Monopolistic Competition and Optimum Product Diversity Under Firm Heterogeneity*

Swati Dhingra
CEP, London School of Economics & CEPR

John Morrow
CEP, London School of Economics

This Draft: June 30, 2014

Abstract

Empirical work has drawn attention to the high degree of productivity differences within industries, and its role in resource allocation. This paper examines the allocational efficiency of such markets. Productivity differences introduce two new margins of potential inefficiency: selection of the right distribution of firms and allocation of the right quantities across firms. We show that these considerations impact welfare and policy analysis. Market power across firms leads to distortions in resource allocation. Demand-side elasticities determine how resources are misallocated and when increased competition from market expansion provides welfare gains.

JEL Codes: F1, L1, D6.
Keywords: Efficiency, Productivity, Social welfare, Demand elasticity, Markups.

Acknowledgments. We thank Bob Staiger for continued guidance and Steve Redding for encouragement. We are grateful to G Alessandria, C Arkolakis, R Armenter, A Bernard, S Chatterjee, D Chor, S Durlauf, C Engel, T Fally, R Feenstra, K Head, W Keller, J Lin, E Ornelas, G Ottaviano, M Parenti, N Pavcnik, T Sampson, D Sturm, J Thisse, J Van Reenen, A Weinberger, B Zissimos and M Zhu for insightful comments, K Russ and A Rodriguez-Clare for AEA discussions and T Besley for advice. This paper has benefited from helpful comments of participants at AEA 2011 and 2013, CEPR-ERWIT, CEPR-IO, Columbia, Davis, DIME-ISGEP 2010, ETSG 2012, Georgetown, Harvard KSG, HSE St Petersburg, ISI, FIW, LSE, Louvain, Mannheim, Maryland, NBER, Oxford, Philadelphia Fed, Princeton, Toronto, Virginia Daarden, Wisconsin and Yale. Swati thanks the IES (Princeton) for their hospitality. We acknowledge the financial support from Portuguese national funds by FCT (Fundacao para a Ciencia e a Tecnologia) project PTDC/EGE-ECO/122115/2010. A preliminary draft was a dissertation chapter at Wisconsin in 2010.

*The first line is the title of Dixit and Stiglitz (1977). Contact: s.dhingra@lse.ac.uk; j.morrow1@lse.ac.uk.

1 Introduction

Empirical work has drawn attention to the high degree of heterogeneity in firm productivity, and the constant reallocation of resources across different firms.[1] The focus on productivity differences has provided new insights into market outcomes such as industrial productivity, firm pricing and welfare gains from policy changes.[2] When firms differ in productivity, the distribution of resources across firms also affects the allocational efficiency of markets. In a recent survey, Syverson (2011) notes the gap between social benefits and costs across firms has not been adequately examined, and this limited understanding has made it difficult to implement policies to reduce distortions (pp. 359). This paper examines allocational efficiency in markets where firms differ in productivity. We focus on three key questions. First, does the market allocate resources efficiently? Second, what is the nature of distortions, if any? Third, can economic integration reduce distortions through increased competition?

Symmetric firm models explain when resource allocation is efficient by examining the trade-off between quantity and product variety in imperfectly competitive markets.[3] When firms differ in productivity, we must also ask which types of firms should produce and which should be shut down. Firm differences in productivity introduce two new margins of potential inefficiency: selection of the right distribution of firms and allocation of the right quantities across firms. For example, it could be welfare-improving to skew resources towards firms with lower costs (to conserve resources) or towards firms with higher costs (to preserve variety). Furthermore, differences in market power across firms lead to new trade-offs between variety and quantity. These considerations impact optimal policy rules in a fundamental way, distinct from markets with symmetric costs. One contribution of the paper is to understand how these considerations affect welfare and policy analysis.

A second contribution of the paper is to show when increased competition improves welfare and efficiency. When market allocations are inefficient, increased competition (from trade or growth) may exacerbate distortions and lead to welfare losses (Helpman and Krugman 1985). A second-best world offers no guarantee of welfare gains from trade. But, by creating larger, more competitive markets, trade may reduce the distortions associated with imperfect competition and provide welfare gains (Krugman 1987). This insight is even more relevant in a heterogeneous cost environment because of new sources of potential inefficiency. We explain when integration provides welfare gains by aligning private and social incentives. As a benchmark,

[1] Example, Bartelsman and Doms (2000); Tybout (2003); Feenstra (2006); Bernard, Jensen, Redding and Schott (2007).

[2] Example, Pavcnik (2002); Asplund and Nocke (2006); Foster et al. (2001); Melitz and Redding (2012).

[3] Example, Spence (1976); Venables (1985); Mankiw and Whinston (1986); Stiglitz (1986).

we show integration with large world markets provides a policy option to correct distortions.[4]

To understand efficiency in general equilibrium, we examine resource allocation in the standard setting of a monopolistically competitive industry with heterogeneous firm productivity and free entry (e.g. Melitz 2003). We begin our analysis by considering constant elasticity of substitution (CES) demand. In this setting, we show market allocations are efficient, despite differences in firm productivity. This is striking, as it requires the market to induce optimal resource allocations across aggregate variety, quantity and productivity. As in symmetric firm models, there are two sources of potential inefficiency: the inability of firms to appropriate the full consumer surplus and to account for business stealing from other firms. CES demand uniquely ensures these two externalities exactly offset each other. Firm heterogeneity does not introduce any new distortions because the magnitude of these externalities does not vary across firms. Firms earn positive profits which seems surprising based on the logic of average cost pricing that is designed to return producer surplus to consumers. When productivity differs, the market requires prices above average costs to induce firms to enter and potentially take a loss. Free entry ensures the wedge between prices and average costs exactly finances sunk entry costs, and positive profits are efficient. Therefore, the market implements the first-best allocation and laissez faire industrial policy is optimal.[5]

What induces market efficiency and how broadly does this result hold? We generalize the demand structure to the variable elasticity of substitution form of Dixit and Stiglitz (1977), which provides a rich setting for a wide range of market outcomes (Vives 2001; Zhelobodko, Kokovin, Parenti and Thisse 2012). When demand elasticity varies with quantity and firms vary in productivity, markups vary within a market. This accounts for the stylized facts that firms are rarely equally productive and markups are unlikely to be constant.[6] Introducing this empirically relevant feature of variable elasticities turns out to be crucial in understanding distortions. When elasticities vary, firms differ in market power and market allocations reflect the distortions of imperfect competition. Nonetheless, we show the market maximizes real revenues. This is similar to perfect competition models, but now market power implies private benefits to firms

[4]International integration is equivalent to an expansion in market size (e.g., Krugman 1979). As our focus is on efficiency, we abstract from trade frictions which introduce cross-country distributional issues.

[5]Melitz (2003) considers both variable and fixed costs of exporting. In a separate note, we show that the open Melitz economy is efficient, even with trade frictions. In the presence of fixed export costs, the firms a policymaker would close down in the open economy are exactly those that would not survive in the market. However, a policymaker would not close down firms in the absence of export costs. Thus, the rise in productivity following trade provides welfare gains by optimally internalizing trade frictions.

[6]CES demand provides a useful benchmark by forcing constant markups that ensure market size plays no role in productivity changes. However, recent studies find market size matters for firm size (Campbell and Hopenhayn 2005) and productivity dispersion (Syverson 2004). Foster, Haltiwanger and Syverson (2008) show that "profitability" rather than productivity is more important for firm selection, suggesting a role for richer demand specifications. For further discussion, see Melitz and Trefler (2012).

are perfectly aligned with social benefits only under CES demand. More generally, the appropriability and business stealing effects need not exactly offset each other, and firm heterogeneity introduces a new source of potential inefficiency. When firm differ in productivity, entry of an additional variety shifts business across the entire distribution of firms and induces distortions relative to optimal allocations.

The pattern of distortions is determined by two elasticities: the demand elasticity, which measures market incentives through markups, and the elasticity of utility, which measures social incentives through a firm's contribution to welfare. We show that the way in which these incentives differ characterizes the precise nature of misallocations. This also yields two new insights relating productivity differences to misallocations. First, differences in market power across firms imply misallocations are not uniform: some firms over-produce while others underproduce within the same market. For instance, the market may favor excess entry of low productivity firms, thereby imposing an externality on high productivity firms who end up producing too little. Second, differences in market power impact economy-wide outcomes. The distribution of markups affects ex ante profitability, and therefore the economy-wide trade-off between aggregate quantity and variety. This is in sharp contrast to symmetric firm markets, where markups (or demand elasticities) do not matter for misallocations, as emphasized by Dixit and Stiglitz (1977) and Vives (2001). Differences in productivity underline the importance of demand elasticity for allocational efficiency, and complement the message of Weyl and Fabinger (2012) and Parenti et al. (2014) that richer demand systems enable a better understanding of market outcomes.

As misallocations vary by firm productivity, one potential policy option that does not require firm-level information is international integration. The idea of introducing foreign competition to improve efficiency goes back at least to Melvin and Warne (1973). We show that market integration always provides welfare gains when private and social incentives are aligned, which again is characterized by the demand elasticity and the elasticity of utility. This result ties the Helpman-Krugman characterization of gains from trade to the welfare approach of Spence-Dixit-Stiglitz. Symmetric firm models with CES demand provide a lower bound for the welfare gains from integration. Gains from trade under aligned preferences are higher due to selection of the right distribution of firms and and allocation of the right quantities across firms. As a benchmark for understanding efficiency gains, we follow the literature on imperfect competition in large markets and examine whether integration with large global markets leads to allocative efficiency (Vives 2001, Chapter 6). Integration with large markets will push outcomes towards a new concept, the "CES limit", where firms converge to charging constant markups. Unlike a perfectly competitive limit (Hart 1985), productivity dispersion and market power persist in the

CES limit. Yet the market is efficient and integration with large global markets is therefore a first-best policy to eliminate the distortions of imperfect competition. However, as the limit may require a market size which is unattainable even in fully integrated world markets, integration may be an incomplete tool to reduce distortions.

Related Work. Our paper is related to work on firm behavior and welfare in industrial organization and international economics. As mentioned earlier, the trade-off between quantity and variety occupies a prominent place in the study of imperfect competition. We contribute to this literature by studying these issues in markets where productivity differences are important. To highlight the potential scope of market imperfections, we consider variable elasticity of substitution (VES) demand. In contemporaneous work, Zhelobodko et al. (2012)demonstrate the richness and tractability of VES market outcomes under various assumptions such as multiple sectors and vertical differentiation.[7] The focus on richer demand systems is similar to Weyl and Fabinger (2012) who characterize several industrial organization results in terms of pass-through rates. Unlike these papers, we examine the efficiency of market allocations, so our findings depend on both the elasticity of utility and the demand elasticity. To the best of our knowledge, this is the first paper to show market outcomes with heterogeneous firms are first-best under CES demand.[8]

The findings of our paper are also related to a tradition of work on welfare gains from trade. Helpman and Krugman (1985) and Dixit and Norman (1988) examine when trade is beneficial under imperfect competition. We generalize their finding and link it to model primitives of demand elasticities, providing new results even in the symmetric firm literature. In recent influential work, Arkolakis et al. (2012a,b) show richer models of firm heterogeneity and variable markups are needed for these microfoundations to affect welfare gains from trade. In line with this insight, we generalize the demand structure and show that firm heterogeneity and variable

[7]While VES utility does not include the quadratic utility of Melitz and Ottaviano (2008) and the translog utility of Feenstra (2003), Zhelobodko et al. show it captures the qualitative features of market outcomes under these forms of non-additive utility.

[8]We consider this to be the proof of a folk theorem which has been "in the air." Matsuyama (1995) and Bilbiie, Ghironi and Melitz (2006) find the market equilibrium with symmetric firms is socially optimal only when preferences are CES. Epifani and Gancia (2011) generalize this to multiple sectors while Eckel (2008) examines efficiency when firms affect the price index. Within the heterogeneous firm literature, Baldwin and Robert-Nicoud (2008) and Feenstra and Kee (2008) discuss certain efficiency properties of the Melitz economy. In their working paper, Atkeson and Burstein (2010) consider a first order approximation and numerical exercises to show productivity increases are offset by reductions in variety. We provide an analytical treatment to show the market equilibrium implements the unconstrained social optimum. Helpman, Itskhoki and Redding (2011) consider the constrained social optimum. Their approach differs because the homogeneous good fixes the marginal utility of income. Our work is closest to Feenstra and Kee who focus on the CES case. Considering 48 countries exporting to the US in 1980-2000, they also estimate that rise in export variety accounts for an average 3.3 per cent rise in productivity and GDP for the exporting country.

markups matter for both welfare gains and allocational efficiency.[9] As in Melitz and Redding (2013), we find that the cost distribution matters for the magnitude of welfare gains from integration. Building on Bernard, Eaton, Jensen and Kortum (2003), de Blas and Russ (2010) also examine the role of variable markups in welfare gains but do not consider efficiency. We follow the direction of Tybout (2003) and Katayama, Lu and Tybout (2009) who suggest the need to map productivity gains to welfare and optimal policies.

The paper is organized as follows. Section 2 recaps the standard monopolistic competition framework with firm heterogeneity. Section 3 contrasts efficiency of CES demand with inefficiency of VES demand and Section 4 characterizes the distortions in resource allocation. Section 5 examines welfare gains from integration, deriving a limit result for large markets. Section 6 concludes.

2 Model

We adopt the VES demand structure of Dixit and Stiglitz within the heterogeneous firm framework of Melitz. Monopolistic competition models with heterogeneous firms differ from earlier models with product differentiation in two significant ways. First, costs of production are unknown to firms before sunk costs of entry are incurred. Second, firms are asymmetric in their costs of production, leading to firm selection based on productivity. This Section lays out the model and recaps the implications of asymmetric costs for consumers, firms and equilibrium outcomes.

2.1 Consumers

We explain the VES demand structure and then discuss consumer demand. The exposition for consumer demand closely follows Zhelobodko et al. (2012) which works with a similar setting and builds on work by Vives (2001).

An economy consists of a mass L of identical workers, each endowed with one unit of labor and facing a wage rate w normalized to one. Workers have identical preferences for a differentiated good. The differentiated good is made available as a continuum N of horizontally differentiated varieties indexed by $i \in [0,N]$. Given prices p_i for the varieties, every worker chooses quantity q_i for each of the varieties to maximize her utility subject to her budget constraint.

[9] For instance, linear VES demand and Pareto cost draws fit the gravity model, but firm heterogeneity still matters for market efficiency. More generally, VES demand is not nested in the Arkolakis et al. models and does not satisfy a log-linear relation between import shares and welfare gains, as illustrated in the Online Appendix.

Preferences over differentiated goods take the general VES form:

$$U(\mathbf{q}) \equiv \int_0^N u(q_i)di \tag{1}$$

where $u(\cdot)$ is thrice continuously differentiable, strictly increasing and strictly concave on $(0,\infty)$, and $u(0)$ is normalized to zero. The concavity of u ensures consumers love variety and prefer to spread their consumption over all available varieties. Here $u(q_i)$ denotes utility from an individual variety i. Under CES preferences, $u(q_i) = q_i^\rho$ as specified in Dixit-Stiglitz and Krugman (1980).[10]

For each variety i, VES preferences induce an inverse demand $p(q_i) = u'(q_i)/\delta$ where δ is the consumer's budget multiplier. As u is strictly increasing and concave, for any fixed price vector the consumer's maximization problem is concave. The necessary condition which determines the inverse demand is sufficient, and has a solution provided inada conditions on u.[11] Multiplying both sides of the inverse demand by q_i and aggregating over all i, the budget multiplier is $\delta = \int_0^N u'(q_i) \cdot q_i di$. The consumer budget multiplier δ will act as a demand shifter and the inverse demand will inherit the properties of the marginal utility $u'(q_i)$. In particular, the inverse demand elasticity $|d\ln p_i / d\ln q_i|$ equals the elasticity of marginal utility $\mu(q_i) \equiv |q_i u''(q_i)/u'(q_i)|$, which enables us to characterize market allocations in terms of demand primitives. Under CES preferences, the elasticity of marginal utility is constant and the inverse demand elasticity does not respond to consumption ($|d\ln p_i / d\ln q_i| = \mu(q_i) = 1-\rho$). When $\mu'(q_i) > 0$, the inverse demand of a variety becomes more elastic as its consumption increases. The opposite holds for $\mu'(q_i) < 0$, where the demand for a variety becomes less elastic as its price rises.

The inverse demand elasticity summarizes market demand, and will enable a characterization of market outcomes. A policymaker maximizes utility, and is not concerned with market prices. Therefore, we define the elasticity of utility $\varepsilon(q_i) \equiv u'(q_i)q_i/u(q_i)$, which will enable a characterization of optimal allocations. The elasticity of utility can be understood as follows. The real expenditure on variety i is $u'(q_i)q_i$ and the contribution of variety i to welfare is $u(q_i)$. Therefore, $1-\varepsilon(q_i) = \left(u(q_i) - u'(q_i)q_i\right)/u(q_i)$ denotes the proportion of social benefits not captured by real expenditure when introducing variety i. Under CES preferences, the elasticity of utility is constant and $1-\varepsilon(q_i) = 1-\rho$. For $(1-\varepsilon(q_i))' < 0$, the welfare contribution of a

[10] The specific CES form in Melitz is $U(\mathbf{q}) \equiv \left(\int (q_i^\rho di\right)^{1/\rho}$ but the normalization of the exponent $1/\rho$ in Equation (1) will not play a role in allocation decisions.

