Review of Economic Studies (2024) **91**, 1373–1405 https://doi.org/10.1093/restud/rdad061

Advance access publication 9 June 2023

The Darwinian Returns to Scale

David Rezza Baqaee
UCLA
Emmanuel Farhi[†]
Harvard
and
Kunal Sangani
Harvard

First version received August 2021*; Editorial decision March* 2023*; Accepted June* 2023 *(Eds.)*

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 nonparametrically calibrate demand curves with data from Belgian manufacturing firms and quantify our results. We find that mild increasing returns at the microlevel can catalyze large increasing returns at the macrolevel. Between 70 and 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.

Key words: Efficiency, Market size, Returns to scale

JEL codes: E0, L1, O4

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

[†]Emmanuel Farhi tragically passed away in July 2020. Emmanuel was a one-in-a-lifetime collaborator and friend.

The editor in charge of this paper was Thomas Chaney.

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 microlevel (measured by the average ratio of marginal to average cost) can catalyze large increasing returns at the macrolevel.[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 microlevel and macrolevel 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 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; 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/anticompetitive 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

1. See Basu and Fernald (1997) on how aggregation can amplify microreturns 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.
3. The preferences we use, which Matsuyama and Ushchev (2017) call homothetic with a single aggregator, 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.

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/anticompetitive effect, which have been studied in detail in previous work, have theoretically ambiguous effects on welfare. The selection effect (2) results from the fact that, as the market expands, the minimum level of profitability a firm must have to survive can change. This mechanism 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/anticompetitive effect (3) results from the fact that firms' markups may change as the market expands. Of the three channels, the pro/anticompetitive 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 (procompetitive effect). Recent studies exploring the pro/anticompetitive 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 nonparametrically 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.

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.

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 procompetitive 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 quasirents, 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 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 Supplementary 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.

1.1. *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

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 total factor productivity 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.

(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 under (nonhomothetic) 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 procompetitive effects in an open economy with an export margin following shocks to iceberg trade costs. They find that procompetitive effects on welfare are zero when preferences are homothetic and mildly reduce, rather than increase, welfare for important classes of nonhomothetic 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 procompetitive) are generically zero following a change in market size. Nevertheless, our findings on the procompetitive 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 nonparametric model. Using this strategy, we quantify the importance of the Darwinian, selection, and procompetitive 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 nonparametric 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.

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 2(a)). Klenow and Willis (2016) preferences cannot simultaneously match the distributions of pass-throughs and markups (see Supplementary 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.

1.2. *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 nonparametrically. Section 7 is a quantitative application. Section 8 summarizes extensions, and Section 9 concludes. The Supplementary Appendix contains all the proofs.

2. MODEL SETUP

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

2.1. *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

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

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,

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

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 $P^Y Y = I = 1$.[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.

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).

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

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.

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.

2.2. *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 equation (1), the price elasticity of demand facing a variety of type θ, denoted σ_θ, is given by

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

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]

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

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

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

13. See Matsuyama and Ushchev (2017) for a proof.
14. For an extension with oligopolistic competition, see Supplementary 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.* Johnson and Myatt, 2006; Ravn *et al.*, 2006; Gilchrist *et al.*, 2017). Since our objective is to compare long-run steady states, we abstract from these considerations.
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.

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.

2.3. *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)–(6).

2.4. *Notation*

Denote the sales share density by

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

This is a density because it is always nonnegative 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

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

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 $\frac{-\sigma_\theta}{\rho_\theta}\frac{\partial \log \mu_\theta}{\partial \theta} + (\frac{\sigma_\theta}{\rho_\theta} - 1)\frac{\partial \log A_\theta}{\partial \theta} - \frac{\partial \log f_{o,\theta}}{\partial \theta} > 0$, where ρ_θ is the pass-through function defined in terms of primitives by equation (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.

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.

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,

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

where the markup function is given by equation (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" ($\mu_{\theta}'\left(\frac{p}{P}\right) < 0$ 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,

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

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).

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 "nonappropriability": 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 $1/\delta_{\theta}$.

