Monopolistic Competition and Optimum Product Diversity Under Firm Heterogeneity*

Swati Dhingra
CEP, London School of Economics

John Morrow
CEP, London School of Economics

This Draft: November 23, 2012

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 sources 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 George Alessandria, Costas Arkolakis, Roc Armenter, Andy Bernard, Satyajit Chatterjee, Davin Chor, Steve Durlauf, Charles Engel, Thibault Fally, Rob Feenstra, Keith Head, Wolfgang Keller, Jim Lin, Emanuel Ornelas, Gianmarco Ottaviano, Mathieu Parenti, Nina Pavcnik, Andres Rodriguez-Clare, Tom Sampson, Daniel Sturm, Jacques Thisse, John Van Reenen, Ben Zissimos and Mian Zhu for insightful comments, Katheryn Russ for AEA discussion and Tim Besley for advice. This paper has benefited from helpful comments of participants at AEA 2011, DIME-ISGEP 2010, ETSG 2012, HSE St Petersburg, ISI, FIW, LSE, Louvain, Mannheim, Oxford, Philadelphia Fed, Princeton and Wisconsin. Swati thanks the IES (Princeton) for their hospitality. 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.

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 sources 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] E.g., Bartelsman and Doms (2000); Tybout (2003); Bernard, Jensen, Redding and Schott (2007).
[2] E.g., Pavcnik (2002); Asplund and Nocke (2006); Foster et al. (2001); Melitz and Redding (2012).
[3] E.g., 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. Firm heterogeneity does not introduce any new distortions, but firms earn positive profits. This result seems surprising, based on the logic of average cost pricing which 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 forthcoming). 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 are perfectly aligned with social benefits only under CES demand. More generally, market power induces distortions relative to optimal allocations and demand-side elasticities determine these distortions.

The pattern of distortions is determined by two elasticities: the demand elasticity, which

[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. 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).

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 under-produce 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) 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. 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. (forthcoming) 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 markups matter for both welfare gains and allocational efficiency.[9] 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.

[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. 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.

[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.

2 Model

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. We adopt the VES demand structure of Dixit and Stiglitz and the heterogeneous firm framework of Melitz, and refer to this setting as the Dixit-Stiglitz-Melitz framework. In this Section, we briefly recap the implications of asymmetric costs for consumers, firms and equilibrium outcomes.

2.1 Consumers

A mass L of identical consumers in an economy are each endowed with one unit of labor and face a wage rate w normalized to one. Preferences are identical across all consumers. Let M_e denote the mass of entering varieties and $q(c)$ denote the quantity consumed of variety c by each consumer. A consumer has preferences over differentiated goods $U(M_e, q)$ which take the general VES form:

$$U(M_e, q) \equiv M_e \int u(q(c))dG. \tag{1}$$

Here u denotes utility from an individual variety and $\int u(q)dG$ denotes utility from a unit bundle of differentiated varieties. Under CES preferences, $u(q) = q^\rho$ as specified in Dixit-Stiglitz and Krugman (1980).[10] More generally, we assume preferences satisfy usual regularity conditions which guarantee well defined consumer and firm problems.

Definition 1. (Regular Preferences) u satisfies the following: $u(0)$ is normalized to zero, u is twice continuously differentiable, increasing and concave, $(u'(q) \cdot q)'$ is strictly decreasing in quantity, and the elasticity of marginal utility $\mu(q) \equiv |qu''(q)/u'(q)|$ is less than one.

For each variety c, VES preferences induce an inverse demand $p(q(c)) = u'(q(c))/\delta$ where δ is a 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 $q(c)$ and aggregating over all c, the

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

[11] Utility functions not satisfying inada conditions are permissible but may require parametric restrictions to ensure existence. We will assume inada conditions on utility and revenue, though they are not necessary for all results.

budget multiplier is $\delta = M_e \int_0^{c_d} u'(q(c)) \cdot q(c) dG$.

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. Each firm produces a single variety, so the mass of entering firms is the mass of entering varieties M_e. Upon entry, each firm receives a unit cost c drawn from a distribution G with continuously differentiable pdf g.[12]

After entry, should a firm produce, it incurs a fixed cost of production f. Each firm faces an inverse demand of $p(q(c)) = u'(q(c))/\delta$ and acts as a monopolist of variety c. Post entry, the profit of firm c is $\pi(c)$ where $\pi(c) \equiv \max_{q(c)}[p(q(c)) - c]q(c)L - f$. The regularity conditions guarantee the monopolist's FOC is optimal and the quantity choice is determined by the equality of marginal revenue and marginal cost. Specifically, $p + q \cdot u''(q)/\delta = c$ and the markup rate is $(p(c) - c)/p(c) = -qu''(q)/u'(q)$. This shows that the elasticity of marginal utility summarizes the inverse demand elasticity as

$$\mu(q) \equiv |qu''(q)/u'(q)| = |d \ln p(q)/d \ln q| = (p(c) - c)/p(c).$$

2.3 Market Equilibrium

Profit maximization implies firms produce if they can earn non-negative profits. We denote the cutoff cost level of firms that are indifferent between producing and exiting from the market as c_d. The cutoff cost c_d is fixed by the zero profit condition, $\pi(c_d) = 0$. Since firms with cost draws higher than the cutoff level do not produce, the mass of producers is $M = M_e G(c_d)$.

