Monopolistic Competition and Optimum Product Diversity Under Firm Heterogeneity

Swati Dhingra
Centre for Economic Performance, LSE

John Morrow
Centre for Economic Performance, LSE

This Draft: July 31, 2012

Abstract

A fundamental question in theories of imperfect competition is whether the market allocates resources efficiently. We generalize the Spence-Dixit-Stiglitz framework to heterogeneous firms, addressing when the market provides optimal quantities, variety and productivity. This approach yields several insights about allocational efficiency under firm heterogeneity. First, constant elasticity of substitution (CES) demand ensures market allocations are efficient, despite differences in firm productivity. Second, when demand elasticities vary, market allocations are not efficient and reflect the distortions of imperfect competition. These distortions are not uniform within a market: some firms over produce while others under produce, and the pattern is determined by two demand side elasticities. Third, market imperfections derive from insufficient competition. Integration with large markets can achieve allocational efficiency in the absence of domestic policy.

JEL Codes: F1, L1, D6.
Keywords: Selection, Monopolistic competition, Efficiency, Productivity, Social welfare, Demand elasticity.

Acknowledgments. We thank Bob Staiger for continued guidance and Steve Redding and Katheryn Russ for discussing the paper. 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, Steve Redding, Andres Rodriguez-Clare, Thomas Sampson, Daniel Sturm, Jacques Thisse, John Van Reenen and Mian Zhu for insightful comments. This paper has benefited from helpful comments of participants at AEA 2011, DIME-ISGEP 2010, ISI Delhi, FIW, LSE, Louvain-Core, Oxford, the Philadelphia Fed, Princeton and Wisconsin-Madison. Preliminary draft circulated as "When is Selection on Firm Productivity a Gain from Trade?" was a dissertation chapter at Wisconsin-Madison in 2010. Swati thanks the IES (Princeton) for their hospitality. Contact: s.dhingra@lse.ac.uk and j.morrow1@lse.ac.uk.

1 Introduction

Empirical work has drawn attention to the high degree of heterogeneity in firm productivity.[1] The introduction of firm heterogeneity in monopolistic competition models has provided new insights into the allocation of resources across different firms. A fundamental question in this setting is allocational efficiency. Symmetric firm models explain when market allocations are efficient by examining the trade off between quantity and product variety. When firms are heterogeneous in productivity, we must also ask which types of firms should produce and which should be shut down. 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 how firm heterogeneity affects the efficiency of resource allocation across firms. 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?

We answer these questions in the standard setting of a monopolistically competitive industry with heterogeneous productivity draws and free entry (e.g. Melitz 2003). To allow rich interrelationships between productivity and markups, we consider the general class of variable elasticity demand systems, introduced by Dixit and Stiglitz (1977). The Dixit-Stiglitz model of monopolistic competition and the Melitz approach to firm heterogeneity are standard tools for addressing firm behavior in general equilibrium. This setting therefore provides a theoretical benchmark to understand distortions in resource allocations across firms. It also accounts for the stylized facts that firms are rarely equally productive and markups are unlikely to be constant.[2]

When demand elasticity varies with quantity and firms vary in productivity, markups vary within a market. These considerations impact optimal policy rules in a fundamental way, distinct from markets with symmetric costs or constant markups. There are two new sources of potential inefficiency: selection of the right distribution of firms and allocation of the right quantities across firms with different costs. 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). The relative position of a firm in the cost distribution matters, and one contribution of the paper is to show how the interplay of differences in productivity and variable markups affects welfare and policy analysis.

As inefficiencies arise due to imperfect competition across firms, we might expect increased competition to improve efficiency. International integration expands market size and provides op-

[1]For surveys, see Bartelsman and Doms (2000); Tybout (2003); Bernard, Jensen, Redding and Schott (2007).