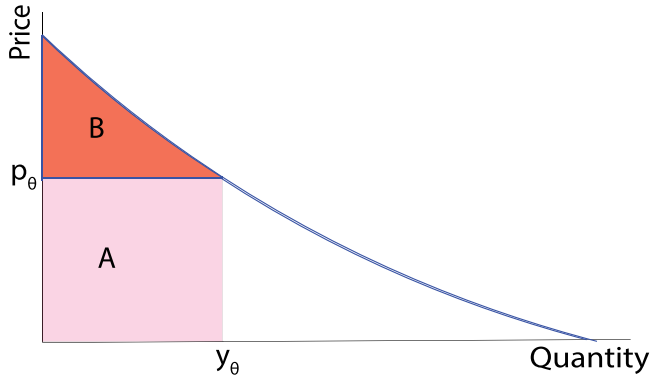

FIGURE 1

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

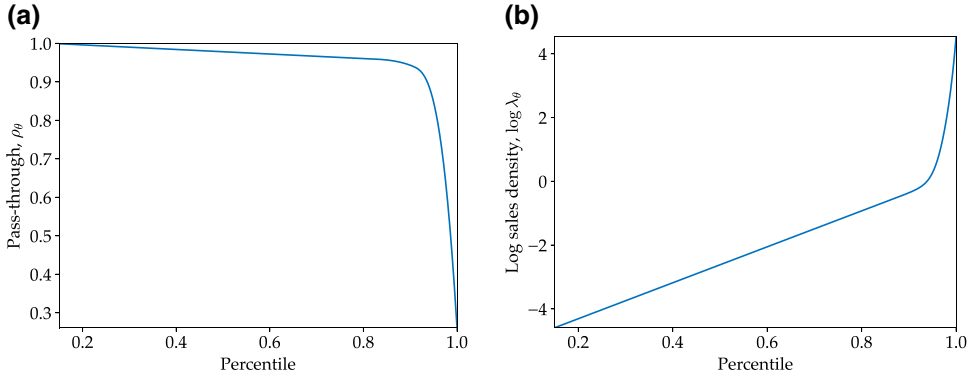

FIGURE 2

Pass-throughs and sales share density as a function of firm type θ. (a) Pass-through ρ_θ. (b) Log sales share density $\log \lambda_\theta$

3.2. *Welfare*

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

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

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 equation (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 equation (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, for example $\bar{A} = \mathbb{E}_\lambda[A_\theta]$. 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 equation (10) to equation (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.

3.3.1. 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, equation (11) holds as an equality and so the CES model has efficient entry.

3.3.2. 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 equation (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, equation (12) holds as an equality and so the CES model has efficient exit.

3.3.3. 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

20. Equivalently, Lemma 1 can be understood through the lens of the nonappropriability 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.

21. In reality, there may be other distortions that make it suboptimal 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.

expansion of more productive firms increases welfare, but this is only because "high productivity" proxies for "high markup."[22] In a CES model, equation (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.

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 $L_1, L_2, \ldots, L_N$. The market equilibrium if these N countries trade freely is the same as the market equilibrium in a single, closed economy with size $L_1 + L_2 + \cdots + L_N$; 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 preexisting 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

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 equation (10), which consider a reallocation towards firms with higher *levels* of productivity A_θ as improving efficiency. See Section 3.2.

23. Although we focus on changes in market size in the body of the paper, in Supplementary 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 Supplementary Appendix F.2 for a derivation.

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]

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

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

$$(\textit{Selection Effect})\ \xi^{\theta^*} = \left(\mathbb{E}_\lambda[\delta_\theta] - \delta_{\theta^*}\right) \lambda_{\theta^*} \gamma_{\theta^*} \left(\mathbb{E}_\lambda\left[\frac{\sigma_{\theta^*}}{\sigma_\theta}\right] - 1\right) \lesseqgtr 0,$$

$$(\textit{Pro/Anticompetitive Effect})\ \xi^{\mu} = \mathbb{E}_\lambda\left[(1-\rho_\theta)\,\sigma_\theta \left(1 - \frac{\mathbb{E}_\lambda[\delta_\theta]}{\mu_\theta}\right)\right] \mathbb{E}_\lambda\left[\frac{1}{\sigma_\theta}\right] \lesseqgtr 0,$$

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 equation (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

25. Supplementary 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).

26. In terms of primitives, this is

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

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 equation (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 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.

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

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

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 procompetitive 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 equation (16) and overhead costs are zero, the change in welfare from an increase in market size is given by*

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

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 Section 1, 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 equation (17) appears because the reallocations caused by the Darwinian effect save on labor, and these extra resources are funneled into additional entry.[30]

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

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

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

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

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

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).

29. Supplementary 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 equation (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.

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.