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, $\int \pi(c) dG = f_e$. The next two Sections examine the efficiency properties of this Dixit-Stiglitz-Melitz 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

[12] Some additional regularity conditions on G are required for existence of a market equilibrium in Melitz.

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. Nonetheless, 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 all VES demand structures except CES. Following Dixit and Stiglitz, we start with an exposition of efficiency under CES demand and then discuss market inefficiency under VES demand.

3.1 Welfare under Isoelastic Demand

A policymaker maximizes individual welfare U as given in Equation (1).[13] The policymaker is unconstrained and chooses the mass of entrants, quantities and types of firms that produce. At the optimum, zero quantities will be chosen for varieties above a cost threshold c_d. Therefore, all optimal allocational decisions can be summarized by quantity $q(c)$, potential variety M_e and productivity c_d. 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

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

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. Proposition 1 shows the market provides the first-best quantity, variety and productivity.

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

The proof of Proposition 1 differs from standard symmetric firm monopolistic competition results because optimal quantity varies non-trivially with unit cost, variety and cutoff productivity. We discuss the rationale for optimality below.

In symmetric firm models with CES demand, firms charge positive markups which result in lower quantities than those implied by marginal cost pricing. However, the markup is constant so the market price (and hence marginal utility) is proportional to unit cost, ensuring proportionate reduction in quantity from the level that would be observed under marginal cost pricing (Baumol and Bradford 1970). Moreover, free entry ensures price equals average cost so profits exactly finance the fixed cost of production. The market therefore induces firms to indirectly

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

internalize the effects of higher variety on consumer surplus, resulting in an efficient market equilibrium (Grossman and Helpman 1993).

With heterogeneous firms, markups continue to be constant, which implies profits are heterogeneous. One might imagine enforcing average cost pricing across different firms would induce an efficient allocation but, average cost pricing is too low to compensate firms because it will not cover ex ante entry costs. Instead, the market ensures prices above average costs at a level that internalizes the losses faced by exiting firms. Post entry, surviving firms charge prices higher than average costs ($p(c) \geq [cq(c) + f/L]/q(c)$) which compensates them for the possibility of paying f_e to enter and then being too unproductive to survive. CES demand ensures that c_d and M_e are at optimal levels that fix $p(c_d)$, thereby fixing absolute prices to optimal levels. The marginal entrant imposes a business stealing externality on other firms, but also does not account for the variety gain and productivity loss from its entry. These effects exactly offset each other, and wages induced by the market exactly reflect the shadow value of resources at the optimal allocation.

The way in which CES preferences cause firms to optimally internalize aggregate economic conditions can be made clear by defining the elasticity of utility $\varepsilon(q) \equiv u'(q) \cdot q/u(q)$ and the 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, there is a multiplier λ which encapsulates the shadow cost of labor. The social surplus is $u(q) - \lambda cq$ and the optimal quantities ensure $u'(q(c)) = \lambda c$. Therefore, 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, a 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. A direct implication of Proposition 1 is that laissez faire industrial policy is optimal under constant elasticity demand. In the next subsection, we examine the role of variable elasticities on market efficiency.[14]

3.2 Welfare beyond Isoelastic Demand

Efficiency of the market equilibrium in a Dixit-Stiglitz-Melitz framework is tied to CES demand. To highlight this, we consider the general class of variable elasticity of substitution (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.*[15]

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.*

[14]The CES efficiency result may seem surprising in the context of Dixit and Stiglitz (1977) who find that market allocations are second-best but not first-best. Dixit and Stiglitz consider two sectors (a differentiated goods sector and a homogeneous goods sector) and assume a general utility function to aggregate across these goods. This causes the markups charged in the homogeneous and differentiated goods to differ, leading to inefficient market allocations. In keeping with Melitz, we consider a single sector to develop results for market efficiency in terms of markups.

[15]CES demand is necessary but not sufficient for efficiency. To see this, extend the CES demand of Melitz to CES-Benassy preferences $U(M_e, c_d, q) \equiv \nu(M_e) \int_0^{c_d} q(c)^\rho g(c) dc$. Here u is CES but varieties and the unit bundle are valued differently through $\nu(M_e)$. Market allocations under CES-Benassy are the same as CES. However, firms do not fully internalize consumers' taste for variety, leading to suboptimal allocations. 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 are optimal only if taste for variety exactly equals the markup to cost ratio, and Helpman and Krugman (1985) and Feenstra and Kee (2008) derive a GDP function for this economy.

This result shows that the market resource allocation is generally not aligned with the social optimum under VES demand. The market and efficient allocations are solutions to:

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

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

For CES demand, $u(q) = q^\rho$ while $u'(q)q = \rho q^\rho$ implying revenue maximization is perfectly aligned with welfare maximization. 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. This leads us to an examination of the nature of misallocations induced by the market.