[2]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 evidence, see Melitz and Trefler (2012).

portunities to correct the distortions of market power. This idea of introducing foreign competition to improve efficiency goes back at least to Melvin and Warne (1973). As is well understood in this literature, increased competition (from trade or growth) does not guarantee welfare gains, and may exacerbate distortions (Helpman and Krugman 1985). This insight becomes more relevant in a heterogeneous cost environment because of new sources of potential inefficiency. As a benchmark, we examine whether integration with large world markets provides a policy option to correct distortions in the absence of domestic policies.[3]

We begin our analysis of market distortions by considering constant elasticity of substitution (CES) demand. We show that when firms vary in productivity, market allocations are efficient. 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. With productivity differences, 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. As markups do not vary across firms, the monopolistic production levels are not skewed across firms. 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 reflect the optimal shadow value of labor. Therefore, the market implements the first-best allocation and laissez faire industrial policy is optimal.[4]

What induces market efficiency and how broadly does this result hold? We generalize the demand structure to the variable elasticity of substitution (VES) form of Dixit and Stiglitz which permits variable markups and provides a rich setting for a wide range of market outcomes (Vives 2001; Zhelobodko, Kokovin, Parenti and Thisse forthcoming). Within this setting, 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.

The pattern of distortions is determined by two demand side elasticities: the inverse demand elasticity, which measures market incentives, and the elasticity of utility ($d\ln u(q)/d\ln q$), which measures the contribution of a firm's production to welfare. Misalignment of these elasticities determines the bias in market allocations: some firms over-produce while others under-produce within the same market. For instance, the market may favor excess entry of low productivity firms,

[3]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.

[4]Melitz (2003) considers both variable and fixed costs of exporting. We show that the open Melitz economy is efficient, even in the presence of 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.

thereby imposing an externality on high productivity firms who end up producing too little. Additionally, the markup distribution affects ex ante profitability, and therefore the trade-off between aggregate quantity and variety depends on both the elasticity of utility and the inverse demand elasticity.[5]

Although our results characterize the bias in resource allocation, this leaves open the question of feasible policy options. Integration with international markets introduces foreign competition and can potentially mitigate distortions. To capture the role of integration as a policy tool, we examine the effects of integration with large markets. Such integration will push outcomes towards a new concept, the "monopolistically competitive limit", in which the economy continues to exhibit heterogeneous firms who possess market power and differ in size. This shows that productivity dispersion can persist in large markets, in contrast to the perfectly competitive limit of Hart (1985). As in the perfectly competitive limit, the monopolistically competitive limit is efficient and integration with large global markets is therefore a first-best policy to eliminate the distortions of imperfect competition. However, as the monopolistically competitive limit may require a market size which is unattainable even in fully integrated world markets, integration may be an incomplete tool to reduce distortions. When markups vary, integration with small markets cannot generally replace domestic industrial policy.

The paper is organized as follows. Section 2 relates this paper to previous work and Section 3 recaps the standard monopolistic competition framework with firm heterogeneity. Section 4 contrasts the efficiency of CES demand with inefficiency of VES demand and Section 5 characterizes the bias in resource allocation in a VES economy. Section 6 examines how integration can eliminate distortions, deriving a limit result for large markets. Section 7 concludes.

2 Related Work

Our paper is related to work on welfare gains in industrial organization and international economics. The trade-off between quantity and variety occupies a prominent place in the industrial organization literature (e.g., Mankiw and Whinston 1986). We contribute to this literature by studying the effects of firm heterogeneity. The analysis is motivated by efficiency properties which have been studied at length in symmetric firm models of monopolistic competition.[6] To the best of our knowledge, this is the first paper to show market outcomes with heterogeneous firms are first best.[7] Efficiency of market allocations implies that exogenous "shocks" (such as changes in

[5] These findings are in sharp contrast to symmetric firm models, where the elasticity of utility completely determines the bias in market allocations and the inverse demand elasticity does not matter for misallocations, as emphasized by Dixit and Stiglitz (1977) and Vives (2001).

[6] For example, Spence (1976); Dixit and Stiglitz (1977); Venables (1985); Epifani and Gancia (2011); Behrens and Murata (2012).