[11] Additional assumptions to guarantee existence and uniqueness of the market equilibrium are in a separate note available online. Utility functions not satisfying inada conditions are permissible but may require parametric restrictions to ensure existence.

variety relative to expenditure is more elastic when its consumption is low. For $(1-\varepsilon(q_i))' > 0$, the welfare contribution of a variety is more sensitive when more of it is consumed. We discuss the interpretation of these elasticities in more detail.

2.1.1 Interpretation of Elasticities

Zhelobodko et al. (2012) show that the elasticity of marginal utility $\mu(q_i)$ can also be interpreted in terms of substitution across varieties. For symmetric consumption levels ($q_i = q$), this elasticity equals the inverse of the elasticity of substitution between any two varieties. For $\mu'(q) > 0$, higher consumption per variety or fewer varieties for a given total quantity, induces a lower elasticity of substitution between varieties. Consumers perceive varieties as being less differentiated when they consume more, but this relationship does not carry over to heterogeneous consumption levels.

For symmetric consumption levels, Vives (2001) points out that $1-\varepsilon(q)$ is the degree of preference for variety as it measures the proportion of the utility gain from adding a variety, holding quantity per firm fixed. Extrapolating to heterogeneous varieties, $1-\varepsilon(q)$ measures the relative contribution of variety to total utility from adding another variety, holding the average quantity level q and the dispersion of quantities across varieties fixed. If $1-\varepsilon(q) = 0$, there is no preference for variety, and the composition of consumption is irrelevant for welfare. If $1-\varepsilon(q) = 1$, utility depends only on variety, not quantity per variety. For $(1-\varepsilon(q))' > 0$, consumers have a higher preference for variety when they consume more per variety. This can be explained in a framework following Kuhn and Vives (1999). Utility can be re-written to explicitly account for taste for variety, $U \equiv NqV(q)$ for q such that $\int q_i V(q_i)di = qV(q) \equiv (Q/N)V(Q/N)$ where Q is total quantity. Holding average quantity q fixed, adding a variety increases utility by $dU/dN = qV(q)$. This gain consists of a pure variety effect on welfare, holding total quantity fixed: $dU/dN = QV'(q)\left(-Q/N^2\right)$. Utility also rises due to an increase in total quantity, holding variety fixed: $dU/dQ = [V(Q/N) + QV'(Q/N)/N](dQ/dN)$. Since the total quantity increase is $dQ/dN = q$, the output effect is given by $dU/dQ = V(q)q[1+V'(q)q/V(q)]$. The two effects add up to give the total effect of adding a variety at constant quantity per firm. The ratio of the variety effect to the total utility gain from adding a variety equals $1-\varepsilon(q) = -V'(q)q/V(q)$ at the average quantity q.

2.2 Firms

There is a continuum of firms which may enter the market for differentiated goods, by paying a sunk entry cost of $f_e > 0$. The mass of entering firms is denoted by M_e. Firms are monop-

olistically competitive and each firm produces a single unique variety. A firm faces an inverse demand of $p(q_i) = u'(q_i)/\delta$ for variety i. It acts as a monopolist of its unique variety but takes aggregate demand conditions δ as given. Upon entry, each firm receives a unit cost $c \geq 0$ drawn from a distribution G with continuously differentiable pdf g. Each variety can therefore be indexed by the unit cost c of its producer.

After entry, should a firm produce, it incurs a fixed cost of production $f > 0$. Profit maximization implies firms produce if they can earn non-negative profits net of the fixed costs of production. A firm with cost draw c chooses its quantity $q(c)$ to $\max_{q(c)}[p(q(c)) - c]q(c)L$ and $q(c) > 0$ if $\pi(c) = \max_{q(c)}[p(q(c)) - c]q(c)L - f > 0$. To ensure the firm's quantity FOC is optimal, we assume marginal revenue is strictly decreasing in quantity and the elasticity of marginal utility $\mu(q) = |qu''(q)/u'(q)|$ is less than one. A firm chooses its quantity to equate marginal revenue and marginal cost ($p + q \cdot u''(q)/\delta = c$), and concavity of the firm problem ensures low cost firms supply higher quantities and charge lower prices.

The markup charged by a firm with cost draw c is $(p(c) - c)/p(c) = -q(c)u''(q(c))/u'(q(c))$. This shows that the elasticity of marginal utility $\mu(q)$ summarizes the markup:

$$\mu(q(c)) = |q(c)u''(q(c))/u'(q(c))| = (p(c) - c)/p(c).$$

When $\mu'(q) > 0$, low cost firms supply higher quantities at higher markups.

2.3 Market Equilibrium

Profits fall with unit cost c, and the cutoff cost level of firms that are indifferent between producing and exiting from the market is denoted by c_d. The cutoff cost c_d is fixed by the zero profit condition, $\pi(c_d) = 0$. Firms with cost draws higher than the cutoff level earn negative profits and do not produce. The mass of producing firms in equilibrium is therefore $M = M_e G(c_d)$.

In summary, each firm faces a two stage problem: in the second stage it maximizes profits given a known cost draw, and in the first stage it decides whether to enter given the expected profits in the second stage. To study the Chamberlinian tradeoff between quantity and variety, we maintain the standard free entry condition imposed in monopolistic competition models. Specifically, ex ante average profit net of sunk entry costs must be zero, $\int_0^{c_d} \pi(c) dG = f_e$. This free entry condition along with the consumer's budget constraint ensures that the resources used by firms equal the total resources in the economy, $L = M_e \left[\int_0^{c_d} (cq(c)L + f) dG + f_e\right]$.

2.4 Social Optimum

To assess the efficiency of resource allocation in the market equilibrium, we now describe the policymaker's optimal allocation. A policymaker maximizes individual welfare U as given in Equation (1) by choosing the mass of entrants, quantities and types of firms that produce.[12] The policymaker can choose any allocation of resources that does not exceed the total resources in the economy. However, she faces the same entry process as for the market: a sunk entry cost f_e must be paid to get a unit cost draw from $G(c)$. Fixed costs of production imply that the policymaker chooses zero quantities for varieties above a cost threshold. Therefore, all optimal allocation decisions can be summarized by quantity $q(c)$, potential variety M_e and a productivity cutoff c_d. The policymaker chooses $q(c)$, c_d and M_e to

$$\max M_e \int_0^{c_d} u(q(c))dG \text{ where } L \geq M_e \left\{ \int_0^{c_d} [cq(c)L + f]dG + f_e \right\}.$$

Our approach for arriving at the optimal allocation is to think of optimal quantities $q^{\text{opt}}(c)$ as being determined implicitly by c_d and M_e so that per capita welfare can be written as

$$U = M_e \int_0^{c_d} u(q^{\text{opt}}(c))dG. \tag{2}$$

Optimal quantities ensure marginal utility equals social marginal cost of a variety, $u'(q^{\text{opt}}(c)) = \lambda c$ where λ is the resource multiplier for fixed c_d and M_e. Note that $q(c)$ is a function of λc that maximizes U and depends on both the distribution of costs and aggregate entry decisions. Fixing the optimal λ and showing sufficiency of such candidate quantity functions is handled using variational calculus techniques in the Appendix. After solving for each q^{opt} conditional on c_d and M_e, Equation (2) can be maximized in c_d and M_e. Of course, substantial work is involved in showing sufficiency, but we relegate this to the Appendix. The next two Sections compare the market and optimal allocations in this framework.

3 Market Efficiency

Having described an economy consisting of heterogeneous, imperfectly competitive firms, we now examine efficiency of market allocations. Outside of cases in which imperfect competition leads to competitive outcomes with zero profits, one would expect the coexistence of positive markups and positive profits to indicate inefficiency through loss of consumer surplus. Nonethe-

[12] Free entry implies zero expected profits, so the focus is on consumer welfare.

less, this Section shows that CES demand under firm heterogeneity exhibits positive markups and profits for surviving firms, yet it is allocationally efficient. However, this is a special case. Private incentives are not aligned with optimal production patterns for any VES demand structure except CES. Following Dixit and Stiglitz, we start with efficiency under CES demand and then explain market inefficiency under VES demand. We then discuss the externalities arising in the market and the reasons for efficiency under CES demand.

3.1 Market and Optimal Allocations

Proposition 1 shows the market provides the first-best quantity, variety and productivity. The proof of Proposition 1 differs from symmetric firm monopolistic competition results because optimal quantity varies non-trivially with unit cost, variety and cutoff productivity. The main finding is that laissez faire industrial policy is optimal under CES demand.

Proposition 1. *Every market equilibrium of a CES economy is socially optimal.*

Proposition 1 shows that the market allocation is optimal under CES demand and we now contrast the market allocation across symmetric and heterogeneous firms. When firms are symmetric, resource allocation reflects average cost pricing. Firms charge positive markups which result in lower quantities than those implied by marginal cost pricing. Even though firms do not charge marginal costs, their market price (and hence marginal utility) is proportional to marginal cost because markups are constant. This ensures proportionate reductions in quantity from the level that would be observed under marginal cost pricing (Baumol and Bradford 1970). These reduced quantity levels are efficient because the marginal utility of income adjusts to ensure that the ratio of marginal utility to marginal cost of a variety coincides with the social value of labor ($u'(q)/c = \delta/(1-\mu) = \lambda$). Free entry equates price to average cost, and the markup exactly finances the fixed cost of an additional variety. The market therefore induces an efficient allocation.

With heterogeneous firms, markups continue to be constant and marginal utility is proportional to marginal cost. One might infer enforcing average cost pricing across different firms would induce an efficient allocation, as in symmetric firm models. But average cost pricing is too low to compensate firms because it will not cover ex ante entry costs. The market ensures prices above average costs at a level that internalizes the losses faced by exiting firms. Entry is at optimal levels that fix $p(c_d)$, thereby fixing absolute prices to optimal levels. Post entry, surviving firms charge prices higher than average costs ($p(c) \geq [cq(c) + f/L]/q(c)$) and the markups exactly compensate them for the possibility of paying f_e to enter and then being too unproductive to survive.

The way in which CES preferences cause firms to optimally internalize aggregate economic conditions can be made clear through a variety-specific explanation. The elasticity of utility $\varepsilon(q) \equiv u'(q) \cdot q/u(q)$ can be used to define a "social markup" $1-\varepsilon(q)$. We term $1-\varepsilon(q)$ the social markup because it denotes the utility from consumption of a variety net of its resource cost. At the optimal allocation, the multiplier λ encapsulates the social value of labor and the social surplus from a variety is $u(q)-\lambda cq$. At the optimal quantity, $u'(q(c)) = \lambda c$ and the social markup is

$$1-\varepsilon(q) = 1-u'(q)\cdot q/u(q) = (u(q)-\lambda cq)/u(q). \quad \text{(Social Markup)}$$

For any optimal allocation, the quantity that maximizes social benefit from variety c solves

$$\max_q (u(q)/\lambda - cq)L - f = \frac{1-\varepsilon(q^{\text{opt}}(c))}{\varepsilon(q^{\text{opt}}(c))} cq^{\text{opt}}(c)L - f.$$

In contrast, the incentives that firms face in the market are based on the private markup $\mu(q) = (p(q)-c)/p(q)$, and firms solve:

$$\max_q (p(q)q - cq)L - f = \frac{\mu(q^{\text{mkt}}(c))}{1-\mu(q^{\text{mkt}}(c))} cq^{\text{mkt}}(c)L - f.$$

Since ε and μ depend only on the primitive $u(q)$, we can examine what demand structures would make the economy optimally select firms. Clearly, if private markups $\mu(q)$ coincide with social markups $1-\varepsilon(q)$, "profits" will be the same at every unit cost. Examining CES demand, we see precisely that $\mu(q) = 1-\varepsilon(q)$ for all q. Thus, CES demand incentivizes exactly the right firms to produce. Since the optimal set of firms produce under CES demand, and private and social profits are the same, market entry will also be optimal. As entry M_e and the cost cutoff c_d are optimal, the competition between firms aligns the budget multiplier δ to ensure optimal quantities.

Efficiency of the market equilibrium in our framework is tied to CES demand. To highlight this, we consider the general class of VES demand specified in Equation (1). Direct comparison of FOCs for the market and optimal allocation shows constant markups are necessary for efficiency. Therefore, within the VES class, optimality of market allocations is unique to CES preferences.

Proposition 2. *Under VES demand, a necessary condition for the market equilibrium to be socially optimal is that u is CES.*

Proof. Online Appendix. □

Under general VES demand, market allocations are not efficient and do not maximize individual welfare. Proposition 3 shows that the market instead maximizes aggregate real revenue ($M_e \int u'(q(c)) \cdot q(c) dG$) generated in the economy.

Proposition 3. *Under VES demand, the market maximizes aggregate real revenue.*

Proposition 3 shows decentralized profit maximization coincides with centralized revenue maximization. While firms have no individual influence over entry M_e or consumers' marginal utility of income δ, they do have decentralized control over quantities $q(c)$ and the decision whether to produce at all. A shadow value of labor $\hat{\delta}$ from a policymaker who wishes to maximize real revenue acts exactly like δ, since firms solve $\max_q L\left[u'(q)/\delta - c\right]q$ while the policymaker solves $\max_q L\left[u'(q) - \hat{\delta}c\right]q$ and clearly this results in the same (individual) quantity and production decisions at $\delta = \hat{\delta}$. Therefore decentralized profit maximization coincides with centralized revenue maximization if the marginal utility of income and shadow value of labor happen to coincide, conditional on equivalent entry. That $\delta = \hat{\delta}$ happens in the marketplace comes not from firms (who take δ as exogenous), but from consumers who internalize aggregate firm decisions and identify their marginal utility of income with the real value of their labor. That entry in the market matches entry chosen by a revenue maximizing policymaker comes from the ex ante decisions of firms which aggregates market outcomes through rational expectations.

This result shows that the market and optimal allocations are generally not aligned under VES demand. The market and optimal allocations are solutions to:

$$\max M_e \int_0^{c_d} u'(q(c)) \cdot q(c) dG \text{ where } L \geq M_e \left\{ \int_0^{c_d} [cq(c)L + f] dG + f_e \right\} \qquad \text{Market}$$

$$\max M_e \int_0^{c_d} u(q(c)) dG \qquad \text{where } L \geq M_e \left\{ \int_0^{c_d} [cq(c)L + f] dG + f_e \right\} \qquad \text{Optimum}$$

For CES demand, $u(q) = q^\rho$ while $u'(q)q = \rho q^\rho$ implying revenue maximization is perfectly aligned with welfare maximization. The CES result is therefore a limiting case of allocations under VES demand. Outside of CES, quantities produced by firms are too low or too high and in general equilibrium, this implies productivity of operating firms is also too low or too high. Market quantity, variety and productivity reflect distortions of imperfect competition. To understand these distortions, the next sub-section explains the externalities arising in the market and the subsequent Section examines the nature of misallocations.

3.2 Understanding Externalities

Although straightforward, the variety-level explanation of comparing private and social markups obscures the externalities at play in firm decisions. The market results in revenue-maximizing allocations that reflect externalities arising from private incentives. This sub-section discusses market externalities and the reasons for CES efficiency when firms differ in productivity.

Under symmetric firms, Mankiw and Whinston (1986) show that there are two market externalities. First, firms cannot capture the entire surplus generated by their production, and this lack of appropriability discourages firm entry. This is summarized by the elasticity of utility which measures the proportion of utility from a variety not captured by the real revenues ($1-\varepsilon(q)=1-u'(q)q/u(q)$). Second, firms do not internalize the downward pressure imposed by their production on prices of other firms, and this business stealing effect tends to encourage too much entry. This externality is summarized by the inverse demand elasticity $\mu(q)$. Under CES demand, the appropriability externality exactly counteracts the business stealing externality and there is no incentive to deviate from optimal entry (Grossman and Helpman 1993).

Our setting differs from standard symmetric firm models in two respects. First, firms are heterogeneous so the market must ensure an optimal selection of firms for production and the optimal distribution of quantities across these firms. Second, wages are determined endogenously and the marginal utility of income is not fixed by an outside good (as is typical in symmetric firm models). We therefore generalize the efficiency analysis from Vives (2001) to heterogeneous firms and endogenous marginal utility of income. To understand the potential sources of inefficiency, we now examine how a decline in firm entry affects the real expenditure needed to maintain welfare.