4.2.3. Pro/Anticompetitive 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/Anticompetitive 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*

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

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 *procompetitive effect* of scale. In this example, the procompetitive effect is captured entirely by $\rho < 1$: markups decrease since each firm's price rises relative to the aggregate price index. As equation (18) makes clear, the welfare impact of these procompetitive 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.

4.3.1. 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.* quasirents). 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.

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

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

$$\text{(Darwinian Effect) } \zeta^{\epsilon} = (\bar{\mu} - 1)\, Cov_{\lambda}\left[\sigma_{\theta}, \frac{1}{\mu_{\theta}}\right] \geq 0,$$

$$\text{(Selection Effect) } \zeta^{\theta^*} = \lambda_{\theta^*}\gamma_{\theta^*}\left(\frac{\bar{\mu}}{\mu_{\theta^*}} - 1\right)\left(\mathbb{E}_{\lambda}\left[\frac{\sigma_{\theta^*}}{\sigma_{\theta}}\right] - 1\right) \geq 0,$$

$$\text{(Procompetitive Effect) } \zeta^{\mu} = -\mathbb{E}_{\lambda}\left[\frac{\bar{\mu} - 1}{\sigma_{\theta}}\right]\mathbb{E}_{\lambda}\left[(\sigma_{\theta} - 1)\,(1 - \rho_{\theta})\right] \leq 0, \quad if\, \rho_{\theta} \leq 1.$$

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 firms. The Darwinian and selection effects are mitigated by the third effect, which captures firms' markup adjustments (the procompetitive effect). The procompetitive 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 procompetitive effect. If firms are homogeneous, then only the procompetitive 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.

4.3.2. 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 equation (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 ($\rho_{\theta} < 1$),

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 equation (9) are zero. Our decomposition of

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 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 equation (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 Supplementary 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*

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

where $\psi_e = \Delta f_e/(\Delta f_e + (1-G(\theta^*))\mathbb{E}[f_{o,\theta}])$ *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 equation (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 procompetitive effect of the entry tax. Together with the welfare

efficiency in (9) does not impose these restrictions. This also separates the decomposition in equation (9) from decompositions that do not capture how entering/exiting varieties affect welfare, such as Petrin and Levinsohn (2012) and Baqaee and Farhi (2019).

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

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 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.

5.1. *Darwinian effect*

Consider again the economy in Corollary 1, where there are no overhead costs and preferences are given by equation (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 equation (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.

5.2. *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/anticompetitive effects are still absent.

Corollary 5 (Darwinian and selection effect)**.** *When preferences are given by equation (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.

5.3. *Pro/Anticompetitive 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/Anticompetitive 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 standalone 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. *Nonparametric 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,*

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

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.*

The intuition for Proposition 4 follows from equation (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

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 microlevel moments and use the calibrated model to perform counterfactual exercises.

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

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 and 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*

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

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

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

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 equation (21), we start by writing the relationship between marginal cost changes and changes in firms' markups μ_θ and sales λ_θ:

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

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

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

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 equation (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 equation (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 equation (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 equation (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 equation (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 equation (20) precludes this possibility, we relax this restriction by adding variation in markups orthogonal to firm size in Supplementary 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 Supplementary Appendix A.

6.2.1. 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 to 2007, which contains information on prices and sales, collected by Statistics Belgium. Using exchange rate shocks as instruments for changes in

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 $\mathbb{E}_\lambda[\delta_\theta]$.

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 Supplementary Figure A.2 in Appendix A.

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

Finally, we use firm exit rates by age reported by Sterk *et al.* (2021). The exit rate for new entrants is about 15 percentage points higher than mature firms, so we set $\theta^* = 0.15$.

6.2.2. 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.* (2023) 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 Supplementary 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 procompetitive effects, do not vary significantly with $\bar{\mu}$. For completeness, in Supplementary 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.

6.2.3. Calibrated statistics. Figure 2(a) and (b) displays pass-throughs, ρ_θ, and log sales, $\log \lambda_\theta$, as a function of type θ. See Supplementary Appendix A for details about how we construct these figures. Figure 2(a) shows that pass-throughs decrease from 1 for the smallest firms to about 0.3 for the largest firms. Figure 2(b) 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 3(a) 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), Grassi *et al.* (2022), and Forlani *et al.* (2023).[39] If we used Klenow and Willis (2016) preferences instead (and continued 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 Supplementary Appendix N for more detail). Second, our calibrated markups are positively correlated with firm output and

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. Supplementary Appendix D provides results using this assumption. The calibrated elasticities are different, but the overall message does not change.