[7] We consider this to be the proof of a folk theorem. The idea of efficiency in Melitz 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. Within the heterogeneous firm literature, Baldwin and Robert-Nicoud (2008)

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).

To highlight the potential scope of market imperfections, we generalize the well known CES demand framework to VES demand. In contemporaneous work, Zhelobodko, Kokovin, Parenti and Thisse (forthcoming) develop complementary results for market outcomes under VES demand and demonstrate its richness and tractability under various assumptions such as multiple sectors and vertical differentiation.[8] The focus on variable markups is similar to de Blas and Russ (2010) who build on the baseline framework of Bernard, Eaton, Jensen and Kortum (2003) to understand pricing behavior of heterogeneous firms. Unlike Zhelobodko et al. and de Blas and Russ, we are interested in the role of markups in determining market distortions.

We show that the market maximizes aggregate real revenue in a VES economy. Helpman and Krugman (1985) provide a GDP function for symmetric firms while Feenstra and Kee (2008) derive one for the Melitz model, holding aggregate demand conditions fixed. In contrast, we consider heterogeneous firms in general equilibrium. We also study the limiting behavior of a VES economy. A large literature examines whether monopolistic competition arises as a limit to oligopolistic pricing and when monopolistic competition converges to perfect competition in symmetric firm models (Vives 2001, Chapter 6). We examine when market expansion leads to efficiency. The monopolistically competitive limit is first-best despite positive markups and firm heterogeneity.

The findings of our paper are related to an emerging literature on welfare gains in new trade models. Generalizing Krugman (1980) to heterogeneous firms, Melitz shows that opening to trade raises welfare through reallocation of resources towards high productivity firms. In recent influential work, Arkolakis, Costinot and Rodriguez-Clare (2012a) and Arkolakis, Costinot, Donaldson and Rodriguez-Clare (2012b) show that introducing firm heterogeneity or variable markups does not change the mapping between trade data and welfare gains from trade. We focus instead on efficiency of resource allocations and show that firm heterogeneity and variable markups matter for allocational efficiency.[9] Our work is in line with Tybout (2003) and Katayama, Lu and Tybout (2009) who point to the limitations of the empirical literature in mapping observed productivity

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 that 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 in the presence of a homogeneous good. Their approach differs because the homogeneous good fixes the marginal utility of income.

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

[9]For instance, linear VES demand and Pareto cost draws fit the gravity framework, but firm heterogeneity and variability of markups still matter for market efficiency in this setting. Further, our VES demand framework is not nested within the assumptions of Arkolakis et al. (2012a,b), as illustrated in the Appendix.

gains to welfare and optimal policies.

3 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. Third, 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.

3.1 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 conditions:

1. $u(0)$ is normalized to zero.
2. u is twice continuously differentiable, increasing and concave.
3. $(u'(q) \cdot q)'$ is strictly decreasing in quantity.
4. The elasticity of marginal utility $\mu(q) \equiv |qu''(q)/u'(q)|$ is less than one.

For each variety indexed by 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

[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.

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

3.2 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 faces a cost function $\mathrm{TC}(q(c)) \equiv cq(c) + f$ where f denotes the fixed cost of production. 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 given by

$$p + q \cdot u''(q)/\delta = c. \quad \text{(MR=MC)}$$

$MR = MC$ ensures that the markup rate is $(p(c) - c)/p(c) = -qu''(q)/u'(q) = \mu(q(c))$. Therefore, the elasticity of marginal utility summarizes the inverse demand elasticity as $\mu(q) \equiv |qu''(q)/u'(q)| = |d \ln p(q)/d \ln q|$.

3.3 Market equilibrium

Profit maximization implies that 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 (ZPC), $\pi(c_d) = 0$. 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. 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,

$$\int \pi(c) dG = f_e. \quad \text{(FE)}$$

The next two Sections examine the efficiency properties of this framework.

[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.

