^{\epsilon}+\xi^{\theta^*}+\xi^{\mu}+\left(\mathbb{E}_\lambda[\delta_\theta]-\delta_{\theta^*}\right)\lambda_{\theta^*}\gamma_{\theta^*}\right]\right)\psi_e\, d\tau, \tag{19}$$

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

Whether an entry tax increases welfare depends on the sign of the term in parentheses in (19). This term is more likely to be positive—and an entry tax is more likely to be welfare-enhancing—if entry is excessive ($\mathbb{E}_\lambda[\delta_\theta] < \bar{\mu}$), if selection is too tough ($\mathbb{E}_\lambda[\delta_\theta] < \delta_{\theta^*}$), or if the beneficial reallocations from entry given by ξ^{ϵ}, ξ^{θ^*}, and ξ^{μ} are small. We call $-\xi^{\epsilon}\psi_e d\tau$ the Darwinian effect of the entry tax, $-(\xi^{\theta^*}+(\mathbb{E}_\lambda[\delta_\theta]-\delta_{\theta^*})\lambda_{\theta^*}\gamma_{\theta^*})\psi_e d\tau$ the selection effect of the entry tax, and $-\xi^{\mu}\psi_e d\tau$ the pro-competitive effect of the entry tax. Together with the welfare effect due to the initial wedge on entry efficiency, $(1-\mathbb{E}_\lambda[\delta_\theta]/\bar{\mu})\psi_e d\tau$, these forces sum to the total effect of an entry tax on welfare.

An immediate implication of Proposition 3 is that excessive entry (as defined in Lemma 1) is not a sufficient condition for an entry tax to be welfare-increasing. For example, if the beneficial reallocations from entry ($\xi^{\epsilon}+\xi^{\theta^*}+\xi^{\mu}$) are sufficiently large, then attempting to

[33]For more discussion of first-best policy, see Appendix G.1, where we characterize the policy that achieves first-best and calculate the distance of the decentralized equilibrium to the efficient frontier.

correct for excessive entry with an entry tax will actually be welfare-reducing because the economy loses the beneficial cross-sectional reallocations associated with entry.

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

Darwinian effect. Consider again the economy in Corollary 1, where there are no overhead costs and preferences are given by (16). In this example, the entry tax has no effect on firms' markups or on selection.

Corollary 4 (Darwinian Effect)**.** *When preferences are given by* (16) *and overhead costs are zero, the change in welfare from a marginal tax on entry is positive if, and only if,*

$$\mathbb{E}_\lambda [\delta_\theta] < (1 - \xi^\epsilon)\, \bar{\mu}.$$

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

Selection effect. Suppose we retain complete pass-through preferences, but now allow for nonzero overhead costs, as in Corollary 2. The economy now features both Darwinian and selection effects, but pro-/anti-competitive effects are still absent.

Corollary 5 (Darwinian and Selection Effect)**.** *When preferences are given by* (16) *and overhead costs are nonzero, the change in welfare from a marginal tax on entry is positive if, and only if,*

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

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

Pro/anti-competitive effect. Finally, consider an economy with homogeneous firms, as in Corollary 3. In this economy, entry has no Darwinian or selection effects, since firms are identical.

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

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

6 Calibration Strategy

In this section, we discuss how to map our model to data. We first show how data on firm pass-throughs, sales, and exit rates can be used to calibrate the model, without imposing a functional form on preferences or on the distribution of firm productivities. We then implement our approach using Belgian data and compare the calibrated model's match to untargeted moments. The demand system we calibrate is potentially useful for other applications, since it can simultaneously match realistic pass-through, markup, and sales distributions. We provide stand-alone code on our websites for evaluating our demand system. In Section 7, we use the calibrated model to consider counterfactuals where we change market size or introduce an entry tax, in line with our theoretical results.[34]

6.1 Non-Parametric Calibration Approach

The model has many degrees of freedom, so in order to take the model to data, we impose the following restrictions on overhead costs $f_{o,\theta}$ and expenditure share functions s_θ.

Assumption 1. *Firms have identical overhead costs $f_{o,\theta} = f_o$, and expenditure share functions s_θ take the form,*

$$s_\theta(\frac{p_\theta}{P}) = s(\frac{1}{B_\theta}\frac{p_\theta}{P}) = s(\frac{1}{A_\theta B_\theta}\frac{\mu_\theta}{P}), \tag{20}$$

where B_θ are type-specific quality shifters.

Allowing for unobserved quality shifters B_θ is important since two firms that charge the same price in the data can have very different sales. If there were no quality shifters, one could identify $s(\cdot)$ by simply plotting price against sales in the cross-section. In practice, this is untenable because the prices firms report are not directly comparable to each other.

Proposition 4 shows that we can identify types from observables under Assumption 1.

Proposition 4 (Identification of Firm Types)**.** *Suppose Assumption 1 holds. Then, sales λ_θ and profitability X_θ are strictly increasing in the product of physical productivity and quality, $A_\theta B_\theta$. Furthermore, any two firms with identical sales also have identical pass-throughs ρ_θ, markups μ_θ, and consumer surplus ratios δ_θ.*

[34] To test the model, one would like to observe the response of an economy to exogenous shocks to market size. In the absence of well-identified shocks to market size, our approach is to calibrate our model to match micro-level moments and use the calibrated model to perform counterfactual exercises.

The intuition for Proposition 4 follows from (20): since "quality-adjusted" prices p_θ/B_θ are strictly decreasing in $A_\theta B_\theta$ and pass-throughs are greater than zero, firm sales must be strictly increasing in $A_\theta B_\theta$.[35] Moreover, since a higher $A_\theta B_\theta$ enlarges the quality-adjusted production possibilities set, with constant overhead costs, profitability X_θ must also be increasing in $A_\theta B_\theta$.

Proposition 4 implies that firms can be ordered by sales, profitability, and $A_\theta B_\theta$ interchangeably. Hence, we can identify firms' types by their rank in the sales distribution. Accordingly, we set each firm's type to be the fraction of firms with less sales, so that the distribution of types $G(\theta)$ is uniform over $[0, 1]$.

Once we identify firms' types, we can proceed to identify the sufficient statistics necessary to calculate the comparative statics in Sections 4-5. Since pass-throughs are related to the second derivative of the residual expenditure function $s(\cdot)$, we can solve a set of differential equations to recover markups and consumer surplus ratios up to boundary conditions. Proposition 5 shows how we calculate these statistics given data on the firms' sales, pass-throughs, exit rates by age, and values for the aggregate markup and average consumer surplus ratio.

Proposition 5 (Calibration of Sufficient Statistics)**.** *Suppose Assumption 1 holds. Given an aggregate markup $\bar{\mu}$ and data on pass-throughs ρ_θ and sales λ_θ, markups are given by the solution to*

$$\frac{d\log\mu_\theta}{d\theta} = (\mu_\theta - 1)\frac{1-\rho_\theta}{\rho_\theta}\frac{d\log\lambda_\theta}{d\theta} \qquad s.t. \qquad \mathbb{E}_\lambda[\mu_\theta^{-1}]^{-1} = \bar{\mu}. \tag{21}$$

Given the above inputs and an average consumer surplus ratio $\bar{\delta}$, consumer surplus ratios are given by the solution to

$$\frac{d\log\delta_\theta}{d\theta} = \left(\frac{\mu_\theta}{\delta_\theta} - 1\right)\frac{d\log\lambda_\theta}{d\theta} \qquad s.t. \qquad \mathbb{E}_\lambda[\delta_\theta] = \bar{\delta}. \tag{22}$$

Given firm exit rates by age, θ^* is the difference between the first-year exit rate and the exit rate of mature firms. The overhead cost is $f_o = \lambda_{\theta^*}(1-1/\mu_{\theta^*})/(1-\theta^*)$, the entry cost is $\Delta f_e = \mathbb{E}_\lambda[1-1/\mu_\theta] - (1-\theta^*)f_o$, and the hazard rate of profitability at the cutoff is given by $\gamma_{\theta^*} = \rho_{\theta^*}/(1-\theta^*)(\partial\log\lambda_\theta/d\theta|_{\theta^*})^{-1}$.

The intuition for these results follows. First, to get (21), we start by writing the relationship between marginal cost changes and changes in firms' markups μ_θ and sales λ_θ:

$$d\log\mu_\theta = (\rho_\theta - 1)\,d\log mc_\theta, \qquad \text{and} \qquad d\log\lambda_\theta = (1-\sigma_\theta)\rho_\theta\,d\log mc_\theta.$$

The first equation uses the fact that $d\log p_\theta = \rho_\theta d\log mc_\theta$, and the second equation uses the fact that $d\log p_\theta y_\theta = (1-\sigma_\theta)d\log p_\theta$.

Under Assumption 1, all firms face the same residual expenditure function (up to quality shifters B_θ). Thus, we can use these same equations to characterize how markups and sales

[35] The condition in Footnote 16 guarantees that $\rho_\theta > 0$ for all θ.

change as we vary productivity/quality in the cross-section of firms:

$$\frac{d\log\mu_\theta}{d\theta} = (1-\rho_\theta)\frac{d\log(A_\theta B_\theta)}{d\theta}, \qquad \text{and} \qquad \frac{d\log\lambda_\theta}{d\theta} = \frac{\rho_\theta}{\mu_\theta - 1}\frac{d\log(A_\theta B_\theta)}{d\theta}.$$

Here, we use the fact that differences in quality across firms are isomorphic to differences in physical productivity in terms of firms' resulting markups and sales (as can be seen from (20)). We do not need to identify physical productivity and quality separately, and we refer to their product $A_\theta B_\theta$ as a variety's productivity for simplicity. The second equation also uses the Lerner condition to substitute $\sigma_\theta - 1 = 1/(\mu_\theta - 1)$.

Combining these two equations yields (21). Intuitively, the distribution of sales and the distribution of pass-throughs govern the distribution of markups in the model. Incomplete pass-through ($\rho_\theta < 1$) in the data means that, as we increase a firm's productivity (and hence decrease its marginal cost), the firm's markup increases. Thus, in the model, firms with higher $A_\theta B_\theta$ have higher markups. The rate at which markups increase in the cross-section is pinned down by the rate at which sales increase in the cross-section, since sales are monotonically increasing in productivity. Once we solve for markups using (21), we can use either of the differential equations that relate markups and sales to productivity to back out $A_\theta B_\theta$ with boundary condition $A_{\theta^*}B_{\theta^*}$, which we can normalize to one.

Next, the differential equation (22) for consumer surplus ratios can be derived by differentiating (8). As a variety's sales increase, the rate at which the total area under the demand curve for that variety ($\delta_\theta\lambda_\theta$) increases is inversely related to the elasticity of the demand curve (in particular, $d(\delta_\theta\lambda_\theta) = \mu_\theta d\lambda_\theta$). For example, when demand curves are locally perfectly elastic, $\mu_\theta = 1$, the area under the demand curve increases one-for-one with sales. Combining this with the product rule ($\lambda_\theta d\delta_\theta = d(\delta_\theta\lambda_\theta) - \delta_\theta d\lambda_\theta$) implies that consumer surplus ratios increase with sales when $\mu_\theta > \delta_\theta$.

Finally, since $G(\theta^*) = \theta^*$ is the share of firms that exit upon realizing their type, we can identify the cutoff type θ^* by taking the difference between exit rates of entrants and mature firms. Given the cutoff type θ^*, calculating the remaining statistics is straightforward: we can normalize the initial mass of entrants $M = 1$ and market size $L = 1$ without loss, and calculate overhead costs from the selection condition (5), entry costs from the free entry condition (6), and the hazard rate of profitability from pass-throughs and the sales distribution.

Before moving forward, we discuss two features of the restrictions assumed in Proposition 5. First, the assumption that all firms lie on the same residual expenditure function (up to quality shifters B_θ) means that pass-throughs in the time series, which capture how a firm changes its markup if its marginal cost changes, are equal to cross-sectional pass-throughs, which capture how firms' markups vary with productivity/quality in the cross-section. This restriction allows us to use data on pass-through of marginal cost to prices to calibrate how firms' markups vary in the cross-section. Second, as shown in Proposition 5, the restriction on

residual expenditure functions in (20) implies a one-to-one mapping between firms' sales and markups. In the data, there is substantial heterogeneity in markups even conditional on size. While (20) precludes this possibility, we relax this restriction by adding variation in markups orthogonal to firm size in Appendix L.

Nevertheless, the preferences we calibrate are less constrained than previous work since we do not use off-the-shelf functional forms for either demand curves or the distribution of firm productivities. This means that we can match data on the distribution of firm sales and pass-throughs by size exactly.[36]

6.2 Calibration Implementation

We implement Proposition 5 using data on firm pass-throughs, the distribution of firm sales, and exit rates by firm age. We refer readers interested in a more detailed description of our data sources to Appendix A.

Data sources. For pass-throughs ρ_θ, we use estimates of pass-throughs by firm size for manufacturing firms in Belgium from Amiti et al. (2019). They use administrative firm-product level data (Prodcom) from 1995–2007, which contains information on prices and sales, collected by Statistics Belgium. Using exchange rate shocks as instruments for changes in marginal cost, and controlling for changes in competitors' prices, they identify partial equilibrium pass-throughs by firm size under assumptions consistent with our model. Their estimates are shown in Figure A.2 in Appendix A.

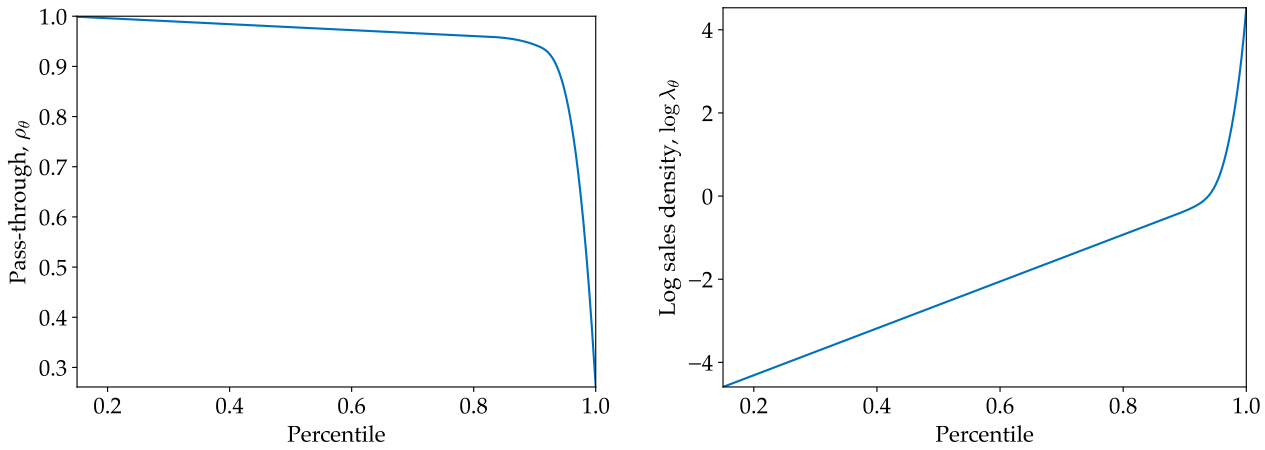
For sales λ_θ, we use the sales distribution for the universe of Belgian manufacturing firms from VAT declarations. The cumulative sales share distribution is shown in Figure A.1 in Appendix A.[37,38]

Finally, we use firm exit rates by age reported by Pugsley et al. (2018). The exit rate for new entrants is about 15 percentage points higher than mature firms, so we set $\theta^* = 0.15$.

[36] In principle, one could alternatively use estimates of markups μ_θ or consumer surplus ratios δ_θ in conjunction with sales λ_θ to calibrate the model. We instead rely on pass-throughs since estimating markups and consumer surplus ratios is more difficult, typically requiring production function estimation for markups and experimental evidence for consumer surplus ratios. The downside is that calibrating the model using pass-throughs ρ_θ requires outside information to pin down boundary conditions $\bar{\mu}$ and $\mathbb{E}_\lambda[\delta_\theta]$.

[37] The Prodcom sample used by Amiti et al. (2019) does not include firms with less than 1 million euros in sales. Since Amiti et al. (2019) find that the average pass-through for the smallest 75% of firms in Prodcom is 0.97, when we merge their pass-through estimates with the firm sales distribution, we assume the smallest firm has a pass-through of one and interpolate pass-throughs for the set of firms with sales under 1 million euros.

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



(a) Pass-through ρ_θ

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

Figure 2: Pass-throughs and sales share density as a function of firm type θ.

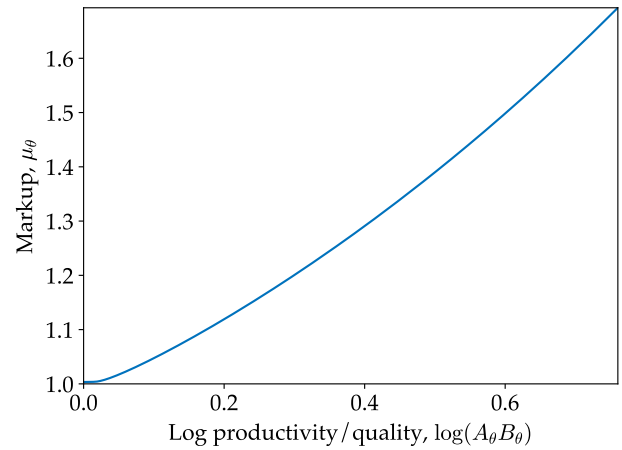
Boundary conditions. Our results require taking a stand on two boundary conditions: the aggregate markup $\bar{\mu}$ and the average consumer surplus ratio $\mathbb{E}_\lambda[\delta_\theta]$. Recent work estimating markups of Prodcom firms by Forlani et al. (2022) finds an average markup of 1.091, so we choose $\bar{\mu} = 1.09$. We focus on two benchmark calibrations of $\mathbb{E}_\lambda[\delta_\theta]$: (1) efficient entry $\mathbb{E}_\lambda[\delta_\theta] = \bar{\mu}$ (see Lemma 1), and (2) efficient selection $\mathbb{E}_\lambda[\delta_\theta] = \delta_{\theta^*}$ (see Lemma 2).

In Appendix B, we show that the level of aggregate increasing returns to scale is sensitive to the choice of $\bar{\mu}$, but the relative contributions of technical and allocative efficiency, and of the Darwinian, selection, and pro-competitive effects, do not vary significantly with $\bar{\mu}$. For completeness, in Appendix B we vary both $\mathbb{E}_\lambda[\delta_\theta]$ and $\bar{\mu}$ along a two-dimensional grid and show that the results we report in the main text are representative of broader patterns.

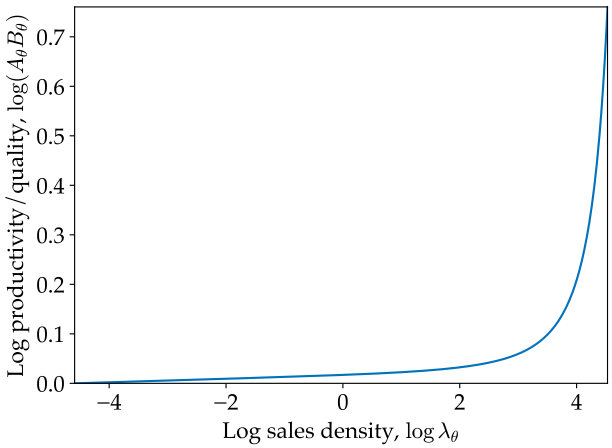
Calibrated statistics. Figures 2a and 2b display pass-throughs, ρ_θ, and log sales, $\log \lambda_\theta$, as a function of type θ. See Appendix A for details about how we construct these figures. Figure 2a shows that pass-throughs decrease from 1 for the smallest firms to about 0.3 for the largest firms. Figure 2b shows that sales are initially increasing exponentially (linear in logs), but become super-exponential towards the end reflecting a high degree of concentration in the tail.