4 Market Distortions and Variable Elasticities

Although we have identified the conflict between private markups $\mu(q)$ captured by firms and social markups $1 - \varepsilon(q)$ that would maximize welfare as the source of distortions, we have not investigated the nature of these distortions. In this Section, we characterize how the market allocates resources relative to the social optimum in terms of markups. Specifically, the bias in market quantity, productivity and variety is determined by how private and social markups vary with quantity ($\mu'(q)$ and $(1 - \varepsilon(q))'$). We start with a discussion of markup and quantity patterns, and then show that different markup patterns induce very different biases in market allocations. We summarize the pattern of distortions and discuss empirical evidence for different demand characteristics. To highlight the importance of firm heterogeneity and variable markups, we finally compare our results with distortions under symmetric firms.

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. For CES demand, markups are constant ($\mu' = 0$).

The sign of $(1-\varepsilon(q))'$ determines how social markups vary with quantity. When it is positive $(1-\varepsilon(q))' > 0$, social markups are higher at higher levels of quantity. As above, this implies a negative correlation between social markups $1-\varepsilon$ and unit costs c. Conversely, when $(1-\varepsilon(q))' < 0$, the "boutique" varieties which are consumed in small quantities provide relatively higher social markups. Under CES preferences, $(1-\varepsilon(q))'$ is again zero.

To bring out the distinction in distortions for different markup patterns, Definition 2 below characterizes preferences as aligned when private and social markups move in the same direction and misaligned when they move in different directions.

Definition 2. 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,[16] HARA and generalized CES (proposed by Dixit and Stiglitz).[17]

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 Quantity, Productivity and Entry Distortions

We now characterize the misallocations by demand characteristics. The distortions in quantity, productivity and entry are discussed in turn.

4.2.1 Quantity Bias

Quantity distortions across firms depend on whether private and social incentives are aligned or misaligned. We show that when private and social markups are misaligned, market quantities

[16] 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.

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

$q^{\text{mkt}}(c)$ 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:

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

When incentives are misaligned, market and optimal quantities are too high or too low across all varieties. In particular, when $\mu' > 0 > (1-\varepsilon)'$, the market over-rewards firms producing higher quantities and all firms over-produce $q^{\text{mkt}}(c) > q^{\text{opt}}(c)$. When $\mu' < 0 < (1-\varepsilon)'$, market production is too low ($q^{\text{mkt}}(c) < q^{\text{opt}}(c)$). Therefore, firms are either over-rewarded ($\mu' > 0$) for producing q or under-rewarded ($\mu' < 0$), 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. 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 low cost firms ($q^{\text{mkt}} > q^{\text{opt}}$ for low c and $q^{\text{mkt}} < q^{\text{opt}}$ for high c). The market over-rewards low cost firms who impose an externality on high cost firms. When $\mu', (1-\varepsilon)' < 0$, the bias is reversed and quantities are biased towards high cost firms. Therefore, when private and social markups are aligned, the market under or over produces quantity, depending on a firm's costs. Proposition 4 summarizes the bias in market quantities.

Proposition 4. *When preferences are misaligned, $q^{\text{mkt}}(c)$ and $q^{\text{opt}}(c)$ never cross:*
1. If $\mu' > 0 > (1-\varepsilon)'$, market quantities are too high: $q^{\text{mkt}}(c) > q^{\text{opt}}(c)$.
2. If $\mu' < 0 < (1-\varepsilon)'$, 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).
1. If $\mu' > 0$ and $(1-\varepsilon)' > 0$, $q^{\text{mkt}}(c) > q^{\text{opt}}(c)$ for $c < c^*$ and $q^{\text{mkt}}(c) < q^{\text{opt}}(c)$ for $c > c^*$.
2. If $\mu' < 0$ and $(1-\varepsilon)' < 0$, $q^{\text{mkt}}(c) < q^{\text{opt}}(c)$ for $c < c^*$ and $q^{\text{mkt}}(c) > q^{\text{opt}}(c)$ for $c > c^*$.

This shows the misallocation in production differs across firms, and variable demand elasticities characterize the pattern of misallocations.