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

4 Efficiency in a VES Economy

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

4.1 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.*

Proof. See Appendix. □

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

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

finance the fixed cost of production. The market therefore induces firms to indirectly 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 ignores its effect on resource costs, but this is exactly offset by the variety gain and productivity loss from its entry. The market thereby ensures resource costs 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

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

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:

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

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 in greater detail.[14]

4.2 Welfare beyond isoelastic demand

Efficiency of the market equilibrium in a Dixit-Stiglitz-Melitz economy 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. See 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 ($M_e \int u'(q(c)) \cdot q(c) \cdot LdG$) generated in the economy.

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

Proof. See Appendix. □

Proposition 3 shows that market resource allocation is generally not aligned with the social optimum under VES demand. The market and efficient allocations are solutions to:

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

[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 optimality of market allocations. To see this, extend the CES demand of Melitz to CES-Benassy preferences $U(M_e, c_d, q) \equiv \nu(M_e) \int_0^{c_d} q(c)^\rho g(c) dc$. In this example, 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 ($\nu_B = (1-\rho)/\rho$).

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 bias in resource allocations.

5 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 characterize distortions. 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.

5.1 Markup and Quantity Patterns

The pattern of markups across firms in a VES economy is determined by μ' and $(1-\varepsilon)'$. When $\mu'(q) > 0$, 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). This means high 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 richer VES demand brings out the distinction between $\mu' > 0$ and $\mu' < 0$, which is crucial in understanding distortions across firms.

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.

We show the relationship between markups and quantity characterize market distortions in a VES economy. 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

[16]The expo-power utility form 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.

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

5.2 Quantity, Productivity and Entry Distortions

We now characterize the bias in market allocations compared to the optimal allocation by demand characteristics. The biases in quantity, productivity and entry are discussed in turn.

5.2.1 Quantity Biases

Quantity distortions across firms depend on whether private and social markups move together as quantities change. We will say that private and social incentives are *partially aligned* when μ' and $(1-\varepsilon)'$ have the same sign. Conversely, incentives are *misaligned* when μ' and $(1-\varepsilon)'$ have different signs. We show that when private and social markups are misaligned, market quantities $q^{\text{mkt}}(c)$ are uniformly too high or low relative to optimal quantities $q^{\text{opt}}(c)$. In contrast, when private and social markups are partially aligned, whether quantities are over produced or under produced depends on each firm's cost.

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

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

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 biased 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 biased 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 partially 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 $(1-\varepsilon)'$ and μ' have different signs, $q^{\text{mkt}}(c)$ 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 $(1-\varepsilon)'$ and μ' have the same sign 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 $c > c^*$.*

Proof. See Appendix. □

Proposition 4 characterizes the bias in quantities and highlights the importance of variable elasticities and heterogeneity in determining which firms receive a higher or lower than optimal share of resources in the market.

5.2.2 Productivity Cutoff Biases

The bias in firm selection is determined by the relation between social markups and quantity. Proposition 5 shows that productivity in the market 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:*

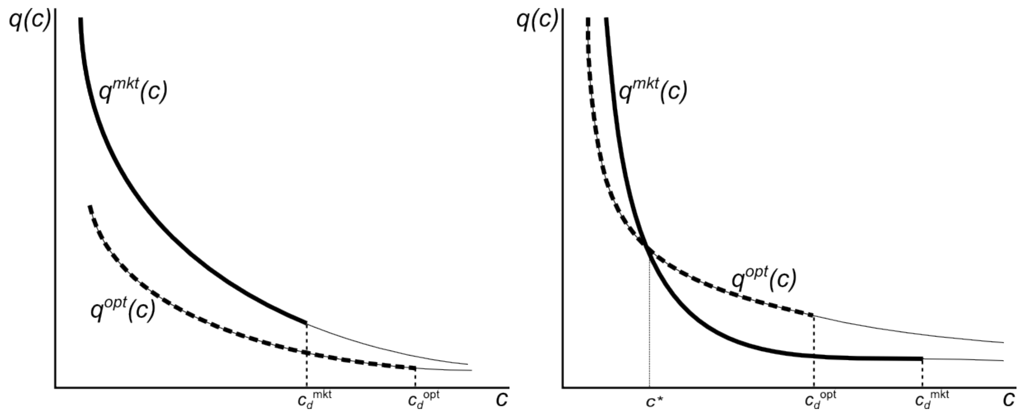
1. *If $(1-\varepsilon)' > 0$, market productivity is too low: $c_d^{\text{mkt}} > c_d^{\text{opt}}$.*