We are interested in the trade-off between variety $N=M_eG(c_d)$ and quantities $q(c)$, formulated as a uniform scaling of quantities $s(N)$ that maintains consumer welfare when variety changes for a given distribution of producers. To monetize this trade-off, we define an expenditure function $e\left(p(c,N),N,U^{\text{mkt}}\right)$ at the market level of welfare, U^{mkt}, and prices $p(c,N)$ that support a uniform scaling of quantities $s(N)$ as above. As real incomes are $\delta=e$, this necessitates

$$p(c,N)=u'\left(s(N)q(c)\right)/\delta(N)$$

and consequently at market prices (where $s(N) = 1$), the change in real expenditure is

$$\begin{aligned} d\ln e/d\ln N &= 1 + d\ln \int_0^{c_d} u'\left(s\left(N\right)q\left(c\right)\right)s\left(N\right)q\left(c\right)dG\left(c\right)/d\ln N. \\ &= 1 + s'\left(N\right)N\int_0^{c_d} u'\left(q\left(c\right)\right)q\left(c\right)\left[1-\mu\left(q\left(c\right)\right)\right]dG\left(c\right)/\left(\delta/N\right) \end{aligned}$$

which consists of the direct effect of entry on expenditure through a change in variety and the indirect effects through quantity and price per firm. In particular, $s'\left(N\right) = -1/N\overline{\varepsilon}$ where $\bar{\varepsilon} \equiv \int_0^{c_d} u'\left(q\right)qdG/\int_0^{c_d} u\left(q\right)dG$.[13]

Letting $\tilde{\mu} \equiv \int_0^{c_d} u'\left(q\right)q\mu\left(q\right)dG/\int_0^{c_d} u'\left(q\right)qdG$, the change in real expenditure is therefore

$$d\ln e/d\ln N = \left[1-\overline{\varepsilon}-\tilde{\mu}\right]/\overline{\varepsilon}.$$

When firm are symmetric, $d\ln e/d\ln N = \left[1-\varepsilon-\mu\right]/\varepsilon$ for ε and μ evaluated at the market quantity. This highlights two externalities arising in the market. First, firms are unable to appropriate the full consumer surplus through revenues as measured by $(1-\varepsilon)$. Lower entry requires higher real expenditure to maintain welfare because consumers have a taste for variety. Second, firms do not account for the effect of their sales on the demand for other firms' products. This business stealing externality is measured by μ. Lower entry reduces business stealing and requires less real expenditure to maintain welfare. Under symmetric firms and CES demand, the market allocation is efficient because the appropriability externality balances the business stealing externality ($1-\varepsilon-\mu = 0$), leading to optimal entry and production.

When firms differ in productivity, the change in real expenditure needed to maintain welfare upon entry is

$$\frac{d\ln e}{d\ln N} = \frac{1}{\overline{\varepsilon}}\left[\underbrace{-(1-\bar{\varepsilon})}_{\text{Appropriability}} + \underbrace{\bar{\mu}}_{\text{Business Stealing}} + \underbrace{\int_0^{c_d}\left(\mu(q)-\bar{\mu}\right)\frac{u'(q)q}{\int u'(q)qdG}dG}_{\text{Business Shifting}}\right].$$

for $\bar{\mu} \equiv \int_0^{c_d} u\left(q\right)\mu\left(q\right)dG/\int_0^{c_d} u\left(q\right)dG$. As earlier, the first and second terms measure the appropriability externality and the business stealing externality. With heterogeneous firms, these two externalities are represented by the average across all varieties. The third term represents the business shifting effect of entry. It consists of the revenue-weighted average of the deviation in business stealing across firms $(\mu - \bar{\mu})$ and summarizes whose business suffers upon entry.

[13] This is because the change in welfare ($0 = 1 + d\ln\int_0^{c_d} u\left(s\left(N\right)q\left(c\right)\right)dG\left(c\right)/d\ln N$) gives $0 = 1 + s'\left(N\right)N\int_0^{c_d} u'\left(q\left(c\right)\right)q\left(c\right)dG\left(c\right)/\int_0^{c_d} u\left(q\left(c\right)\right)dG\left(c\right)$.

Under CES demand or symmetric firms, all firms charge the same markup and business shifting does not arise. More generally, business shifting arises when firms differ in productivity. This leads us to an examination of the distribution of misallocations induced by the market.

4 Market Distortions and Variable Elasticities

Having identified externalities, we characterize how the market allocates resources relative to the social optimum. In their symmetric firm setting, Dixit and Stiglitz (1977) examine when the market under-produces and over-produces. They find that the bias in market allocation is determined by how the elasticity of utility varies with quantity $(1-\varepsilon(q))'$. When firms differ in productivity, we show that the variation in the inverse demand elasticity $\mu'(q)$ also matters for the bias in market allocations.

We start with a discussion of markup and quantity patterns and then discuss how these demand patterns determine misallocations in symmetric firm models. Under firm heterogeneity, different demand patterns induce different misallocations. We first summarize the misallocations by demand patterns and then discuss empirical evidence for different demand elasticities. Finally, we consider extensions of the basic framework to understand the robustness of the misallocations.

4.1 Markup and Quantity Patterns

We will show that the relationship between markups and quantity characterizes distortions. It is therefore useful to define preferences by the signs of $\mu'(q)$ and $(1-\varepsilon(q))'$. When $\mu'(q)>0$, private markups are positively correlated with quantity. This is the case studied by Krugman (1979): firms are able to charge higher markups when they sell higher quantities. Our regularity conditions guarantee low cost firms produce higher quantities (Section 3.1), so low cost firms have both high q and high markups. When $\mu'(q)<0$, small "boutique" firms charge higher markups. Similarly, the sign of $(1-\varepsilon(q))'$ determines how social markups vary with quantity. For CES demand, private and social markups are constant ($\mu'=0$, $(1-\varepsilon)'=0$).

To bring out the distinction in distortions for different markup patterns, Definition 1 below characterizes preferences as aligned when private and social markups move in the same direction and misaligned when they move in different directions.

Definition 1. Private and social incentives are *aligned* when μ' and $(1-\varepsilon)'$ have the same sign. Conversely, incentives are *misaligned* when μ' and $(1-\varepsilon)'$ have different signs.

To fix ideas, Table 1 summarizes μ' and $(1-\varepsilon)'$ for commonly used utility functions. Among the forms of $u(q)$ considered are expo-power,[14] HARA and generalized CES (proposed by Dixit and Stiglitz).[15]

Table 1: Private and Social Markups for Common Utility Forms

	$(1-\varepsilon)' < 0$	$(1-\varepsilon)' > 0$
$\mu' > 0$	Generalized CES ($\alpha > 0$): $(q+\alpha)^\rho$	CARA, Quadratic HARA ($\alpha > 0$): $\frac{(q/(1-\rho)+\alpha)^\rho - \alpha^\rho}{\rho/(1-\rho)}$ Expo-power ($\alpha > 0$): $\frac{1-\exp\left(-\alpha q^{1-\rho}\right)}{\alpha}$
$\mu' < 0$	HARA ($\alpha < 0$): $\frac{(q/(1-\rho)+\alpha)^\rho - \alpha^\rho}{\rho/(1-\rho)}$ Expo-power ($\alpha < 0$): $\frac{1-\exp\left(-\alpha q^{1-\rho}\right)}{\alpha}$	Generalized CES ($\alpha < 0$): $(q+\alpha)^\rho$

4.2 Misallocations under Symmetric Firms

Dixit and Stiglitz examine how the market allocation deviates from the optimal allocation. They find that the elasticity of utility determines the bias in production and entry. We state their result below and discuss how productivity differences affect distortions subsequently.

Proposition 4. *Under symmetric firms, the pattern of misallocation is as follows:*
1. If $(1-\varepsilon)' < 0$*, market quantities are too high and market entry is too low.*
2. If $(1-\varepsilon)' > 0$*, market quantities are too low and market entry is too high.*

Proof. Dixit and Stiglitz (1977). □

Variation in the elasticity of utility summarizes the difference between the lack of appropriability and business stealing because $\varepsilon' q/\varepsilon = 1-\varepsilon-\mu$. When $(1-\varepsilon)' > 0$, the business stealing externality outweighs the appropriability externality. Firms ignore the negative effect of entry on prices and the market provides too much variety. When $(1-\varepsilon)' < 0$, the business stealing externality is smaller and the market provides too little variety. Under symmetric firms, the business shifting effect is irrelevant and the variation in firm markups $\mu'(q)$ does not affect the bias in market allocations.

The symmetric firm case simplifies the analysis of misallocations as the tradeoff is between two decisions: quantity and entry. In contrast, determining misallocations across heterogeneous

[14] The expo-power utility was proposed by Saha (1993) and recently used by Holt and Laury (2002) and Post, Van den Assem, Baltussen and Thaler (2008) to model risk aversion empirically.

[15] The parameter restrictions are $\rho \in (0,1)$, $\alpha > q/(\rho-1)$ for HARA and $\alpha > -q$ for Generalized CES.

firms is less obvious because quantities vary by firm productivity, and this variation depends on entry and selection. Further, the business shifting effect depends on the distribution of markups and can have different signs depending on the variation in private and social markups. The next sub-section explains these misallocations for heterogeneous firms. Examining misallocations across the entire distribution of firms reveals two substantive results. First, as we might expect, the misallocation of resources across firms differs by productivity. An interesting finding is that this heterogeneity in misallocation can be severe enough that some firms over-produce while others under-produce. For example, as we will show below, when $\mu' > 0$ and $(1-\varepsilon)' > 0$, excess production by small firms imposes an externality on large firms. Large firms produce below their optimal scale and too many small firms enter the market. In this case, the market diverts resources away from large firms towards small firms. Second, accounting for firm heterogeneity shows that both the elasticity of utility and the inverse demand elasticity determine resource misallocations. When firms are symmetric, only the elasticity of utility determines misallocations and the inverse demand elasticity does not matter (Proposition 4). The presence of firm heterogeneity fundamentally changes the qualitative analysis. When markups vary, firms with different productivity levels charge different markups. This creates a new externality and affects the quantity and entry decisions. Therefore, firm heterogeneity and variable markups alter the standard policy rules for correcting misallocation of resources.

4.3 Quantity, Productivity and Entry Distortions

We now characterize the misallocations by demand characteristics. The distortions in quantity, productivity and entry are discussed in turn. The sign of the bias in market outcomes depends on both μ' and $(1-\varepsilon)'$.

4.3.1 Quantity Bias

Quantity distortions across firms depend on whether private and social incentives are aligned or misaligned. We show that when private and social incentives are misaligned, market quantities $q^{\text{mkt}}(c)$ are uniformly too high or low relative to optimal quantities $q^{\text{opt}}(c)$. In contrast, when private and social markups are aligned, whether firms over-produce or under-produce depends on their productivity.

The relationship between market and optimal quantities is fixed by FOCs for revenue maximization and welfare maximization. The market chooses $[1-\mu(q^{\text{mkt}})]u'(q^{\text{mkt}}) = \delta c$, while the optimal quantity is given by $u'(q^{\text{opt}}) = \lambda c$. Therefore, the relationship of market and optimal

quantities is

$$\frac{\text{Firm MB}}{\text{Social MB}} = \frac{\left[1-\mu\left(q^{\text{mkt}}\right)\right]\cdot u'\left(q^{\text{mkt}}\right)}{u'\left(q^{\text{opt}}\right)} = \frac{\delta c}{\lambda c} = \frac{\text{Firm MC}}{\text{Social MC}}.$$

The ratio of real revenue to welfare δ/λ depends on entry, productivity and the distribution of quantities. It summarizes the industry-wide distortions through the lack of appropriability and business stealing across all varieties. The variety-specific externality arises from business shifting which is captured by $\mu(q^{\text{mkt}}(c))$.

When incentives are *misaligned*, market and optimal quantities are too high or too low across all varieties and the direction of this bias is similar to the symmetric firm case. In particular, when $(1-\varepsilon)' < 0 < \mu'$, the market over-rewards firms producing higher quantities and all firms over-produce $q^{\text{mkt}}(c) > q^{\text{opt}}(c)$. When $(1-\varepsilon)' > 0 > \mu'$, market production is too low ($q^{\text{mkt}}(c) < q^{\text{opt}}(c)$). Therefore, firms are either over-rewarded for producing q or under-rewarded, and quantities are distorted in the same direction for all firms.

When incentives are *aligned*, the gap between the market and social cost of resources (δ and λ) is small enough that quantities are not uniformly distorted across all firms. The business shifting effect can dominate the average appropriability and business stealing effects, leading to differences in production bias across firms. Quantities are equal for some c^* where $1-\mu\left(q^{\text{mkt}}(c^*)\right) = \delta/\lambda$. For all other varieties, quantities are still distorted. When $\mu', (1-\varepsilon)' > 0$, market production is biased towards high cost firms ($q^{\text{mkt}} < q^{\text{opt}}$ for low c and $q^{\text{mkt}} > q^{\text{opt}}$ for high c). The market shifts business away from low cost firms and over-rewards high cost firms. When $\mu', (1-\varepsilon)' < 0$, the bias is reversed and low cost firms over-produce. Therefore, when private and social markups are aligned, whether the market under or over produces depends on a firm's costs. Proposition 5 summarizes the bias in market quantities.

Proposition 5. *When preferences are misaligned, $q^{\text{mkt}}(c)$ and $q^{\text{opt}}(c)$ never cross:*
1. If $(1-\varepsilon)' < 0 < \mu'$, market quantities are too high: $q^{\text{mkt}}(c) > q^{\text{opt}}(c)$.
2. If $(1-\varepsilon)' > 0 > \mu'$, market quantities are too low: $q^{\text{mkt}}(c) < q^{\text{opt}}(c)$.
In contrast, when preferences are aligned and $\inf_q \varepsilon(q) > 0$, $q^{\text{mkt}}(c)$ and $q^{\text{opt}}(c)$ have a unique crossing c^* (perhaps beyond market and optimal cost cutoffs).
3. If $(1-\varepsilon)' > 0$ and $\mu' > 0$, $q^{\text{mkt}}(c) < q^{\text{opt}}(c)$ for $c < c^*$ and $q^{\text{mkt}}(c) > q^{\text{opt}}(c)$ for $c > c^*$.
4. If $(1-\varepsilon)' < 0$ and $\mu' < 0$, $q^{\text{mkt}}(c) > q^{\text{opt}}(c)$ for $c < c^*$ and $q^{\text{mkt}}(c) < q^{\text{opt}}(c)$ for $c > c^*$.

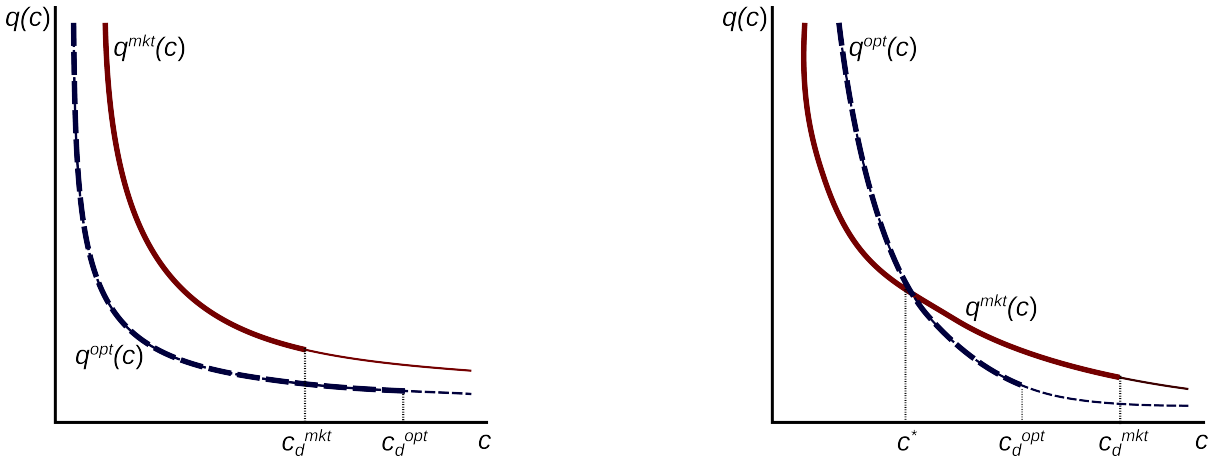
4.3.2 Productivity Bias

The distortion in firm selection is determined by the relation between the elasticity of utility and quantity. Proposition 6 shows that market productivity is either too low or high, depending on whether social markups are increasing or decreasing. We use this result now to depict the pattern of misallocation graphically, and discuss the result further below.

Proposition 6. *Market productivity is too low or high, as follows:*
1. If $(1-\varepsilon)' > 0$*, market productivity is too low:* $c_d^{\text{mkt}} > c_d^{\text{opt}}$.
2. If $(1-\varepsilon)' < 0$*, market productivity is too high:* $c_d^{\text{mkt}} < c_d^{\text{opt}}$.

Propositions 5 and 6 show the market misallocates resources across firms, and variable demand elasticities characterize the pattern of these misallocations. Figure 1 illustrates the bias in firm-level production for aligned and misaligned preferences when private markups increase in quantity. For ease of reference, Table 2 summarizes the misallocations by demand characteristics.[16] A discussion of the externalities at play in the results follow in the next sub-section.

Figure 1: Bias in Firm Production by Preferences



(a) Misaligned: $\mu' > 0 > (1-\varepsilon)'$ (b) Aligned: $\mu' > 0$ and $(1-\varepsilon)' > 0$

[16] Table 2 characterizes the qualitative role of demand elasticities in misallocations. Using a quantitative measure of distortions reiterates their importance. The loss from misallocations can be summarized by the difference between social and market "profits", evaluated at optimal allocations. This measure consists of the difference between average social markup and average private markup $(1-\bar{\varepsilon}-\bar{\mu})$, and the covariance between social and private markups $\text{Cov}(1-\varepsilon,\mu)$. The covariance component shows that the distribution of markups matters for quantifying distortions, except when firms are symmetric or markups are constant (leading to zero covariance).