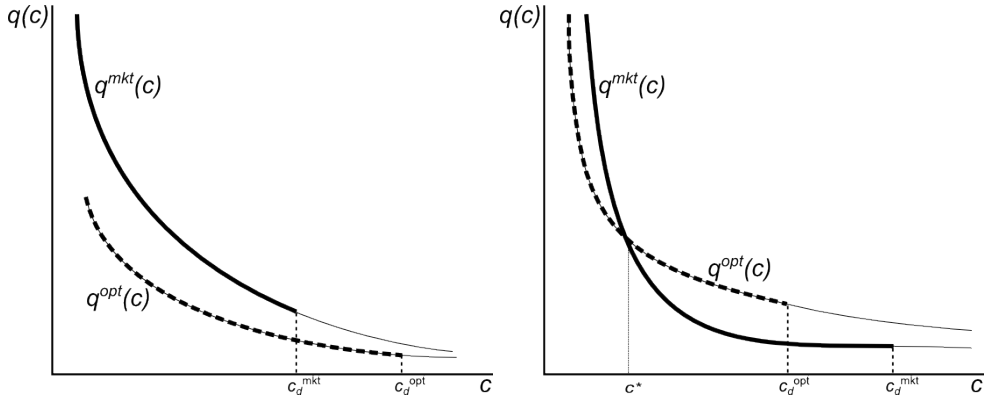
4.2.2 Productivity Bias

The distortion in firm selection is determined by the relation between social markups and quantity. Proposition 5 shows that market productivity is either too low or high, depending on whether social markups are increasing or decreasing. Revenue of the cutoff productivity firm is proportional to $u'(q)q$ while its contribution to utility is $u(q)$. Therefore, the gap in productivity cutoffs is determined by $\varepsilon(q)$ and the market bias depends on $\varepsilon'(q)$. Increasing social markups $(1-\varepsilon)' > 0$ encourage higher optimal quantity at lower costs. In general equilibrium, this translates into a lower cost cutoff at the optimum, so market costs are too high.

Proposition 5. *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 4 and 5 explain how the market misallocates resources across firms. Figure 1 illustrates the bias in firm-level production for aligned and misaligned preferences when private markups increase in quantity.

Figure 1: Bias in Firm Production by Preferences



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

4.2.3 Entry Bias

Although a comparison of market entry to optimal entry in this setting is generally hard to make, Proposition 6 establishes their relative levels when private and social markups are aligned. Market entry is too low when private markups are increasing and market entry is too high when private markups are decreasing. When incentives are misaligned, quantity and productivity distortions have opposing effects on entry so the entry bias depends on the magnitudes of exogenous parameters.

Proposition 6. *The market over or under produces varieties, as follows:*
1. If $(1-\varepsilon)', \mu' < 0$*, the market has too much entry:* $M_e^{\text{mkt}} > M_e^{\text{opt}}$.
2. If $(1-\varepsilon)', \mu' > 0$*, the market has too little entry:* $M_e^{\text{mkt}} < M_e^{\text{opt}}$*. (Assuming* $\mu' q/\mu \leq 1$*).*

4.2.4 Empirical Evidence for Demand Characteristics

This Section has shown that the underlying demand structure can lead to very different distortions. For ease of reference, Table 2 summarizes the misallocations by demand characteristics.

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: $c_d^{\text{mkt}} < c_d^{\text{opt}}$ Entry Ambiguous	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 Low: $c_d^{\text{mkt}} > c_d^{\text{opt}}$ Entry Too Low: $M_e^{\text{mkt}} < M_e^{\text{opt}}$
$\mu' < 0$	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 High: $c_d^{\text{mkt}} < c_d^{\text{opt}}$ Entry Too High: $M_e^{\text{mkt}} > M_e^{\text{opt}}$	Quantities Too Low: $q^{\text{mkt}}(c) < q^{\text{opt}}(c)$ Productivity Too Low: $c_d^{\text{mkt}} > c_d^{\text{opt}}$ Entry Ambiguous

As the pattern of misallocation depends on how private and social markups vary with quantity, a natural question is whether empirical work can identify which case in Table 2 is relevant. Systematic empirical evidence on the relationship between markups and quantities is sparse (Weyl and Fabinger 2012). However, existing studies suggest that the relationship differs across markets, and therefore we cannot restrict attention to a single case. For example, De Loecker, Goldberg, Khandelwal and Pavcnik (2012) directly estimate the cross-sectional relationship for large Indian manufacturers and find private markups are increasing in quantity $\mu'(q) > 0$.[18] With direct information on prices and costs, Cunningham (2011) instead finds evidence for decreasing private markups among drugstore products in the US. Social markups

[18]The bulk of empirical work on pass-through rates and firm selection also suggests private markups increase with quantities. However, some studies suggest markups decrease with quantities as they find a rise in markups after entry (Zhelobodko et al. forthcoming).

are rarely observable, and there is lack of consensus on how they respond to quantity (Vives 2001). Spence suggests social markups decrease with quantity while Dixit and Stiglitz propose increasing social markups. Therefore, we cannot rule out specific cases without further empirical investigation of the market under consideration.[19]

4.3 Comparison with Symmetric Firms

In the remainder of this Section, we compare the bias in market allocations under symmetric and heterogeneous firms. Dixit and Stiglitz find that only the elasticity of utility matters for quantity misallocation and the elasticity of demand is not relevant for determining efficiency of production levels. We state their result below and discuss how productivity differences affect distortions and efficiency analysis.