Figure 3a shows the results from solving the differential equation (21). Our calibrated markups are increasing and convex in log productivity. While the average markup level is pinned down by our choice of $\bar{\mu}$, the distribution of markups is not targeted. Nevertheless, the markups we back out are consistent with direct estimates. First, we find that markups range from close to one to about 1.7 for the largest firms. This range of markups is broadly consistent with previous estimates of firm markups by De Loecker et al. (2020), Ridder et al. (2021), and Forlani et al. (2022).[39] If we used Klenow and Willis (2016) preferences instead (and continued

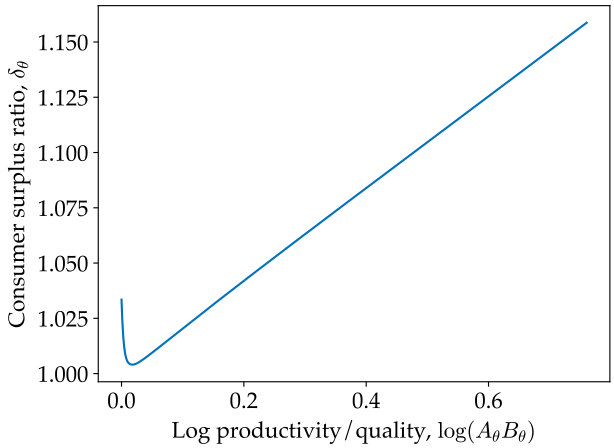
[39]Using the production function approach to estimate markups for French manufacturing firms, Ridder et al.



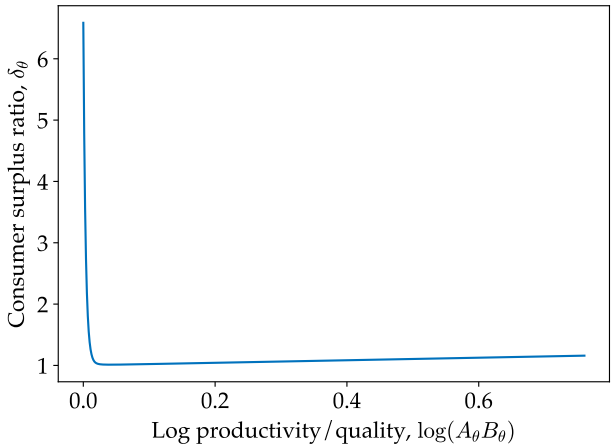
(a) Markup μ_θ



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



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



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

Figure 3: Markups and consumer surplus ratios with $\bar{\mu} = 1.090$.

to match the distribution of pass-throughs from Amiti et al. 2019), we would instead estimate markups on the order of 100 for large firms (as opposed to 1.7 in our calibration; see Appendix N for more detail). Second, our calibrated markups are positively correlated with firm output and sales. This positive covariance between markups and firm size is consistent with evidence from Burstein et al. (2020), Ridder et al. (2021), and De Loecker et al. (2016).

Figure 3b shows the distribution of log productivity/unobserved quality. As with the sales density, the productivity density is also initially exponential, and becomes super exponential in the tail. Since price elasticities are decreasing in θ, productivity has to change by more than sales in the cross-section to allow firms to get large. Figures 3c and 3d show the consumer surplus ratio δ_θ for the efficient selection case ($\delta_{\theta^*} = \mathbb{E}_\lambda[\delta_\theta]$) and the efficient entry case

(2021) find that 10th percentile of firm markups is between 0.91–0.97 and the 90th percentile of firm markups is between 1.36–2.97. Similarly, Forlani et al. (2022) (using Belgian manufacturing firms) and De Loecker et al. (2020) (using public U.S. firms) find that the majority of firm markups are between 1 and 2.

($\bar{\mu} = \mathbb{E}_\lambda[\delta_\theta]$). Figure B.1 in Appendix B plots the residual demand curve and shows that it has a distinctly non-isoelastic shape, indicating substantial departures from CES.

7 Quantitative Results

In this section, we use the calibrated model to calculate how changes in market size and a marginal tax on entry affect welfare. We decompose welfare gains into changes in technical and allocative efficiency—i.e., gains holding the allocation of resources fixed and gains due to the reallocation of resources—and further decompose allocative efficiency changes into the Darwinian, selection, and pro-competitive margins. As extensions, we compare macro and micro returns to scale and illustrate how increases in market size affect industrial concentration.

Welfare effect of a market expansion. Table 1 reports the elasticity of consumer welfare to market size, following Theorem 1. The response of welfare is decomposed into changes due to technical efficiency and allocative efficiency, and the allocative effect is further disaggregated into the Darwinian, selection, and pro-competitive effects.

	Efficient selection $\mathbb{E}_\lambda[\delta_\theta] = \delta_{\theta^*}$	Efficient entry $\mathbb{E}_\lambda[\delta_\theta] = \bar{\mu}$
Welfare: $d \log Y$	0.259	0.278
Technical efficiency	0.033	0.090
Allocative efficiency	0.225	0.188
Darwinian effect	0.235	0.631
Selection effect	0.000	-0.344
Pro-competitive effect	-0.010	-0.099
Real GDP per capita	0.043	0.043
Aggregate markup	0.494	0.494

Table 1: The elasticity of welfare, real GDP per capita, and aggregate markup to population.

We start by discussing the case with efficient selection first ($\mathbb{E}_\lambda[\delta_\theta] = \delta_{\theta^*}$). The elasticity of per-capita consumer welfare to population is 0.259. Only around a tenth of the overall effect is due to the technical efficiency effect (0.033), while changes in allocative efficiency (0.225) account for around nine-tenths of the overall effect. That is, the increase in market size brings about substantial benefits from reallocation, and the gains from these improvements are much larger than direct gains from technical efficiency.

The change in allocative efficiency from the Darwinian effect is large and positive at 0.235. The selection and pro-competitive effects are insignificant in comparison. The change

in allocative efficiency from the selection effect is zero by construction, since the surplus associated with exiting varieties is equal to the average consumer surplus. The change in allocative efficiency from the pro-competitive effect is slightly negative at −0.010.[40]

The elasticity of real GDP per capita is much smaller than the elasticity of welfare to market size at 0.043. Changes in real GDP only reflect the decrease in markups of continuing varieties due to the pro-competitive effect and do not capture changes in consumer surplus due to entry and exit.

The aggregate markup increases with market size. Recall from the discussion of Proposition 1 that whether the aggregate markup increases depends on whether the Darwinian effect and the selection effect, which both increase the aggregate markup, dominate the pro-competitive effect, which reduces firms' markups. In our calibration, the Darwinian effect plays the dominant role in increasing the aggregate markup. Accordingly, the share of income earned by production labor falls as market size grows.

Next, consider the case with efficient entry. The elasticity of welfare with respect to population shocks is now slightly higher at 0.278. The technical efficiency effect is now 0.090, reflecting the fact that $\mathbb{E}_\lambda[\delta_\theta] = \bar{\mu} = 1.09$. The allocative efficiency effect is still much more important than the technical efficiency effect at 0.188.

The Darwinian effect is now much larger at 0.631. The main reason for the increase is because $(\mathbb{E}_\lambda[\delta_\theta] - 1)$ is now 0.090 instead of 0.033. This implies that entry is more valuable. Since the labor saved by the Darwinian effect is funneled into more entry, this makes the Darwinian effect more beneficial. The selection effect is now non-zero and negative at −0.344. The reason for this can be seen from Figure 3d, which shows that the consumer surplus ratio at the cutoff is much higher than average. Hence, as the cutoff increases in response to toughening competition, socially valuable firms are forced to exit. Finally, the pro-competitive effect is still negative and larger in magnitude at −0.099. The pro-competitive effect is now more negative because entry was initially excessive in the efficient selection case, so reductions in markups had a beneficial effect on entry efficiency. Since we are now imposing entry efficiency, this

[40] To understand the pro-competitive effect, we rewrite ξ^μ as

$$\xi^\mu = \underbrace{\left(1 - \frac{\mathbb{E}_\lambda[\delta_\theta]}{\bar{\mu}}\right)\mathbb{E}_\lambda\left[1 - \rho_\theta\right]}_{\text{Effect on entry efficiency}} + \underbrace{\mathbb{E}_\lambda\left[\delta_\theta - 1\right]\left(\mathbb{E}_\lambda\left[\frac{1}{\sigma_\theta}\right]Cov_\lambda\left[\rho_\theta, \sigma_\theta\right] - \mathbb{E}_\lambda\left[1 - \rho_\theta\right]Cov_\lambda\left[\sigma_\theta, \frac{1}{\mu_\theta}\right]\right)}_{\text{Effect on cross-sectional misallocation}}.$$

The first term is similar to the procompetitive effect with homogeneous firms in Corollary 3 and captures the fact that a larger market size leads firms to cut their markups (since $\rho_\theta < 1$), which improves welfare when entry is initially excessive. The second term is due to cross-sectional heterogeneity in markups. The first covariance accounts for the fact that firms with different markups may cut their markups by different amounts, and is positive if high-markup firms have lower pass-throughs (as in our calibration). The second covariance accounts for the fact that, for a given change in prices, firms with high price elasticities and thus low markups expand more than firms with low price elasticities, which exacerbates cross-sectional misallocation. In this empirical calibration, this final covariance dominates the other terms. This is why the overall sign of the pro-competitive effect is negative.

effect no longer operates, and the overall contribution of changing markups to welfare is more negative.

As mentioned when discussing our choice of boundary conditions above, the level of aggregate increasing returns to scale is sensitive to our choice of the aggregate markup $\bar{\mu}$. However, in Appendix B we show that the relative contributions of allocative and technical efficiency to aggregate returns to scale are similar across values of $\bar{\mu}$ from 1.05 to 1.15. Moreover, the Darwinian effect plays the dominant role in driving aggregate increasing returns across the grid of boundary conditions we consider for $\bar{\mu}$ and $\mathbb{E}_\lambda[\delta_\theta]$.

How important can selection be? An important theme in the literature has been to emphasize the role of the selection margin (increases in the productivity/quality cutoff) as a driver of productivity and welfare gains. However, in our baseline results, the selection margin is either neutral (when $\delta_{\theta^*} = \mathbb{E}_\lambda[\delta_\theta]$) or deleterious (when $\mathbb{E}_\lambda[\delta_\theta] = \bar{\mu}$). 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 boundary conditions. We set $\delta_{\theta^*} = 1$, which implies that the residual demand curve for infra-marginal firms is perfectly horizontal. In other words, the marginal firms produce no excess consumer surplus for the household. This maximizes the benefits of the selection margin for welfare.

The results, however, are quantitatively very similar to those in Table 1. Specifically, the welfare effect is 0.259 with an allocative efficiency effect of 0.225. The contribution of the selection effect is positive, but negligible, at 0.002, and the overwhelming force remains the Darwinian effect (0.232). These results suggest that the role played by the selection margin is not an anomaly resulting from our choice of initial conditions.

How important is heterogeneity? To emphasize the interaction of heterogeneity and inefficiency, we compare our model to a model with homogeneous firms. We set firms' pass-through equal to the average (sales-weighted) pass-through in the data, and use the same average markup and consumer surplus ratio as in Table 1. Table 2 shows the results.

	$\delta = \delta_{\theta^*}$	$\delta = \mu$
Welfare: $d \log Y$	0.061	0.090
Technical efficiency	0.033	0.090
Allocative efficiency	0.027	0.000
Real GDP per capita	0.043	0.043
Average markup	-0.043	-0.043

Table 2: The elasticity of welfare, real GDP per capita, and aggregate markup to population for homogeneous firms.

The most striking difference is that both the elasticity of welfare to market size and changes in allocative efficiency are much smaller, due to the absence of the Darwinian effect. In a model with homogeneous firms, the sole source of inefficiency comes from excessive or insufficient entry (see Corollary 3). Thus, when entry is efficient (the second column), there are no changes in allocative efficiency at all. Even when entry is not efficient, changes in allocative efficiency—which are due solely to the pro-competitive effect—are fairly small. Moreover, since only the pro-competitive effect remains, the homogeneous firm model predicts a falling, rather than rising, aggregate markup when the market expands.

Are there larger increasing returns at the macro vs. micro levels? The micro returns to scale is the ratio of average cost to marginal cost minus one, $(ac_\theta/mc_\theta - 1)$, where a value of zero means constant returns to scale. The (harmonic) average of micro returns to scale across surviving producers is thus $1/\mathbb{E}_\lambda[1/(ac_\theta/mc_\theta - 1)] = \bar{\mu} - 1$.

Hence, average micro returns to scale are $\bar{\mu} - 1 = 0.09$. Increasing returns at the aggregate level are much larger: between 0.259 and 0.278. This means that even small technological increasing returns at the micro level can give rise to large increasing returns to scale at the aggregate level. Once again, the interaction of inefficiency and heterogeneity is key. If the economy were efficient, macro and micro returns would be identical, and if the economy had homogeneous firms, the difference between macro and micro returns would be much smaller.

Implications for industrial concentration. Our results suggest that the beneficial reallocations associated with a larger market may come hand-in-hand with increased concentration. Figure 4 shows the Lorenz curve for the distribution of sales as the market size increases.[41] Quantitatively, as the market expands, the concentration of sales rises.[42]

Furthermore, and more importantly from a welfare perspective, when markups covary negatively with pass-throughs (which is the case in our calibration), then an increase in market size always leads high-markup firms to expand in employment terms relative to low-markup firms (see Appendix F equation 29). In fact, an increase in market size causes firms with low price-elasticities and pass-throughs to expand even in per capita terms if $\sigma_\theta \rho_\theta < 1$. This inequality also holds in our calibration for the very largest firms.

Welfare effect on an entry tax. Table 3 shows the effect of an entry tax on welfare using Proposition 3. Note that resources are held fixed, so all changes in welfare arise from changes in allocative efficiency. We decompose the change in welfare into the effect due to the initial

[41] To produce these figures, we compute the equilibrium allocation nonlinearly by solving a system of differential equations. See Appendix B for details.

[42] See recent work by Matsuyama and Ushchev (2022) who show that this is a generic phenomenon when pass-throughs are decreasing in quantity. Figure B.3 in Appendix B also shows that the concentration of employment rises with market size in our quantitative calibration.

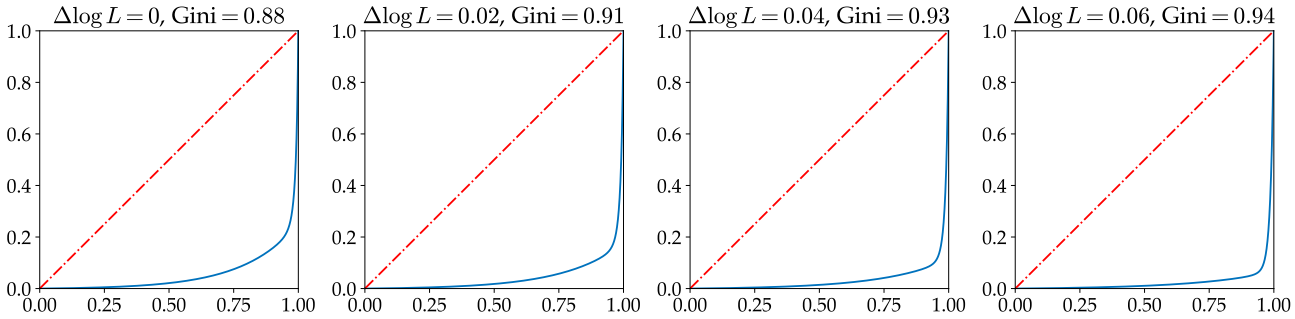


Figure 4: Each panel depicts the Lorenz curve for the sales distribution for different values of the market size parameter L. The dotted red line indicates the line of perfect equality.

wedge on entry efficiency, and the Darwinian, selection, and pro-competitive effects of the entry tax described in Proposition 3. The last row of the table re-computes the welfare effect of an entry tax in a model with homogeneous firms calibrated to have a pass-through equal to the average sales-weighted pass-through.

	Efficient selection $\mathbb{E}_\lambda[\delta_\theta] = \delta_{\theta^*}$	Efficient entry $\mathbb{E}_\lambda[\delta_\theta] = \bar{\mu}$
Welfare: $d \log Y$	-0.155	-0.161
Effect due to initial wedge on entry efficiency	0.052	0.000
Darwinian effect of entry tax	-0.215	-0.579
Selection effect of entry tax	0.000	0.328
Pro-competitive effect of entry tax	0.009	0.091
Welfare with homog. firms	0.027	0.000

Table 3: Welfare effect of an entry tax, following Proposition 3.

For both choices of the boundary conditions, we find that the entry tax is welfare-reducing (and an entry subsidy is welfare-enhancing). Since the tax reduces entry, the Darwinian effect operates in reverse, as loosening competition reallocates resources to low-markup firms and exacerbates misallocation. While the selection and pro-competitive effects are (weakly) beneficial, losses due to Darwinian reallocations outweigh these benefits. In contrast, when firm heterogeneity is excluded from the model, the entry tax is beneficial or has no effect.

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

8 Extensions

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

Other generalizations of CES preferences. In Appendix I, we also derive our results using a different generalization of CES preferences (called HDIA preferences by Matsuyama and Ushchev, 2017). The Kimball (1995) demand system is a special case of these preferences.

Theorem 2 in Appendix I shows that the response of welfare to an increase in market size under HDIA preferences is

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

where $\mathbb{E}_\lambda[\delta_\theta]$, ξ^ϵ, ξ^{θ^*}, and ξ^μ are the same as in the main text. The change in allocative efficiency under HDIA preferences features a multiplier effect. This is because these preferences have an additional feedback loop between reductions in the price index P and increases in welfare Y. Appendix I calibrates the HDIA model and shows that the elasticity of welfare to market size under HDIA preferences is slightly larger than our results in the main text.

Nonlinear response. One might worry that the reallocative effects in our quantitative model could peter out quickly if we kept increasing the size of the market. Table B.1 and Figure B.2 in Appendix B present nonlinear results and show that the forces identified for small shocks by Theorem 1 continue to apply for large shocks.

Optimal policy and distance to the efficient frontier. In the main text, we focus exclusively on comparative statics of the decentralized equilibrium. For completeness, in Appendix G, we characterize the policy that implements the first-best. By numerically implementing the first-best policy, we find that losses due to distortions in the decentralized equilibrium are between 5.9–7.2% depending on boundary conditions. Therefore, changes in allocative efficiency can be large even when the decentralized equilibrium is not far from the frontier.

Proposition 6 in Appendix G also provides an analytical approximation of the distance to the efficient frontier as we move away from the CES benchmark. We show that, to a second-order, the distance to the frontier is given by

$$\log\frac{Y^{opt}}{Y} \approx \underbrace{\frac{1}{2}\left(\mathbb{E}_\lambda\left[\delta_\theta\right] - 1\right) Cov_\lambda\left[\sigma_\theta, \log\frac{1}{\mu_\theta}\right]}_{\text{Relative production inefficiency}} + \underbrace{\frac{1}{2}\mathbb{E}_\lambda\left[\sigma_\theta\right]\left(\frac{\mathbb{E}_\lambda\left[\delta_\theta\right]}{\bar{\mu}} - 1\right)^2}_{\text{Entry inefficiency}} + \underbrace{\frac{1}{2}\left(\mathbb{E}_\lambda\left[\delta_\theta\right] - \delta_{\theta^*}\right)^2 \lambda_{\theta^*}\gamma_{\theta^*}\frac{\sigma_{\theta^*}}{\delta_{\theta^*}}}_{\text{Selection inefficiency}}.$$

The three terms, which are all positive, correspond to how the three margins of inefficiency (relative production, entry, and selection) contribute to overall misallocation.

Variation in markups and pass-throughs unrelated to size. In our calibration, we assume that markups and pass-throughs vary only as a function of firm size. In practice, firms' markups also vary for reasons unrelated to size. Appendix L shows how our results change if there is variation in pass-throughs and price elasticities (and hence markups) unrelated to size. We find that this additional variation strengthens the Darwinian effect. A back-of-the-envelope exercise suggests that additional heterogeneity in markups does not significantly change our results.

Chaney (2008) entry. In the main text, we assume there is an unbounded mass of potential entrants that enter the market until expected profits equal the fixed cost of entry. In Appendix J, we consider an alternative entry technology where the mass of potential entrants is finite and proportional to population, as in Chaney (2008). We show that the Darwinian effect persists in this version of the model.

9 Conclusion

In this paper, we analyze the origins of aggregate increasing returns to scale. We find that changes in allocative efficiency—i.e., changes in welfare due to the reallocation of resources—constitute the majority of gains from an increase in market size. That is, intensifying competition in a larger market reallocates resources across uses in a way that improves efficiency.

In particular, the lion's share of efficiency gains come from a force we call the Darwinian effect, which reallocates resources to high-markup firms and alleviates cross-sectional misallocation. This effect is distinct from two forces often studied in the literature—an increase in market size may toughen selection, and an increase in market size may lead firms to reduce their markups—which we find are either minor or deleterious for welfare.

In addition to improving the cross-sectional allocation of resources, Darwinian reallocations increase the economy's aggregate markup, decrease the share of income earned by production labor, and lead to an increased concentration of sales and employment in large firms. In our calibrated model, an increase in market size improves efficiency precisely because it increases industrial concentration and redistributes resources to large, high-markup firms. Our analysis raises the possibility that beneficial reallocations from globalization come hand-in-hand with increases in concentration and aggregate markups.