39. Using the production function approach to estimate markups for French manufacturing firms, Grassi *et al.* (2022) find that 10th percentile of firm markups is between 0.91 and 0.97 and the 90th percentile of firm markups is between 1.36 and 2.97. Similarly, Forlani *et al.* (2023) (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.

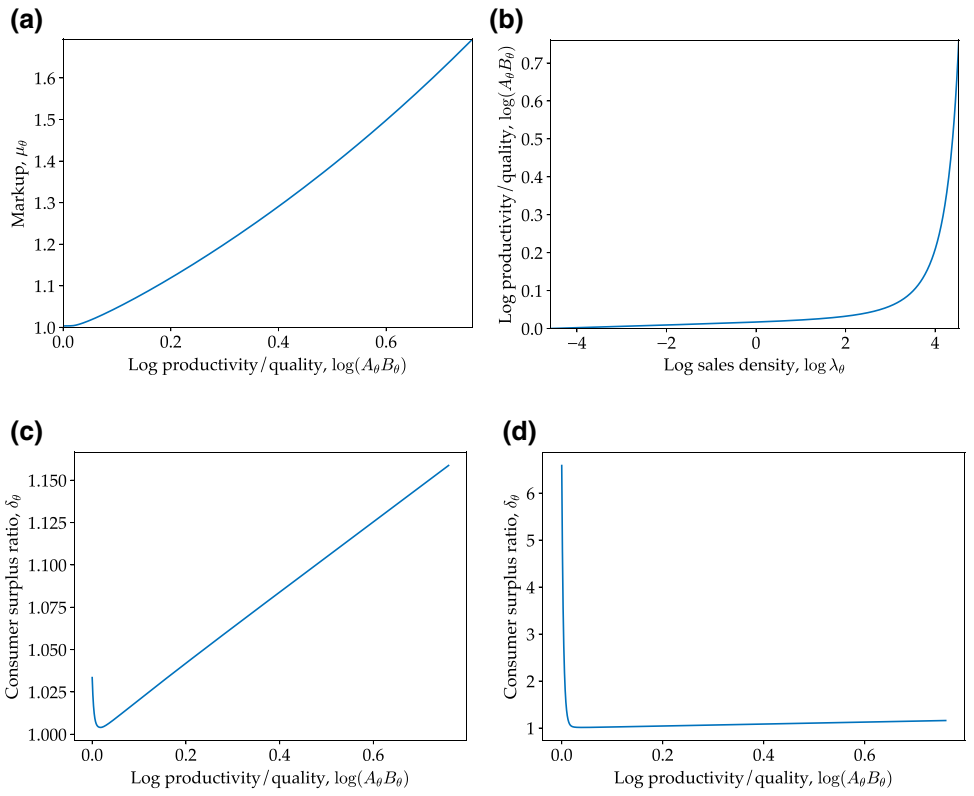

FIGURE 3

Markups and consumer surplus ratios with $\bar{\mu} = 1.090$. (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)

sales. This positive covariance between markups and firm size is consistent with evidence from Burstein *et al.* (2020), Grassi *et al.* (2022), and De Loecker *et al.* (2016).

Figure 3(b) 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. Figure 3(c) and (d) shows the consumer surplus ratio δ_θ for the efficient selection case ($\delta_{\theta^*} = \mathbb{E}_\lambda[\delta_\theta]$) and the efficient entry case ($\bar{\mu} = \mathbb{E}_\lambda[\delta_\theta]$). Supplementary Figure B.1 in Appendix B plots the residual demand curve and shows that it has a distinctly nonisoelastic 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—that is 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 procompetitive margins. As extensions, we compare macro- and micro-returns to scale and illustrate how increases in market size affect industrial concentration.

TABLE 1
The elasticity of welfare, real GDP per capita, and aggregate markup to population

	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
Procompetitive effect	−0.010	−0.099
Real GDP per capita	0.043	0.043
Aggregate markup	0.494	0.494

7.1. *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 procompetitive effects.

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 procompetitive 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 procompetitive 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 procompetitive effect and do not capture changes in consumer surplus due to entry and exit.

40. To understand the procompetitive 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 procompetitive effect is negative.