Table 2: Distortions by Demand Characteristics

	$(1-\varepsilon)' < 0$	$(1-\varepsilon)' > 0$
$\mu' > 0$	Quantities Too High: $q^{\text{mkt}}(c) > q^{\text{opt}}(c)$ Productivity Too High: $c_d^{\text{mkt}} < c_d^{\text{opt}}$	Quantities High-Cost Skewed: $q^{\text{mkt}}(c) < q^{\text{opt}}(c)$ for $c < c^*$ $q^{\text{mkt}}(c) > q^{\text{opt}}(c)$ for $c > c^*$ Productivity Too Low: $c_d^{\text{mkt}} > c_d^{\text{opt}}$
$\mu' < 0$	Quantities Low-Cost Skewed: $q^{\text{mkt}}(c) > q^{\text{opt}}(c)$ for $c < c^*$ $q^{\text{mkt}}(c) < q^{\text{opt}}(c)$ for $c > c^*$ Productivity Too High: $c_d^{\text{mkt}} < c_d^{\text{opt}}$	Quantities Too Low: $q^{\text{mkt}}(c) < q^{\text{opt}}(c)$ Productivity Too Low: $c_d^{\text{mkt}} > c_d^{\text{opt}}$

4.3.3 Understanding Externalities and Productivity

While Proposition 6 follows from a general equilibrium analysis, the decision to introduce a marginal variety can be intuitively explained as follows. Under increasing social markups $(1-\varepsilon)' > 0$, the lack of appropriability of a marginal variety is lower than its business stealing effect. This encourages production of the marginal variety and the cost cutoff in the market is too high. Although the marginal variety steals business and shifts business across varieties, its impact is small and the bias in the cost cutoff is determined by the elasticity of utility. We now illustrate this reasoning in a similar fashion as for entry in Section 3.1.

We are interested in the trade-off between productivity c_d and quantities $q(c)$ for a uniform scaling of quantities that maintains consumer welfare when c_d changes, holding M_e fixed. At market prices, evaluating the change in real expenditure to maintain consumer welfare upon a rise in the cost cutoff yields

$$\begin{aligned} d\ln e/d\ln c_d &= d\ln \int_0^{c_d} u'\left(s(N)\,q(c)\right) s(N)\,q(c)\,dG(c)\,/d\ln c_d \\ &= c_d g(c_d)\left[M_e s'(N)\left[1-\tilde{\mu}\right] + u(q(c_d))\varepsilon(q(c_d))\right]/\bar{\varepsilon}\int_0^{c_d} u(q(c))dG(c). \end{aligned}$$

Defining $x_d \equiv x(q(c_d))$, the change in real expenditure is[17]

$$\frac{d\ln e}{d\ln c_d} = \frac{u_d c_d g(c_d)}{\bar{\varepsilon}\int u(q)dG}\left[\underbrace{-(1-\varepsilon_d)}_{\text{Appropriability}} + \underbrace{\mu_d}_{\text{Business Stealing}} + \underbrace{\int (\mu(q)-\mu_d)\frac{u'(q)q}{\int u'(q)qdG}dG}_{\text{Business Shifting}}\right]$$

As earlier, the change in real expenditure highlights the lack of appropriability, business stealing and business shifting. The marginal firm is unable to appropriate the full surplus it generates, and this appropriability externality is measured by $(1-\varepsilon(q(c_d)))$. The marginal variety steals business from other firms ($\mu(q(c_d))$) and shifts business across them ($\mu(q)-\mu(q(c_d))$). Under CES demand, the business stealing externality exactly outweighs the appropriability externality, and there is no business shifting. More generally, the externalities differ and their net effect on the change in real expenditure can be signed. The change in real expenditure needed to maintain welfare upon a rise in the cost cutoff is

$$d\ln e/d\ln c_d = \left(u_d c_d g(c_d)/\int u dG\right)\left[-(1-\tilde{\varepsilon}-\tilde{\mu})+(\varepsilon_d-\tilde{\varepsilon})\right]/\bar{\varepsilon}.$$

The sign of the first term in square brackets is the sign of $(1-\varepsilon)'$. The second term also takes the sign of $(1-\varepsilon)'$ because the marginal firm makes the lowest quantity. Although the marginal firm shifts business, this impact is smaller and the change in real expenditure needed to maintain welfare is determined by the elasticity of utility.

This analysis also highlights that a comparison of the mass of entrants in the market and the optimum is generally hard to make. The change in real income needed to maintain welfare upon a fall in entry is $d\ln e/d\ln N = -(1-\bar{\varepsilon}-\bar{\mu})+\int(\mu(q(c))-\bar{\mu})\frac{u'(q(c)))q(c)}{\int u'(q(c)))q(c)dG}dG$. Unlike the business shifting effect of the marginal variety, business shifting from entry need not be dominated by the net effect from the appropriability externality and the business stealing externality. The first term in $d\ln e/d\ln N$ summarizes the tradeoff between the appropriability externality and the business stealing effect and takes the sign of $(1-\varepsilon)'$. The second term summarizes the business shifting effect and depends on the sign of μ'. Consider the case with aligned preferences and increasing markups. Then the first term is positive and the second term is negative. The business shifting effect dampens the other two externalities and lower real expenditure is needed to maintain welfare upon a fall in entry.[18] As the externalities move in opposite direc-

[17] At the market allocation, $M_e s'(N) = -u(q(c_d))/\bar{\varepsilon}\int_0^{c_d} u(q(c))\,dG(c)$ because the change in welfare is $0 = 1 + s'(N)N\int_0^{c_d} u'(q(c))\,q(c)\,dG(c)/\int_0^{c_d} u(q(c))\,dG(c)$.

[18] This is consistent with early insights from Vickers (1995) and Vives (2001) arguing that an increase in entry hurts high productivity firms more than low productivity firms, and cost asymmetries lead to an improvement in

tions, the bias in potential entry M_e and available variety $M_e G(c_d)$ cannot be determined without further information on demand and cost parameters. The net effect of the three externalities and hence the bias in potential entry depends on the relative magnitudes of demand and cost parameters including the cost distribution $G(c)$.[19] While firm heterogeneity makes entry distortions dependent on the cost distribution, the bias in quantity and productivity can be unambiguously inferred from the demand-side elasticities. In the remainder of this Section, we first examine the robustness of these findings under alternative modeling assumptions and then discuss empirical work on estimating the demand-side elasticities.

4.4 Extensions of the Basic Framework

As many different fields of economics (such as macroeconomics and urban economics) use monopolistically competitive models, we extend our basic framework to different modelling assumptions used in these fields to discuss the robustness of CES efficiency and misallocations under VES demand. Details are in an online Appendix and a summary of four key extensions is provided here.

First, suppose the costs of production of a firm vary with its scale of production. To account for non-constant marginal costs, let the variable cost of production be $\omega(q) \cdot cq$ and assume $2\omega' + \omega'' q > 0$ for all feasible quantities to ensure strict concavity of the firm problem. The market maximizes aggregate revenue under non-constant marginal costs. As firms account for the interdependence between their unit costs and quantity, CES demand ensures the same trade-off between different externalities and leads to efficient allocations (as shown in an online Appendix). Under VES demand, the bias in quantity and productivity are the same as Propositions 5 and 6.

Second, let firms choose their advertising technology as in Arkolakis (2010). A firm can reach a fraction $n(c)$ of consumers by spending $\left[1-(1-n(c))^{1-\theta}\right] L^{\vartheta} f/(1-\theta)$ units of labor for $\theta, \vartheta \in [0,1]$. The production cost f therefore varies with the fraction of consumers that a firm chooses to reach. The market allocates resources efficiently under CES demand when the costs of commencing production are akin to advertising costs. The market maximizes aggregate revenue and the quantity and productivity distortions are the same as earlier. A new insight is gained from this richer model of fixed costs. The market does not choose the optimal levels of advertising. When $(1-\varepsilon)' > 0$, low cost firms under-advertise and reach too few consumers ($n(c)$ is too low). High cost firms over-advertise and their $n(c)$ is too high. For $(1-\varepsilon)' < 0$, low

the entry externality.

[19] Focusing on a linear demand setting with Pareto cost draws, Nocco, Ottaviano and Salto (2013) find that the mass of firms cannot be unambiguously ranked.

cost firms in the market over-advertise while high cost firms under-advertise.

Third, the efficiency and misallocation results are robust to introducing multiple sectors, conditional on the resource allocation for the sector. Following Zhelobodko et al. (2012), let the multi-sector utility function be $U(q_0, Q)$ where q_0 is a homogeneous numeraire good and $Q \equiv M_e \int u(q)dG$ is the sub-utility from differentiated goods. Conditional on a resource allocation of $(1-q_0)$ towards differentiated goods, the bias in quantity and productivity is the same as earlier.

The market allocation within the differentiated goods sector is efficient under CES demand. This however does not imply that the market and the optimum have the same level of $(1-q_0)$. For instance, in the Cobb-Douglass specification of Dixit and Stiglitz (1977), $U(q_0, Q) = q_0^{1-\gamma} Q^{\gamma}$, the optimal allocation for the homogeneous good is $q_0^{\text{opt}} = 1-\gamma$ while the market allocation is $q_0^{\text{mkt}} = (1-\gamma)/(1-\gamma+\gamma\bar{\varepsilon})$. The markups charged in the homogeneous and the differentiated goods sectors differ, leading to inefficient market allocations. The markup for the homogeneous good is one and the marginal utility of income is fixed by the homogeneous good. Marginal cost pricing ($p = c$) therefore aligns the markups across the two sectors. Thus, Dixit and Stiglitz suggest marginal cost pricing and lumpsum entry subsidies to induce optimal allocations across sectors. In keeping with Melitz, we consider a single sector and find resources are optimally allocated in the market. In a single sector economy, prices are proportional to marginal costs ($p = \delta c$) but the marginal utility of income δ is no longer fixed by the homogeneous good. Market allocations are first best as the marginal utility of income reflects the social cost of resources.[20]

For completeness, we finally note that the CES demand of Melitz is also necessary for efficiency under the CES-Benassy class of preferences. Benassy (1996) points out that the "taste for variety" under Dixit-Stiglitz preferences is closely linked to the degree of market power of competitors. Taste for variety can be disentangled from market power through Benassy preferences $U(M_e, c_d, q) \equiv \nu(M_e) \int_0^{c_d} q(c)^{\rho} g(c)dc$ which value quantity and variety differently through $\nu(M_e)$. Following Benassy (1996) and Alessandria and Choi (2007), when $\nu(M_e) = M_e^{\rho(\nu_B+1)}$, these preferences disentangle "taste for variety" ν_B from the markup to cost ratio $(1-\rho)/\rho$. Market allocations under CES-Benassy are the same as CES. However, firms do not fully internalize consumers' taste for variety, leading to suboptimal allocations. Market allocations are optimal only if taste for variety exactly equals the markup to cost ratio.[21]

As the underlying demand structure can lead to very different distortions, the remainder of

[20] In related work, Behrens et al. (forthcoming) examine efficiency in a multi-sector model with constant absolute risk aversion (CARA) preferences.

[21] Helpman and Krugman (1985) and Feenstra and Kee (2008) derive a GDP function for this economy, and Cole and Davies (forthcoming) highlight variety distortions by introducing existence values for variety.

this Section discusses empirical evidence for different demand parameters.

4.5 Empirical Evidence for Demand Characteristics

The pattern of misallocation depends on demand-side elasticities. A natural question is whether empirical work can identify which case in Table 2 is relevant. Although the elasticity of utility is typically unobservable, the inverse demand elasticity (or firm markups) has been a subject of research in industrial organization. A large empirical literature in industrial organization shows a high level of markup dispersion across plants, and finds much larger markup dispersion within industries rather than across industries (example Klette 1999; Nishimura et al. 1999). The empirical relationship between markups and quantities is largely in line with increasing markups though there are industries which show decreasing markups.

The empirical literature can be broadly classified into papers that use price-cost margins to measure markups and those that use variants of the Hall methodology to estimate markups.[22] In a series of influential papers, Roberts and Supina (1996, 2001) use physical output, revenue, and input expenditures to measure price-cost margins for a number of U.S. manufactured products and show the majority of products exhibit increasing markups. Focusing on products with little scope for vertical differentiation, they document a high and persistent level of price dispersion across plants for most products. They find markups increase with plant size and often monotonically across quartiles of plant size for six of the thirteen products (polyester blend fabrics, bread, coffee, oak flooring, softwood plywood, newsprint). Two products (cotton sheeting, gasoline) show no significant change in markups with plant size. For the remaining four products (hardwood plywood, vans, corrugated boxes and concrete), markups decrease significantly with increases in plant size across the whole size distribution. One concern with the latter finding is that decreasing markups might be driven by the decision of large plants to operate in larger, more competitive markets, as shown by Syverson (2004) for ready-mixed concrete.

Studies based on the Hall methodology largely find a positive relationship between markups and quantity. In a careful study using data on physical quantities, De Loecker et al. (2012) find markups are positively correlated with firm productivity of large Indian manufacturers during 1989-2003. De Loecker and Warzynski (2012) estimate a positive correlation between markups and productivity for Slovenian manufacturing firms during 1994-2000 and Dhyne et al. (2011) also find markups are positively related to firm productivity for Belgian bread manufacturers during 1995-2009. On the other hand, a highly-cited study by Klette (1999) shows Norwegian

[22]The Hall methodology estimates the price-cost markup as the slope coefficient from a regression of output growth on the share-weighted rate of input growth. A discussion of this approach is provided in Tybout (2003) and De Loecker and Goldberg (2013).

firms with higher markups tend to have lower productivity.[23]

While the empirical literature largely finds increasing firm markups, social markups are rarely observable and early papers on monopolistic competition express a lack of consensus on how they respond to quantity. Spence (1976) suggests social markups increase with quantity while Dixit and Stiglitz propose decreasing social markups. Vives (2001) discusses three reasons for considering increasing private and social markups as the normal case (Chapter 6). First, for symmetric consumption, this would imply that consumers have an increasing preference for variety and a higher inverse demand elasticity at a higher output per variety. Second, aligned preferences are theoretically appealing because the elasticity of $1-\varepsilon$ equals the elasticity of μ in the limit as q approaches zero under a relatively mild assumption. Finally, commonly-used preferences exhibit aligned preferences with increasing markups. For instance, $(1-\varepsilon)' > 0$ whenever $\mu' > 0$ in the HARA class (as shown in Table 1). Moreover, the generalized CES example of Dixit and Stiglitz for decreasing markups is not continuous at zero when it is appropriately normalized to ensure $u(0) = 0$. While we cannot rule out specific cases without further empirical investigation, the assumption of increasing private and social markups has appealing properties for theoretical work.[24]

5 Efficiency and Market Size

Having discussed misallocations, this Section examines welfare and efficiency from integration with world markets. The existence of gains from international trade is one of the "most fundamental results" in economics (Costinot and Rodriguez-Clare (2013)). Increases in market size encourage competition, so we might expect that integration would reduce market power and improve welfare. However, the following insight of Helpman and Krugman (1985) (pp. 179) is relevant:

> Unfortunately imperfect competition, even if takes as sanitized a form as monop-

[23] A separate literature provides evidence for increasing markups by estimating the price response to exchange rate fluctuations. The typical estimate for exchange rate pass through is less than one, which suggests increasing markups (because the pass-through rate corresponds to $(1-\mu)/(1-\mu+\mu' q/\mu)$). A discussion of this literature is provided in Goldberg and Knetter (1997) and more recently in Klenow and Malin (2010).

[24] While private markups can be estimated using pricing and production data, distinguishing increasing and decreasing social markups is more challenging as they are unlikely to be directly observable. Consequently, for standard firm level data sets, policy inferences require more structure on demand. One approach is to use flexible demand systems that leave determination of the four cases up to the data. For example, the VES form $u(q) = aq^{\rho} + bq^{\gamma}$ allows all sign combinations of $\varepsilon'(q)$ and $\mu'(q)$ (Online Appendix). This form overlaps with the adjustable pass-through demand system (Bulow and Pfleiderer 1983; Weyl and Fabinger 2012). If sufficient data is available, another approach is to recover $\varepsilon(q)$ from price and quantity data using $\varepsilon(q) = p(q)q/\int p(q)dq$ or from markup and quantity data using $\ln \varepsilon(q)/q = \int_0^q -(\mu(t)/t)\,dt - \ln\left[\int_0^q \exp\{\int_0^s -(\mu(t)/t)\,dt\}\,ds\right]$.

olistic competition, does not lead the economy to an optimum. As a result there is no guarantee that expanding the economy's opportunities, through trade or anything else, necessarily leads to a gain. We cannot prove in general that countries gain from trade in the differentiated products model.

Building on this insight, we address two related questions. First, we examine when market expansion provides welfare gains. Having characterized distortions, we first show that welfare gains are related to the demand-side elasticities mentioned earlier. Next, we examine efficiency in large markets to understand the potential of market expansion in eliminating distortions. We show large integrated markets can eliminate distortions, while preserving firm heterogeneity. Finally, we discuss the role of firm heterogeneity and variable elasticities for quantitative work measuring the welfare gains from international trade.