References

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

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

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

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

Atkeson, Andrew and Ariel Burstein, "Pricing-to-market, trade costs, and international relative prices," *American Economic Review*, 2008, *98* (5), 1998–2031.

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

Baily, Martin Neil, Charles Hulten, and David Campbell, "Productivity Dynamics in Manufacturing Plants," *Brookings Papers on Economic Activity*, 1992, *23*, 187–267.

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

_ and _, "Entry vs. Rents: Aggregation with Economies of Scale," Technical Report 27140, National Bureau of Economic Research 2020.

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

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

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

Burstein, Ariel and Gita Gopinath, "International prices and exchange rates," *Handbook of International Economics*, 2014, *4*, 391–451.

_ , Vasco M. Carvalho, and Basile Grassi, "Bottom-up Markup Fluctuations," Technical Report 27958, National Bureau of Economic Research 2020.

Chaney, Thomas, "Distorted gravity: the intensive and extensive margins of international trade," *American Economic Review*, 2008, *98* (4), 1707–21.

_ and Ralph Ossa, "Market size, division of labor, and firm productivity," *Journal of International Economics*, 2013, *90* (1), 177–180.

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

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

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

Edmond, Chris, Virgiliu Midrigan, and Daniel Yi Xu, "Competition, markups, and the gains from international trade," *American Economic Review*, 2015, *105* (10), 3183–3221.

_ , _ , and _ , "How costly are markups?," Technical Report, National Bureau of Economic Research 2018.

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

Fally, Thibault, "Generalized separability and integrability: Consumer demand with a price aggregator," *Journal of Economic Theory*, 2022, *203* (105471).

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

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

Forlani, Emanuele, Ralf Martin, Giordano Mion, and Mirabelle Muuls, "Unraveling Firms: Demand, Productivity and Markups Heterogeneity," June 2022. Working paper.

Foster, Lucia, John C. Haltiwanger, and C. J. Krizan, *New Developments in Productivity Analysis*, University of Chicago Press,

Gilchrist, Simon, Raphael Schoenle, Jae Sim, and Egon Zakrajšek, "Inflation dynamics during the financial crisis," *American Economic Review*, 2017, *107* (3), 785–823.

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

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

Johnson, Justin P. and David P. Myatt, "On the simple economics of advertising, marketing, and product design," *American Economic Review*, 2006, *96* (3), 756–784.

Kehrig, Matthias and Nicolas Vincent, "The micro-level anatomy of the labor share decline," *The Quarterly Journal of Economics*, 2021, *136* (2), 1031–1087.

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

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

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

Lipsey, Richard G. and Kelvin Lancaster, "The general theory of second best," *The Review of*

Economic Studies, 1956, *24* (1), 11–32.

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

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

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

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

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

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

_ and _ , "Selection and Sorting of Heterogeneous Firms through Competitive Pressures," 2022. Working paper.

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

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

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

Melitz, Marc J. and Saso Polanec, "Dynamic Olley-Pakes productivity decomposition with entry and exit," *The RAND Journal of Economics*, 2015, *46* (2), 362–375.

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

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

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

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

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

Petrin, Amil and James Levinsohn, "Measuring aggregate productivity growth using plant-level data," *The RAND Journal of Economics*, 2012, *43* (4), 705–725.

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

Ravn, Morten, Stephanie Schmitt-Grohé, and Martin Uribe, "Deep habits," *The Review of Economic Studies*, 2006, *73* (1), 195–218.

Ridder, Maarten De, Basile Grassi, and Giovanni Morzenti, "The hitchhiker's guide to markup estimation," 2021. Working paper.

Spence, Michael, "Product selection, fixed costs, and monopolistic competition," *The Review*

of economic studies, 1976, *43* (2), 217–235.

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

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

Vives, Xavier, *Oligopoly pricing: old ideas and new tools*, MIT press, 2001.

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

We use information from VAT declaration in Belgium for the year 2014 to recover the sales distribution of Belgian manufacturers. Table A.1 displays the underlying data.

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

Table A.1: Firm size distribution for manufacturing firms from VAT declarations in Belgium for 2014.

With some abuse of notation, let $\theta \in [0, 1]$ be the fraction of observations up to some size. Then the cumulative share of sales for firms up to some cutoff θ (i.e. the "Share of sales" column in Table A.1) is defined as

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

where $\lambda(\theta)$ is the sales share density. We fit a smooth curve to $\Lambda(\theta)$ of the form $\exp(c_0 + c_1\theta + c_2\theta^{c_3})$, displayed in Figure A.1. Then, we compute the sales share density $\lambda(\theta)$, which is given by

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

For pass-throughs ρ_θ, 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 sales above 1 million euros, which means that the sample is not representative of all manufacturers. The estimates are in Table A.2.

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	0.33897530	0.8689
300	0.88848094	0.28832632	0.40832230	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 A.2: 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 $\lambda^P(\theta)$ be the sales share density of Prodcom firms of type θ. We again compute $\lambda^P(\theta)$ by the same method above, but using only the sample of firms in Prodcom.

The variable $\alpha(\theta)$ satisfies

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

We fit a flexible spline function to $\alpha(\theta)$, shown in Figure A.2. To recover the pass-throughs $\rho(\theta)$, we write

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

Hence, we can recover the pass-through function via

$$\rho(\theta) = \frac{\left(\int_0^\theta \lambda^P(x)dx\right)}{\lambda^P(\theta)}\frac{d\alpha}{d\theta} + \alpha(\theta).$$

Since we have $\lambda^P(\theta)$ and $\alpha(\theta)$, we can solve for pass-throughs $\rho(\theta)$ as a function of the number of employees.

Finally, we merge our pass-through information from Prodcom with the sales density from VAT declarations by assuming that the pass-through 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 Proposition 5, solving the differential equations numerically using the Runge-Kutta algorithm on a large grid.

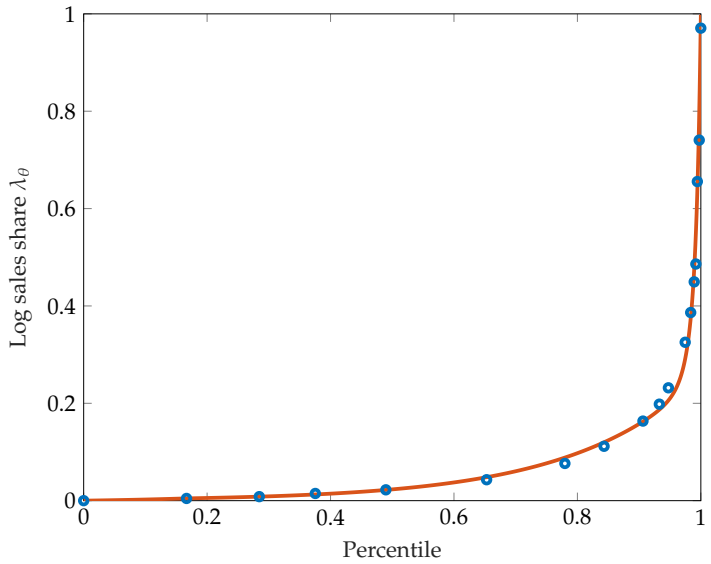


Figure A.1: Cumulative sales share distribution. The blue dots are cumulative sales share for firms smaller than the percentile given by the x-axis in Prodcom. The solid red line is a fitted spline.

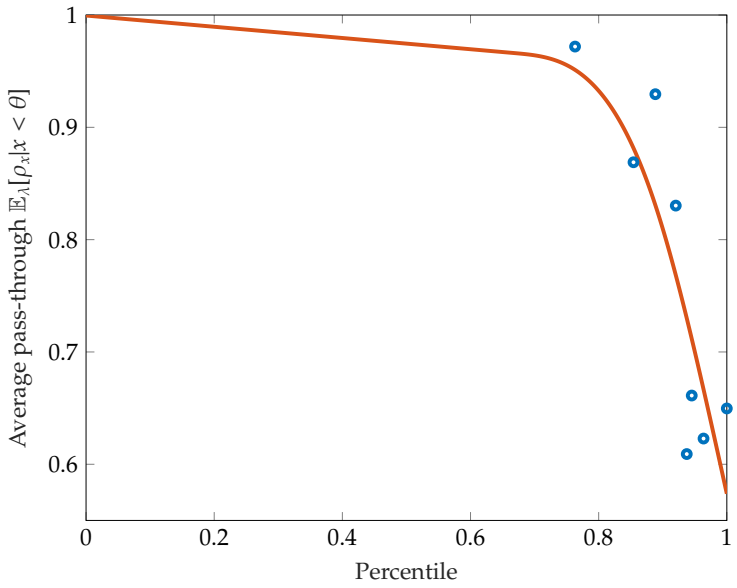
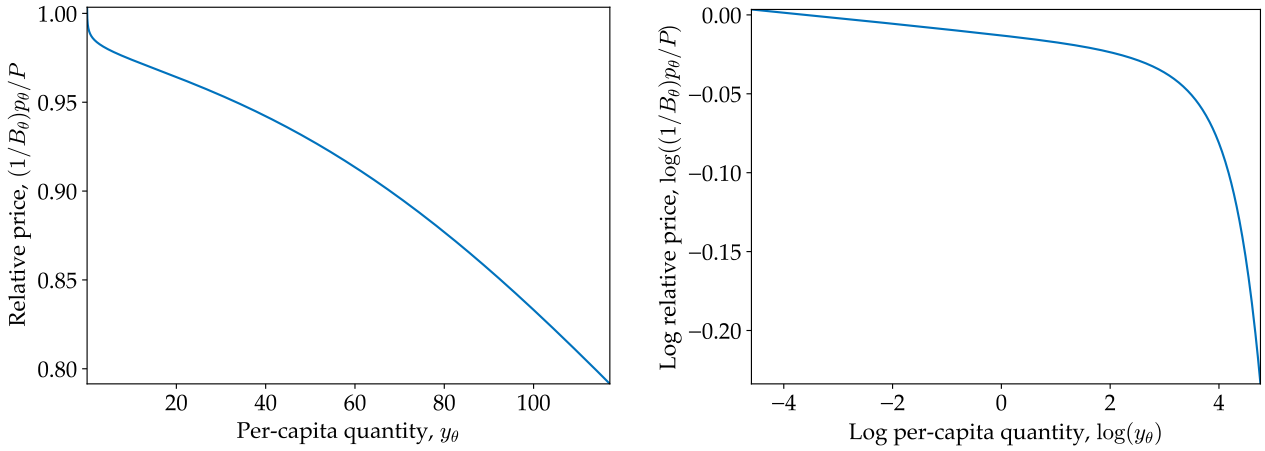


Figure A.2: The blue dots are estimates from Amiti et al. (2019) of the average sales-weighted pass-through by size percentile. The red line is a fitted spline.

Appendix B Additional Tables and Figures

Figure B.1 plots the residual demand curve in linear and log-log terms under the calibration with $\bar{\mu} = 1.09$ and efficient selection. (The residual demand curves are qualitatively similar under the efficient entry case.) As mentioned in the main text, Figure B.1a shows that our estimate has a distinctly non-isoelastic shape, indicating substantial departures from CES.



(a) Inverse residual demand curve

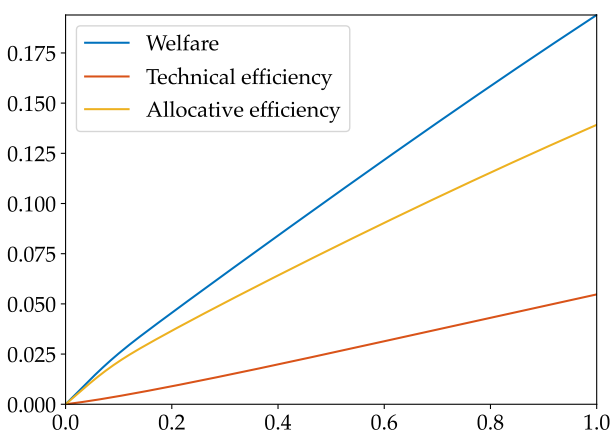
(b) Log-log inverse residual demand curve

Figure B.1: Residual demand curve (quality-adjusted price against quantity) for the efficient selection case with $\bar{\mu} = 1.09$. The results for the efficient entry case are similar.

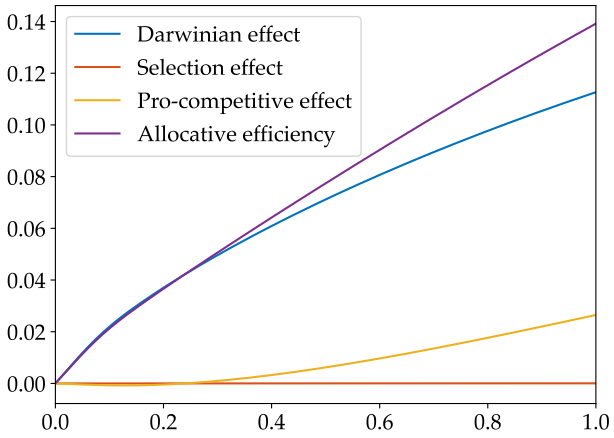
Figure B.2 shows the nonlinear response of welfare, and its decomposition following Theorem 1, for non-infinitesimal changes in market size. We compute this decomposition by numerically solving the system of ordinary differential equations in Appendix C and cumulative (i.e. integrating) the first-order changes.

Figure B.2 shows cumulated changes in welfare and each channel for the calibration with efficient selection $\mathbb{E}_\lambda[\delta_\theta] = \delta_{\theta^*}$. The first panel shows that even though their relative importance decreases slightly with the size of the shock, changes in allocative efficiency continue to dwarf changes in technical efficiency even for large shocks. The second panel shows that as the population grows, changes in allocative efficiency due to the pro-competitive channel start to account for a non-trivial part of overall changes in allocative efficiency. This happens because as we increase population, the harmonic average of markups increases due to the Darwinian effect. This means that entry becomes more excessive, and hence that reallocations triggered by individual markup reductions improve allocative efficiency more.

Table B.1 reports the average elasticity of welfare and real GDP per capita to population for a large shock $\Delta \log L = 0.5$. To calculate the response of the model to large shocks, we use the series of differential equations in Appendix C and cumulate over a series of changes starting at the initial equilibrium. Although the model is far from being log-linear, the qualitative conclusions are unchanged.



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



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

Figure B.2: Decomposition of changes in welfare and allocative efficiency following Proposition 1, for large shocks. These graphs show the case with $\bar{\mu} = 1.09$ and efficient selection.

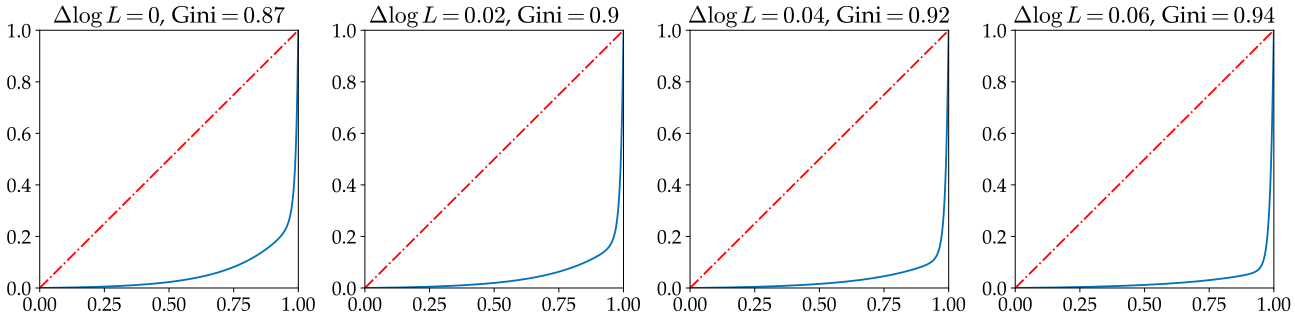


Figure B.3: Each panel depicts the Lorenz curve for the employment distribution for different values of the market size parameter L. The dotted red line indicates the line of perfect equality.

Figure 4 in the main text shows that, in our quantitative calibration, the concentration of sales increases with market size. Figure B.3 plots Lorenz curves for the employment distribution as we increase market size. The initial concentration of employment is lower than that of sales (Gini coefficients of 0.87 versus 0.88) and rises as the market expands.

Tables B.2 and B.3 test the robustness of our results over a two-way grid of boundary conditions for $\bar{\mu}$ and δ_{θ^*}. In Table B.2, we report the response of welfare and changes in allocative efficiency to market size, following Theorem 1, for different boundary conditions. Although the magnitude of $d \log Y / d \log L$ changes as we change the boundary conditions, the contribution of allocative efficiency to the overall total is at least 50% of the overall effect. Table B.3 breaks down the overall effect on allocative efficiency into the different margins of adjustment (Darwinian, selection, and pro-competitive). The Darwinian effect is always responsible for the bulk of the positive effect. As mentioned, for a given $\bar{\mu}$, the selection effect is strongest when δ_{θ^*} is lowest, but even for $\delta_{\theta^*} = 1$, the selection effect is negligible.

	Efficient selection $\mathbb{E}_\lambda[\delta_\theta] = \delta_{\theta^*}$	Efficient entry $\mathbb{E}_\lambda[\delta_\theta] = \bar{\mu}$
Welfare: $\Delta \log Y$	0.206	0.199
Technical efficiency	0.051	0.065
Allocative efficiency	0.155	0.134
Darwinian effect	0.142	0.211
Selection effect	0.000	-0.065
Pro-competitive effect	0.012	-0.012
Real GDP per capita	0.097	0.097
Aggregate markup	0.153	0.153

Table B.1: The average elasticity of welfare and real GDP per capita to population for a large shock $\Delta \log L = 0.5$.