The aggregate markup increases with market size. As we discussed after Proposition 1, the aggregate markup increases if the Darwinian and the selection effects, 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 nonzero and negative at -0.344. The reason for this can be seen from Figure 3(d), 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 procompetitive effect is still negative and larger in magnitude at -0.099. The procompetitive 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 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 Supplementary 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 $\mathbb{E}_\lambda[\delta_\theta]$.

7.2. *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 inframarginal 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.

7.3. *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)

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

	$\delta = \delta_{\theta*}$	$\delta = \mu$
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

pass-through in the data, and use the same average markup and consumer surplus ratio as in Table 1. Table 2 shows the results.

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 procompetitive effect—are fairly small. Moreover, since only the procompetitive effect remains, the homogeneous-firm model predicts a falling, rather than rising, aggregate markup when the market expands.

7.4. *Are there larger increasing returns at the macrolevel versus microlevels?*

The microreturns 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 microreturns 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 microlevel 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 returns and micro returns would be identical, and if the economy had homogeneous firms, the difference between macro returns and micro returns would be much smaller.

7.5. *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

41. To produce these figures, we compute the equilibrium allocation nonlinearly by solving a system of differential equations. See Supplementary 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. Supplementary Figure B.3 in Appendix B also shows that the concentration of employment rises with market size in our quantitative calibration.

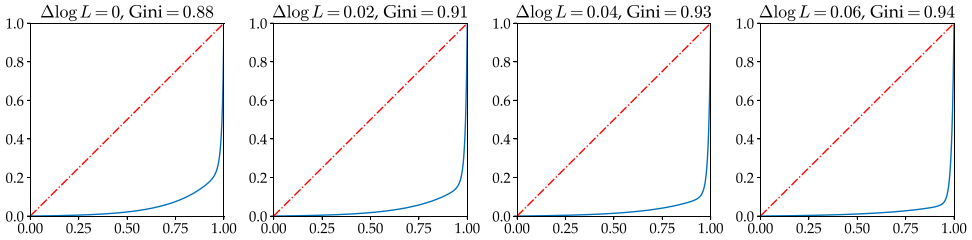

FIGURE 4

Lorenz curve for the sales distribution for different values of the market size parameter L. The dotted line indicates the line of perfect equality

TABLE 3

Welfare effect of an entry tax, following Proposition 3

	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
Procompetitive effect of entry tax	0.009	0.091
Welfare with homog. firms	0.027	0.000

(see Supplementary 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.

7.6. *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 wedge on entry efficiency, and the Darwinian, selection, and procompetitive 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.

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 procompetitive 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.

8.1. *Other generalizations of CES preferences*

In Supplementary 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.

Supplementary 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. Supplementary 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.

8.2. *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. Supplementary 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.

8.3. *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 Supplementary 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 and 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.

Supplementary 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.

8.4. *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. Supplementary 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.

8.5. *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 Supplementary 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—that is 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.

Acknowledgments. We thank the editor and four anonymous reviewers. 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.

Supplementary Data

Supplementary data are available at *Review of Economic Studies* online.

Data Availability Statement

The data and code underlying this research are available on Zenodo at https://doi.org/10.5281/zenodo.7926854.

REFERENCES

AGHION, P., BERGEAUD, A., BOPPART, T., *et al.* (2019), "Missing Growth from Creative Destruction", *American Economic Review*, **109**, 2795–2822.

AMITI, M., ITSKHOKI, O. and KONINGS, J. (2019), "International Shocks, Variable Markups, and Domestic Prices", *The Review of Economic Studies*, **86**, 2356–2402.

ARKOLAKIS, C., COSTINOT, A., DONALDSON, D., *et al.* (2019), "The Elusive Pro-Competitive Effects of Trade", *The Review of Economic Studies*, **86**, 46–80.

ASPLUND, M. and NOCKE, V. (2006), "Firm Turnover in Imperfectly Competitive Markets", *The Review of Economic Studies*, **73**, 295–327.

AUTOR, D., DORN, D., KATZ, L. F., *et al.* (2020), "The Fall of the Labor Share and the Rise of Superstar Firms", *The Quarterly Journal of Economics*, **135**, 645–709.

BAILY, M. N., HULTEN, C. and CAMPBELL, D. (1992), "Productivity Dynamics in Manufacturing Plants", *Brookings Papers on Economic Activity*, **23**, 187–267.

BAQAEE, D. R. and FARHI, E. (2019), "Productivity and Misallocation in General Equilibrium" (Technical Report, National Bureau of Economic Research).