5.1 Integration, Market Size and Efficiency

We begin with the equivalence between market expansion and trade. Proposition 7 shows an economy can increase its market size by opening to trade with foreign markets. The market equilibrium between freely trading countries of sizes $L_1, ..., L_n$ is identical to the market equilibrium of a single autarkic country of size $L = L_1 + ... + L_n$, echoing Krugman (1979). This result is summarized as Proposition 7.

Proposition 7. *Free trade between countries of sizes $L_1, ..., L_n$ has the same market outcome as a unified market of size $L = L_1 + ... + L_n$.*

Proof. Online Appendix and Krugman (1979). □

Proposition 7 implies that the market distortions detailed in Section 5 persist in integrated markets. Resource allocation in an integrated market is suboptimal, except under CES demand. When markups vary, marginal revenues do not correspond to marginal utilities so market allocations are not aligned with efficient allocations. This is particularly important when considering trade as a policy option, as it implies that opening to trade may take the economy further from the social optimum. For example, market expansion from trade may induce exit of low productivity firms from the market when it is optimal to keep more low productivity firms with the purpose of preserving variety.

Helpman and Krugman (1985) provide sufficient conditions for welfare gains from trade. They show when productivity and variety do not decline after integration, then there are gains

from trade.[25] In terms of primitives, we find integration is always beneficial when preferences are aligned. This is true for any cost distribution, but requires a regularity condition for decreasing private markups ($2+\mu''q/\mu'(1-\mu)\geq 0$). We summarize this in Proposition 8.

Proposition 8. *Market expansion increases welfare when preferences are aligned. (Provided* $2+\mu''q/\mu'(1-\mu)\geq 0$ *whenever* $\mu'<0$*).*

The economic reasoning for Proposition 8 follows from similar responses of the two demand-side elasticities to changes in quantity. An increase in market size increases competition and reduces per capita demand for each variety. When preferences are aligned, demand shifts alter the private and social markups in the same direction. The market therefore incentivizes firms towards the right allocation and provides higher welfare. Building on this result, Bykadorov et al. (2014) show that aligned preferences are necessary and sufficient for welfare gains from trade under symmetric firms and variable marginal costs.

The role of aligned markups in firm survival highlights how trade increases welfare. When aligned markups increase with quantity, a rise in market size forces out the least productive firms. Since social markups are positively correlated with quantity, the least productive firms also contribute relatively little to welfare and their exit is beneficial. When markups decrease with quantity, small "boutique" firms contribute at a higher rate to welfare and are also able to survive after integration by charging higher markups. Integration enables the market to adapt their production in line with social incentives, leading to welfare gains from trade.

While integration can increase welfare, a more ambitious question is: can we ever expect trade to eliminate the distortions of imperfect competition? Following Stiglitz (1986), we study market and optimal outcomes as market size becomes arbitrarily large. Since small markets have insufficient competition, looking at large markets allows us to understand where market expansion is headed and when international trade enables markets to eventually mitigate distortions.

5.2 Efficiency in Large Markets

We examine when integrating with large global markets enables a small economy to overcome its market distortions. From a theoretical perspective, we term a large market the limit of the economy as the mass of workers L approaches infinity, and in practice we might expect that sufficiently large markets approximate this limiting case.[26]

[25] Specifically, let w denote the wage and $C(w,q)=w(c+f/q)$ denote the average unit cost function for producing q units of variety c. When firms are symmetric in c, trade is beneficial as long as variety does not fall ($M_e\geq M_e^{\text{aut}}$) and average unit cost of the autarky bundle is lower ($C(w,q)\cdot q^{\text{aut}}\leq C(w,q^{\text{aut}})\cdot q^{\text{aut}}$).

[26] How large markets need to be to justify this approximation is an open quantitative question.

Large markets enable us to understand whether competition can eliminate distortions. For instance, when firms are symmetric, large markets eliminate distortions as *per capita* fixed costs fall to zero. This is because free entry leads to average cost pricing ($p = c + f/qL$), so the per capita fixed costs summarize market power. As market size grows arbitrarily large and per capita fixed costs fall to zero, markups disappear leading to perfect competition and efficient allocations in large markets.

Building on this reasoning, we develop the large market concept in two directions to understand the sources of inefficiency. First, we tie the conditions for efficiency to demand primitives, taking into account endogeneity of allocations. In the simple example above, this amounts to determining how f/qL changes with market size under different model primitives. Second, we examine whether productivity differences are compatible with large markets. When firms are heterogeneous, simply knowing per capita fixed costs does not explain the distribution of productivity, prices and quantity. At least three salient outcomes can occur. One outcome is that competitive pressures might weed out all firms but the most productive. This occurs for instance when marginal revenue is bounded, as when u is quadratic or CARA (e.g. Behrens and Murata 2012). It may also happen that access to large markets allows even the least productive firms to amortize fixed costs and produce. To retain the fundamental properties of monopolistic competition under productivity differences, we chart out a third possibility between these two extremes: some, but not all, firms produce. To do so, we maintain the previous regularity conditions for a market equilibrium. In order to aid the analysis, we make three assumptions on demand at small quantities. The first assumption enables a clear distinction between the three salient outcomes in large markets.

Assumption (Interior Markups)**.** *The inverse demand elasticity and elasticity of utility are bounded away from* 0 *and* 1 *for small quantities. Formally,* $\lim_{q\to 0}\mu(q)$ *and* $\lim_{q\to 0}\varepsilon(q) \in (0,1)$.

The assumption of interior markups guarantees that as the quantity sold from a firm to a consumer becomes small (as happens for all positive unit cost firms), markups remain positive ($\mu > 0$) and prices remain bounded ($\mu < 1$). It also guarantees that the added utility provided per labor unit at the optimum converges to a non-zero constant (e.g., Solow 1998, Kuhn and Vives 1999). An example of a class of utility functions satisfying interior markups is the expo-power utility where $u(q) = [1 - \exp\left(-\alpha q^{1-\rho}\right)]/\alpha$ for $\rho \in (0,1)$. It nests CES preferences for $\alpha = 0$.

When markups are interior, there is a sharp taxonomy of what may happen to the distribution of costs, prices and total quantities ($Lq(c)$), as shown in Proposition 10 in the Appendix. In words, Proposition 10 shows that when markups are interior and the cost cutoff converges,

one of three things must happen. 1) Only the lowest cost firms remain and prices go to zero (akin to perfect competition), while the lowest cost firms produce infinite total quantities. 2) Post-entry, all firms produce independent of cost while prices become unbounded and the total quantities produced become negligible, akin to a "rentier" case where firms produce little after fixed costs are incurred. 3) The cost cutoff converges to a positive finite level, and a non-degenerate distribution of prices and total quantities persists. Although each of these possibilities might be of interest, we focus on the case when the limiting cost draw distribution exhibits heterogeneity ($\lim_{L\to\infty} c_d^{\text{mkt}} > 0$) but fixed costs still play a role in determining which firms produce ($\lim_{L\to\infty} c_d^{\text{mkt}} < \infty$). We therefore make the following assumption, which by Proposition 10 will guarantee non-degenerate prices and total quantities:

Assumption (Interior Convergence)**.** *In the large economy, the market and optimal allocations have a non-degenerate cost distribution in which some but not all entrants produce.*

Under interior markups and convergence, the economy converges to a monopolistically competitive limit distinct from the extremes of a perfectly competitive limit or a rentier limit. As the economy grows, each worker consumes a negligible quantity of each variety. At these low levels of quantity, the inverse demand elasticity does not vanish and firms can still extract a positive markup μ. This is in sharp contrast to a competitive limit, in which firms are left with no market power and μ drops to zero. Similarly, the social markup $(1-\varepsilon)$ does not drop to zero in the monopolistically competitive limit, so each variety contributes at a positive rate to utility even at low levels of quantity. The monopolistically competitive limit is therefore consistent with positive markups which become more uniform with increased market size.

In fact, this monopolistically competitive limit has a sharper characterization very close to the conditions which characterize a finite size market under CES demand (including efficiency). We therefore refer to it as a "CES limit" and introduce one last regularity condition to obtain this result.

Assumption (Market Identification)**.** *Quantity ratios distinguish price ratios for small q:*

$$\textit{If } \kappa \neq \tilde{\kappa} \textit{ then } \lim_{q\longrightarrow 0} p(\kappa q)/p(q) \neq \lim_{q\longrightarrow 0} p(\tilde{\kappa} q)/p(q).$$

Market identification guarantees production levels across firms can be distinguished if the firms charge distinct prices as quantities sold become negligible. Combining these three assumptions of interior markups, convergence and identification ensures the large economy goes to the CES limit, summarized as Proposition 9. The intuition for the role of these assumptions follows. As market size grows large, $q \longrightarrow 0$ so under Interior Markups, $(p-c)/p =$

$\mu(q) \longrightarrow \mu(0)$ and, finite but non-zero markups can persist in the large economy. Since profits are $\mu(q)/(1-\mu(q)) \cdot Lcq$, whether a particular firm survives in the large economy depends on how variable costs Lcq evolve with market size. Clearly, if variable costs diverge to zero for a firm with cost c, that firm must eventually exit, while if variable costs diverge to infinity, the firm must eventually enter. To arrive at the CES limit, necessarily variable costs must converge to a positive level, which requires convergence of the total quantity sold, Lq. However, since firms are embedded in a heterogeneous environment where aggregate conditions impact firm behavior, the pointwise convergence of markups $\{\mu(q(c))\}$ is not sufficient to guarantee that total quantities $\{Lq(c)\}$ are well behaved in aggregate. What is sufficient is that prices $\{p(c)\}$ can distinguish firms as market size grows large, thus the Market Identification condition.[27]

Proposition 9. *Under the above assumptions, as market size approaches infinity, outcomes approach the CES limit. This limit has the following characteristics:*

1. *Prices, markups and expected profits converge to positive constants.*
2. *Per capita quantities $q(c)$ go to zero, while aggregate quantities $Lq(c)$ converge.*
3. *Relative quantities $Lq(c)/Lq(c_d)$ converge to $(c/c_d)^{-1/\alpha}$ with $\alpha = \lim_{q \longrightarrow 0} \mu(q)$.*
4. *The entrant per worker ratio M_e/L converges.*
5. *The market and socially optimal allocations coincide.*

Proposition 9 shows that integration with large markets can push economies based on variable elasticity demand to the CES limit. In this limit, the inverse demand elasticity and the elasticity of utility become constant, ensuring the market outcome is socially optimal. Firms charge constant markups which exactly cross-subsidize entry of low productivity firms to preserve variety. This wipes out the distortions of imperfect competition as the economy becomes large. While dealing with the assumptions of the market equilibrium is somewhat delicate (see Appendix), we can explain Proposition 9 intuitively in terms of our previous result that CES preferences induce efficiency. In large markets, the quantity $q(c)$ sold to any individual consumer goes to zero, so markups $\mu(q(c))$ converge to the same constant independent of c.[28] This convergence to constant markups aligns perfectly with those generated by CES preferences with an exponent equal to $1 - \lim_{q \longrightarrow 0} \mu(q)$. Thus, large markets reduce distortions until market allocations are perfectly aligned with socially optimal objectives.

It is somewhat remarkable that the large market outcome, which exhibits cost differences and remains imperfectly competitive, is socially optimal. Such persistence of imperfect competition is consistent with the observation of Samuelson (1967) that "the limit may be at an

[27] From a technical standpoint, this guarantees entry is well behaved, avoiding pathological sequences of potential equilibria as market size grows large.

[28] The rate at which markups converge depends on c and is in any case endogenous (see Appendix).

irreducible positive degree of imperfection" (Khan and Sun 2002). Perloff and Salop (1985) also note that the markup disappears if the utility from a variety is bounded, but unbounded entry may not eliminate the markup when this condition is not met. We show that is precisely what happens at the CES limit. While the CES limit is optimal despite imperfect competition, it is an open empirical question whether markets are sufficiently large for this to be a reasonable approximation to use in lieu of richer variable elasticity demand. When integrated markets are small, variable markups are crucial in understanding distortions and additional gains can be reaped by using domestic policy in conjunction with trade policy.

5.3 Quantitative Literature on Welfare Gains from Trade

A growing body of work seeks to quantify the gains from international trade. New quantitative trade models typically estimate welfare gains from trade under CES demand. In an influential paper, Arkolakis et al. (2012a) show that welfare in a model with heterogeneous firms can be summarized by two sufficient statistics: the share of expenditure on domestically produced goods and the elasticity of trade with respect to trade costs. As these sufficient statistics are common to heterogeneous and representative firm models, welfare gains estimated from import shares and constant trade elasticities using trade data are the same across heterogeneous and representative firm models. However, the two models only deliver the same estimates for welfare gains when the underlying structural parameters for preferences and technology differ across the models. We use this insight of Melitz and Redding (2013) to explain the relevance of our optimality results for the quantitative literature on gains from trade.

Melitz and Redding find that the heterogeneous firm model of Melitz provides quantitatively higher gains from trade than an equivalent representative firm model when the structural parameters are the same across these models. As they mention, this can be understood by appealing to the social optimality results for CES demand (Proposition 1). Consider initial equilibria in the heterogeneous and homogeneous firm models that feature identical aggregate statistics and welfare. In the homogeneous firm model, unit cost is exogenously fixed, and hence remains unchanged when the economy opens to trade. In the heterogeneous firm model, the cost distribution changes when the economy opens to trade. In a companion note (Dhingra and Morrow 2014), we show that the open economy equilibrium with trade frictions is efficient under CES demand. Since the policymaker chooses to change the cost cutoff in an open economy, the open economy market allocation must yield higher welfare than any other feasible allocation (where the unit cost is unchanged). The allocation where the unit cost does not change is identical to the open economy equilibrium in the homogeneous firm model. Therefore the open economy equilibrium in the heterogeneous firm model must yield higher welfare than the open econ-

omy equilibrium in the homogeneous firm model. This shows that a quantitative trade model with the same structural parameters across models will provide higher welfare gains in a setting with firm heterogeneity. The optimality of market allocations ensures that firm heterogeneity increases the magnitude of welfare gains from trade.

Departing from CES preferences, market allocations are no longer optimal. This raises the question of the role played by firm heterogeneity in altering the magnitude of welfare gains from trade. While we do not model trade costs, Proposition 8 shows market expansion through trade provides higher welfare gains when firms differ in productivity. Under aligned preferences and the regularity condition ($2+\mu''q/\mu'(1-\mu)\geq 0$), we discuss when models with firm heterogeneity and variable elasticities provide higher welfare gains from trade than representative firm models.

For a given change in real income, the welfare gains from trade depend on the different assumptions on demand and firm costs. Welfare is $U = M_e \int u(q)dG = \delta/\bar{\varepsilon}$ where the average elasticity of utility is $\bar{\varepsilon} \equiv \int \varepsilon u dG / \int u dG$. An increase in market size increases real income at the rate of the average markup ($d\ln\delta/d\ln L = \int \mu pq dG / \int pq dG \equiv \tilde{\mu}$). The change in average elasticity can be decomposed into the change in $\varepsilon(q)$ given $u/\int u dG$, and the change in the weights $u/\int u dG$. Let $x_d \equiv x(q(c_d))$, then the change in the average elasticity of utility is

$$\frac{d\ln\bar{\varepsilon}}{d\ln L} = \int \frac{\varepsilon' u}{\bar{\varepsilon}\int u dG}\frac{d\ln q}{d\ln L}dG + \underbrace{\int \frac{u'\varepsilon - u'\bar{\varepsilon}}{\bar{\varepsilon}\int u dG}\frac{d\ln q}{d\ln L}dG + \frac{u_d}{\bar{\varepsilon}\int u dG}(\varepsilon_d - \bar{\varepsilon})\, c_d g(c_d)\frac{d\ln c_d}{d\ln L}}_{\text{Reallocation across heterogeneous firms}}.$$

The first term denotes the change due to a fall in quantity per firm, holding fixed the share of each variety in the average elasticity. The second and third terms denote the change in the average elasticity of utility due to a reallocation of resources across heterogeneous firms. Reallocation of resources across firms changes the share of each variety in the average elasticity of utility through $\left(u(q)/\int_0^{c_d} u(q)dG\right)'$. Using this decomposition, we can explain the role of variable elasticities and firm heterogeneity in welfare gains from trade.

For a given change in real income ($d\ln\delta/d\ln L = \tilde{\mu}$), we decompose the gains from trade into gains for a representative firm and gains due to differences in firm productivity. Defining the market outcome of a representative firm as the revenue-weighted average of heterogeneous firms, the gains from trade for a given change in real income are:

$$\frac{d\ln U}{d\ln L} = \underbrace{\tilde{\mu}\int \frac{1-\varepsilon}{\mu}\frac{\varepsilon u}{\bar{\varepsilon}\int u dG}dG}_{\text{CES}} + \underbrace{\tilde{\mu}\int\left(\frac{1-\varepsilon+\mu' q/(1-\mu)}{\mu+\mu' q/(1-\mu)} - \frac{1-\varepsilon}{\mu}\right)\frac{\varepsilon u}{\int \varepsilon u dG}dG}_{\text{VES \& Representative Firm}}$$
$$+ \underbrace{\tilde{\mu}\int \frac{\varepsilon-\bar{\varepsilon}}{\mu+\mu' q/(1-\mu)}\frac{\varepsilon u}{\bar{\varepsilon}\int u dG}dG}_{\text{Quantity Reallocation}} + \underbrace{\frac{u_d}{\int u dG}\frac{c_d g(c_d)}{\bar{\varepsilon}(1-\mu_d)}(\varepsilon_d-\bar{\varepsilon})(\tilde{\mu}-\mu_d)}_{\text{Firm Selection}}$$

The first line contains the gains from trade for a representative firm. The first component is the welfare gain when firm markups are constant and the second component shows how welfare gains change when markups vary with quantity. Under CES demand, the welfare gain is the revenue-weighted average of $1-\varepsilon$. VES demand adds the second component which is positive when markups are increasing and negative when markups are decreasing with quantity.