Proposition 7. *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). □

In terms of determining misallocations, the symmetric firm case simplifies the analysis as we need only compare two decisions, q and M_e. In contrast, determining misallocations across heterogeneous firms is less obvious because quantities vary by firm productivity. Further, the bias in quantities and productivity can have opposing implications for the bias in firm entry. For instance, when firms produce too little quantity, there is downward pressure on wages and high cost firms are able to survive in the market. A higher cost cutoff in turn bids up wages, so firm quantities and the cost cutoff have opposite effects on the ex ante profitability of 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, when $\mu' < 0$ and $(1-\varepsilon)' > 0$, excess production by medium-sized firms imposes an externality on large and

[19] 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 $\ln \varepsilon(q)/q = \int_0^q -(\mu(t)/t)\,dt - \ln\left[\int_0^q \exp\left\{\int_0^s -(\mu(t)/t)\,dt\right\} ds\right]$.

small firms. Large firms produce below their optimal scale and small firms are deterred from entering. In this case, the market diverts resources away from small and large firms towards medium-sized firms. Second, accounting for firm heterogeneity shows both the elasticity of utility and the inverse demand elasticity determine resource misallocations. Under symmetric firms, only the elasticity of utility determines misallocations and the inverse demand elasticity does not matter. Specifically, Proposition 7 does not depend on $\mu'(q)$. The presence of firm heterogeneity fundamentally changes the qualitative analysis. When markups vary, firms with different productivity levels charge different markups. This affects their quantity decisions as well as their incentives to enter. Therefore, firm heterogeneity and variable markups alter the standard policy rules for correcting misallocation of resources by the market.[20]

5 Efficiency and Market Size

Increases in market size encourage competition, so we might expect that integrated markets 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 monopolistic 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 are able to show that welfare gains are related to the demand-side elasticities discussed earlier. To understand the potential of market expansion in eliminating distortions, we examine efficiency in large markets. Large integrated markets can eliminate distortions, while preserving firm heterogeneity.

[20]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).

5.1 Integration, Market Size and Efficiency

We begin with the equivalence between market expansion and trade. Proposition 8 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 8.

Proposition 8. *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 8 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.[21] 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. We summarize this result in Proposition 9.

Proposition 9. *Market expansion increases welfare when preferences are aligned. (Provided $(\mu q)'' \leq 0$ whenever $\mu' < 0$.)*

The economic reasoning for Proposition 9 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 private and social markups in the same direction. The market therefore incentivizes firms towards the right allocation and provides higher welfare.

[21] 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 ($C(w,q) \cdot q^{\text{aut}} \leq C(w, q^{\text{aut}}) \cdot q^{\text{aut}}$).

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.[22]

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

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

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 the CES 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 12 in the Appendix. In words, Proposition 12 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 12 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 10. 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.[23]

Proposition 10. *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.

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

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 10 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 10 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.[24] 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 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.

5.2.1 CES Efficiency with Trade Frictions

We have examined how opening to trade with small and large markets affects distortions. Conceptualizing integration as access to new markets enables us to provide a theoretical benchmark. A more realistic scenario however is one with partial trade liberalization where international trade entails additional costs. In this sub-section, we introduce trade frictions as in Melitz and

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

show that the CES economy continues to be efficient. We then argue that trade frictions introduce distributional issues, which we do not address in this paper.

Let $\tau \geq 1$ denote the iceberg trade cost and $f_x \geq 0$ denote the fixed cost of exporting goods abroad. When $\tau = 1$ and $f_x = 0$, the economy faces no trade frictions in integrating with world markets. Proposition 1 shows that the autarkic and integrated market allocations are efficient under CES demand. This implies that a world planner would never levy trade taxes even when it could collect tax revenues by choosing $\tau > 1$ or $f_x > 0$. The CES efficiency result is therefore robust to endogenously chosen trade frictions. As Proposition 11 below shows, CES demand ensures the market picks the right allocations even in the presence of exogenous trade frictions.[25]

Proposition 11. *Every market equilibrium of identical open Melitz economies with trade frictions is socially optimal.*

Proof. Online Appendix. □

Proposition 11 is striking in that the differences in firm costs do not generate inefficiencies despite heterogeneity of profits and the different effects that trade frictions will have on firm behavior. Furthermore, selection of firms performs the function of allocating additional resources optimally without any informational requirements. Under CES demand, laissez faire industrial policy is optimal for the world economy.[26]

The CES efficiency results of Propositions 1 and 11 imply that the higher productivity cutoff of an open Melitz economy is not optimal in autarky. This seems counter-intuitive, as Melitz shows that trade provides productivity and welfare gains by reallocating resources towards low cost firms. Why then is the lower cost cutoff of the open economy inefficient in autarky? Proposition 11 shows trade frictions make a new mix of productivity and variety efficient. The market minimizes losses from trade frictions by weeding out high cost firms. Conditional on trade costs, market selection of firms is optimal. In autarky, choosing a productivity cutoff that corresponds to a higher level of frictions would provide productivity gains at the expense of too little variety, and would decrease welfare.[27]

[25] Technically, we need to be careful in specifying the policymaker's objective function in the presence of multiple countries. Formal details are in the Online Appendix and we note here that the policymaker maximizes per capita world welfare.

[26] However, terms of trade externalities may exist and lead to a breakdown of laissez faire policies (Demidova and Rodriguez-Clare 2009). Moreover, Chor (2009) considers policy intervention in the presence of multinationals and a homogeneous goods sector.

[27] Another implication of market efficiency is that exogenous "shocks" (such as changes in trade frictions) affect world welfare only through their direct effect on welfare. As market allocations maximize world welfare, the indirect effects can be ignored when studying the impact of exogenous shocks on welfare under CES demand (for example, Atkeson and Burstein 2010).