BASU, S. and FERNALD, J. G. (1997), "Returns to Scale in US Production: Estimates and Implications", *Journal of Political Economy*, **105**, 249–283.

BILBIIE, F. O., GHIRONI, F. and MELITZ, M. J. (2012), "Endogenous Entry, Product Variety, and Business Cycles", *Journal of Political Economy*, **120**, 304–345.

BILBIIE, F. O., GHIRONI, F. and MELITZ, M. J. (2019), "Monopoly Power and Endogenous Product Variety: Distortions and Remedies", *American Economic Journal: Macroeconomics*, **11**, 140–74.

BURSTEIN, A., CARVALHO, V. M. and GRASSI, B. (2020), "Bottom-up Markup Fluctuations" (Technical Report 27958, National Bureau of Economic Research).

CHANEY, T. (2008), "Distorted Gravity: The Intensive and Extensive Margins of International Trade", *American Economic Review*, **98**, 1707–21.

CHANEY, T. and OSSA, R. (2013), "Market Size, Division of Labor, and Firm Productivity", *Journal of International Economics*, **90**, 177–180.

CORCOS, G., GATTO, M. D., MION, G., *et al.* (2012), "Productivity and Firm Selection: Quantifying the New Gains from Trade", *The Economic Journal*, **122**, 754–798.

DE LOECKER, J., EECKHOUT, J. and UNGER, G. (2020), "The Rise of Market Power and the Macroeconomic Implications", *The Quarterly Journal of Economics*, **135**, 561–644.

DE LOECKER, J., GOLDBERG, P. K., KHANDELWAL, A. K., *et al.* (2016), "Prices, Markups, and Trade Reform", *Econometrica*, **84**, 445–510.

GRASSI, B., DE RIDDER, M. and MORZENTI, G. (eds) (2022), "DP17532 The Hitchhiker's Guide to Markup Estimation" (CEPR Press Discussion Paper No. 17532). https://cepr.org/publications/dp17532

DHINGRA, S. and MORROW, J. (2019), "Monopolistic Competition and Optimum Product Diversity under Firm Heterogeneity", *Journal of Political Economy*, **127**, 196–232.

DIXIT, A. K. and STIGLITZ, J. E. (1977), "Monopolistic Competition and Optimum Product Diversity", *The American Economic Review*, **67**, 297–308.

EDMOND, C., MIDRIGAN, V. and XU, D. Y. (2015), "Competition, Markups, and the Gains from International Trade", *American Economic Review*, **105**, 3183–3221.

EDMOND, C., MIDRIGAN, V. and XU, D. Y. (2018), "How Costly are Markups?" (Technical Report, National Bureau of Economic Research).

EPIFANI, P. and GANCIA, G. (2011), "Trade, Markup Heterogeneity and Misallocations", *Journal of International Economics*, **83**, 1–13.

FEENSTRA, R. C. (2018), "Restoring the Product Variety and Pro-Competitive Gains from Trade with Heterogeneous Firms and Bounded Productivity", *Journal of International Economics*, **110**, 16–27.

FEENSTRA, R. C. and WEINSTEIN, D. E. (2017), "Globalization, Markups, and US Welfare", *Journal of Political Economy*, **125**, 1040–1074.

FORLANI, E., MARTIN, R., MION, G., *et al.* (2023), "Unraveling Firms: Demand, Productivity and Markups Heterogeneity", *The Economic Journal*, uead031. doi:10.1093/ej/uead031

FOSTER, L., HALTIWANGER, J. C. and KRIZAN, C. J. (2001), *New Developments in Productivity Analysis* (University of Chicago Press).

GILCHRIST, S., SCHOENLE, R., SIM, J., *et al.* (2017), "Inflation Dynamics during the Financial Crisis", *American Economic Review*, **107**, 785–823.

HELPMAN, E. and KRUGMAN, P. R. (1985), *Market Structure and Foreign Trade: Increasing Returns, Imperfect Competition, and the International Economy* (Cambridge, MA: MIT Press).

JOHNSON, J. P. and MYATT, D. P. (2006), "On the Simple Economics of Advertising, Marketing, and Product Design", *American Economic Review*, **96**, 756–784.