The second line consists of the gains from trade arising due to differences in firm productivity. The first component of the second line is the welfare gain from changes in *relative* quantities across firms. When firms differ in productivity, market size affects their output levels differently and resources are reallocated across firms. For aligned preferences, quantity reallocation increases the welfare gains from trade under the regularity condition. The second component shows the welfare gains from firm selection. Aligned preferences ensure the market selects the right firms as it expands and leads to higher welfare gains.

Under aligned preferences, reallocation of resources across heterogeneous firms increases the welfare gains from trade beyond those arising in a representative firm model. As most empirical studies are consistent with increasing markups ($\mu' > 0$), structural estimates based on CES demand therefore provide a lower bound ($1-\tilde{\varepsilon}$) for the potential gains from trade. For a given change in real income, accounting for firm heterogeneity and increasing markups would reveal higher welfare gains from trade. The magnitude of these additional gains depends on the markup variation (through $\varepsilon(q(c)) - \varepsilon(q(c_d))$ and $\mu'(q(c))$) and on the productivity distribution (through $g(c_d)$).

6 Conclusion

This paper examines the efficiency of market allocations when firms vary in productivity and markups. Considering the Spence-Dixit-Stiglitz framework, the efficiency of CES demand is valid even with productivity differences across firms. This is because market outcomes maximize revenue, and under CES demand, private and social incentives are perfectly aligned.

Generalizing to variable elasticities of substitution, firms differ in market power which affects the trade-off between quantity, variety and productivity. Unlike symmetric firm models, the market distortions depend on the elasticity of demand and the elasticity of utility. Under CES demand, these two elasticities are constant and miss out on meaningful trade-offs. When these elasticities vary, the pattern of misallocations depends on how demand elasticities change with quantities, so policy analysis should ascertain these elasticities and take this information into account. While the modeling framework we consider provides a theoretical starting point to understand distortions across firms, enriching the model with market-specific features can yield better policy insights. Neary and Mrazova (2013) and Parenti et al. (2014) provide further generalizations of demand and costs, and Bilbiie et al. (2006) and more recently Opp et al. (2013) consider dynamic misallocations. Future work can also provide guidance on the design of implementable policies to realize further welfare gains.

We focus on international integration as a key policy tool to realize potential gains. Market expansion does not guarantee welfare gains under imperfect competition. As Dixit and Norman (1988) put it, this may seem like a "sad note" on which to end. But we find that integration provides welfare gains when the two demand-side elasticities ensure private and social incentives are aligned. Integrating with large markets also holds out the possibility of approaching the CES limit, which induces constant markups and therefore an efficient outcome. Even though integration can cause market and social objectives to perfectly align, "How Large is Large?" is an open question. Further work might quantify these relationships and thereby exhibit the scope of integration as a tool to improve the performance of imperfectly competitive markets.

References

Alessandria, G. and H. Choi, "Do Sunk Costs of Exporting Matter for Net Export Dynamics?," *The Quarterly Journal of Economics*, 2007, *122* (1), 289–336.

Arkolakis, C., A. Costinot, and A. Rodriguez-Clare, "New trade models, same old gains?," *American Economic Review*, 2012, *102* (1), 94–130.

_ , _ , D. Donaldson, and A. Rodriguez-Clare, "The Elusive Pro-Competitive Effects of Trade," *Working Paper*, 2012.

Arkolakis, Costas, "Market Penetration Costs and the New Consumers Margin in International Trade," *Journal of Political Economy*, 2010, *118* (6), 1151–1199.

Asplund, M. and V. Nocke, "Firm turnover in imperfectly competitive markets," *The Review of Economic Studies*, 2006, *73* (2).

Atkeson, A. and Burstein, "Innovation, Firm Dynamics, and international Trade," *Journal of Political Economy*, 2010, *118* (3), 433–484.

Baldwin, R. E. and F. Robert-Nicoud, "Trade and growth with heterogeneous firms," *Journal of International Economics*, 2008, *74* (1), 21–34.

Bartelsman, E. J. and M. Doms, "Understanding productivity: Lessons from longitudinal microdata," *Journal of Economic literature*, 2000, *38* (3).

Baumol, W. J. and D. F. Bradford, "Optimal Departures From Marginal Cost Pricing," *The American Economic Review*, 1970, *60* (3), 265–283.

Behrens, K., G. Mion, Y. Murata, and J. Südekum, "Trade, wages, and productivity," *International Economic Review*, forthcoming.

Behrens, Kristian and Yasusada Murata, "Trade, competition, and efficiency," *Journal of International Economics*, 2012, *87* (1), 1–17.

Benassy, J. P., "Taste for variety and optimum production patterns in monopolistic competition," *Economics Letters*, 1996, *52* (1), 41–47.

Bernard, A. B., J. B. Jensen, S. J. Redding, and P. K. Schott, "Firms in International Trade," *The Journal of Economic Perspectives*, 2007, *21* (3), 105–130.

_ , J. Eaton, J. B. Jensen, and S. Kortum, "Plants and Productivity in International Trade," *American Economic Review*, 2003.

Bilbiie, F. O., F. Ghironi, and M. J. Melitz, "Monopoly power and endogenous variety in dynamic stochastic general equilibrium: distortions and remedies," *manuscript, University of Oxford, Boston College, and Princeton University*, 2006.

Bulow, J. I. and P. Pfleiderer, "A note on the effect of cost changes on prices," *The Journal of Political Economy*, 1983, *91* (1), 182–185.

Bykadorov, Igor, Alexey Gorn, Sergey Kokovin, and Evgeny Zhelobodko, "Losses from trade in Krugman's model: almost impossible," *Working Paper*, 2014.

Campbell, J. R. and H. A. Hopenhayn, "Market Size Matters," *Journal of Industrial Economics*, 2005, *53* (1), 1–25.

Cole, Matthew T. and Ronald B. Davies, "Royale with Cheese: The Effect of Globalization on the Variety of Goods," *Review of Development Economics*, forthcoming.

Costinot, Arnaud and Andrés Rodriguez-Clare, "Trade Theory with Numbers: Quantifying the Consequences of Globalization," in "In: Helpman, E.(Ed.), Handbook of international economics" Citeseer 2013.

de Blas, B. and K. Russ, "Understanding Markups in the Open Economy under Bertrand Competition," *NBER Working Papers*, 2010.

Dhyne, Emmanuel, Amil Petrin, and Frederic Warzynski, "Prices, Markups and Quality at the Firm-Product Level," Technical Report, Mimeo, University of Minnesota 2011.

Dixit, A. K. and J. E. Stiglitz, "Monopolistic Competition and Optimum Product Diversity," *The American Economic Review*, 1977, *67* (3), 297–308.

_ **and V. Norman**, *Theory of international trade*, Cambridge Univ. Press, 1988.

Eckel, Carsten, "Globalization and specialization," *Journal of International Economics*, May 2008, *75* (1), 219–228.

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

Feenstra, R. and H. L. Kee, "Export variety and country productivity: Estimating the monopolistic competition model with endogenous productivity," *Journal of International Economics*, 2008, *74* (2), 500–518.

Feenstra, R. C., "A homothetic utility function for monopolistic competition models, without constant price elasticity," *Economics Letters*, 2003, *78* (1), 79–86.

_ , "New Evidence on the Gains from Trade," *Review of World Economics*, 2006, *142* (4), 617–641.

Foster, L., J. C. Haltiwanger, and C. J. Krizan, "Aggregate productivity growth. Lessons from microeconomic evidence," in "New developments in productivity analysis," University of Chicago Press, 2001.

_ **, J. Haltiwanger, and C. Syverson**, "Reallocation, firm turnover, and efficiency: Selection on productivity or profitability?," *American Economic Review*, 2008, *98* (1), 394–425.

Goldberg, P. and Michael Knetter, "Goods prices and exchange rates: what have we learned," *Journal of Economic Literature*, 1997, *5862*.

Grossman, Gene M. and Elhanan Helpman, *Innovation and Growth in the Global Economy*, MIT Press, 1993.

Hart, O. D., "Monopolistic competition in the spirit of Chamberlin: A general model," *The Review of Economic Studies*, 1985, *52* (4), 529.

Helpman, E. and P. R. Krugman, *Market Structure and Foreign Trade: increasing returns, imperfect competition, and the international economy*, MIT Press, 1985.

_ **, O. Itskhoki, and S. J. Redding**, "Trade and Labor Market Outcomes," *NBER Working Paper 16662*, 2011.

Holt, C. A. and S. K. Laury, "Risk aversion and incentive effects," *American Economic Review*, 2002, *92* (5), 1644–1655.

Katayama, H., S. Lu, and J. R. Tybout, "Firm-level productivity studies: illusions and a solution," *International Journal of Industrial Organization*, 2009, *27* (3), 403–413.

Khan, M. A. and Y. Sun, "Non-cooperative games with many players," *Handbook of Game Theory with Economic Applications*, 2002, *3*, 1761–1808.

Klenow, Peter J. and Benjamin A. Malin, "Microeconomic Evidence on Price-Setting," *Handbook of Monetary Economics*, 2010, *3*, 231–284.

Klette, Tor J., "Market power, scale economies and productivity: estimates from a panel of establishment data," *The Journal of Industrial Economics*, 1999, *47* (4), 451–476.

Krugman, P., "Increasing Returns, Monopolistic Competition, and International Trade," *Journal of International Economics*, 1979, *9* (4), 469–479.

Krugman, P. R., "Is free trade passé?," *The Journal of Economic Perspectives*, 1987, *1* (2).

Krugman, Paul R., "Scale Economies, Product Differentiation, and the Pattern of Trade," *American Economic Review*, 1980, *70* (5), 950–959.

Kuhn, K. U. and X. Vives, "Excess entry, vertical integration, and welfare," *The Rand Journal of Economics*, 1999, *30* (4), 575–603.

Loecker, Jan De and Frederic Warzynski, "Markups and Firm-Level Export Status," *The American Economic Review*, 2012, *102* (6), 2437–2471.

_ **and Pinelopi K. Goldberg**, "Firm Performance in a Global Market," *The Annual Review of Economics*, 2013.

_ **,** _ **, Amit K. Khandelwal, and Nina Pavcnik**, "Prices, markups and trade reform," Technical Report, National Bureau of Economic Research 2012.

Mankiw, N. G. and M. D. Whinston, "Free entry and social inefficiency," *The RAND Journal of Economics*, 1986, pp. 48–58.

Matsuyama, Kiminori, "Complementarities and Cumulative Processes in Models of Monopolistic Competition," *Journal of Economic Literature*, June 1995, *33* (2), 701–729.

Melitz, M. J. and S. J. Redding, "Heterogeneous Firms and Trade," *Handbook of International Trade (commissioned)*, August 2012.

Melitz, Marc and Daniel Trefler, "Gains from Trade when Firms Matter," *Journal of Economic Perspectives*, 2012, *26.*

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

_ **and Gianmarco I. P. Ottaviano**, "Market Size, Trade, and Productivity," *Review of Economic Studies*, October 2008, *75* (1), 295–316.

_ **and Stephen J. Redding**, "Firm Heterogeneity and Aggregate Welfare," Technical Report, National Bureau of Economic Research 2013.

Melvin and R. D. Warne, "Monopoly and the theory of international trade," *Journal of International Economics*, 1973, *3* (2), 117–134.

Neary, J. Peter and Monika Mrazova, "Not so demanding: Preference structure, firm behavior, and welfare," Technical Report, University of Oxford, Department of Economics 2013.

Nishimura, Kiyohiko G., Yasushi Ohkusa, and Kenn Ariga, "Estimating the mark-up over marginal cost: a panel analysis of Japanese firms 1971–1994," *International Journal of Industrial Organization*, 1999, *17* (8), 1077–1111.

Nocco, Antonella, Gianmarco I. P. Ottaviano, and Matteo Salto, "Monopolistic Competition and Optimum Product Selection: Why and How Heterogeneity Matters," *CEP Discussion Paper*, April 2013, *1206.*

Opp, Marcus M., Christine A. Parlour, and Johan Walden, "A Theory of Dynamic Resource Misallocation and Amplification," *Available at SSRN 2049447*, 2013.

Parenti, Mathieu, Philip Ushchev, and Jacques-François Thisse, "Toward a theory of monopolistic competition," 2014.

Pavcnik, N., "Trade Liberalization, Exit, and Productivity Improvements: Evidence from Chilean Plants," *The Review of Economic Studies*, 2002, *69* (1), 245–276.

Perloff, Jeffrey M. and Steven C. Salop, "Equilibrium with Product Differentiation," *The Review of Economic Studies*, 1985, *52* (1).

Post, T., M. J. Van den Assem, G. Baltussen, and R. H. Thaler, "Deal or no deal? Decision making under risk in a large-payoff game show," *The American Economic Review*, 2008, *98* (1), 38–71.

Roberts, Mark J. and Dylan Supina, "Output price, markups, and producer size," *European Economic Review*, 1996, *40* (3), 909–921.

_ **and** _ , "Output price and markup dispersion in micro data: The roles of producer heterogeneity and noise," *Advances in Applied Microeconomics*, 2001, *9*, 1–36.

Rudin, W., *Principles of mathematical analysis*, McGraw-Hill New York, 1964.

Saha, A., "Expo-power utility: A'flexible'form for absolute and relative risk aversion," *American Journal of Agricultural Economics*, 1993, pp. 905–913.

Samuelson, P. A., "The monopolistic competition revolution," *Monopolistic competition theory: studies in impact*, 1967, pp. 105–38.

Solow, R. M., *Monopolistic competition and macroeconomic theory*, Cambridge University Press, 1998.

Spence, M., "Product Selection, Fixed Costs, and Monopolistic Competition," *The Review of Economic Studies*, 1976, *43* (2), 217–235.

Stiglitz, J. E., "Towards a more general theory of monopolistic competition," *Prices, competition and equilibrium*, 1986, p. 22.

Syverson, C., "Market Structure and Productivity: A Concrete Example," *Journal of Political Economy*, 2004, *112* (6), 1181–1222.

_ , "What Determines Productivity?," *Journal of Economic Literature*, 2011, *49* (2).

Troutman, J. L., *Variational calculus and optimal control: Optimization with elementary convexity*, New York: Springer-Verlag, 1996.

Tybout, J. R., "Plant-and firm-level evidence on "new" trade theories," *Handbook of International Trade*, 2003, *1*, 388–415.

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

Vickers, John, "Concepts of competition," *Oxford Economic Papers*, 1995, pp. 1–23.

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

Weyl, E. G. and M. Fabinger, "Pass-through as an Economic Tool," *University of Chicago, mimeo*, September 2012.

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

A Appendix: Proofs

A.1 A Folk Theorem

In this context, we need to define the policy space. Provided M_e and $q(c)$, and assuming without loss of generality that all of $q(c)$ is consumed, allocations are determined. The only question remaining is what class of $q(c)$ the policymaker is allowed to choose from. A sufficiently rich class for our purposes is $q(c)$ which are positive and continuously differentiable on some closed interval and zero otherwise. This follows from the basic principle that a policymaker will utilize low cost firms before higher cost firms. Formally, we restrict q to be in sets of the form

$$\mathscr{Q}_{[0,c_d]} \equiv \{q \in \mathscr{C}^1, > 0 \text{ on } [0,c_d] \text{ and } 0 \text{ otherwise}\}.$$

We maintain Melitz's assumptions which imply a unique market equilibrium, and use the following shorthand throughout the proofs: $G(x) \equiv \int_0^x g(c)dc$, $R(x) \equiv \int_0^x c^{\rho/(\rho-1)} g(c)dc$.

Proof of Proposition 1. Assume a market equilibrium exists, which guarantees that $R(c)$ is finite for admissible c. First note that at both the market equilibrium and the social optimum, $L/M_e = f_e + fG(c_d)$ implies utility of zero so in both cases $L/M_e > f_e + fG(c_d)$. The policymaker's problem is

$$\max\ M_e L \int_0^{c_d} q(c)^\rho g(c)dc \text{ subject to } f_e + fG(c_d) + L\int_0^{c_d} cq(c)g(c)dc = L/M_e$$

where the maximum is taken over choices of $M_e,\ c_d,\ q \in \mathscr{Q}_{[0,c_d]}$. We will exhibit a globally optimal $q^*(c)$ for each fixed (M_e, c_d) pair, reducing the policymaker's problem to a choice of M_e and c_d. We then solve for M_e as a function of c_d and finally solve for c_d.