Table B.2: Change in log welfare and allocative efficiency for different boundary conditions

	δ_{θ^*}									
$\bar{\mu}$	1	2	3	4	5	6	7	8	9	10
1.05	[0.137, 0.119]	[0.141, 0.112]	[0.144, 0.106]	[0.148, 0.099]	[0.151, 0.092]	[0.155, 0.086]	[0.158, 0.079]	[0.162, 0.073]	[0.165, 0.066]	[0.169, 0.059]
1.06	[0.166, 0.145]	[0.170, 0.138]	[0.173, 0.131]	[0.177, 0.125]	[0.180, 0.118]	[0.184, 0.111]	[0.187, 0.105]	[0.191, 0.098]	[0.194, 0.091]	[0.198, 0.085]
1.07	[0.196, 0.171]	[0.200, 0.164]	[0.203, 0.157]	[0.207, 0.151]	[0.210, 0.144]	[0.214, 0.137]	[0.217, 0.131]	[0.221, 0.124]	[0.224, 0.117]	[0.228, 0.111]
1.08	[0.227, 0.198]	[0.231, 0.191]	[0.234, 0.184]	[0.237, 0.178]	[0.241, 0.171]	[0.244, 0.164]	[0.248, 0.158]	[0.251, 0.151]	[0.255, 0.144]	[0.258, 0.138]
1.09	[0.259, 0.225]	[0.262, 0.219]	[0.265, 0.212]	[0.269, 0.205]	[0.272, 0.199]	[0.276, 0.192]	[0.279, 0.185]	[0.283, 0.178]	[0.286, 0.172]	[0.290, 0.165]
1.10	[0.291, 0.254]	[0.294, 0.247]	[0.298, 0.240]	[0.301, 0.234]	[0.305, 0.227]	[0.308, 0.220]	[0.312, 0.213]	[0.315, 0.207]	[0.319, 0.200]	[0.322, 0.193]
1.11	[0.324, 0.283]	[0.328, 0.276]	[0.331, 0.270]	[0.334, 0.263]	[0.338, 0.256]	[0.341, 0.249]	[0.345, 0.243]	[0.348, 0.236]	[0.352, 0.229]	[0.355, 0.222]
1.12	[0.358, 0.313]	[0.362, 0.307]	[0.365, 0.300]	[0.369, 0.293]	[0.372, 0.286]	[0.376, 0.279]	[0.379, 0.273]	[0.382, 0.266]	[0.386, 0.259]	[0.389, 0.252]
1.13	[0.394, 0.344]	[0.397, 0.338]	[0.401, 0.331]	[0.404, 0.324]	[0.407, 0.317]	[0.411, 0.311]	[0.414, 0.304]	[0.418, 0.297]	[0.421, 0.290]	[0.424, 0.283]
1.14	[0.430, 0.377]	[0.434, 0.370]	[0.437, 0.363]	[0.440, 0.356]	[0.444, 0.349]	[0.447, 0.343]	[0.451, 0.336]	[0.454, 0.329]	[0.457, 0.322]	[0.461, 0.315]
1.15	[0.468, 0.410]	[0.471, 0.403]	[0.475, 0.396]	[0.478, 0.390]	[0.481, 0.383]	[0.485, 0.376]	[0.488, 0.369]	[0.492, 0.362]	[0.495, 0.356]	[0.498, 0.349]

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

Table B.3: Change in allocative efficiency for different boundary conditions

		δ_{θ^*}				
		1	3	5	7	9
$\bar{\mu}$	1.05	[0.122, 0.001, -0.004]	[0.259, -0.118, -0.035]	[0.397, -0.238, -0.066]	[0.534, -0.357, -0.098]	[0.672, -0.477, -0.129]
	1.06	[0.148, 0.001, -0.005]	[0.287, -0.119, -0.036]	[0.426, -0.240, -0.068]	[0.565, -0.360, -0.100]	[0.704, -0.481, -0.131]
	1.07	[0.175, 0.002, -0.006]	[0.315, -0.120, -0.038]	[0.456, -0.242, -0.070]	[0.596, -0.364, -0.101]	[0.736, -0.485, -0.133]
	1.08	[0.203, 0.002, -0.007]	[0.345, -0.121, -0.039]	[0.486, -0.244, -0.071]	[0.628, -0.367, -0.104]	[0.769, -0.490, -0.136]
	1.09	[0.232, 0.002, -0.009]	[0.375, -0.122, -0.041]	[0.518, -0.246, -0.073]	[0.661, -0.370, -0.106]	[0.804, -0.494, -0.138]
	1.10	[0.262, 0.002, -0.011]	[0.406, -0.123, -0.043]	[0.551, -0.248, -0.076]	[0.695, -0.373, -0.108]	[0.839, -0.498, -0.141]
	1.11	[0.293, 0.003, -0.013]	[0.439, -0.124, -0.045]	[0.584, -0.250, -0.078]	[0.730, -0.376, -0.111]	[0.875, -0.502, -0.143]
	1.12	[0.325, 0.003, -0.015]	[0.472, -0.125, -0.048]	[0.619, -0.252, -0.081]	[0.766, -0.379, -0.113]	[0.912, -0.507, -0.146]
	1.13	[0.359, 0.003, -0.017]	[0.507, -0.125, -0.050]	[0.655, -0.254, -0.083]	[0.803, -0.382, -0.116]	[0.951, -0.511, -0.150]
	1.14	[0.393, 0.004, -0.020]	[0.542, -0.126, -0.053]	[0.692, -0.256, -0.086]	[0.841, -0.386, -0.120]	[0.991, -0.515, -0.153]
	1.15	[0.429, 0.004, -0.023]	[0.580, -0.127, -0.056]	[0.730, -0.258, -0.090]	[0.881, -0.389, -0.123]	[1.032, -0.519, -0.157]

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

Appendix C Proofs

In this section, we log-linearize the model and derive Theorem 1. We expand the equilibrium equations presented in Section 2 to the first order in the shocks. We use these equations to prove our results. Under Assumption 1, we can also iterate on them as differential equations to solve for the response of endogenous variables non-linearly in our calibrated model.

Price aggregator. Differentiating the definition of the price aggregator, we find

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

Welfare. Matsuyama and Ushchev (2017) show that the ideal price index is related to the price aggregator P by

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

Differentiating this equation, using the budget constraint $P^Y Y = 1$, and combining with the equation for $d\log P$ above yields

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

Quantities. Differentiating the demand curve facing each variety, we get

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

Markups. Differentiating the markup equation, we get

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

Prices. Differentiating the equation for prices, we find

$$d\log p_\theta = d\log\mu_\theta - d\log A_\theta.$$

Sales shares. Differentiating the sales shares equation, we find

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

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

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

Selection. Differentiating the selection condition, we get

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

We define

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

which allows us to write the selection condition more simply as

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

Entry. Differentiating the free-entry condition yields

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

Proof of Theorem 1. To solve for the change in welfare following a change in market size, $d\log L$, we take the system of log-linearized equations above and set $d\log A_\theta = d\log f_{o,\theta} = d\log f_e = 0$. We get the following system of eight equations:

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

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

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

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

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

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

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

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

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

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

The last equation gives intuition for how the price aggregator moves as the market size increases. An increase in market size lowers the price aggregator due to new entry. Since competition between all varieties is mediated by the price aggregator, this decrease in the price index then affects the relative quantities demanded of each variety, the selection cutoff, and the markup adjustments chosen by each firm.

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

$$d\log Y = (\bar{\delta}-1)\,d\log L + \left(\xi^\epsilon + \xi^{\theta^*} + \xi^\mu\right)\bar{\mu}d\log L.$$

The first term captures the change in welfare due to technical efficiency while the second term captured the change in welfare due to allocative efficiency. This equation gives the result in Theorem 1. ■

Proof of Proposition 1. The aggregate markup is given by,

$$\bar{\mu} = \mathbb{E}_\lambda\left[\mu_\theta^{-1}\right]^{-1}.$$

Log-linearizing, we get:

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

From above, we use:

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

Substituting these into the expression for $d\log\bar{\mu}$ and re-arranging yields Proposition 1. ■

Proof of Proposition 2. The change in real GDP (as measured by statistical agencies) is given by

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

Using $d\log p_\theta = d\log\mu_\theta = (1-\rho_\theta)d\log P = -(1-\rho_\theta)\mathbb{E}_\lambda\left[\frac{1}{\sigma_\theta}\right]\bar{\mu}d\log L$, we get:

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

■

See Appendix E for the proof of Proposition 3.

Proof of Lemma 1. To derive (11), note that the initial allocation of labor allocates a fraction $l = \mathbb{E}[l_\theta] = \mathbb{E}_\lambda[1/\mu_\theta]$ to variable production, and the remainder to entry and overhead. Suppose we take reduce the fraction of labor allocated to variable production (while preserving the proportions of variable production labor allocated across firms) by $d\log l_\theta = d\log l$. Re-allocating that labor to entry and overhead costs allows us to increase consumer welfare by

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

where $d\log l_e$ is the increase in labor allocated to entry. This gain is consumer welfare is offset by a reduction in the per-capita quantity consumed of each variety, equal to $\mathbb{E}_\lambda[d\log y_\theta] =$

$d \log l - d \log M$. Rearranging, we find that the net change in welfare from reducing the fraction of labor allocated to variable production and increasing the allocation to entry is positive if and only if the average consumer surplus ratio exceeds the harmonic average of markups, yielding the condition in (11) above. ■

Proof of Lemma 2. To derive this condition, 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 $d \log y_\theta = 0$. Plugging these perturbations into (9), the overall effect on welfare is $(\mathbb{E}_\lambda[\delta_\theta] - \delta_{\theta^*})\lambda_{\theta^*}[g(\theta^*)/(1-G(\theta^*))]d\theta^*$, which is positive (too little selection) if and only if $\delta_{\theta^*} < \mathbb{E}_\lambda[\delta_\theta]$. ■

Proof of Lemma 3. The intuition is the following. Consider a reduction $d \log l_{\theta'} < 0$ in the fraction of labor allocated to the supply of varieties in $(\theta', \theta' + d\theta')$ and a complementary increase $d \log l_\theta = -(g(\theta')/g(\theta))(l_{\theta'}/l_\theta)d \log l_{\theta'} > 0$ in the fraction of labor allocated to the supply of varieties in $(\theta, \theta + d\theta')$, which, using the fact that $l_{\theta'}/l_\theta = (\lambda_{\theta'}/\mu_{\theta'})/(\lambda_\theta/\mu_\theta)$, can be rewritten as $d \log l_\theta = -(g(\theta')/g(\theta))(\lambda_{\theta'}/\mu_{\theta'})/(\lambda_\theta/\mu_\theta)d \log l_{\theta'} > 0$. This leads to a decrease $d \log y_{\theta'} = d \log l_{\theta'} < 0$ in the quantity of the former varieties and an increase $d \log y_\theta = -(g(\theta')/g(\theta))(\lambda_{\theta'}/\mu_{\theta'})/(\lambda_\theta/\mu_\theta)d \log l_{\theta'} > 0$ in the quantity of the latter varieties. The net effect on welfare is $g(\theta')\lambda_{\theta'}d \log y_{\theta'}d\theta' + g(\theta)\lambda_\theta d \log y_\theta d\theta' = -(\mu_\theta/\mu_{\theta'} - 1)\lambda_{\theta'}g(\theta')d\theta' d \log l_{\theta'}$, which is positive if and only $\mu_\theta > \mu_{\theta'}$.

■

Proof of Proposition 4. Under Assumption 1, sales densities λ_θ are given by:

$$\lambda_\theta = (1-G(\theta^*))Mp_\theta y_\theta = (1-G(\theta^*))Ms(\frac{1}{A_\theta B_\theta}\frac{\mu_\theta}{P}).$$

Taking the partial derivative of λ_θ with respect to A_θ and B_θ yields:

$$\frac{\partial \log \lambda_\theta}{\partial \log A_\theta} = \frac{\partial \log \lambda_\theta}{\partial \log B_\theta} = \rho_\theta(\sigma_\theta - 1).$$

Since $\rho_\theta > 0$ and $\sigma_\theta > 1$, sales are strictly increasing in $A_\theta B_\theta$.

Profitability X_θ is

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

Hence, under Assumption 1, the partial derivative of X_θ with respect to A_θ and B_θ is:

$$\frac{\partial \log X_\theta}{\partial \log A_\theta} = \frac{\partial \log X_\theta}{\partial \log B_\theta} = \sigma_\theta - 1.$$

Again, since $\sigma_\theta > 1$, profitability is strictly increasing in $A_\theta B_\theta$.

To show that any two firms with identical sales also have identical pass-throughs, markups, and consumer surplus ratios, note also that since $s(\cdot)$ is strictly decreasing in its argument, there is a one-to-one mapping between sales and a firm's quality-adjusted relative price, $\frac{1}{B_\theta}\frac{p_\theta}{P}$. Thus, for any two firms with sales λ_θ and $\lambda_{\theta'}$, if $\lambda_\theta = \lambda_{\theta'}$, then $\frac{1}{B_\theta}\frac{p_\theta}{P} = \frac{1}{B_{\theta'}}\frac{p_{\theta'}}{P}$. Finally, since under Assumption 1, pass-throughs, markups, and consumer surplus ratios can all be expressed in terms of the function $s(\cdot)$ and a firm's relative price $\frac{1}{B_\theta}\frac{p_\theta}{P}$, these firms must also have identical pass-throughs, markups, and consumer sales ratios. ■

Proof of Proposition 5. The proof of (21) follows directly from combining the two differential equations relating markups and sales to $A_\theta B_\theta$ provided in the main text.

For (22), recall

$$\delta_\theta = \frac{s_\theta(\frac{p_\theta}{P}) + \int_{p_\theta/P}^{\infty} \frac{s_\theta(\xi)}{\xi} d\xi}{s_\theta(\frac{p_\theta}{P})}.$$

Differentiating the numerator, we find that the change in the total surplus from a variety follows:

$$d \log \left[s_\theta(\frac{p_\theta}{P}) + \int_{p_\theta/P}^{\infty} \frac{s_\theta(\xi)}{\xi} d\xi \right] = -\frac{\sigma_\theta}{\delta_\theta} d \log \frac{p_\theta}{P} = \frac{\sigma_\theta}{\sigma_\theta - 1} \frac{1}{\delta_\theta} d \log \lambda_\theta.$$

Meanwhile, the denominator is simply proportional to revenues, so $d \log s_\theta = d \log \lambda_\theta$. Hence we get,

$$d \log \delta_\theta = \left(\frac{\mu_\theta}{\delta_\theta} - 1 \right) d \log \lambda_\theta.$$

For the overhead cost, recall from the selection condition,

$$\left(1 - \frac{1}{\mu_{\theta^*}}\right) \frac{\lambda_{\theta^*}}{f_{o,\theta}} = (1 - G(\theta^*)) \frac{M}{L}.$$

Normalizing $M = 1$ and $L = 1$, applying uniform overhead costs $f_{o,\theta} = f_o$, and using $G(x) = x$, we get

$$(1 - \theta^*) f_o = \left(1 - \frac{1}{\mu_{\theta^*}}\right) \lambda_{\theta^*}.$$

For the entry cost, recall from the free entry condition,

$$\frac{L}{M} \mathbb{E}_\lambda \left[1 - \frac{1}{\mu_\theta} \right] = f_e \Delta + (1 - G(\theta^*)) \, \mathbb{E} \left[f_{o,\theta} \right].$$

Thus,

$$\mathbb{E}_{\lambda}\left[1-\frac{1}{\mu_{\theta}}\right]=f_{e}\Delta+(1-\theta^{*})\,f_{o}.$$

■

Appendix D Implementation Using 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 D.1 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 D.2. The basic qualitative message of our previous results in Table 1 is unchanged, and the Darwinian effects are still overwhelmingly the dominant force in the model.

No of Employees	No of Products	No of firms
5	1.3636364	22
10	2.0550459	109
20	2.2004950	404
30	2.4203297	728
50	2.4203895	873
75	2.3727506	389
100	3.2946860	207
200	3.2250000	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 D.1: Number of products on average from Prodcom sample in 2014.

	Efficient selection $\mathbb{E}_\lambda[\delta_\theta] = \delta_{\theta^*}$	Efficient entry $\mathbb{E}_\lambda[\delta_\theta] = \bar{\mu}$
Welfare: $d \log Y$	0.169	0.266
Technical efficiency	0.042	0.090
Allocative efficiency	0.127	0.176
Darwinian effect	0.121	0.260
Selection effect	0.000	-0.062
Pro-competitive effect	0.006	-0.021
Real GDP per capita	0.030	0.030
Aggregate markup	0.211	0.211

Table D.2: The elasticity of welfare and real GDP per capita to population following Propositions 1 and 2 for heterogeneous firms case using product-level data.

Appendix E Welfare Response to an Entry Tax

This appendix presents the proof of Proposition 3, which characterizes the response of welfare to a marginal tax on entry.

We modify our setup to allow for an entry tax. As in the main text, demand curves are given by

$$y_\theta = \frac{I}{p_\theta} s_\theta(\frac{p_\theta}{P}).$$

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

$$\int_\Theta p_\theta y_\theta dF(\theta) = I = w + g, \qquad \text{and} \qquad \Lambda_L = \frac{w}{w+g}. \tag{23}$$

We use the wage as the numeraire, normalizing $w = 1$ throughout.

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

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

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

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

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

$$\tau f_e \Delta M = gL.$$

We combine this equation with (23) to solve for the labor share in terms of the entry tax,

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

We log-linearize the above conditions at the point where the tax is initially zero, and hence $\tau = 0, \Lambda_L = 1$. The response of welfare to a change in the entry tax is thus described by the

fixed point of the following system of equations:

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

where we define the share of fixed costs spent on entry ψ_e as

$$\psi_e = \frac{\Delta f_e}{\Delta f_e + \left(1-G(\theta^*)\right)\mathbb{E}\left[f_{o,\theta}\right]}.$$

Solving the fixed point yields,

$$
\begin{aligned}
d\log Y = &\left[1 - \frac{\mathbb{E}_\lambda\left[\delta_\theta\right]}{\bar{\mu}} - \left(\mathbb{E}_\lambda\left[\delta_\theta\right]-\delta_{\theta^*}\right)\lambda_{\theta^*}\gamma_{\theta^*}\mathbb{E}_\lambda\left[\frac{\sigma_{\theta^*}}{\sigma_\theta}\right]\right.\\
&\left.-\mathbb{E}_\lambda\left[\frac{1}{\sigma_\theta}\right]\mathbb{E}_\lambda\left[\left(1-\rho_\theta\right)\left[1-\left(\mathbb{E}_\lambda\left[\delta_\theta\right]-1\right)\left(\sigma_\theta-1\right)\right]\right] - \left(\mathbb{E}_\lambda\left[\delta_\theta\right]-1\right)\left(\mathbb{E}_\lambda\left[\frac{1}{\sigma_\theta}\right]\mathbb{E}_\lambda\left[\sigma_\theta\right]-1\right)\right]\psi_e d\tau.
\end{aligned}
$$

We use the definitions of ξ^ϵ, ξ^{θ^*}, and ξ^μ in the main text to simplify this expression to the result in Proposition 3.

Appendix F Decomposing Technical and Allocative Efficiency

In this appendix, we explicitly solve out for $\partial \log \mathcal{Y}/\partial \mathcal{X}$ and $\partial \mathcal{X}/\partial \log L$ in equation (14). We also discuss an alternative decomposition of the elasticity of welfare to market size in terms of the decentralized equilibrium's distance to the efficient frontier.

F.1 Decomposition Using Changes in the Allocation Matrix

In this section, we show how the allocation of labor changes in response to market size and how those changes impact welfare. Combining the change in welfare due to changes in the allocation with changes in the allocation due to market size yields Theorem 1.

We start with (14), which decomposes the change in welfare into a change in technical efficiency, holding the allocation of resources across uses constant, and a change in allocative efficiency.

$$d \log Y = \underbrace{\frac{\partial \log \mathcal{Y}}{\partial \log L} d \log L}_{\text{technical efficiency}} + \underbrace{\frac{\partial \log \mathcal{Y}}{\partial \mathcal{X}} \frac{\partial \mathcal{X}}{\partial \log L} d \log L}_{\text{allocative efficiency}}. \tag{24}$$

For simplicity, we fix the entry cost f_e and the overhead cost f_o (which, for the purpose of this appendix, we assume is identical across firms), and only consider changes in market size. The allocation vector

$$\mathcal{X} = \left\{ \frac{L^{\text{entry}}}{L}, \frac{L^{\text{overhead}}}{L}, \frac{L^{\text{variable}}}{L}, \left\{ \frac{l_\theta}{L^{\text{variable}}} \right\} \right\},$$

includes the fraction of total labor used for entry, overhead, variable production, and the fraction of labor used for variable production per variety of type θ. (Following the notation in the main text, we use l_θ to denote the per-capita production labor used for each variety with type θ.) For an allocation to be feasible, it must satisfy the constraints,

$$L^{\text{variable}} = \int_\Theta (l_\theta \cdot L) dF(\theta), \qquad \text{and} \qquad L = L^{\text{entry}} + L^{\text{overhead}} + L^{\text{variable}}. \tag{25}$$

We will also limit our focus to a subset of feasible allocations that have a single selection cutoff θ^* whereby firms with types $\theta \geq \theta^*$ produce and those with types $\theta < \theta^*$ do not produce, where the cumulative distribution of firm types follows $G(\theta)$, and where entry and overhead labor are allocated to all types for which the allocation of variable production labor is strictly positive. This implies that resources used for entry and overhead costs are given by

$$L^{\text{entry}} = M f_e \qquad \text{and} \qquad L^{\text{overhead}} = M \left(1 - G(\theta^*)\right) f_o.$$

Accordingly, the measure of varieties with type θ is $dF(\theta) = M g(\theta) \mathbf{1}_{(\theta \geq \theta^*)} d\theta$.

In order to expand (24), we will need the partial derivative of welfare with respect to

changes in each element of the allocation matrix $\mathcal{X}$ and how changes in the allocation matrix $\mathcal{X}$ are bound by our feasibility constraints. Let us start by log-linearizing the feasibility constraints in (25). First, for the constraint that labor allocated to variable production of each variety must total the overall labor allocated to variable production, we get:

$$d\log\frac{L^{\text{variable}}}{L} = d\log L + d\log\frac{L^{\text{entry}}}{L} - \lambda_{\theta^*}\frac{\bar{\mu}}{\mu_{\theta^*}}\frac{g(\theta^*)}{1-G(\theta^*)}d\theta^* + \int_{\theta^*}^{1}\lambda_\theta\frac{\bar{\mu}}{\mu_\theta}d\log l_\theta\frac{g(\theta)}{1-G(\theta^*)}d\theta. \quad (26)$$

Second, for the feasibility constraint on total labor usage, we get

$$d\log\frac{L^{\text{variable}}}{L} = -\left[\bar{\mu}-1\right]d\log\frac{L^{\text{entry}}}{L} + \bar{\mu}\left(1-\frac{1}{\mu_{\theta^*}}\right)\lambda_{\theta^*}\frac{g(\theta^*)}{1-G(\theta^*)}d\theta^*. \quad (27)$$

Note that here we directly use changes in the selection cutoff θ^* as a proxy for the share of labor spent on overhead costs. Since the set of feasible allocations we consider requires $L^{\text{overhead}} = M(1-G(\theta^*))f_o$, for a given share of labor allocated to entry, there is a one-to-one mapping between the selection cutoff and the share of labor allocated to overhead costs.