KEHRIG, M. and VINCENT, N. (2021), "The Micro-Level Anatomy of the Labor Share Decline", *The Quarterly Journal of Economics*, **136**, 1031–1087.
KIMBALL, M. (1995), "The Quantitative Analytics of the Basic Neomonetarist Model", *Journal of Money, Credit and Banking*, **27**, 1241–77.
KLENOW, P. J. and WILLIS, J. L. (2016), "Real Rigidities and Nominal Price Changes", *Economica*, **83**, 443–472.
KRUGMAN, P. R. (1979), "Increasing Returns, Monopolistic Competition, and International Trade", *Journal of International Economics*, **9**, 469–479.
LIPSEY, R. G. and LANCASTER, K. (1956), "The General Theory of Second Best", *The Review of Economic Studies*, **24**, 11–32.
MANKIW, N. G. and WHINSTON, M. D. (1986), "Free Entry and Social Inefficiency", *RAND Journal of Economics*, **17**, 48–58.
MATSUYAMA, K. and USHCHEV, P. (2017), "Beyond CES: Three Alternative Classes of Flexible Homothetic Demand Systems" (Working Paper, Northwestern University).
MATSUYAMA, K. and USHCHEV, P. (2020a), "Constant Pass-Through".
MATSUYAMA, K. and USHCHEV, P. (2020b), "When Does Procompetitive Entry Imply Excessive Entry?" (CEPR Discussion Paper No. DP14991). https://ssrn.com/abstract=3650105
MATSUYAMA, K. and USHCHEV, P. (2022), "Selection and Sorting of Heterogeneous Firms through Competitive Pressures" (Working Paper, Northwestern University).
MAYER, T., MELITZ, M. J. and OTTAVIANO, G. I. (2014), "Market Size, Competition, and the Product Mix of Exporters", *American Economic Review*, **104**, 495–536.
MELITZ, M. J. (2003), "The Impact of Trade on Intra-Industry Reallocations and Aggregate Industry Productivity", *Econometrica*, **71**, 1695–1725.
MELITZ, M. J. and OTTAVIANO, G. I. P. (2008), "Market Size, Trade, and Productivity", *The Review of Economic Studies*, **75**, 295–316.
MELITZ, M. J. and POLANEC, S. (2015), "Dynamic Olley-Pakes Productivity Decomposition with Entry and Exit", *The RAND Journal of Economics*, **46**, 362–375.
MELITZ, M. J. and REDDING, S. J. (2015), "New Trade Models, New Welfare Implications", *American Economic Review*, **105**, 1105–46.
MRÁZOVÁ, M. and NEARY, J. P. (2017), "Not So Demanding: Demand Structure and Firm Behavior", *American Economic Review*, **107**, 3835–74.
MRÁZOVÁ, M. and NEARY, J. P. (2019), "IO For Export(s)".
OLLEY, G. S. and PAKES, A. (1996), "The Dynamics of Productivity in the Telecommunications Equipment Industry", *Econometrica*, **64**, 1263–1297.
PAVCNIK, N. (2002), "Trade Liberalization, Exit, and Productivity Improvements: Evidence from Chilean Plants", *The Review of Economic Studies*, **69**, 245–276.
PETRIN, A. and LEVINSOHN, J. (2012), "Measuring Aggregate Productivity Growth Using Plant-Level Data", *The RAND Journal of Economics*, **43**, 705–725.
STERK, V., SEDLÁČEK, P. and PUGSLEY, B. (2021), "The Nature of Firm Growth", *American Economic Review*, **111**, 547–579.
RAVN, M., SCHMITT-GROHÉ, S. and URIBE, M. (2006), "Deep Habits", *The Review of Economic Studies*, **73**, 195–218.
SPENCE, M. (1976), "Product Selection, Fixed Costs, and Monopolistic Competition", *The Review of Economic Studies*, **43**, 217–235.
TREFLER, D. (2004), "The Long and Short of the Canada-US Free Trade Agreement", *American Economic Review*, **94**, 870–895.
VENABLES, A. J. (1985), "Trade and Trade Policy with Imperfect Competition: The Case of Identical Products and Free Entry", *Journal of International Economics*, **19**, 1–19.
VIVES, X. (2001), *Oligopoly Pricing: Old Ideas and New Tools* (Cambridge, MA: MIT Press).
ZHELOBODKO, E., KOKOVIN, S., PARENTI, M., *et al.* (2012), "Monopolistic Competition: Beyond the Constant Elasticity of Substitution", *Econometrica*, **80**, 2765–2784.