Finding $q^*(c)$ for M_e, c_d fixed. For convenience, define the functionals $V(q), H(q)$ by

$$V(q) \equiv L\int_0^{c_d} v(c,q(c))dc, \qquad H(q) \equiv L\int_0^{c_d} h(c,q(c))dc$$

where $h(c,x) \equiv xcg(c)$ and $v(c,x) \equiv x^\rho g(c)$. One may show that $V(q) - \lambda H(q)$ is strictly concave $\forall \lambda$.[29] Now for fixed (M_e, c_d), consider the problem of finding q^* given by

$$\max_{q \in \mathscr{Q}_{[0,c_d]}} V(q) \text{ subject to } H(q) = L/M_e - f_e - fG(c_d). \tag{3}$$

Following Troutman (1996), if some q^* maximizes $V(q) - \lambda H(q)$ on $\mathscr{Q}_{[0,c_d]}$ for some λ and satisfies the constraint then it is a solution to Equation (3). For any λ, a sufficient condition for some q^* to be a global maximum on $\mathscr{Q}_{[0,c_d]}$ is

$$D_2 v(c,q^*(c)) \quad = \quad \lambda D_2 h(c,q^*(c)). \tag{4}$$

This follows because (4) implies for any such q^*, $\forall \xi$ s.t. $q^* + \xi \in \mathscr{Q}_{[0,c_d]}$ we have $\delta V(q^*;\xi) = \lambda \delta H(q^*;\xi)$ (where δ denotes the Gateaux derivative in the direction of ξ) and q^* is a global max since $V(q) - \lambda H(q)$ is strictly concave. Condition (4) is $\rho q^*(c)^{\rho-1} g(c) = \lambda c g(c)$ which implies $q^*(c) = (\lambda c/\rho)^{1/(\rho-1)}$.[30] From above, this q^* serves as a solution to $\max V(q)$ provided that $H(q^*) = L/M_e - f_e - fG(c_d)$. This will be satisfied by an appropriate λ since for fixed λ

[29] Since h is linear in x, H is linear and since v is strictly concave in x (using $\rho < 1$) so is V.

[30] By abuse of notation we allow q^* to be ∞ at $c = 0$ since reformulation of the problem omitting this single point makes no difference to allocations or utility which are all eventually integrated.

we have

$$H(q^*) \;=\; L\int_0^{c_d}(\lambda c/\rho)^{1/(\rho-1)}cg(c)dc = L(\lambda/\rho)^{1/(\rho-1)}R(c_d)$$

so choosing λ as $\lambda^* \equiv \rho\left(L/M_e - f_e - fG(c_d)\right)^{\rho-1}/L^{\rho-1}R(c_d)^{\rho-1}$ makes q^* a solution. In summary, for each (M_e, c_d) a globally optimal q^* satisfying the resource constraint is

$$q^*(c) \;=\; c^{1/(\rho-1)}\left(L/M_e - f_e - fG(c_d)\right)/LR(c_d) \tag{5}$$

which must be >0 since $L/M_e - f_e - fG(c_d)$ must be >0 as discussed at the beginning.
Finding M_e for c_d fixed. We may therefore consider maximizing $W(M_e, c_d)$ where

$$W(M_e, c_d) \equiv M_e L\int_0^{c_d} q^*(c)^\rho g(c)dc = M_e L^{1-\rho}[L/M_e - f_e - fG(c_d)]^\rho R(c_d)^{1-\rho}. \tag{6}$$

Direct investigation yields a unique solution to the FOC of $M_e^*(c_d) = (1-\rho)L/(f_e + fG(c_d))$ and $d^2W/d^2M_e < 0$ so this solution maximizes W.
Finding c_d. Finally, we have maximal welfare for each fixed c_d from Equation (6), explicitly $\tilde{W}(c_d) \equiv W(M_e^*(c_d), c_d)$. We may rule out $c_d = 0$ as an optimum since this yields zero utility. Solving this expression and taking logs shows that

$$\ln\tilde{W}(c_d) \;=\; \ln\rho^\rho(1-\rho)^{1-\rho}L^{2-\rho} + (1-\rho)\left[\ln R(c_d) - \ln\left(f_e + fG(c_d)\right)\right].$$

Defining $B(c_d) \equiv \ln R(c_d) - \ln\left(f_e + fG(c_d)\right)$ we see that to maximize $\ln\tilde{W}(c_d)$ we need maximize only $B(c_d)$. In order to evaluate critical points of B, note that differentiating B and rearranging using $R'(c_d) = c_d^{\rho/(\rho-1)}g(c_d)$ yields

$$B'(c_d) = \left\{c_d^{\rho/(\rho-1)} - R(c_d)f/\left[f_e + fG(c_d)\right]\right\}/g(c_d)R(c_d). \tag{7}$$

Since $\lim_{c_d\longrightarrow 0} c_d^{\rho/(\rho-1)} = \infty$ and $\lim_{c_d\longrightarrow\infty} c_d^{\rho/(\rho-1)} = 0$ while $R(c_d)$ and $G(c_d)$ are bounded, there is a positive interval $[a,b]$ outside of which $B'(x) > 0$ for $x \le a$ and $B'(x) < 0$ for $x \ge b$. Clearly $\sup_{x\in(0,a]}B(x), \sup_{x\in[b,\infty)}B(x) < \sup_{x\in[a,b]}B(x)$ and therefore any global maximum of B occurs in (a,b). Since B is continuously differentiable, a maximum exists in $[a,b]$ and all maxima occur at critical points of B. From Equation (7), $B'(c_d) = 0$ iff $R(c_d)/c_d^{\rho/(\rho-1)} - G(c_d) = f_e/f$. For c_d that satisfy $B'(c_d) = 0$, M_e^* and q^* are determined and inspection shows the entire system corresponds to the market allocation. Therefore B has a unique critical point, which is a global maximum that maximizes welfare.

A.2 VES Market Allocation

Proof of Proposition 3. Consider a policymaker who faces a utility function $v(q) \equiv u'(q)\,q$. Provided $v(q)$ satisfies the regularity conditions used in the proof of optimality, it follows that the conditions below characterize the unique constrained maximum of $LM_e \int_0^{c_d} u'(q(c))\,q(c)dG$, where δ denotes the Lagrange multiplier:

$$\begin{aligned} u''(q(c))\,q(c) + u'(q(c)) &= \delta c, \\ u'(q(c_d))\,q(c_d)/\left(c_d q(c_d) + f/L\right) &= \delta, \\ \int_0^{c_d} u'(q(c))\,q(c)dG/\left(\int_0^{c_d} \left[cq(c) + f/L\right] dG + f_e/L\right) &= \delta, \\ M_e\left(\int_0^{c_d} Lcq(c) + f\,dG + f_e\right) &= L. \end{aligned}$$

Comparing these conditions, we see that if δ is the same as under the market allocation, the first three equations respectively determine each firm's optimal quantity choice, the ex post cost cutoff, and the zero profit condition while the fourth is the resource constraint and must hold under the market allocation. Therefore if this system has a unique solution, the market allocation maximizes $LM_e \int_0^{c_d} u'(q(c))\,q(c)dG$. Since these conditions completely characterize every market equilibrium, the assumed uniqueness of the market equilibrium guarantees such a unique solution.

A.3 Static Distortion Results

Proof of Proposition 5. The result relies on the following relationship we first prove:

$$\overline{\sigma} \equiv \sup_{c \le c_d^{\text{mkt}}} \varepsilon\left(q^{\text{mkt}}(c)\right) > \delta/\lambda > \inf_{c \le c_d^{\text{opt}}} \varepsilon\left(q^{\text{opt}}(c)\right) \equiv \underline{\sigma}. \tag{8}$$

To see this recall $\delta = M_e^{\text{mkt}} \int_0^{c_d^{\text{mkt}}} u'\left(q^{\text{mkt}}(c)\right) q^{\text{mkt}}(c)dG$ so $\overline{\sigma} > \delta/\lambda$ because

$$\delta/\overline{\sigma} = M_e^{\text{mkt}} \int_0^{c_d^{\text{mkt}}} \left(\varepsilon\left(q^{\text{mkt}}(c)\right)/\overline{\sigma}\right) u\left(q^{\text{mkt}}(c)\right) dG < M_e^{\text{mkt}} \int_0^{c_d^{\text{mkt}}} u\left(q^{\text{mkt}}(c)\right) dG \tag{9}$$

and λ is the maximum welfare per capita so $\lambda > M_e^{\text{mkt}} \int_0^{c_d^{\text{mkt}}} u\left(q^{\text{mkt}}(c)\right) dG > \delta/\overline{\sigma}$. A similar argument shows $\lambda\underline{\sigma} < \delta$, giving Equation (8). Now note that

$$\left[u''\left(q^{\text{mkt}}(c)\right) q^{\text{mkt}}(c) + u'\left(q^{\text{mkt}}(c)\right)\right]/\delta = c, \qquad u'\left(q^{\text{opt}}(c)\right)/\lambda = c. \tag{10}$$

And it follows from Equations (10) we have

$$\left[1-\mu\left(q^{\text{mkt}}(c)\right)\right]\cdot u'\left(q^{\text{mkt}}(c)\right)/u'\left(q^{\text{opt}}(c)\right)=\delta/\lambda. \tag{11}$$

Suppose $\mu'>0>(1-\varepsilon)'$, and it is sufficient to show $\inf_{c\leq c_d^{\text{mkt}}} 1-\mu\left(q^{\text{mkt}}(c)\right)\geq\overline{\sigma}$, since then Equations (8) and (11) show that $u'\left(q^{\text{mkt}}(c)\right)<u'\left(q^{\text{opt}}(c)\right)$ which implies $q^{\text{mkt}}(c)>q^{\text{opt}}(c)$. Since $\mu'>0>(1-\varepsilon)'$ and by assumption $\lim_{c\to 0} q^{\text{mkt}}(c)=\infty=\lim_{c\to 0} q^{\text{opt}}(c)$,

$$\inf_{c\leq c_d^{\text{mkt}}} 1-\mu\left(q^{\text{mkt}}(c)\right)=\lim_{q\to\infty} 1-\mu(q)=\lim_{q\to\infty}\varepsilon(q)+\varepsilon'(q)\,q/\varepsilon(q)\geq\lim_{q\to\infty}\varepsilon(q)=\overline{\sigma}.$$

Similarly, if $\mu'<0<(1-\varepsilon)'$ one may show that $\sup_{c\leq c_d^{\text{mkt}}} 1-\mu\left(q^{\text{mkt}}(c)\right)\leq\underline{\sigma}$, implying from Equations (8) and (11) that $q^{\text{mkt}}(c)<q^{\text{opt}}(c)$.

Now consider the cases when μ' and ε' have different signs, and since $\inf_q\varepsilon(q)>0$, from above in both cases it holds that $\inf_{q>0} 1-\mu(q)=\inf_{q>0}\varepsilon(q)$ and $\sup_{q>0} 1-\mu(q)=\sup_{q>0}\varepsilon(q)$. The arguments above have shown that $\sup_{q>0}\varepsilon(q)>\delta/\lambda>\inf_{q>0}\varepsilon(q)$ and therefore

$$\sup_{q>0} 1-\mu(q)>\delta/\lambda>\inf_{q>0} 1-\mu(q).$$

It follows from Equation (11) that for some c^*, $1-\mu\left(q^{\text{mkt}}(c^*)\right)=\delta/\lambda$ and therefore $u'\left(q^{\text{mkt}}(c^*)\right)=u'\left(q^{\text{opt}}(c^*)\right)$ so $q^{\text{mkt}}(c^*)=q^{\text{opt}}(c^*)$. Furthermore, $q^{\text{mkt}}(c)$ is strictly decreasing in c so with $\mu'\neq 0$, c^* is unique. Returning to Equation (11), using the fact that $q^{\text{mkt}}(c)$ is strictly decreasing in c also shows the relative magnitudes of $q^{\text{mkt}}(c)$ and $q^{\text{opt}}(c)$ for $c\neq c^*$.

Proof of Proposition 6. For $\alpha\in[0,1]$, define $v_\alpha(q)\equiv\alpha u'(q)\,q+(1-\alpha)\,u(q)$ and also define $w(q)\equiv u'(q)\,q-u(q)$ so $v_\alpha(q)=u(q)+\alpha w(q)$. Consider the continuum of maximization problems (indexed by α) defined as:

$$\max_{M_e,c_d,q(c)} LM_e\int_0^{c_d} v_\alpha(q(c))\,dG \text{ subject to } L\geq M_e\left(\int_0^{c_d} Lcq(c)+f\,dG+f_e\right). \tag{12}$$

Let the Lagrange multiplier associated with each α in Equation (12) be written as $\beta(\alpha)$. By appealing to the envelope theorem and differentiating (12) in M_e we have $\beta(\alpha)=M_e\int_0^{c_d} v_\alpha(q(c))\,dG$ and that $d\beta/d\alpha=M_e\int_0^{c_d} w(q(c))\,dG=M_e\int_0^{c_d} u(q(c))\,[\varepsilon(q)-1]\,dG<0$. The conditions char-

acterizing the solution to every optimum also imply

$$\beta(\alpha) = v_\alpha(q(c_d))/(c_d q(c_d) + f/L),$$

whereby we arrive at

$$\begin{aligned} dv_\alpha(q(c_d))/d\alpha &= (d\beta/d\alpha)(v_\alpha(q(c_d))/\beta) + \beta((dc_d/d\alpha)q(c_d) + c_d(dq(c_d)/d\alpha)) \\ &= w(q(c_d)) + v'_\alpha(q(c_d))(dq(c_d)/d\alpha) \\ &= w(q(c_d)) + \beta c_d(dq(c_d)/d\alpha) \end{aligned}$$

so cancellation and rearrangement, using the expressions for β, $d\beta/d\alpha$ above shows

$$\begin{aligned} \beta q(c_d)(dc_d/d\alpha) &= w(q(c_d)) - (v_\alpha(q(c_d))/\beta)(d\beta/d\alpha) \\ &= w(q(c_d)) - \left(v_\alpha(q(c_d))/M_e \int_0^{c_d} v_\alpha(q(c))\,dG\right) \cdot M_e \int_0^{c_d} w(q(c))\,dG. \end{aligned}$$

We conclude that $dc_d/d\alpha \gtrless 0$ when $w(q(c_d)) \int_0^{c_d} v_\alpha(q(c))\,dG \gtrless v_\alpha(q(c_d)) \int_0^{c_d} w(q(c))\,dG$. Expanding this inequality we have (suppressing $q(c)$ terms in integrands):

$$w(q(c_d)) \int_0^{c_d} u dG + \alpha w(q(c_d)) \int_0^{c_d} w dG \gtrless u(q(c_d)) \int_0^{c_d} w dG + \alpha w(q(c_d)) \int_0^{c_d} w dG.$$

Cancellation and expansion again show this is equivalent to

$$u'(q(c_d))\,q(c_d) \int_0^{c_d} u dG \gtrless u(q(c_d)) \int_0^{c_d} u' q(c) dG.$$

Finally, this expression can be rewritten $\varepsilon(q(c_d)) \gtrless \int_0^{c_d} \varepsilon(q(c))\,u(q(c))\,dG / \int_0^{c_d} u(q(c))\,dG$ and since $q(c)$ is strictly decreasing in c, we see $dc_d/d\alpha \gtrless 0$ when $\varepsilon' \lesseqgtr 0$. Note that Equation (12) shows $\alpha = 0$ corresponds to the social optimum while $\alpha = 1$ corresponds to the market equilibrium. It follows that when $\varepsilon' < 0$ that $dc_d/d\alpha > 0$ so we have $c_d^{\text{mkt}} > c_d^{\text{opt}}$ and vice versa for $\varepsilon' > 0$.

A.4 Welfare Gains from Trade

The sufficient condition for gains from trade follows from differentiating $U = M_e \int u(q) dG = \delta/\bar{\varepsilon}$ where the average elasticity of utility is $\bar{\varepsilon} \equiv \int \varepsilon u dG / \int u dG$. An increase in market size

raises the marginal utility of income at the rate of average markups $d\ln\delta/d\ln L = \int \mu pqdG/\int pqdG \equiv \tilde{\mu}$. From $d\ln\delta/d\ln L$ and $d\ln\bar{\varepsilon}/d\ln L$, the change in welfare is

$$\frac{d\ln U}{d\ln L} = \tilde{\mu}\left[1+\int \frac{1-\mu-\bar{\varepsilon}}{\mu+\mu' q/(1-\mu)}\frac{\varepsilon u}{\bar{\varepsilon}\int udG}dG\right] + \left[\frac{u_d}{\int udG}\frac{c_d g(c_d)}{\bar{\varepsilon}(1-\mu_d)}(\varepsilon_d-\bar{\varepsilon})(\tilde{\mu}-\mu_d)\right].$$

When preferences are aligned, the second term in square brackets is positive because μ and $(1-\varepsilon)$ move in the same direction. Change in the cost cutoff therefore has a positive effect on welfare, irrespective of the cost distribution $G(c)$. The first term in square brackets is also positive when preferences are aligned, given the regularity condition $(2+\mu'' q/\mu'(1-\mu)\geq 0)$.