Next, we need each of the partial derivatives of welfare with respect to changes in the allocation matrix. As exposited in the discussion of (14), we have that

$$\begin{aligned}
\frac{\partial\log\mathcal{Y}}{\partial\log L^{\text{entry}}} &= \mathbb{E}_\lambda[\delta_\theta],\\
\frac{\partial\log\mathcal{Y}}{\partial\theta^*} &= -\delta_{\theta^*}\lambda_{\theta^*}\frac{g(\theta^*)}{1-G(\theta^*)}\\
\frac{\partial\log\mathcal{Y}}{\partial\log L^{\text{variable}}} &= 1,\\
\frac{\partial\log\mathcal{Y}}{\partial\log l_\theta} &= \frac{p_\theta y_\theta}{I}.
\end{aligned}$$

where the third line comes from the homotheticity of preferences and the final line applies Shephard's Lemma.

Plugging in each of these and the two feasibility constraints (26) and (27) into our equation for welfare, we get:

$$\begin{aligned}
d\log Y = &\underbrace{\left(\mathbb{E}_\lambda\left[\delta_\theta\right]-1\right)d\log L}_{\text{technical efficiency}}\\
&\underbrace{+\left(\mathbb{E}_\lambda\left[\delta_\theta\right]-\bar{\mu}\right)d\log\frac{L^{\text{entry}}}{L} + \left[\bar{\mu}-\delta_{\theta^*}\right]\lambda_{\theta^*}\frac{g(\theta^*)}{1-G(\theta^*)}d\theta^* + Cov_\lambda\left[-\frac{\bar{\mu}}{\mu_\theta}, d\log\frac{l_\theta}{L^{\text{variable}}}\right]}_{\text{allocative efficiency}}. \quad (28)
\end{aligned}$$

Equation (28) decomposes the change in allocative efficiency following a change in market size into three components: (1) the change in welfare due to the change in the share of labor allocated to entry, (2) the change in welfare due to the change in the selection cutoff, and (3) the change in welfare due to reallocations of production labor across types. (Note that changes in the share of labor allocated to variable production are subsumed in (1) and (2) because of the feasibility constraint linking the allocation of labor across entry, overhead, and production.)

It is worth emphasizing that term (3) fully captures the effect of changes in the cross-sectional allocation on welfare. Note that this is not the same as changes in the dispersion of wedges, which is a commonly used statistic for the degree of misallocation. As term (3) shows, allocative efficiency can change even if wedges are held constant.

We now consider the form these allocative efficiency terms take in the decentralized equilibrium. The change in the share of labor allocated to variable entry and the change in the selection cutoff in the decentralized equilibrium can be derived by rearranging the log-linearized equilibrium conditions in Appendix C to get

$$d\log\frac{L^{\text{entry}}}{L}=\left(\lambda_{\theta^*}\gamma_{\theta^*}\bar{\mu}\left(\mathbb{E}_\lambda\left[\frac{\sigma_{\theta^*}}{\sigma_\theta}\right]-1\right)+\left[\mathbb{E}_\lambda\left[\sigma_\theta-1\right]-\mathbb{E}_\lambda\left[\left(\sigma_\theta-1\right)\left(1-\rho_\theta\right)\right]\right]\mathbb{E}_\lambda\left[\frac{1}{\sigma_\theta}\right]\bar{\mu}-1\right)d\log L,$$

and

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

As for term (3) which describes changes in cross-sectional efficiency, we can write:

$$\begin{aligned}&Cov_\lambda\left(-\bar{\mu}/\mu_\theta,d\log l_\theta\right)\\&=\left(\underbrace{Cov_\lambda\left[\bar{\mu}/\mu_\theta,\sigma_\theta\right]}_{>0}+\underbrace{\left(-\mathbb{E}_{\lambda\sigma}\left[1-\rho_\theta\right]\right)Cov_\lambda\left[\bar{\mu}/\mu_\theta,\sigma_\theta\right]}_{<0}+\underbrace{\mathbb{E}_\lambda\left[\sigma_\theta\right]Cov_{\lambda\sigma}\left[\bar{\mu}/\mu_\theta,\rho_\theta\right]}_{>0\text{ (in data)}}\right)\mathbb{E}_\lambda\left[\frac{1}{\sigma_\theta}\right]\bar{\mu}d\log L\end{aligned}$$

This equation shows that there are three types of reallocations that take place. The first reallocation is the Darwinian effect: decreases in the price index reallocate resources to high-markup firms. The second reallocation is due to price adjustments: if all firms cut their prices by the same amount (in percentages), firms that are more elastic will expand relative to firms with inelastic demand. The third reallocation is due to heterogeneous price adjustments: firms with low pass-throughs cut their markups more, and hence expand relative relative to high-passthrough firms. The first effect unambiguously reallocates to high-markup firms, the second effect unambiguously reallocates to low-markup firms, and the third effect depends on how pass-throughs and markups covary.

By grouping terms, we see that if markups negatively covary with pass-throughs (as they

do in our calibration), the net effect is always a reallocation of labor to high-markup firms:

$$Cov_\lambda\left(-\bar{\mu}/\mu_\theta, d\log l_\theta\right) = \left(\underbrace{\mathbb{E}_{\lambda\sigma}\left[\rho_\theta\right] Cov_\lambda\left[\bar{\mu}/\mu_\theta, \sigma_\theta\right]}_{>0} + \underbrace{\mathbb{E}_\lambda\left[\sigma_\theta\right] Cov_{\lambda\sigma}\left[\bar{\mu}/\mu_\theta, \rho_\theta\right]}_{>0 \text{ (in data)}}\right) \mathbb{E}_\lambda\left[\frac{1}{\sigma_\theta}\right] \bar{\mu} d\log L \tag{29}$$

Plugging in this expressions for $Cov_\lambda\left(-\bar{\mu}/\mu_\theta, d\log l_\theta\right)$, as well as the above expressions for $d\log L^{\text{entry}}/L$ and $g(\theta^*)/(1-G(\theta^*))d\theta^*$, into (28) yields Theorem 1.

F.2 Changes in the Distance to the Efficient Frontier

Our notion of allocative efficiency compares changes in welfare due to reallocations against the benchmark where the allocation of resources is held constant. A different notion of allocative efficiency that is also used in the literature measures changes in the distance to the efficient frontier, which is given by

$$\begin{aligned} d\log Y = & \underbrace{\frac{d\log Y^{opt}}{d\log L} d\log L}_{\text{Change in welfare at frontier}} + \underbrace{\frac{d\log Y/Y^{opt}}{d\log L} d\log L}_{\text{Change in distance to frontier}} \\ & = (\mathbb{E}_{\lambda^{opt}}[\delta_\theta^{opt}] - 1) d\log L + \left(\frac{d\log Y}{d\log L} - (\mathbb{E}_{\lambda^{opt}}[\delta_\theta^{opt}] - 1)\right) d\log L, \end{aligned}$$

where $\mathbb{E}_{\lambda^{opt}}[\delta_\theta^{opt}]$ is the sales-weighted average of consumer surplus ratios at the efficient point.[43] Note that changes in this measure of allocative efficiency depend on whether the change in welfare at the decentralized equilibrium, which we characterize in Theorem 1, is greater or smaller than $(\mathbb{E}_{\lambda^{opt}}[\delta_\theta^{opt}] - 1)$.

[43] The response of welfare to changes in market size at the efficient frontier can be derived from Theorem 1; we additionally confirm that $\frac{d\log Y^{opt}}{d\log L} = (\mathbb{E}_{\lambda^{opt}}[\delta_\theta^{opt}] - 1) d\log L$ using the policy that implements the first-best from Appendix G.

Appendix G Distance to Efficient Frontier

In this appendix, we characterize the distance to the efficient frontier, that is the amount of misallocation in the decentralized equilibrium compared to the first-best allocation. Note that this is different from how the distance to the efficient frontier changes with market size, which we analyze in Appendix F.2.

In Appendix G.1, we characterize the policy that implements the first-best equilibrium and use it to compute the distance to the frontier in our calibration. In Appendix G.2, we provide an analytical second-order approximation of the distance to the frontier around the CES benchmark, which decomposes the contributions of the different margins of inefficiency to overall misallocation.

G.1 Optimal Policy and Quantitative Results

Suppose there is a social planner who can implement the efficient allocation by regulating markups and imposing sales taxes. Theorem 1 from Baqaee and Farhi (2020) shows that the planner can implement the first-best by setting markups according to the consumer surplus each firm generates $\mu_\theta^{opt} = \delta_\theta$ and setting sales taxes to be the reciprocal of markups $\tau_\theta^{opt} = 1/\mu_\theta$. Intuitively, the markups provide socially optimal incentives along the extensive margin, and the output taxes undo the inefficiencies brought about by dispersed markups.[44]

We numerically implement the first-best policy in order to compute the distance to the efficient frontier for our calibrated model. The results in Table G.1 provide exact results (they are not calculated using the approximation in Proposition 6 below) for the distance to the frontier under the boundary conditions used in the main text, for both for the case with heterogeneous firms and for the case with homogeneous firms.

	Efficient selection $\bar{\delta} = \delta_{\theta^*}$	Efficient entry $\bar{\delta} = \bar{\mu}$
Heterogeneous firms	0.059	0.072
Homogeneous firms	0.022	0.000

Table G.1: Distance to the efficient frontier $\log(Y^{opt}/Y)$.

With heterogeneous firms, we find that the distance to the efficient frontier is around 6–7%. While these numbers are sizable, one might think that they are not large enough. Indeed, in Section 7, we saw in the decentralized equilibrium, cumulated changes in allocative efficiency are large relative to cumulated changes in technical efficiency even for large increases in

[44] See Edmond et al. (2018) for an alternative implementation of the optimal allocation using taxes. Bilbiie et al. (2019) also consider related issues in a dynamic context.

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.[45]

With homogeneous firms, the distance to the frontier is zero when $\delta = \mu$ since in that case entry, which is the only margin that can be distorted, is efficient. Otherwise the distance to the frontier is somewhat smaller than with heterogeneous firms. 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.

G.2 Analytical Second-Order Approximation

In this subsection, we provide an analytical expression for the social costs of the distortions caused by monopolistic competition around the efficient CES benchmark. As we show below, the proof of this result makes use of the optimal policy described in Appendix G.1.

We index the demand system $s_{\theta,t}$ by some parameter t, where $t = 0$ gives a CES form for $s_\theta(\cdot)$, and moving from $t = 0$ perturbs the residual expenditure function away from CES 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 6 (Distance to Frontier)**.** *The difference between welfare at the first-best allocation and the decentralized equilibrium can be approximated around $t = 0$ by*

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

where the remainder term is order t^3.

The first term captures distortions in the relative sizes of existing firms. It is related to heterogeneity in markups μ_θ and is also increasing in the average consumer surplus ratio $\mathbb{E}_\lambda[\delta_\theta]$.

[45] This discussion goes back to our definition of changes in allocative efficiency as the changes in welfare that arise from the reallocation of resources as opposed to the change in the distance to the efficient frontier discussed in Footnote 24.

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

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

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

Proof of Proposition 6. Denote welfare at the efficient frontier Y^{opt} and for a given t denote welfare at the decentralized equilibrium $Y(t)$. For some infinitesimal dt, the distance to the efficient frontier $\mathcal{L}$ to a second order is

$$\begin{aligned}\mathcal{L} = \log Y^{opt} - \log Y(dt) &\approx \log Y(0) - \left[\log Y(0) + \left.\frac{d\log Y}{\partial t}\right|_{t=0} dt + \frac{1}{2}\left.\frac{d^2\log Y}{dt^2}\right|_{t=0}(dt)^2\right] \\ &= -\frac{1}{2}\left.\frac{d^2\log Y}{dt^2}\right|_{t=0}(dt)^2 \\ &\approx -\frac{1}{2}\left.\frac{d\log Y}{dt}\right|_{t=dt} dt.\end{aligned}$$

The second line uses the fact that, by the Envelope Theorem, the first derivative of Y with respect to t at the efficient point is equal to zero. The third line uses a first-order expansion of $\frac{d\log Y}{dt}|_{t=dt}$ to substitute for $\frac{d^2\log Y}{dt^2}|_{t=0}dt$. Intuitively, we can take the first-order effect of moving toward the efficient frontier at the decentralized equilibrium and divide by two, since we know the derivative once we reach the efficient point is zero and the average of two first-order approximations yields a second-order approximation.

To get the derivative $\frac{d\log Y}{dt}$, we use the fact that the distance to the frontier is given by integrating changes in welfare from the decentralized equilibrium at t (given by markups μ_θ in the decentralized equilibrium and sales taxes $\tau_\theta = 1$) to the efficient allocation (which can be implemented using $\mu^{opt}_\theta = \delta_\theta$ and $\tau^{opt}_\theta = 1/\mu_\theta$):

$$\log\frac{Y^{opt}}{Y(t)} = \int_{(\mu_\theta(t),1)}^{(\delta_\theta(t),1/\delta_\theta(t))}\left[\frac{\partial\log Y}{\partial\log\mu_\theta}\frac{\partial\log\mu_\theta}{\partial\nu} + \frac{\partial\log Y}{\partial\log\tau_\theta}\frac{\partial\log\tau_\theta}{\partial\nu}\right]d\nu,$$

where $d\nu$ increments changes in the policy from $(\mu_\theta(t), 1)$ to $(\delta_\theta(t), 1/\delta_\theta(t))$.

Taking the derivative with respect to t and applying the Envelope Theorem yields

$$\frac{d\log Y(t)}{dt} = \left(\frac{\partial \log Y}{\partial \log \mu_\theta}\frac{d\log \mu_\theta}{dt} + \frac{\partial \log Y}{\partial \log \tau_\theta}\frac{d\log \tau_\theta}{dt}\right)\Bigg|_{(\mu_\theta(t),1)}$$

Hence, we proceed in two steps. In the first step, we rewrite our system of equilibrium equations allowing for exogenous markups and sales taxes, and log-linearize these equations to get how welfare responds to changes in markups and sales taxes. Second, we specialize these equations to the decentralized equilibrium and apply changes in markups and taxes toward the efficient point (i.e., $-\frac{d\log \mu_\theta}{dt}$ and $-\frac{d\log \tau_\theta}{dt}$) to get $\frac{d\log Y}{dt}$. We use this to solve for the distance to the efficient frontier $\mathcal{L}$.

Step 1:

We rewrite our system of equilibrium equations allowing for these sales taxes and exogenous markups. Now, the equilibrium equations include our definition of the labor share,

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

the implicit definition of the price aggregator P,

$$M\int_{\theta^*}^1 s(\frac{p_\theta}{P})g(\theta)d\theta = 1,$$

demand curves per variety,

$$y_\theta = \frac{1}{p_\theta \Lambda_L}s(\frac{p_\theta}{P}),$$

prices,

$$p_\theta = \frac{\mu_\theta \tau_\theta}{A_\theta},$$

our definition of welfare,

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

sales shares,

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

the selection condition,

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

and the free entry condition,

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

The log-linearized system of equations, allowing for exogenous changes in markups $d\log\mu_\theta$ and sales taxes $d\log\tau_\theta$, is

$$\begin{aligned}
d\log\Lambda_L &= \left(1-\frac{\lambda_{\theta^*}}{\tau_{\theta^*}\Lambda_L}\right)\frac{g(\theta^*)}{1-G(\theta^*)}d\theta^* + \mathbb{E}_\lambda\left[\frac{1}{\tau_\theta\Lambda_L}\left(d\log\lambda_\theta - d\log\tau_\theta\right)\right].\\
0 &= d\log M - \lambda_{\theta^*}\frac{g(\theta^*)}{1-G(\theta^*)}d\theta^* + \mathbb{E}_\lambda\left[\left(1-\sigma_\theta\right)d\log(\frac{p_\theta}{P})\right].\\
d\log y_\theta &= -d\log\Lambda_L - \sigma_\theta d\log\frac{p_\theta}{P} - d\log P.\\
d\log p_\theta &= d\log\mu_\theta + d\log\tau_\theta.\\
d\log Y &= \left(\mathbb{E}_\lambda[\delta_\theta]-1\right)d\log M - \lambda_{\theta^*}\left(\delta_{\theta^*}-1\right)\frac{g(\theta^*)}{1-G(\theta^*)}d\theta^* - \mathbb{E}_\lambda\left[d\log p_\theta\right] - d\log\Lambda_L.\\
d\log\lambda_\theta &= d\log p_\theta + d\log y_\theta + d\log\Lambda_L + \frac{-g(\theta^*)}{(1-G(\theta^*))}d\theta^* + d\log M.\\
0 &= \frac{1}{\mu_{\theta^*}-1}d\log\mu_{\theta^*} + d\log\lambda_{\theta^*} - d\log\tau_{\theta^*} + \left[\frac{\partial\log\left(1-\frac{1}{\mu_\theta}\right)\frac{\lambda_\theta}{\tau_\theta}\frac{1}{f_{o,\theta}}}{\partial\theta}\right]d\theta^*\\
&\quad + \frac{g(\theta^*)}{(1-G(\theta^*))}d\theta^* - d\log\Lambda_L - d\log M + d\log L.\\
0 &= d\log L + \mathbb{E}_{\lambda(1-1/\mu)(1/\tau)}\left[\frac{1}{\mu_\theta-1}d\log\mu_\theta - d\log\tau_\theta + d\log p_\theta + d\log y_\theta\right].
\end{aligned}$$

Step 2:

In the second step, we apply these formulas at the decentralized equilibrium, where $\tau_\theta = 0$ and $\mu_\theta = \sigma_\theta/(\sigma_\theta - 1)$. We then apply changes in markups and taxes toward the efficient point.