Modeling trade between equally sized countries makes the role of trade frictions clear cut. When countries differ in size, trade frictions introduce cross-country distributional issues which obscure the pure efficiency question. Specifically, consider two countries of different sizes with CES demand. Market allocations are efficient when these countries trade with each other and face no trade frictions. These market allocations maximize social welfare with equal Pareto weights assigned to every individual in the two countries. Introducing trade frictions will continue to induce efficient market allocations, but with unequal Pareto weights. Let ω^{mx} denote the Pareto weight on welfare of country m from consuming goods of country x. Following Proposition 8, ω^{mx} can be defined to ensure the market allocation is an interior solution to:

$$\max_{q,c_d,M_e} \sum_x \sum_m \omega^{mx} M_e^x \int_0^{c_d^{mx}} u'(q^{mx}(c)) \cdot q^{mx}(c) L^m dG \qquad \text{where}$$
$$L^x \geq M_e^x \left\{ \sum_m \int_0^{c_d^{xm}} [\tau^{xm} c q^{xm}(c) L^m + f^{xm}] dG + f_e \right\} \quad \text{for each } x.$$

This shows the market is implicitly favoring certain consumers, so that resource allocation reflects distributional outcomes in addition to cost competitiveness. As our focus is on efficiency, we model the stylized case of frictionless trade and consider more general demand structures which can explain a greater range of market outcomes. The cross-country distribution of welfare gains is important but beyond the focus of this study. We leave this avenue to future research and conclude in the next Section.

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 and trade frictions. 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 nature of market distortions depends 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. 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.

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, 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.

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

Chor, D., "Subsidies for FDI: Implications from a model with heterogeneous firms," *Journal of International Economics*, 2009, *78* (1), 113–125.

Cunningham, T., "Comparisons and Choice," *Working Paper*, 2011.

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

Demidova, S. and A. Rodriguez-Clare, "Trade policy under firm-level heterogeneity in a small economy," *Journal of International Economics*, 2009, *78* (1), 100–112.

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.

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.

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.

Grossman, G. M. and E. 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.

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

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

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

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

Loecker, J. De, P. K. Goldberg, A. K. Khandelwal, and N. Pavcnik, "Prices, Markups and Trade Reform," *Working Paper*, March 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., "The Impact of Trade on Intra-Industry Reallocations and Aggregate Industry Productivity," *Econometrica*, 2003, *71* (6), 1695–1725.

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

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

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.

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.

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, E., S. Kokovin, M. Parenti, and J. F. Thisse, "Monopolistic competition in general equilibrium: Beyond the CES," *Econometrica*, forthcoming.

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 con-

cave $\forall\lambda$.[28] 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 $q^*(c) = (\lambda c/\rho)^{1/(\rho-1)}$.[29] 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 we have

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

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

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

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.

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

[29] 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.

Solving this expression and taking logs shows that

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

Defining $B(c_d) \equiv \ln R(c_d) - \ln \left(f_e + f G(c_d) \right)$ 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

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

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 $x \leq a$ and $B'(x) < 0$ for $x \geq b$. 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 $LM_e \int_0^{c_d} u'\left(q(c)\right) q(c) dG$, where δ denotes the Lagrange multiplier:

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

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 4. The result relies on the following relationship we first prove:

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

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 there-

fore

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

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 5. 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 $d\beta/d\alpha = M_e \int_0^{c_d} w(q(c))\, dG = M_e \int_0^{c_d} u(q(c))\left[\varepsilon(q) - 1\right] dG < 0$. The conditions characterizing 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

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

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

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

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):

$$w\left(q(c_d)\right)\int_0^{c_d} udG + \alpha w\left(q(c_d)\right)\int_0^{c_d} wdG \gtrless u\left(q(c_d)\right)\int_0^{c_d} wdG + \alpha w\left(q(c_d)\right)\int_0^{c_d} wdG.$$

Cancellation and expansion again show this is equivalent to

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

Finally, this expression can be rewritten $\varepsilon\left(q(c_d)\right) \gtrless \int_0^{c_d} \varepsilon\left(q(c)\right)u\left(q(c)\right)dG / \int_0^{c_d} u\left(q(c)\right)dG$ 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 $\varepsilon' > 0$.

Proof of Proposition 6. For any preferences v, defining $\varepsilon_v(q) \equiv v'(q)q/v(q)$ and $\mu_v(q) \equiv -v''(q)q/v'(q)$ it holds that at any social optimum that

$$1/M_e = \int_0^{c_d} cq(c)/\varepsilon_v\left(q(c)\right)dG(c)$$

Defining $B_v(c) \equiv cq(c)/\varepsilon_v\left(q(c)\right)$ which is the integrand of the equation above, we have

$$B_v'(c) = q(c)/\varepsilon_v\left(q(c)\right) + c\left(dq(c)/dc\right)\left[1 - \varepsilon_v'\left(q(c)\right)q(c)/\varepsilon_v\left(q(c)\right)\right]/\varepsilon_v\left(q(c)\right). \tag{13}$$