Proof of Proposition 8. Following the discussion above, it is sufficient to show that for $\gamma(c) \equiv \varepsilon\left(\mu+\mu' q/\left(1-\mu\right)\right)^{-1}$,

$$1+\int \frac{1-\mu-\bar{\varepsilon}}{\mu+\mu' q/(1-\mu)}\frac{\varepsilon u}{\bar{\varepsilon}\int udG}dG = \int \left[1-\bar{\varepsilon}+\mu' q/\left(1-\mu\right)\right]\frac{\gamma u}{\bar{\varepsilon}\int udG}dG \geq 0. \tag{13}$$

This clearly holds for $\mu' \geq 0$, and for the other case where preferences are aligned, we have $\mu' < 0 < \varepsilon'$. Expanding Equation (13) for $\bar{\gamma} \equiv \int \gamma \cdot (u/\int udG)\, dG$ shows that

$$\int \left[1-\bar{\varepsilon}+\mu' q/\left(1-\mu\right)\right]\frac{\gamma u}{\bar{\varepsilon}\int udG}dG = \left[1-\bar{\varepsilon}-\overline{\mu}\right]\bar{\gamma}/\bar{\varepsilon}+1+\int\left[\overline{\mu}-\mu\right]\frac{\gamma u}{\bar{\varepsilon}\int udG}dG.$$

Since $\varepsilon' > 0$, $1-\varepsilon-\mu > 0$ and $\left[1-\bar{\varepsilon}-\overline{\mu}\right]\bar{\gamma}/\bar{\varepsilon}+1 > 0$. Therefore, it is sufficient to show that $\int\left[\overline{\mu}-\mu\right]\frac{\gamma u}{\bar{\varepsilon}\int udG}dG > 0$. This sufficient condition is equivalent to

$$\int \mu \frac{u}{\int udG}dG \geq \int \mu \frac{\gamma u}{\bar{\gamma}\int udG}dG \tag{14}$$

Since $\int \gamma(c)\cdot(u/\bar{\gamma}\int udG)\,dG = 1$ and $d\mu/dc > 0$, it follows that if $d\gamma/dc < 0$, then Equation (14) holds by stochastic dominance. As $d\gamma/dc < 0$ iff $d\gamma/dq > 0$, we examine the sign of $d\gamma/dq$ below.

$$\begin{aligned}\text{sign}\left\{d\gamma/dq\right\} &= \text{sign}\left\{d\ln\varepsilon\left(\mu+\mu' q/\left(1-\mu\right)\right)^{-1}/d\ln q\right\} \\ &= \text{sign}\left\{-\left(2+\mu'' q/\mu'\left(1-\mu\right)\right)\mu' q + \left(\varepsilon' q/\varepsilon - \mu' q/\left(1-\mu\right)\right)\left(\mu+\mu' q/\left(1-\mu\right)\right)\right\}.\end{aligned}$$

The additional hypothesis that $2+\mu'' q/\mu'(1-\mu)\geq 0$ guarantees each term above is positive, so $d\gamma/dq > 0$ and we conclude Equation (14) holds, giving the result.

A.5 Results Regarding the Impact of Large Markets

To arrive at the large market result, we first state Lemmas characterizing convergence in the large market and then show market allocations coincide with optimal allocations. Detailed proofs of the Lemmas are in the Online Appendix.

Lemma. *As market size becomes large:*

1. *Market revenue is increasing in market size and goes to infinity.*
2. *At the optimum, utility per capita is increasing in market size and goes to infinity.*
3. *Market entry goes to infinity.*

Proof. Online Appendix. □

Lemma. *For all market sizes and all positive marginal cost ($c > 0$) firms:*

1. *Profits ($\pi(c)$) and social profits ($\varpi(c) \equiv (1-\varepsilon(c))/\varepsilon(c) \cdot cq(c)L - f$) are bounded.*
2. *Total quantities ($Lq(c)$) in the market and optimal allocation are bounded.*

Proof. Online Appendix. □

Proposition 10. *Assume markups are interior. Then under the market allocation:*

1. $\lim_{L\to\infty} c_d^{\text{mkt}} = \infty$ *iff* $\lim_{L\to\infty} p\left(c_d^{\text{mkt}}\right) = \infty$ *iff* $\lim_{L\to\infty} Lq\left(c_d^{\text{mkt}}\right) = 0$.
2. $\lim_{L\to\infty} c_d^{\text{mkt}} = 0$ *iff* $\lim_{L\to\infty} p\left(c_d^{\text{mkt}}\right) = 0$ *iff* $\lim_{L\to\infty} Lq\left(c_d^{\text{mkt}}\right) = \infty$.
3. $\lim_{L\to\infty} c_d^{\text{mkt}} \in (0,\infty)$ *iff* $\lim_{L\to\infty} p\left(c_d^{\text{mkt}}\right) \in (0,\infty)$ *iff* $\lim_{L\to\infty} Lq\left(c_d^{\text{mkt}}\right) \in (0,\infty)$.

Similarly, under the optimal allocation:

1. $\lim_{L\to\infty} c_d^{\text{opt}} = \infty$ *iff* $\lim_{L\to\infty} u\circ q\left(c_d^{\text{opt}}\right)/\lambda q\left(c_d^{\text{opt}}\right) = \infty$ *iff* $\lim_{L\to\infty} Lq\left(c_d^{\text{opt}}\right) = 0$.
2. $\lim_{L\to\infty} c_d^{\text{opt}} = 0$ *iff* $\lim_{L\to\infty} u\circ q\left(c_d^{\text{opt}}\right)/\lambda q\left(c_d^{\text{opt}}\right) = 0$ *iff* $\lim_{L\to\infty} Lq\left(c_d^{\text{opt}}\right) = \infty$.
3. $\lim_{L\to\infty} c_d^{\text{opt}} \in (0,\infty)$ *iff* $\lim_{L\to\infty} u\circ q\left(c_d^{\text{opt}}\right)/\lambda q\left(c_d^{\text{opt}}\right) \in (0,\infty)$ *iff* $\lim_{L\to\infty} Lq\left(c_d^{\text{opt}}\right) \in (0,\infty)$.

Proof. Note the following zero profit relationships that hold at the cost cutoff c_a, suppressing the market superscripts throughout we have:

$$u'(q(c_d))/\delta - f/\left[Lq(c_d)\cdot\mu\circ q(c_d)/(1-\mu\circ q(c_d))\right] = c_d, \tag{15}$$

$$Lc_d q(c_d)\cdot\mu\circ q(c_d)/(1-\mu\circ q(c_d)) = f. \tag{16}$$

First, if $\lim_{L\to\infty} Lq(c_d) = 0$, Equation (16) implies $c_d\cdot\mu\circ q(c_d)/(1-\mu\circ q(c_d)) \longrightarrow \infty$. Clearly $q(c_d)\longrightarrow 0$ and since $\lim_{q\to 0}\mu(q)\in(0,1)$, $\mu\circ q(c_d)/(1-\mu\circ q(c_d))$ is bounded, and therefore $c_d\longrightarrow\infty$. Now suppose $c_d\longrightarrow\infty$ and since $c_d \le u'(q(c_d))/\delta$, $u'(q(c_d))/\delta\longrightarrow\infty$. Finally, if $u'(q(c_d))/\delta\longrightarrow\infty$, since $\delta\longrightarrow\infty$, necessarily $q(c_d)\longrightarrow 0$ so we find $\mu\circ q(c_d)/(1-\mu\circ q(c_d))$

is bounded. It follows from Equation (16) that $Lc_d q(c_d)$ is bounded, so from Equation (15), $Lq(c_d) \cdot u'(q(c_d))/\delta$ is bounded so $Lq(c_d) \longrightarrow 0$.

If $\lim_{L\to\infty} Lq(c_d) = \infty$, $q(c_d) \longrightarrow 0$ so from $\lim_{q\to 0}\mu(q) \in (0,1)$, $\mu \circ q(c_d)/(1-\mu \circ q(c_d))$ is bounded. Therefore from Equation (16), $c_d \longrightarrow 0$. Now assume $c_d \longrightarrow 0$ so from (16), $Lq(c_d) \cdot \mu \circ q(c_d)/(1-\mu \circ q(c_d)) \longrightarrow \infty$ which implies with Equation (15) that $u'(q(c_d))/\delta \longrightarrow 0$. Finally, if $u'(q(c_d))/\delta \longrightarrow 0$, (15) shows $c_d \longrightarrow 0$.

The second set of equivalences follows from examining the conditions for a firm at the limiting cost cutoff $c_d^\infty \in (0,\infty)$. The argument for the optimal allocation is similar. □

Lemma. *Assume interior convergence. Then as market size grows large:*

1. *In the market, $p(c)$ converges in $(0,\infty)$ for $c > 0$ and $Lq(c_d)$ converges in $(0,\infty)$.*
2. *In the optimum, $u \circ q(c)/\lambda q(c)$ and $Lq(c_d)$ converge in $(0,\infty)$ for $c > 0$.*

Proof. Online Appendix. □

Lemma. *Assume interior convergence and large market identification. Then for the market and social optimum, $Lq(c)$ converges for $c > 0$.*

Proof. Online Appendix. □

Lemma. *At extreme quantities, social and private markups align as follows:*

1. *If $\lim_{q\to 0} 1-\varepsilon(q) < 1$ then $\lim_{q\to 0} 1-\varepsilon(q) = \lim_{q\to 0}\mu(q)$.*
2. *If $\lim_{q\to\infty} 1-\varepsilon(q) < 1$ then $\lim_{q\to\infty} 1-\varepsilon(q) = \lim_{q\to\infty}\mu(q)$.*

Proof. Online Appendix. □

Lemma. *Assume interior convergence and large market identification. As market size grows large:*

1. *$q(c)/q(c_d) \longrightarrow (c/c_d)^{-1/\alpha}$ with $\alpha = \lim_{q\to 0}\mu(q)$.*
2. *The cost cutoffs for the social optimum and market converge to the same value.*
3. *The entrant per worker ratios M_e/L converge to the same value.*

Proof. Define $\Upsilon(c/c_d)$ by (the above results show this limit is well defined)

$$\Upsilon(c/c_d) \equiv \lim_{q\to 0} u'(\Upsilon(c/c_d)q)/u'(q) = c/c_d.$$

We will show in fact that $\Upsilon(c/c_d) = (c/c_d)^{-\alpha}$. It follows from the definition that Υ is weakly decreasing, and the results above show Υ is one to one, so it is strictly decreasing. Define $f_q(z) \equiv u'(zq)/u'(q)$ so $\lim_{q\to 0} f_q(z) = \Upsilon^{-1}(z)$ for all $\Upsilon^{-1}(z) \in (0,1)$. Note

$$f_q'(z) = u''(zq)q/u'(q) = -\mu(zq) \cdot u'(zq)/zu'(q)$$

so since $\lim_{q\to 0}\mu(zq)=\mu^{\infty}\in(0,1)$ and $\lim_{q\to 0}u'(zq)/zu'(q)=\Upsilon^{-1}(z)/z$, we know that $\lim_{q\to 0}f_q'(z)=-\mu^{\infty}\Upsilon^{-1}(z)/z$. On any strictly positive closed interval I, μ and $u'(zq)/zu'(q)$ are monotone in z so $f_q'(z)$ converges uniformly on I as $q\longrightarrow 0$. Rudin (1964) (Thm 7.17) shows

$$\lim_{q\longrightarrow 0}f_q'(z)=d\lim_{q\longrightarrow 0}f_q(z)/dz=-\mu^{\infty}\Upsilon^{-1}(z)/z=d\Upsilon^{-1}(z)/dz. \tag{17}$$

We conclude that $\Upsilon^{-1}(z)$ is differentiable and thus continuous. Given the form deduced in (17), $\Upsilon^{-1}(z)$ is continuously differentiable. Since $d\Upsilon^{-1}(z)/dz=1/\Upsilon'\circ\Upsilon^{-1}(z)$, composing both sides with $\Upsilon(z)$ and using (17) we have $\Upsilon'(z)=-\Upsilon(z)/\mu^{\infty}z$. Therefore Υ is CES, in particular $\Upsilon(z)=z^{-1/\mu^{\infty}}$.

Finally, let c_{∞}^{opt} and c_{∞}^{mkt} be the limiting cost cutoffs as $L\longrightarrow\infty$ for at the social optimum and market, respectively. Letting $q^{\text{opt}}(c)$, $q^{\text{mkt}}(c)$ denote the socially optimal and market quantities, we know from above that for all $c>0$:

$$q^{\text{opt}}(c)/q^{\text{opt}}\left(c_d^{\text{opt}}\right)\longrightarrow\left(c_{\infty}^{\text{opt}}/c\right)^{1/\alpha},\quad q^{\text{mkt}}(c)/q^{\text{mkt}}\left(c_d^{\text{mkt}}\right)\longrightarrow\left(c_{\infty}^{\text{mkt}}/c\right)^{1/\alpha}. \tag{18}$$

Now consider the conditions involving f_e, $\int_0^{c_d^{\text{mkt}}}\pi(c)dG=f_e=\int_0^{c_d^{\text{opt}}}\varpi(c)dG$. Expanding,

$$L\int_0^{c_d^{\text{mkt}}}\frac{\mu\circ q^{\text{mkt}}(c)}{1-\mu\circ q^{\text{mkt}}(c)}cq^{\text{mkt}}(c)dG-fG(c_d^{\text{mkt}})=L\int_0^{c_d^{\text{opt}}}\frac{1-\varepsilon\circ q^{\text{opt}}(c)}{\varepsilon\circ q^{\text{opt}}(c)}cq^{\text{opt}}(c)dG-fG(c_d^{\text{opt}}).$$

It necessarily follows that

$$\begin{aligned}&\lim_{L\longrightarrow\infty}L\int_0^{c_d^{\text{mkt}}}\mu\circ q^{\text{mkt}}(c)/\left(1-\mu\circ q^{\text{mkt}}(c)\right)\cdot cq^{\text{mkt}}(c)dG-fG(c_d^{\text{mkt}})=\\&\lim_{L\longrightarrow\infty}L\int_0^{c_d^{\text{opt}}}\left(1-\varepsilon\circ q^{\text{opt}}(c)\right)/\varepsilon\circ q^{\text{opt}}(c)\cdot cq^{\text{opt}}(c)dG-fG(c_d^{\text{opt}}).\end{aligned} \tag{19}$$

Using Equation (18), we see that $Lq^{\text{opt}}(c)$ and $Lq^{\text{mkt}}(c)$ converge uniformly on any strictly positive closed interval. Combined with the fact that $\lim_{q\to 0}\mu(q)=\lim_{q\to 0}1-\varepsilon(q)$, we see from Equation (19) the limits of the $\mu/(1-\mu)$ and $(1-\varepsilon)/\varepsilon$ terms are equal and factor out of Equation (19), leaving

$$\begin{aligned}&\lim_{L\longrightarrow\infty}Lc_{\infty}^{\text{mkt}}q^{\text{mkt}}(c_{\infty}^{\text{mkt}})\int_0^{c_d^{\text{mkt}}}(c/c_{\infty}^{\text{mkt}})(c/c_d^{\text{mkt}})^{-1/\alpha}dG-fG(c_d^{\text{mkt}})=\\&\lim_{L\longrightarrow\infty}Lc_{\infty}^{\text{opt}}q^{\text{opt}}(c_{\infty}^{\text{opt}})\int_0^{c_d^{\text{opt}}}(c/c_{\infty}^{\text{opt}})(c/c_d^{\text{opt}})^{-1/\alpha}dG-fG(c_d^{\text{opt}}).\end{aligned}$$

Noting $f\left(1-\mu^{\infty}\right)/\mu^{\infty} = Lc_{\infty}^{\text{mkt}} q^{\text{mkt}}(c_{\infty}^{\text{mkt}}) = Lc_{\infty}^{\text{opt}} q^{\text{opt}}(c_{\infty}^{\text{opt}})$, we therefore have

$$\lim_{L\longrightarrow\infty} \int_0^{c_d^{\text{mkt}}} (c/c_{\infty}^{\text{mkt}})^{1-1/\alpha} (c_{\infty}^{\text{mkt}}/c_d^{\text{mkt}})^{-1/\alpha} dG - G(c_d^{\text{mkt}}) =$$
$$\lim_{L\longrightarrow\infty} \int_0^{c_d^{\text{opt}}} (c/c_{\infty}^{\text{opt}})^{1-1/\alpha} (c_{\infty}^{\text{opt}}/c_d^{\text{opt}})^{-1/\alpha} dG - G(c_d^{\text{opt}})$$

so that finally evaluating the limits, we have

$$\int_0^{c_{\infty}^{\text{mkt}}} \left[(c/c_{\infty}^{\text{mkt}})^{1-1/\alpha} - 1\right] dG = \int_0^{c_{\infty}^{\text{opt}}} \left[(c/c_{\infty}^{\text{opt}})^{1-1/\alpha} - 1\right] dG. \tag{20}$$

Letting $h(w) \equiv \int_0^w \left[(c/w)^{1-1/\alpha} - 1\right] dG$, we see that $h'(w) = \int_0^w (1/\alpha - 1)\, c^{1-1/\alpha} w^{1/\alpha - 2} dG$ and since $\alpha = \mu^{\infty} \in (0,1)$, $h' > 0$. Since h is strictly increasing, there is a unique c_{∞}^{opt}, namely $c_{\infty}^{\text{opt}} = c_{\infty}^{\text{mkt}}$ such that Equation (20) holds. Checking the conditions for L/M_e show they coincide between the market and social optimum as well. □