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

$$\begin{aligned}
d\log\Lambda_L &= \left(1-\lambda_{\theta^*}\right)\frac{g(\theta^*)}{1-G(\theta^*)}d\theta^* + \mathbb{E}_\lambda\left[\left(d\log\lambda_\theta - d\log\tau_\theta\right)\right]\\
0 &= d\log M - \lambda_{\theta^*}\frac{g(\theta^*)}{1-G(\theta^*)}d\theta^* + \mathbb{E}_\lambda\left[\left(1-\sigma_\theta\right)d\log(\frac{p_\theta}{P})\right].\\
d\log y_\theta &= -d\log\Lambda_L - \sigma_\theta d\log\frac{p_\theta}{P} - d\log P.\\
d\log p_\theta &= d\log\mu_\theta + d\log\tau_\theta.
\end{aligned}$$

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

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

$$0 = \frac{1}{\mu_{\theta^*} - 1}d\log\mu_{\theta^*} + d\log\lambda_{\theta^*} - d\log\tau_{\theta^*} + \frac{1}{\gamma_{\theta^*}}\frac{g(\theta^*)}{1 - G(\theta^*)}d\theta^* + \frac{g(\theta^*)}{(1 - G(\theta^*))}d\theta^* - d\log\Lambda_L - d\log M.$$

$$0 = \mathbb{E}_{\lambda(1-1/\mu)}\left[\frac{1}{\mu_\theta - 1}d\log\mu_\theta - d\log\tau_\theta + d\log p_\theta + d\log y_\theta\right] = 0.$$

Solving the fixed point yields,

$$\begin{aligned} d\log\Lambda_L &= -\mathbb{E}_\lambda\left[d\log\tau_\theta\right], \\ d\log P &= \mathbb{E}_\lambda\left[d\log\tau_\theta\right], \\ \frac{g(\theta^*)}{1 - G(\theta^*)}d\theta^* &= \sigma_{\theta^*}\gamma_{\theta^*}\left(d\log\tau_{\theta^*} - \mathbb{E}_\lambda\left[d\log\tau_\theta\right]\right), \end{aligned}$$

and

$$\begin{aligned} d\log Y = {} & \left(\mathbb{E}_\lambda\left[\delta_\theta\right] - \delta_{\theta^*}\right)\lambda_{\theta^*}\gamma_{\theta^*}\sigma_{\theta^*}\left[d\log\tau_{\theta^*} - \mathbb{E}_\lambda\left[d\log\tau_\theta\right]\right] \\ & + \left(\mathbb{E}_\lambda\left[\delta_\theta\right] - 1\right)\mathbb{E}_\lambda\left[\left[\sigma_\theta - \mathbb{E}_\lambda\left[\sigma_\theta\right]\right]\left(d\log\tau_\theta + d\log\mu_\theta\right)\right] \\ & + \mathbb{E}_\lambda\left[\left[\left(\mathbb{E}_\lambda\left[\delta_\theta\right] - 1\right)\mathbb{E}_\lambda\left[\sigma_\theta - 1\right] - 1\right]d\log\mu_\theta\right]. \end{aligned}$$

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

$$\begin{aligned} d\log\mu_\theta &= \log(\frac{\delta_\theta}{\mu_\theta}), \\ d\log\tau_\theta &= -\log\delta_\theta, \\ d\log(\mu_\theta\tau_\theta) &= -\log\mu_\theta. \end{aligned}$$

Near the efficient point, we can also use the approximations,

$$\begin{aligned} \log(\frac{\delta_\theta}{\mu_\theta}) &\approx \frac{\delta_\theta}{\mu_\theta} - 1, \\ \log(\frac{\delta_\theta}{\delta_{\theta^*}}) &\approx \frac{\delta_\theta}{\delta_{\theta^*}} - 1. \end{aligned}$$

Plugging into welfare, we get

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

Expanding the last term,

$$\mathbb{E}_\lambda\left[\sigma_\theta\left(\frac{\mathbb{E}_\lambda\left[\delta_\theta\right]}{\mu_\theta} - 1\right)\right]\mathbb{E}_\lambda\left[\frac{\delta_\theta}{\mu_\theta} - 1\right] = \left(\mathbb{E}_\lambda\left[\delta_\theta\right] Cov_\lambda\left[\sigma_\theta, \frac{1}{\mu_\theta}\right] + \mathbb{E}_\lambda\left[\sigma_\theta\right]\left(\frac{\mathbb{E}_\lambda\left[\delta_\theta\right]}{\bar{\mu}} - 1\right)\right)\left(Cov_\lambda\left[\delta_\theta, \frac{1}{\mu_\theta}\right] + \frac{\mathbb{E}_\lambda\left[\delta_\theta\right]}{\bar{\mu}} - 1\right)$$

Note that the term $\left(\frac{\mathbb{E}_\lambda[\delta_\theta]}{\bar{\mu}} - 1\right)$ is order t and both covariances $Cov_\lambda\left[\sigma_\theta, \frac{1}{\mu_\theta}\right]$ and $Cov_\lambda\left[\delta_\theta, \frac{1}{\mu_\theta}\right]$ are order t^2. Since we are interested in a second-order approximation in t, we drop terms of order t^3 or higher to get:

$$\mathbb{E}_\lambda\left[\sigma_\theta\left(\frac{\mathbb{E}_\lambda\left[\delta_\theta\right]}{\mu_\theta} - 1\right)\right]\mathbb{E}_\lambda\left[\frac{\delta_\theta}{\mu_\theta} - 1\right] \approx \mathbb{E}_\lambda\left[\sigma_\theta\right]\left(\frac{\mathbb{E}_\lambda\left[\delta_\theta\right]}{\bar{\mu}} - 1\right)^2.$$

Hence,

$$d\log Y = \left(\mathbb{E}_\lambda\left[\delta_\theta\right] - \delta_{\theta^*}\right)^2 \lambda_{\theta^*}\gamma_{\theta^*}\frac{\sigma_{\theta^*}}{\delta_{\theta^*}} + \left(\mathbb{E}_\lambda\left[\delta_\theta\right] - 1\right) Cov_\lambda\left[\sigma_\theta, -\log\mu_\theta\right] + \mathbb{E}_\lambda\left[\sigma_\theta\right]\left(\frac{\mathbb{E}_\lambda\left[\delta_\theta\right]}{\bar{\mu}} - 1\right)^2.$$

Note that this expression is equal to $\frac{-d\log Y}{dt}$ from our above expressions, since we have applied changes in markups and sales taxes toward the efficient point (i.e., $-\frac{d\log\mu_\theta}{dt}$ and $-\frac{d\log\tau_\theta}{dt}$).

Finally, the loss function encapsulating the distance to the efficient frontier is

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

Plugging in our expression for $d\log Y$ above and rearranging yields Proposition 6. ■

Appendix H Shocks to Entry and Overhead Costs

In this appendix, we characterize comparative statics with respect to shocks to the fixed entry and overhead costs. For simplicity, we consider the case where overhead costs are identical across firms, $f_{o,\theta} = f_o$. Proposition 7 characterizes the response of welfare to a change in fixed costs of entry and overhead costs.

Proposition 7. *In response to changes in fixed costs of entry $d\log f_e$ and fixed overhead costs $d\log f_o$, changes in consumer welfare are given by*

$$d\log Y = -\underbrace{\left(\mathbb{E}_\lambda\left[\delta_\theta\right]-1\right)\left[\psi_e d\log\left(\Delta f_e\right)+\left(1-\psi_e\right)d\log f_o\right]}_{\textit{technical efficiency}}$$
$$-\underbrace{\left(\xi^{\epsilon}+\xi^{\theta^*}+\xi^{\mu}\right)\bar{\mu}\left[\psi_e d\log\left(\Delta f_e\right)+\left(1-\psi_e\right)d\log f_o\right]}_{\textit{allocative efficiency}}$$
$$+\underbrace{\lambda_{\theta^*}\gamma_{\theta^*}\left(\mathbb{E}_\lambda\left[\delta_\theta\right]-\delta_{\theta^*}\right)\psi_e(d\log f_o - d\log(\Delta f_e))}_{\textit{allocative efficiency}}.$$

where ξ^{ϵ}, ξ^{θ^*}, and ξ^{μ} are given in Theorem 1 and $\psi_e = \Delta f_e/\left(\Delta f_e + (1-G(\theta^*))\,\mathbb{E}\left[f_{o,\theta}\right]\right)$ is the entry cost share of all fixed costs as defined in Proposition 3.

To understand these results, it is useful to observe that the free entry condition is homogeneous of degree one in fixed costs and population. This is because the fixed costs only matter to entering firms on a per capita basis, f_e/L and f_o/L. As a result, joint proportional reductions in fixed costs of entry and fixed overhead costs $d\log f_e = d\log f_o < 0$ have exactly the same effects on entry as equivalent increases in population $d\log L = -d\log f_e = -d\log f_o > 0$. Accordingly, the first two terms of Proposition 7 mirror the technical and allocative efficiency terms in Theorem 1.

The overhead cost plays an additional role in regulating the selection margin. As a result, the third term in Proposition 7 captures whether toughening selection increases or decreases allocative efficiency. If the overhead cost increases, the selection cutoff moves up. This improves welfare if the marginal firm provides less social value than the average firm ($\delta_{\theta^*} < \mathbb{E}_\lambda[\delta_\theta]$), and decreases welfare if the marginal firm provides more social value than the average.

The same intuitions carry over to the response of real GDP per capita to a change in fixed entry and overhead costs.

Proposition 8. *In response to changes in fixed costs of entry $d\log f_e$ and fixed overhead costs $d\log f_o$, changes in real GDP per capita are given by*

$$d\log Q = -\mathbb{E}_\lambda\left[\frac{1}{\sigma_\theta}\right]\mathbb{E}_\lambda\left[1-\rho_\theta\right]\bar{\mu}\left[\psi_e d\log\left(\Delta f_e\right)+\left(1-\psi_e\right)d\log f_o\right].$$

Appendix I Generalized Kimball (HDIA) Preferences

In this appendix, we develop a version of our results using an alternative demand system to the HSA preferences we use in the main text. We use a generalization of Kimball (1995) preferences, called homothetic demand with a direct implicit additivity (HDIA) preferences by Matsuyama and Ushchev (2017). The CES demand system is the only point of union between HDIA preferences and the HSA preferences used in the main text. Nevertheless, our theoretical and quantitative results are quite similar when we use HDIA preferences instead.

This appendix is organized as follows. In Section I.1, we set up the consumer and firm problems and describe firm elasticities, markups, pass-throughs, and consumer surplus ratios in terms of primitives. In Section I.2, we present theoretical results analogous to Theorem 1 and Proposition 2 in the main text. Finally, we show that the system of differential equations used to calibrate the model remain valid under HDIA preferences and provide quantitative results analogous to Table 1 and Table 2. The results are qualitatively and quantitatively similar to those in the main text.

I.1 Setup

Under HDIA preferences, per-capita welfare Y is defined implicitly by

$$\int_{\theta\in\Theta} \Upsilon_\theta(\frac{y_\theta}{Y})dF(\theta) = 1, \tag{30}$$

where y_θ is the per-capita consumption of variety θ, the function Υ_θ is increasing and concave with $\Upsilon_\theta(0) = 0$, the set Θ contains all potential varieties, and $dF(\theta)$ is the measure of varieties of type θ.

Consumers maximize their utility Y subject to the budget constraint

$$\int_{\theta\in\Theta} p_\theta y_\theta dF(\theta) = 1, \tag{31}$$

where p_θ is the price of variety θ. As in the main text, per-capita income is the numeraire.

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

$$\frac{p_\theta}{P} = \Upsilon'_\theta(\frac{y_\theta}{Y}), \tag{32}$$

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

$$P = \frac{\bar{\delta}}{Y}, \qquad \text{and} \qquad \frac{1}{\bar{\delta}} = \int_{\theta\in\Theta} \Upsilon'_\theta(\frac{y_\theta}{Y})\frac{y_\theta}{Y}dF(\theta). \tag{33}$$

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

$$\max_{\text{operate},p_\theta} \begin{cases} \left(p_\theta - \frac{1}{A_\theta}\right) L y_\theta - f_{o,\theta} & \text{if the firm operates} \\ 0 & \text{if the firm does not operate} \end{cases} \tag{34}$$

subject to the household per-capita demand curve in (32).

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

$$\mu_\theta(\frac{y}{Y}) = \frac{1}{1 - \frac{1}{\sigma_\theta(\frac{y}{Y})}}, \tag{35}$$

and the price-elasticity of demand is given by,

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

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

$$\left(p_{\theta^*} - \frac{1}{A_{\theta^*}}\right) L y_{\theta^*} = f_{o,\theta^*}, \tag{36}$$

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

$$\int_{\theta^*}^{1} \left[\left(1 - \frac{1}{\mu_\theta}\right) p_\theta y_\theta w L - f_{o,\theta}\right] g(\theta) d\theta = f_e. \tag{37}$$

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

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

$$\rho_\theta(\frac{y}{Y}) = \frac{1}{1 + \frac{\frac{y}{Y}\mu'_\theta(\frac{y}{Y})}{\mu_\theta(\frac{y}{Y})}\sigma_\theta(\frac{y}{Y})}, \qquad \text{and} \qquad \delta_\theta(\frac{y}{Y}) = \frac{\bar{\delta}\Upsilon_\theta(\frac{y}{Y})}{p_\theta y_\theta} = \frac{\Upsilon_\theta(\frac{y}{Y})}{\frac{y}{Y}\Upsilon'_\theta(\frac{y}{Y})}.$$

Note that by integrating the equation for δ_θ, we can show that the demand index in (33) is simply the sales-weighted average of the consumer surplus ratio, $\bar{\delta} = \mathbb{E}_\lambda[\delta_\theta]$.

The sales density is defined as $\lambda_\theta = (1 - G(\theta^*))Mp_\theta y_\theta$. We again denote the harmonic (sales-weighted) average of markups $\bar{\mu} = \mathbb{E}_\lambda[\mu_\theta^{-1}]^{-1}$.

In equilibrium, consumers maximize utility, firms maximize profits, and resource constraints are satisfied. The equilibrium is defined by the implicit definition of welfare (30), the consumer's demand for each variety (32), the household budget constraint (31), firms' profit-maximizing markups (35), the selection cutoff (36), and the free entry condition (37).

I.2 Response to Change in Market Size

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

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

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

where

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

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

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

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

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

where $d\log Y$ is given by Theorem 2.

Proof of Theorem 2 and Proposition 9. In response to an exogenous change in market size $d\log L$, the following system of log-linearized equations describe the movements of all endogenous variables.

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

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

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

I.2.1 Conditions for a Locally Unique Equilibrium

One implication of Theorem 2 is that the change in welfare is not well defined under HDIA preferences if $\xi^\epsilon + \xi^{\theta^*} + \xi^\mu = 1$. Given this concern, in this subsection, we develop conditions under which the model equilibrium exists and is locally unique, so that the comparative static with respect to market size is well defined.

We first begin with a definition of a feasible set of statistics (sales densities, consumer surplus ratios, markups, pass-throughs, variable profit to overhead cost ratios, and selection

cutoff), then show that a condition on these statistics is sufficient to prove that the equilibrium exists and is locally unique (Proposition 10). Finally, we provide a set of simpler (but stricter) sufficient conditions that guarantee existence and local uniqueness (Corollary 7).

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

1. $\int_{\theta\in\Theta} \lambda_\theta d\theta = 1$ and $\lambda_\theta \geq 0$ for all θ,
2. $\delta_\theta, \mu_\theta \geq 1$ for all θ,
3. $\rho_\theta \geq 0$ for all θ,[46]
4. $X_\theta \geq 0$ and $\frac{\partial \log X_\theta}{\partial \theta} > 0$ for all θ, and
5. $X_{\theta^*} = 0$.

Proposition 10 (Existence and Local Uniqueness)**.** *For any feasible* $\left\{\{\lambda_\theta, \delta_\theta, \mu_\theta, \rho_\theta, X_\theta\}_{\theta\in\Theta}, \theta^*\right\}$*, the equilibrium exists and is locally unique if*

$$-(\mathbb{E}_\lambda[\delta_\theta] - 1) < \xi^\epsilon + \xi^{\theta^*} + \xi^\mu < 1,$$

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

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

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

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

[46]The reader may note that assumption (3) is also sufficient for marginal revenue curves to be downward-sloping.

$$X_\theta = \frac{\lambda_\theta}{f_{o,\theta}}\left(1 - \frac{1}{\mu_\theta}\right).$$

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

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

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

and

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

The requirement $-(\mathbb{E}_\lambda[\delta_\theta] - 1) < \xi^\epsilon + \xi^{\theta^*} + \xi^\mu < 1$ ensures both conditions are met. ■

Corollary 7 lists three stricter conditions that are sufficient (but not necessary) to ensure that the condition on $\left\{\{\lambda_\theta, \delta_\theta, \mu_\theta, \rho_\theta, X_\theta\}_{\theta\in\Theta}, \theta^*\right\}$ from Proposition 10 is met.

Corollary 7 (Sufficient Conditions for Existence and Local Uniqueness)**.** *For a feasible* $\left\{\{\lambda_\theta, \delta_\theta, \mu_\theta, \rho_\theta, X_\theta\}_{\theta\in\Theta}, \theta^*\right\}$*, sufficient conditions for the equilibrium to exist and be locally unique are:*

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

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

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

We can bound the left-hand side:

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

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

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

Again, we can bound the left-hand side:

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

where the second line applies the Cauchy-Schwarz inequality and the third line applies Popoviciu's inequality.[47] Hence, we have $\xi^\epsilon+\xi^{\theta^*}+\xi^\mu<1$ if

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

which is satisfied under assumption (2). For context, in the efficient entry case where $\mathbb{E}_\lambda\left[\delta_\theta\right]=1.09$, assumption (2) implies the price elasticity of demand is at most $\sigma^{\max}=45$. For the efficient selection case where $\mathbb{E}_\lambda\left[\delta_\theta\right]=1.033$, the price elasticity of demand is at most $\sigma^{\max}=122$ (i.e., the lowest markup of any firm is 1.008).

The condition that $\xi^\epsilon+\xi^{\theta^*}+\xi^\mu>-(\mathbb{E}_\lambda[\delta_\theta]-1)$ can be rewritten as:

$$\left(\mathbb{E}_\lambda[\delta_\theta]-\delta_{\theta^*}\right)\lambda_{\theta^*}\gamma_{\theta^*}\left(\sigma_{\theta^*}-\mathbb{E}_\lambda\left[\sigma_\theta^{-1}\right]^{-1}\right)+\mathbb{E}_\lambda\left[\rho_\theta\left(\sigma_\theta-1\right)\right]\left(\mathbb{E}_\lambda[\delta_\theta]-1\right)-\mathbb{E}_\lambda\left[\rho_\theta\right]+\mathbb{E}_\lambda[\delta_\theta]>0.$$

Since $\rho_\theta\in[0,1]$ (assumption (1) and feasibility), a sufficient condition is that

$$\left(\mathbb{E}_\lambda[\delta_\theta]-\delta_{\theta^*}\right)\lambda_{\theta^*}\gamma_{\theta^*}\left(\sigma_{\theta^*}-\mathbb{E}_\lambda\left[\sigma_\theta^{-1}\right]^{-1}\right)+\left(\mathbb{E}_\lambda[\delta_\theta]-1\right)>0,$$

or

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

We can bound the right-hand side,

$$\text{RHS}=\mathbb{E}_\lambda\left[\delta_\theta\right]+\frac{1}{\lambda_{\theta^*}\gamma_{\theta^*}}\frac{\mathbb{E}_\lambda\left[\delta_\theta\right]-1}{\sigma_{\theta^*}-\mathbb{E}_\lambda\left[\sigma_\theta^{-1}\right]^{-1}}$$

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

$$
\begin{aligned}
&> \mathbb{E}_\lambda[\delta_\theta] + \frac{1}{\lambda_{\theta^*}\gamma_{\theta^*}} \frac{\mathbb{E}_\lambda[\delta_\theta] - 1}{\sigma^{\max} - 1} \\
&\geq \mathbb{E}_\lambda[\delta_\theta] + \frac{(\mathbb{E}_\lambda[\delta_\theta] - 1)^2}{4\lambda_{\theta^*}\gamma_{\theta^*}}.
\end{aligned}
$$

When the sales-share of the cutoff firm λ_{θ^*} is small, as it is in our quantitative application, this can be large upper bound for δ_{θ^*}.

■

I.3 Calibration

For calibration, we impose the restriction that the aggregator function is identical across types, $\Upsilon_\theta = \Upsilon$.[48] We also assume that overhead costs are homogenous across firms, $f_{o,\theta} = f_o$, so that the sole source of exogenous variation across firm types is due to differing productivities A_θ. Under this restriction, we can use the cross-sectional variation in pass-throughs and sales shares to solve for markups and consumer surplus ratios, up to boundary conditions.

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

$$
\begin{aligned}
\frac{d\log\mu_\theta}{d\theta} &= (1 - \rho_\theta)\frac{d\log A_\theta}{d\theta}, \\
\frac{d\log\lambda_\theta}{d\theta} &= \frac{\rho_\theta}{\mu_\theta - 1}\frac{d\log A_\theta}{d\theta}.
\end{aligned}
$$

Rearranging yields the differential equation,

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

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

For consumer surplus ratios, recall that we can write

$$
\delta_\theta(\frac{y}{Y}) = \frac{\Upsilon_\theta(\frac{y}{Y})}{\frac{y}{Y}\Upsilon'_\theta(\frac{y}{Y})}.
$$

Differentiating both sides and rearranging, we find a differential equation relating consumer

[48] We could easily allow for the aggregator Υ_θ to differ across types according to a multiplicative demand shifter, as in the calibration in the main text (see (20)).

surplus ratios to markups,

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

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

Table I.1 shows the elasticity of welfare and real GDP per capita to market size. The elasticity of welfare to market size is further decomposed into changes in technical and allocative efficiency, including the three margins of adjustment (entry, exist, and markups) discussed in the main text. The results are quantitatively similar to those in the main text. In particular, the majority of gains from an increase in market size are due allocative efficiency effects arising from entry; the selection and pro-competitive channels have zero or mildly deleterious effects on welfare.

	Efficient selection $\bar{\delta} = \delta_{\theta^*}$	Efficient entry $\bar{\delta} = \bar{\mu}$
Welfare: $d \log Y$	0.303	0.317
Technical efficiency: $d \log Y^{tech}$	0.033	0.090
Allocative efficiency: $d \log Y^{alloc}$	0.269	0.227
Darwinian effect: $d \log Y^{\epsilon} - d \log Y^{tech}$	0.284	1.500
Selection effect: $d \log Y^{\epsilon,\theta^*} - d \log Y^{\epsilon}$	0.000	-1.110
Pro-competitive effect: $d \log Y^{\epsilon,\theta^*,\mu} - d \log Y^{\epsilon,\theta^*}$	-0.015	-0.162
Real GDP per capita	0.052	0.053

Table I.1: The elasticity of welfare and real GDP per capita to population following Theorem 2 and Proposition 9.

Table I.2 replicates the analysis in a setting with homogeneous firms. Again, firm heterogeneity appears to play a significant role. Without heterogeneity, we find that the elasticity of welfare to changes in market size are much smaller than in the calibration with heterogeneous firms.