Equation (13) can be considerably simplified using two relationships. The first is

$$1 - \varepsilon_v'\left(q(c)\right)q(c)/\varepsilon_v\left(q(c)\right) = \varepsilon_v\left(q(c)\right) + \mu_v\left(q(c)\right).$$

The second is that manipulating the necessary conditions shows that $dq(c)/dc = -\left(q(c)/c\right)\cdot\left(1/\mu_v\left(q(c)\right)\right)$. Substituting these relationships into Equation (13) yields

$$B_v'(c) = q(c)/\varepsilon_v\left(q(c)\right)\cdot\left[1 - \left[\varepsilon_v\left(q(c)\right) + \mu_v\left(q(c)\right)\right]/\mu_v\left(q(c)\right)\right] = -q(c)/\mu_v\left(q(c)\right).$$

The policymaker's problem corresponds to $v(q) = u(q)$ while the market allocation is generated by maximizing $v(q) = u'(q)q$ so that (suppressing the c argument to q)

$$1/M_e^{\text{opt}} - 1/M_e^{\text{mkt}} = \int_0^{c_d^{\text{opt}}} cq^{\text{opt}}/\varepsilon\left(q^{\text{opt}}\right)dG(c) - \int_0^{c_d^{\text{mkt}}} cq^{\text{mkt}}/\left[1 - \mu\left(q^{\text{mkt}}\right)\right]dG \tag{14}$$

and similarly (suppressing the c arguments):

$$B_u = cq^{\text{opt}}/\varepsilon\left(q^{\text{opt}}\right), \qquad B'_u = -q^{\text{opt}}/\mu\left(q^{\text{opt}}\right),$$
$$B_{u'q} = cq^{\text{mkt}}/\left[1-\mu\left(q^{\text{mkt}}\right)\right], \; B'_{u'q} = -q^{\text{mkt}}/\left[\mu\left(q^{\text{mkt}}\right)+\mu'\left(q^{\text{mkt}}\right)q^{\text{mkt}}/\left(1-\mu\left(q^{\text{mkt}}\right)\right)\right].$$

Now assume $\varepsilon' < 0 < \mu'$, so by above $c_d^{\text{mkt}} > c_d^{\text{opt}}$ and for the result, from Equation (14) it is sufficient to show that $\int_0^{c_d^{\text{opt}}} B_u(c) - B_{u'q}(c)dG(c) \leq 0$. From above, there is also a c^* such that $q^{\text{mkt}}(c) > q^{\text{opt}}(c)$ for $c < c^*$ and $q^{\text{mkt}}(c) < q^{\text{opt}}(c)$ for $c > c^*$. For $c < c^*$, $B_u(c) - B_{u'q}(c) < 0$ as $q^{\text{mkt}}(c) > q^{\text{opt}}(c)$ and $\varepsilon' < 0$ implies

$$cq^{\text{mkt}}/\left[1-\mu\left(q^{\text{mkt}}\right)\right] > cq^{\text{opt}}/\left[1-\mu\left(q^{\text{opt}}\right)\right] > cq^{\text{opt}}/\varepsilon\left(q^{\text{opt}}\right).$$

For $c \geq c^*$, $B_u(c) \leq B_{u'q}(c)$ as from continuity $B_u(c^*) \leq B_{u'q}(c^*)$, while $\mu' > 0$ implies

$$\begin{aligned}\left(B_u(c) - B_{u'q}(c)\right)' &= -q^{\text{opt}}/\mu\left(q^{\text{opt}}\right) + q^{\text{mkt}}/\left[\mu\left(q^{\text{mkt}}\right)+\mu'\left(q^{\text{mkt}}\right)q^{\text{mkt}}/\left(1-\mu\left(q^{\text{mkt}}\right)\right)\right] \\ &< -q^{\text{opt}}/\mu\left(q^{\text{opt}}\right) + q^{\text{mkt}}/\mu\left(q^{\text{mkt}}\right).\end{aligned}$$

Finally, $\mu'(q)q \leq \mu$ implies $q/\mu(q)$ increases in q. With $q^{\text{mkt}}(c) < q^{\text{opt}}(c)$ for $c > c^*$, this implies $\left(B_u(c) - B_{u'q}(c)\right)' \leq 0$ so $B_u(c) \leq B_{u'q}(c)$ for $c > c^*$. With above, $\int_0^{c_d^{\text{opt}}} B_u(c) - B_{u'q}(c)dG(c) \leq 0$ giving the result. For the case $\varepsilon' > 0 > \mu'$, the argument goes through since $\mu'(q)\,q/\mu(q) \leq 1$.

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 $\bar{\varepsilon} \equiv \int \varepsilon u dG / \int u dG$. Average elasticity of utility changes due to a different cost cutoff and quantity allocations across firms. An increase in market size raises the marginal utility of income at the rate of average markups $d\ln\delta/d\ln L = \int \mu pq dG / \int pq dG \equiv \tilde{\mu}$. From $d\ln\delta/d\ln L$ and $d\ln\bar{\varepsilon}/d\ln L$, the change in welfare is

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

When preferences are aligned, the first 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 second term in square brackets is also positive when preferences are aligned, given regularity conditions in Proposition 9.