	$\delta = \delta_{\theta^*}$	$\delta = \mu$
Welfare: $d \log Y$	0.060	0.090
Technical efficiency: $d \log Y^{tech}$	0.033	0.090
Allocative efficiency: $d \log Y^{alloc}$	0.027	0.000
Real GDP per capita	0.043	0.043

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

Appendix J Chaney (2008) Entry

For the model in the main text, our entry technology follows Melitz (2003) and hence Hopenhayn (1992) entry. That is, there is an unbounded mass of potential entrants, and entry occurs until the expected profits from entering the market are equal to the fixed costs of entry. In this extension, we instead follow the entry technology from Chaney (2008), in which there is a fixed mass of potential entrants that is proportional to the size of the market. We show that an increase in market size continues to generate Darwinian reallocations in this version of the model. Interestingly, the Darwinian effect on welfare also shows up in real GDP, in contrast to the model in the main text.

As in Chaney (2008), the mass of potential entrants is assumed to be proportional to market size, so that

$$M = cL,$$

where c is some constant. The types of the M potential entrants is given by the CDF $G(\theta)$, and each potential entrant perfectly observes its productivity (A_θ), demand curve (given by $s_\theta(\cdot)$), and overhead cost ($f_{o,\theta}$) before deciding whether to produce. As in the main text, we order types θ to be increasing in the ratio of variable profits to overhead costs X_θ.

Many of the remaining equilibrium conditions remain identical to our baseline model, but it is worth first stressing differences in the entry/selection margin and in household income.

The selection margin—which now dictates which firms decide to enter and operate in the market—is characterized by a cutoff firm type θ^* such that

$$\left(1 - \frac{1}{\mu_{\theta^*}}\right) p_{\theta^*} y_{\theta^*} L - f_{o,\theta^*} = 0,$$

all firms with type $\theta \geq \theta^*$ enter and stay in the market, and all potential entrants with $\theta < \theta^*$ decide not to enter.

Since we no longer have free entry, firms will now make positive profits, and households will receive both labor income and dividend income. Income per capita I is given by

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

We now consider how welfare changes in response to a change in market size. Following (9), the change in welfare can be characterized by

$$d \log Y = \underbrace{\left(\mathbb{E}_\lambda \left[\delta_\theta\right] - 1\right) d \log M}_{\text{New varieties}} - \underbrace{\lambda_{\theta^*} \left(\delta_{\theta^*} - 1\right) \frac{g(\theta^*)}{1 - G(\theta^*)} d\theta^*}_{\text{Entry/exit at cutoff}} + \underbrace{\mathbb{E}_\lambda \left[d \log I - d \log p_\theta\right]}_{\text{Prices (relative to income)}}.$$

where we have amended the last term of (9) to no includes changes in household income. (We use the nominal wage as the numeraire, $w = 1$). To provide intuition, we start by simplifying this expression using our entry technology and the selection condition. From the assumption that the mass of potential entrants is proportional to market size ($M = cL$), we have:

$$d \log M = d \log L.$$

By log-linearizing the selection condition, we get that the change in the selection cutoff follows

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

Intuitively, increases in market size L and increases in per-capita income I both mean that fixed costs can be spread over a greater number of units sold, which makes it easier for the marginal firm to survive and decreases the selection cutoff. Similarly, increases in P (softening competition) also decrease the selection cutoff.

Substituting these into our expression for welfare, we get

$$d \log Y = \underbrace{\left(\mathbb{E}_\lambda \left[\delta_\theta \right] - 1 \right) d \log L}_{\text{New varieties}} + \underbrace{\lambda_{\theta^*} \gamma_{\theta^*} \left(\delta_{\theta^*} - 1 \right) \left[(\sigma_{\theta^*} - 1)\, d \log P + d \log L + d \log I \right]}_{\text{Entry/exit at cutoff}}$$
$$+ \underbrace{d \log I - \mathbb{E}_\lambda \left[1 - \rho_\theta \right] d \log P}_{\text{Prices (relative to income)}}.$$

Note that improvements in allocative efficiency that free up labor no longer generate new varieties, as they did in our baseline model, since now the mass of potential varieties is fixed. Hence, all allocative efficiency gains will show up in the latter two terms, either by admitting more entry at the selection cutoff or by decreasing prices relative to income.

Solving the general case produces a somewhat cumbersome expression, but we can gain intuition about how the Darwinian effect shows up in this version of the model by considering an example with zero overhead costs and constant-price-elasticity preferences (identical to those considered in Corollary 1):

$$s_\theta(x) = x^{1 - \sigma_\theta}.$$

Then, the change in the price aggregator and in per-capita income are given by,

$$\mathbb{E}_\lambda \left[\sigma_\theta - 1 \right] d \log P = -d \log L,$$

$$d \log I = \mathbb{E}_\lambda \left[\frac{1}{\mu_\theta} \right] d \log P + \mathbb{E}_\lambda \left[\frac{1}{\sigma_\theta} \right] d \log L$$

$$= \left(\mathbb{E}_\lambda \left[\frac{1}{\sigma_\theta} \right] - \mathbb{E}_\lambda \left[\frac{1}{\mu_\theta} \right] \frac{1}{\mathbb{E}_\lambda \left[\sigma_\theta - 1 \right]} \right) d \log L$$
$$= \frac{1}{\mathbb{E}_\lambda \left[\sigma_\theta - 1 \right]} Cov_\lambda \left[\sigma_\theta, \frac{1}{\mu_\theta} \right] d \log L.$$

So, the change in welfare following an increase in market size is:

$$d \log Y = \underbrace{\left(\mathbb{E}_\lambda \left[\delta_\theta \right] - 1 \right) d \log L}_{\text{Technical efficiency}} + \underbrace{\frac{1}{\mathbb{E}_\lambda \left[\sigma_\theta - 1 \right]} Cov_\lambda \left[\sigma_\theta, \frac{1}{\mu_\theta} \right] d \log L}_{\text{Allocative efficiency}}.$$

Just as in Corollary 1, we have silenced the selection and pro-competitive effects, and we are left with an allocative efficiency gain that is solely due to heterogeneities in price elasticities and markups, namely, the Darwinian effect. As in the model in the main text, the Darwinian effect is unambiguously positive as long as there is non-trivial firm heterogeneity, and occurs because a greater mass of entrants results in a reallocation toward high-markup firms, thereby increasing allocative efficiency and (in this version of the model) decreasing prices relative to income.

It is interesting to note that in this version of the model, the Darwinian effect will also show up in real GDP (while it was absent from real GDP in the model in the main text). The change in real GDP as measured by statistical agencies in this version of the model is

$$d \log Q = \frac{1}{\mathbb{E}_\lambda \left[\sigma_\theta - 1 \right]} Cov_\lambda \left[\sigma_\theta, \frac{1}{\mu_\theta} \right] d \log L.$$

The reason that the Darwinian effect does not show in real GDP in our baseline model is that all the freed up labor from the reallocation to high-markup firms is funneled into new variety creation, and the welfare gains from new variety creation cannot be captured by measuring price changes for continuing varieties. However, in this version of the model, the Darwinian effect instead shows up as increased income, which can be measured.

Appendix K Oligopoly Extension

In this appendix, we consider an extension of the model in which the increased entry from a market expansion may erode firms' oligopoly power. We do so by considering a continuum of "product lines," each of which contains several competing firms. We show that when firm entry generates new product lines, an increase in market size leads to improvements in technical efficiency due to the creation of new product lines and allocative efficiency due to the Darwinian effect, as in our baseline results. However, when firm entry is focused instead on generating more competitors per product line, welfare gains come from the reduction in markups caused by the reduction in oligopoly power.

There is a continuum of product lines with types indexed by $\theta \in [0,1]$. The HSA price aggregator P across product lines is determined implicitly by

$$\int_0^1 s_\theta(\frac{p_\theta}{P}) V g(\theta) d\theta = 1,$$

where V is the mass of product lines, p_θ is the price of product line θ, and the density of product lines is given by $g(\theta)$. Within each product line, $N \geq 1$ identical firms engage in Cournot competition, with firm i in product line θ producing per-capita output $y_{i,\theta}$ and setting price $p_{i,\theta}$. Without loss of generality, we normalize the productivity of all firms to one. The outputs of firms within a product line are perfect substitutes, so that the total (per-capita) product line output y_θ is simply the sum of all firms' outputs:

$$y_\theta = \sum_{i \in \mathcal{I}(\theta)} y_{i,\theta} = N y_{i,\theta},$$

where $I(\theta)$ denotes the set of firms in product line θ.

As in Atkeson and Burstein (2008), each firm internalizes the impact of its price on the price index of its product line (p_θ) and on total product line demand (y_θ), but takes households' aggregate spending and the cross-product line price aggregator P as given.

We use σ_θ to denote the elasticity of total product line consumption to the product line price, and use ρ_θ to describe the curvature of demand facing the total product line, so that

$$\sigma_\theta = -\frac{d \log y_\theta}{d \log p_\theta} = 1 - \frac{\frac{p_\theta}{P} s_\theta'\left(\frac{p_\theta}{P}\right)}{s_\theta\left(\frac{p_\theta}{P}\right)}, \qquad \text{and} \qquad \frac{d \log \sigma_\theta}{d \log(\frac{p_\theta}{P})} = (\sigma_\theta - 1)\frac{1-\rho_\theta}{\rho_\theta}.$$

We assume overhead costs are zero, therefore shutting down any selection margin in this extension. The free entry condition thus requires that a firm's expected profits upon entry equal the fixed cost (where the expectation comes from the fact that a firm does not realize its

product line type θ until after paying the fixed cost):

$$\int_0^1 \left(1 - \frac{1}{\mu_{i,\theta}}\right) p_{i,\theta} y_{i,\theta} L g(\theta) d\theta = f_e.$$

To close the model, we need to describe how the number of firms per product line (N) and the mass of product lines (V) scale with an increase in entry. Since the total mass of firms in the economy is $M = NV$, we allow the rate at which new firm creation results in new product line creation versus heightened competition within product lines to be governed by a parameter α, so that

$$N = M^{\alpha}, \qquad \text{and} \qquad V = M^{1-\alpha}.$$

If $\alpha = 0$, all new entry results in the creation of new product line varieties and no increase in within-product line competition. Setting $\alpha = 0$ and $N = 1$ thus generates a special case of our baseline model (special case, because overhead costs are zero). On the other hand, if $\alpha = 1$, all new entry results in heightened competition within existing product lines, and no new varieties are created.[49]

The solution of the firm's profit maximization problem is

$$p_{i,\theta} = \frac{N\sigma_\theta}{N\sigma_\theta - 1} \cdot w,$$

where w is the firm's marginal cost. Note that firms' markups are decreasing in the number of competitors per product line, N, since the outputs of firms within the same product line are perfect substitutes. Using the expression for markups and the fact that $y_\theta = N y_{i,\theta}$, we can rewrite the free entry condition as

$$\frac{L}{N^2} \int_0^1 \frac{p_\theta y_\theta}{\sigma_\theta} g(\theta) d\theta = f_e.$$

It is helpful to note here that, holding all other factors constant, an increase in the number of competitors per product line (N) has two effects on firms' variable profits: it both reduces the share of product line output that is supplied by a single firm, and it also leads to a decrease in the price charged by a single firm within a product line due to the erosion of oligopoly power within the product line. Hence, holding all other factors constant, N must grow at a rate of $\sqrt{L}$ to maintain the same level of variable profits per firm.

Below, we present analytical results for the elasticity of per-capita welfare to an infinitesimal

[49]The parameter α can be micro-founded by assuming that variety creation has diminishing returns in the total entry costs paid by firms. When variety creation is constant-returns in fixed costs paid, the number of varieties scales exactly with the mass of entrants, and hence $M = V$ and $\alpha = 0$. Diminishing returns instead imply $\alpha > 0$, and the case where the total number of varieties is completely inelastic to fixed costs paid yields $\alpha = 1$.

change in market size ($d \log Y / d \log L$) for the edge cases where $\alpha = 0$ and $\alpha = 1$ to highlight intuition, and then present results from quantitative exercises (meant to be suggestive only). In order to analyze the response of all variables to an infinitesimal change in market size, we assume away the integer constraint on the number of competitors per product line N.

When $\alpha = 0$ and $N = 1$, the change in per-capita welfare following a market expansion is composed of a technical efficiency gain and a change in allocative efficiency due to the Darwinian effect and pro-competitive effect, which are familiar from the main text:

$$d \log Y = \underbrace{(\bar{\delta} - 1)\, d \log L}_{\text{Technical efficiency}} + \Bigg(\underbrace{(\bar{\delta} - 1)\, Cov_\lambda \left[\sigma_\theta, \frac{1}{\mu_\theta} \right]}_{\xi^\epsilon} + \underbrace{\mathbb{E}_\lambda \left[\frac{1}{\sigma_\theta} \right] \mathbb{E}_\lambda \left[(1 - \rho_\theta)\, \sigma_\theta \left(1 - \frac{\bar{\delta}}{\mu_\theta} \right) \right]}_{\xi^\mu} \Bigg) \bar{\mu} d \log L.$$

If instead $\alpha = 1$, we clearly will have no technical efficiency effect nor Darwinian effect. The technical efficiency effect is zero in this case since no new product line varieties are created—all new entry is instead allocated to increasing competition within existing product lines. Similarly, the Darwinian effect is zero because, holding markups constant, entry does not affect the aggregate price index. Hence, the mechanism driving the Darwinian effect (entry drives down the price index, and leads to a reallocation away from high-elasticity/low-markup firms) is silenced. The only effect of market entry on welfare is the pro-competitive effect:

$$d \log Y = \frac{\frac{\mathbb{E}_\lambda \left[(\sigma_\theta - 1) \rho_\theta^{\text{effective}} \right] \mathbb{E}_\lambda \left[(\mu_\theta - 1) \rho_\theta^{\text{effective}} \right]}{\mathbb{E}_\lambda \left[(\sigma_\theta - 1)(\mu_\theta - 1) \rho_\theta^{\text{effective}} \right]} + \mathbb{E}_\lambda \left[1 - \rho_\theta^{\text{effective}} \right]}{\frac{\mathbb{E}_\lambda \left[(\sigma_\theta - 1) \rho_\theta^{\text{effective}} \right] \mathbb{E}_\lambda \left[\frac{1 + \rho_\theta^{\text{effective}}}{\sigma_\theta} \right]}{\mathbb{E}_\lambda \left[(\sigma_\theta - 1)(\mu_\theta - 1) \rho_\theta^{\text{effective}} \right]} + \mathbb{E}_\lambda \left[\frac{\sigma_\theta - 1}{\sigma_\theta} \right] + (N - 1)\, \mathbb{E}_\lambda \left[1 - \rho_\theta^{\text{effective}} \right]} \mathbb{E}_\lambda \left[\frac{1}{\sigma_\theta} \right] d \log L \geq 0,$$

where

$$\rho_\theta^{\text{effective}} = \left[\frac{N \sigma_\theta - 1}{(N - 1)\, \sigma_\theta \rho_\theta + (\sigma_\theta - 1)} \right] \rho_\theta \geq 0,$$

is the partial equilibrium pass-through of a change in the price aggregator to the desired markup of a firm with type θ. (When $N = 1$, $\rho_\theta^{\text{effective}}$ is simply equal to ρ_θ.)

While the full expression is rather unwieldy, it is immediately evident that the pro-competitive effect in this case is always positive: an increase in entry erodes oligopoly pricing power, lowering markups and increasing allocative efficiency. Intuitively, since there are no variety gains from new entry, entry is always socially wasteful. Thus, entry is always excessive as long as markups are greater than one. Since an increase in market size leads all firms to reduce their markups, welfare improves.

For more intuition, consider the case where all firms are homogeneous. Then the change

in per-capita welfare is

$$d \log Y = \frac{1}{2} (\mu - 1) \, d \log L.$$

We can see that the pro-competitive effect is increasing in the initial markup μ. Since the sole gains from an increase in market size come from eroding oligopoly power and reducing markups, when μ is small there is little scope for welfare improvements. The reason that welfare gains are scaled by 1/2 is that, as mentioned above, the free entry condition implies that the number of firms scales with the square-root of market size.[50] In the heterogeneous firm case, the pro-competitive effect also includes reallocations across product lines due to the fact that prices of each product line may be differentially affected by the increase in the number of competitors N, leading to the more complicated general expression.

We now present some stylized quantitative results to illustrate the forces depicted in this appendix. These results are not the product of a full calibration and should not be interpreted as such. Rather, these results illustrate the fact that when an increase in market size creates new product lines, the technical efficiency and Darwinian effects are most important (as in our baseline), but that if this entry is instead directed at creating more competitors in existing product lines, that will put more weight on the pro-competitive effect.

For these results, we assume that $N = 1$ in the initial equilibrium. (Starting at some $N \neq 1$ would mean that the model is no longer consistent with the method we use to back out elasticities, markups, and consumer surplus ratios from the data.)

Table K.1 shows the results for values of α ranging from 0 to 0.2. As α increases, both the technical efficiency and Darwinian effect attenuate, since fewer resources are allocated to the creation of new product line varieties. Meanwhile, the pro-competitive effect, which has a mildly negative effect on welfare at $\alpha = 0$, becomes more positive, since the pro-competitive now also includes the beneficial erosion of oligopoly power.

	$\alpha = 0$	$\alpha = 0.05$	$\alpha = 0.1$	$\alpha = 0.15$	$\alpha = 0.2$
Technical efficiency:	0.090	0.085	0.081	0.076	0.072
Darwinian effect:	0.631	0.422	0.301	0.222	0.168
Pro-competitive effect:	-0.099	-0.043	-0.016	-0.001	0.008

Table K.1: Welfare changes due to technical efficiency and allocative efficiency in response to a change in population in the oligopoly model under different values of α.

[50] In this case with $\alpha = 1$ and homogeneous firms, we can rewrite the free entry condition as

$$\frac{L}{N^2} \frac{py}{\sigma} = \frac{L}{N^2} \frac{w}{\sigma} = f_e.$$

Since σ, w, and f_e do not change with the increase in market size, we must have $d \log L = 2d \log N$.

Appendix L Markup and Pass-through Variation Unrelated to Size

The calibration in the main text assumes that firm markups and pass-throughs vary only as a function of firm size. In practice, other factors unrelated to firm size may also influence markups and pass-throughs, however. In this appendix, we consider how our results change if we allow for additional heterogeneity orthogonal to firm size. First, we consider how our analytical results change if we add variation in firms' pass-throughs and price elasticities (which, by the Lerner condition, lead to variation in firms' markups) due to factors unrelated to size. Second, we recalculate the Darwinian effect when there is additional variation in markups unrelated to size commensurate with markup estimates by De Loecker et al. (2016). Both analytically and quantitatively, this additional heterogeneity increases the magnitude of the Darwinian effect.

L.1 Analytical Results with Variation Unrelated to Size

We start by considering how additional variation in price elasticities and pass-throughs unrelated to size affect our analytical results. Suppose the price elasticity and pass-through of firm i are given by

$$\sigma_i = \underbrace{\mathbb{E}[\sigma|\lambda_i]}_{\sigma_\lambda} + \epsilon_i,$$
$$\rho_i = \underbrace{\mathbb{E}[\rho|\lambda_i]}_{\rho_\lambda} + \nu_i,$$

where ϵ_i and ν_i are orthogonal to λ_i (and hence to σ_λ and ρ_λ), but may be correlated with each other. We can microfound this by perturbing the expenditure share function $s_\theta(\cdot)$ by firm.

Introducing this variation does not change the sales-weighted average elasticity and pass-through, since, due to the law of iterated expectations,

$$\begin{aligned}\mathbb{E}_\lambda[\sigma_i] &= \mathbb{E}[\mathbb{E}[\lambda_i\sigma_i|\lambda_i]]/\mathbb{E}[\lambda_i] \\ &= \mathbb{E}[\lambda_i\sigma_\lambda]/\mathbb{E}[\lambda_i] \\ &= \mathbb{E}_\lambda[\sigma_\lambda].\end{aligned}$$

However, terms that depend on the covariance of elasticities, markups, and/or pass-

throughs may change. For example, by Jensen's inequality,

$$\mathbb{E}_\lambda\left[\frac{1}{\sigma_i}\right] = \mathbb{E}_\lambda\left[\frac{1}{\sigma_\lambda + \varepsilon_i}\right] \geq \mathbb{E}_\lambda\left[\frac{1}{\sigma_\lambda}\right].$$

We can thus consider how ξ^ϵ, ξ^{θ^*}, and ξ^μ change when we allow for this additional variation in elasticities and pass-throughs, which corresponds to the allocative gains from an increase in market size.

Darwinian effect:

$$\begin{aligned}
\xi^\epsilon &= \left(\mathbb{E}_\lambda\left[\delta_i\right] - 1\right) Cov_\lambda\left[\sigma_i, \frac{1}{\mu_i}\right] \\
&= \left(\mathbb{E}_\lambda\left[\delta_i\right] - 1\right)\left(Cov_\lambda\left[\sigma_\lambda + \epsilon_i, -\frac{1}{\sigma_\lambda + \epsilon_i}\right]\right) \\
&= \left(\mathbb{E}_\lambda\left[\delta_i\right] - 1\right)\left(\mathbb{E}_\lambda\left[\sigma_\lambda\right]\mathbb{E}_\lambda\left[\frac{1}{\sigma_\lambda + \epsilon_i}\right] - 1\right) \\
&\geq \left(\mathbb{E}_\lambda\left[\delta_i\right] - 1\right)\left(\mathbb{E}_\lambda\left[\sigma_\lambda\right]\mathbb{E}_\lambda\left[\frac{1}{\sigma_\lambda}\right] - 1\right).
\end{aligned}$$

Hence, the Darwinian effect increases in magnitude when we allow for additional variation in elasticities. Intuitively, this is because the Darwinian effect depends on heterogeneity in price elasticities and markups across firms, and so greater variation leads to a larger Darwinian effect.

Selection effect:

$$\begin{aligned}
\xi^{\theta^*} &= \left(\mathbb{E}_\lambda\left[\delta_i\right] - \delta_{\theta^*}\right)\lambda_{\theta^*}\gamma_{\theta^*}\left(\sigma_{\theta^*}\mathbb{E}_\lambda\left[\frac{1}{\sigma_i}\right] - 1\right) \\
&= \left(\mathbb{E}_\lambda\left[\delta_i\right] - \delta_{\theta^*}\right)\lambda_{\theta^*}\gamma_{\theta^*}\left(\sigma_{\theta^*}\mathbb{E}_\lambda\left[\frac{1}{\sigma_\lambda}\right] - 1\right) + \left(\mathbb{E}_\lambda\left[\delta_i\right] - \delta_{\theta^*}\right)\lambda_{\theta^*}\gamma_{\theta^*}\sigma_{\theta^*}\underbrace{\left(\mathbb{E}_\lambda\left[\frac{1}{\sigma_i}\right] - \mathbb{E}_\lambda\left[\frac{1}{\sigma_\lambda}\right]\right)}_{>0}.
\end{aligned}$$

The change in the selection effect is hard to gauge from this formula alone, since the additional variation in price elasticities in pass-throughs may change which firm type θ is the cutoff firm θ^*. If we hold the price elasticity, consumer surplus ratio, and sales share of the marginal firm type constant, whether additional variation in elasticities increases or decreases the selection effect depends on whether δ_{θ^*} is greater or less than $\mathbb{E}_\lambda[\delta_\theta]$.

Pro-competitive effect:

$$\begin{aligned}
\xi^\mu &= \mathbb{E}_\lambda\left[\left(1 - \rho_i\right)\left(\sigma_i - \mathbb{E}_\lambda\left[\delta_i\right]\left(\sigma_i - 1\right)\right)\right] \\
&= -\left(\mathbb{E}_\lambda\left[\delta_i\right] - 1\right)\mathbb{E}_\lambda\left[\left(1 - \rho_i\right)\sigma_i\right] + \mathbb{E}_\lambda\left[\delta_i\right]\mathbb{E}_\lambda\left[1 - \rho_i\right] \\
&= \left(\mathbb{E}_\lambda\left[\delta_i\right] - 1\right) Cov_\lambda\left(\rho_i, \sigma_i\right) - \left(\mathbb{E}_\lambda\left[\delta_i\right] - 1\right)\left(\mathbb{E}_\lambda\left[1 - \rho_i\right]\mathbb{E}_\lambda\left[\sigma_i\right]\right) + \mathbb{E}_\lambda\left[\delta_i\right]\mathbb{E}_\lambda\left[1 - \rho_i\right] \\
&= \left(\mathbb{E}_\lambda\left[\delta_i\right] - 1\right)\left[Cov_\lambda\left(\rho_\lambda, \sigma_\lambda\right) + Cov_\lambda\left(\nu_i, \epsilon_i\right)\right] - \left(\mathbb{E}_\lambda\left[\delta_i\right] - 1\right)\left(\mathbb{E}_\lambda\left[1 - \rho_\lambda\right]\mathbb{E}_\lambda\left[\sigma_\lambda\right]\right) + \mathbb{E}_\lambda\left[\delta_i\right]\mathbb{E}_\lambda\left[1 - \rho_\lambda\right].
\end{aligned}$$

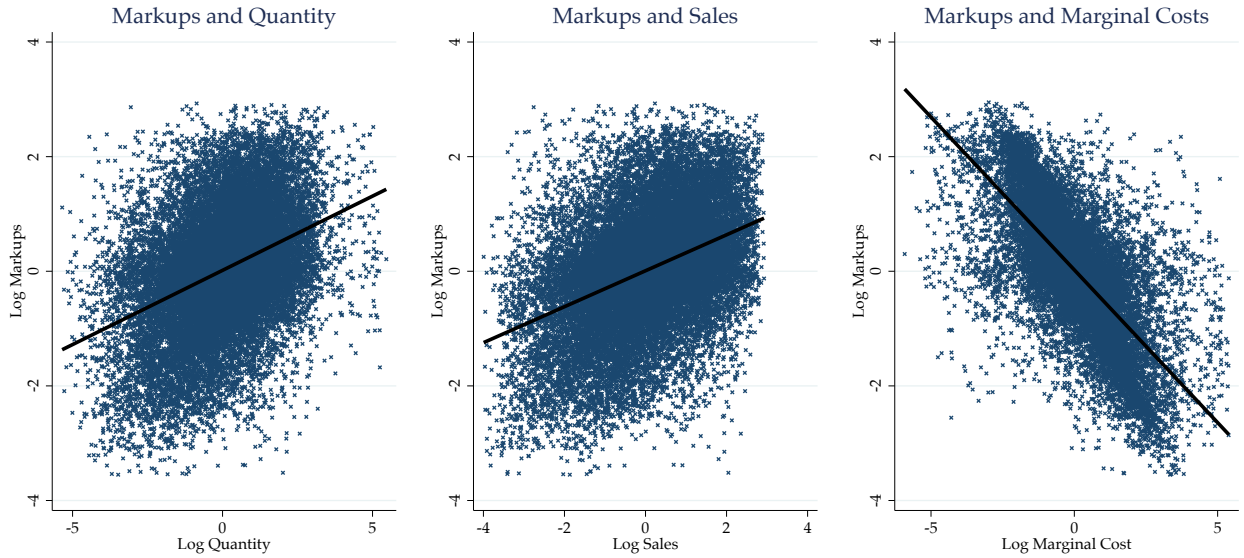


Figure L.1: Scatterplots of product-level markup estimates by De Loecker et al. (2016) for Indian manufacturing firms against output quantity, sales, and estimated marginal costs. As in Figure 1 of De Loecker et al. (2016), plotted variables are demeaned by product-year fixed effects and trimmed at the 3rd and 97th percentiles.

Hence, whether the pro-competitive effect becomes more or less positive depends on the covariance $Cov_\lambda(\nu_i, \epsilon_i)$. If the noise in pass-throughs and price elasticities is positively related, high markup (low elasticity) firms cut their markups more in response to a fall in the aggregate price index, leading to a beneficial reallocation to high-markup firms and an increase in welfare.

L.2 Quantitative Results with Markup Variation Unrelated to Size

We now consider how our quantitative results on the magnitude of the Darwinian effect change if we allow for additional variation in markups unrelated to size consistent with previous empirical work. We use estimates made publicly available by De Loecker et al. (2016), who use price and quantity data to estimate markups at the product level for Indian manufacturing firms. Of course, the geographic and institutional context for Indian manufacturing firms is quite different from our baseline calibration; we present these results only as an illustration of how additional variation in markups would affect our results.

In Figure L.1, we plot log product markups against log output quantity, log sales, and log of estimated marginal costs. All variables are demeaned by product-year fixed effects and trimmed at the 3rd and 97th percentiles (the first panel replicates Figure 1 from De Loecker et al. 2016). Estimated markups are positively correlated with quantity and sales and negatively correlated with estimated marginal costs, as in our calibration. Conditional on sales, there is

also substantial dispersion in the estimated markups. To estimate the portion of variation in markups explained by size, we regress the demeaned log markups on log sales in the data provided by De Loecker et al. (2016). We find that $R^2 = 0.16$, which means that 84% of the variation in markups is due to factors orthogonal to size.

As a back-of-the-envelope exercise, we simulate how adding variation in markups unrelated to size affects the magnitude of the Darwinian effect. We do so by simulating 1 million draws from our firm sales distribution, and adding a random normal error to the markup a firm of that sales share has in our baseline calibration. We choose the standard deviation of the shocks by the simulated method of moments to match the R^2 of the regression of log markups on log sales.[51] Using these markup estimates (and the corresponding price elasticities implied by the Lerner condition), we find that the Darwinian effect is 0.380 when we choose $\mathbb{E}_\lambda[\delta_\theta] = 1.090$. This is moderately larger than our baseline estimate in Table 1, consistent with our analytic results above that additional variation in markups and price elasticities will strengthen the Darwinian effect.

[51] Note that we bound log markups from below by 1.01, to ensure that we do not get markups below one. We find that normal errors in markups with a standard deviation of 0.165 generate an $R^2 = 0.16$ to match the data from De Loecker et al. (2016).

Appendix M The Darwinian Effect under Separable Preferences

A key contribution of our analysis is to isolate the Darwinian effect, which captures how an increase in market size alleviates cross-sectional misallocation. In this appendix, we describe necessary conditions for the Darwinian effect to be positive and discuss cases in which the Darwinian effect will not appear.

The Darwinian effect captures how a fall in the price index affects cross-sectional misallocation, holding firms' markups constant. Let $\mathcal{X}^{\text{variable}} = \{l_\theta / L^{\text{variable}}\}$ denote the fraction of variable production labor allocated to the production of each variety θ. We show in Appendix F that the effect of a change in the cross-sectional allocation of resources on welfare is

$$d \log Y = \frac{\partial \log \mathcal{Y}}{\partial \mathcal{X}^{\text{variable}}} d\mathcal{X}^{\text{variable}} = Cov_\lambda \left[\bar{\mu}/\mu_\theta, d \log l_\theta \right]. \tag{38}$$

That is, a reallocation of labor to firms with initially high markups alleviates cross-sectional misallocation and raises efficiency. This covariance between firms' initial (inverse) markups and the change in their use of labor for variable production is a sufficient statistic for how a change in the cross-sectional allocation of resources affects welfare.

Now, consider a class of preferences in which we can express the per-capita demand curve faced by each variety as

$$y_\theta = Q \, D_\theta(p_\theta / P), \tag{39}$$

where $Q = Q(\{p\}, Y)$ is an aggregate demand shifter that depends on the vector of all firms' prices and money-metric utility Y, and $D_\theta(\cdot)$ is a function that takes the firm's price relative to an aggregate price index, p_θ / P, as an argument and may vary with firm type θ. (Note that P is an aggregate price index, but is not required to be the ideal price index.)

Note that these demand curves are not a complete specification of preferences. Rather, we are considering the case where preferences yield per-capita demand curves that satisfy (39). For example, the class of preferences described by Arkolakis et al. (2019) satisfy (39); as do the HSA, HDIA, and HIIA preferences discussed by Matsuyama and Ushchev (2017), the "Gorman-Pollak" demand system discussed by Fally (2022), and the (non-homothetic) additively separable preferences used by Krugman (1979).[52] Crucially, the demand curves in (39) satisfy a version of the property termed "generalized separability" by Fally (2022), where a variety's demand depends on its price relative to an aggregate price index. Nearly all models in the macroeconomic and international trade literature assume that demand takes a

[52] For example, for HSA preferences, $Q = P$, $D_\theta(x) = s_\theta(x)/x$, and the aggregate price index P is given by (2). For the "Gorman-Pollak" demand system discussed by Fally (2022), take Equation (2) from Fally 2022 and set $F = P$, $H = 1/Q$).

form consistent with (39), as discussed by Burstein and Gopinath (2014) and Arkolakis et al. (2019).

When demand curves follow (39), we can rewrite the covariance in (38) (holding firms' markups constant) as

$$Cov_\lambda\left[\bar{\mu}/\mu_\theta, \frac{d\log l_\theta}{d\log L}\right] = Cov_\lambda\left[\bar{\mu}/\mu_\theta, \frac{\partial \log D_\theta}{\partial \log \frac{p_\theta}{P}}\right]\frac{d\log P}{d\log L}.$$

This covariance is the Darwinian effect, since it captures how increased entry affects the cross-sectional allocation of resources and thus welfare. From the above expression, we see that three conditions are necessary for the Darwinian effect to increase welfare: (1) there must be initial heterogeneity in firms' markups, (2) the aggregate price index must fall in response to the increase in market size, and (3) the partial elasticity of demand with respect to firms' relative prices, $\frac{-\partial \log D_\theta}{\partial \log(p_\theta/P)}$, must be heterogeneous across firms and positively correlated with firms' inverse markups. We call $\frac{-\partial \log D_\theta}{\partial \log(p_\theta/P)}$ the *partial elasticity* of demand facing firm θ because the true elasticity of demand facing the firm may also include changes in Q and P internalized by the firm.[53] This partial elasticity of demand may be heterogeneous across firms either because firms' initial prices vary and D_θ is non-isoelastic, or because D_θ varies with θ. When (1), (2), and (3) hold, the Darwinian effect will exist and be positive.

The third condition explains why nested CES preferences, as in Atkeson and Burstein (2008), will not generate a Darwinian effect. In that case, the demand curve faced by firm i in sector $\mathcal{I}$ is

$$\frac{y_i}{Y_\mathcal{I}} = \left(\frac{Y_\mathcal{I}}{Y}\right)^{-\frac{\eta_1}{\eta_0}}\left(\frac{p_i}{P}\right)^{-\eta_0}, \tag{40}$$

where $Y_\mathcal{I}$ is a CES aggregate of output in sector $\mathcal{I}$, P is the aggregate price index across sectors, and η_0 and η_1 are the within- and across-sector elasticities of substitution. Under these preferences, the partial elasticity of demand with respect to firms' relative prices (η_0) is constant across firms and thus there is no Darwinian effect.

However, if we depart from the knife-edge case in which partial elasticities of demand relative to the aggregate price index are uniform across all firms, increases in market size will generate Darwinian reallocations across firms. Appendix K presents an extension of the model with oligopolistic competition between firms within product lines. In that extension, the Darwinian effect persists as long as the aggregator of product line outputs is non-isoelastic

[53] The elasticity of demand facing the firm is instead

$$\sigma_\theta = \frac{-\partial \log y_\theta}{\partial \log p_\theta} = -\frac{\partial \log Q}{\partial \log p_\theta} - \frac{\partial \log D_\theta}{\partial \log(\frac{p_\theta}{P})}\left(1 - \frac{\partial \log P}{\partial \log p_\theta}\right).$$

Under oligopolistic competition, firms internalize their impact on the aggregate indices Q and P, and hence the partial elasticity of D_θ and the demand elasticity facing the firm may differ. This is why the Atkeson and Burstein (2008) model has heterogeneous markups, but uniform partial elasticities of D_θ.

and an increase in market size decreases the aggregate price index.

Appendix N Klenow-Willis Calibration

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

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

$$\mu_\theta = \mu(\frac{y_\theta}{Y}) = \frac{1}{1 - \frac{1}{\sigma}(\frac{y_\theta}{Y})^{\frac{\epsilon}{\sigma}}}, \tag{41}$$

$$\rho_\theta = \rho(\frac{y_\theta}{Y}) = \frac{1}{1 + \frac{\epsilon}{\sigma - (\frac{y_\theta}{Y})^{\frac{\epsilon}{\sigma}}}} = \frac{1}{1 + \frac{\epsilon}{\sigma}\mu_\theta}. \tag{42}$$

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

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

$$\rho^{\max} = \frac{1}{1 + \epsilon/\sigma}. \tag{43}$$

Amiti et al. (2019) estimate the average pass-through for the smallest 75% of firms in Prodcom is 0.97. In order to match $\rho = 0.97$, we must choose $\epsilon/\sigma \approx 0.03$.

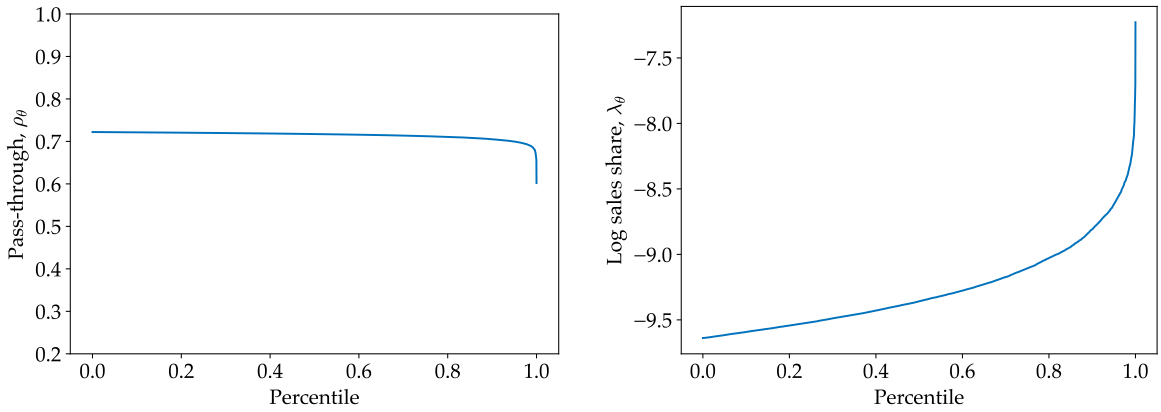
This makes it difficult, however, to match the incomplete pass-throughs estimated for the largest firms. To match a pass-through of $\rho_\theta = 0.3$ with $\epsilon/\sigma = 0.03$, we need a markup of $\mu_\theta \approx 78$ for the largest firms. In contrast, our non-parametric procedure matches the pass-through distribution with realistic markups of around 2 for the largest firms (shown in the main text, Figure 3a). This roughly accords with estimates of markups by De Loecker et al. (2020).

Rather than attempting to match the empirical pass-through distribution, suppose we used a set of parameters from the literature. We adopt the calibration from Appendix D of Amiti et al. (2019): $\sigma = 5$, $\epsilon = 1.6$, and firm productivities are drawn from a Pareto distribution with shape parameter equal to 8.[54] The simulated distributions of firm pass-throughs and sales

[54] We calibrate the model by drawing 10,000 firms and finding a fixed point in output.

shares are shown in Figure N.1. Over the range of drawn productivities, we see little variation in pass-through.

Figure N.1: Pass-through ρ_θ and sales share density $\log \lambda_\theta$ under Klenow and Willis (2016) preferences.



(a) Pass-through ρ_θ

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

Table N.1 shows the response of welfare and real GDP per capita to an increase in market size for Klenow and Willis (2016) preferences, with the results from the main text for comparison. We find that the calibration of Klenow and Willis (2016) preferences attributes nearly all gains to technical efficiency gains, rather than allocative efficiency gains. In particular, the parametric preferences dramatically understate the importance of the Darwinian channel.

	HDIA preferences $\bar{\mu} = 1.090$		Klenow-Willis
	$\bar{\delta} = \delta_{\theta^*}$	$\bar{\delta} = \bar{\mu}$	
Welfare: $d \log Y$	0.303	0.317	0.268
Technical efficiency: $d \log Y^{tech}$	0.033	0.090	0.260
Allocative efficiency: $d \log Y^{alloc}$	0.269	0.227	0.008
Darwinian effect: $d \log Y^{\epsilon} - d \log Y^{tech}$	0.284	1.500	0.009
Selection effect: $d \log Y^{\epsilon,\theta^*} - d \log Y^{\epsilon}$	0.000	-1.110	-0.000
Pro-competitive effect: $d \log Y^{\epsilon,\theta^*,\mu} - d \log Y^{\epsilon,\theta^*}$	-0.015	-0.162	-0.001
Real GDP per capita	0.052	0.053	0.076

Table N.1: Comparison of the elasticity of welfare and real GDP per capita to population in the benchmark and Klenow and Willis (2016) calibrations.