Proof of Proposition 9. Following the discussion above, it is sufficient to show that for $\gamma(c) \equiv (\mu + \mu' q/(1-\mu))^{-1} \cdot (\varepsilon u/\bar{\varepsilon} \int u dG)$,

$$1 + \int \frac{1-\mu-\bar{\varepsilon}}{1-\mu+\mu' q/\mu} \frac{1-\mu}{\mu} \frac{\varepsilon u}{\bar{\varepsilon} \int u dG} dG = \int \left[1 - \bar{\varepsilon} + \mu' q/(1-\mu)\right] \gamma dG \geq 0. \tag{15}$$

This clearly holds for $\mu' \geq 0$, and for the other case where preferences are aligned, we have $\mu' < 0 < \varepsilon'$. Expanding Equation (15) shows that

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

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

$$\int \mu \frac{u}{\int u dG} dG \geq \int \mu \eta \frac{u}{\int u dG} dG \tag{16}$$

where $\eta(c) \equiv \gamma(c) \cdot (\int u dG/u) / \int \gamma$. Since $\int \eta \cdot (u/\int u dG) dG = 1$ and $d\mu/dc > 0$, it follows that if $d\eta/dc < 0$, then Equation (16) holds by stochastic dominance. As $d\eta/dc < 0$ iff $d\eta/dq > 0$, we examine the sign of $d\eta/dq$ below.

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

The additional hypothesis that $(\mu q)'' \leq 0$ guarantees that each term above is positive, so $d\eta/dq > 0$ and we conclude Equation (16) 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 12. *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:

$$u'(q(c_d))/\delta - f/\left[Lq(c_d)\cdot\mu\circ q(c_d)/(1-\mu\circ q(c_d))\right] = c_d, \tag{17}$$

$$Lc_d q(c_d)\cdot\mu\circ q(c_d)/(1-\mu\circ q(c_d)) = f. \tag{18}$$

First, if $\lim_{L\to\infty} Lq(c_d) = 0$, Equation (18) 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 (18) that $Lc_d q(c_d)$ is bounded, so from Equation (17), $Lq(c_d)\cdot u'(q(c_d))/\delta$ 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 (18), $c_d\longrightarrow 0$. Now assume $c_d\longrightarrow 0$ so from (18), $Lq(c_d)\cdot\mu\circ q(c_d)/(1-\mu\circ q(c_d))\longrightarrow\infty$ which implies with Equation (17) that $u'(q(c_d))/\delta\longrightarrow 0$. Finally, if $u'(q(c_d))/\delta\longrightarrow 0$, (17) 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

$$\lim_{q\longrightarrow 0} f_q'(z)=d\lim_{q\longrightarrow 0} f_q(z)/dz=-\mu^\infty\Upsilon^{-1}(z)/z=d\Upsilon^{-1}(z)/dz. \tag{19}$$

We conclude that $\Upsilon^{-1}(z)$ is differentiable and thus continuous. Given the form deduced in (19), $\Upsilon^{-1}(z)$ is continuously differentiable. Since $d\Upsilon^{-1}(z)/dz=1/\Upsilon'\circ\Upsilon^{-1}(z)$, composing both

sides with $\Upsilon(z)$ and using (19) 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 $c > 0$:

$$q^{\text{opt}}(c)/q^{\text{opt}}\left(c_d^{\text{opt}}\right) \longrightarrow \left(c_\infty^{\text{opt}}/c\right)^{1/\alpha}, \quad q^{\text{mkt}}(c)/q^{\text{mkt}}\left(c_d^{\text{mkt}}\right) \longrightarrow \left(c_\infty^{\text{mkt}}/c\right)^{1/\alpha}. \tag{20}$$

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,

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

It necessarily follows that

$$\begin{aligned}
&\lim_{L\longrightarrow\infty} L\int_0^{c_d^{\text{mkt}}} \mu \circ q^{\text{mkt}}(c)/\left(1-\mu \circ q^{\text{mkt}}(c)\right) \cdot c q^{\text{mkt}}(c)dG - fG(c_d^{\text{mkt}}) = \\
&\lim_{L\longrightarrow\infty} L\int_0^{c_d^{\text{opt}}} \left(1-\varepsilon \circ q^{\text{opt}}(c)\right)/\varepsilon \circ q^{\text{opt}}(c) \cdot c q^{\text{opt}}(c)dG - fG(c_d^{\text{opt}}).
\end{aligned} \tag{21}$$

Using Equation (20), 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 (21) the limits of the $\mu/(1-\mu)$ and $(1-\varepsilon)/\varepsilon$ terms are equal and factor out of Equation (21), leaving

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

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

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

so that finally evaluating the limits, we have

$$\int_0^{c_\infty^{\text{mkt}}} \left[(c/c_\infty^{\text{mkt}})^{1-1/\alpha} - 1 \right] dG = \int_0^{c_\infty^{\text{opt}}} \left[(c/c_\infty^{\text{opt}})^{1-1/\alpha} - 1 \right] dG. \tag{22}$$

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 (22) holds. Checking the conditions for L/M_e show they coincide between the market and social optimum as well. □