2. *If* $(1-\varepsilon)' < 0$*, market productivity is too high:* $c_d^{\text{mkt}} < c_d^{\text{opt}}$.

Proof. See Appendix. □

Propositions 4 and 5 explains 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 Resource Allocation for Firm Production in the Market



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

5.2.3 Entry Biases

Although a comparison of market entry to optimal entry is generally hard to make, Proposition 6 establishes their relative levels for the case when private and social markups are partially 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* $\mu'(q)q/\mu \leq 1$*).*

Proof. See Appendix. □

5.2.4 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 bias in market allocations 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 distortions 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 2009). However, existing studies suggest that the relationship differs across markets, and therefore we cannot a priori 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 pharmaceutical products in the US. Social markups 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]

[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 (see Zhelobodko et al. forthcoming).

[19] Distinguishing increasing and decreasing social markups is more challenging as they are unlikely to be directly observable. Furthermore, the welfare implications of a change in trade costs no longer take the simple form provided for CES demand in Arkolakis et al. (2012a). This is shown graphically in the Appendix. 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 $\mu'(q)$ (see Appendix). When $\gamma = 1$, this form generates an adjustable pass-through demand system (Bulow and Pfleiderer 1983; Weyl and Fabinger 2009). If sufficient data is available, another approach is to recover $\varepsilon(q)$ from price and quantity data using $\varepsilon(q) = p(q)q / \int p(q)dq$ or from markup and quantity data using $\ln \varepsilon(q)/q = \int_0^q -(\mu(t)/t)\,dt - \ln\left[\int_0^q \exp\{\int_0^s -(\mu(t)/t)\,dt\}\,ds\right]$.

5.3 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 bias and the elasticity of demand is not relevant for determining efficiency of production levels. We state their result below for comparison with heterogeneous firms.

Proposition 7. *Under symmetric firms, the bias in market allocations 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. See Dixit and Stiglitz (1977). □

In terms of determining the bias, the symmetric firm case simplifies the analysis as we need only compare two decisions, q and M_e. In contrast, determining the bias for heterogeneous firms is less obvious because quantities vary by firm productivity. Further, the biases in quantities and the productivity cutoff 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 the bias in resource allocations across the entire distribution of firms reveals two substantive results. First, as we might expect, the bias in resources allocations across firms differs by productivity. An interesting finding is that this heterogeneity in bias 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 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. Dixit and Stiglitz find that only the elasticity of utility determines the bias in market allocations and the inverse demand elasticity is irrelevant for this purpose. Specifically, their result (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 the bias in resource allocations induced by the market.[20]

[20]Table 2 characterizes the qualitative role of demand elasticities in determining misallocations across firms. Using a quantitative measure of distortions reiterates this finding. The loss from biased market allocations 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 $\mathrm{Cov}(1-\varepsilon,\mu)$. The covariance component shows that the distribution of markups matters for quantifying distortions, except when firms are symmetric or markups are constant.

6 Efficiency and Market Size

As the bias in market allocations varies by firm productivity, a policymaker would potentially need firm-level information to implement policies for improving efficiency. One potential policy option that does not require firm-level information is international integration. We start by showing that international integration with heterogeneous firms is equivalent to an expansion in market size. Increases in market size encourage competition, so we might expect that integrated markets would reduce market power and improve efficiency. 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.

To understand when market expansion might eliminate the distortions of imperfect competition, we examine efficiency in large markets. After establishing the equivalence of integration and market expansion, we show large integrated markets can eliminate distortions, while preserving firm heterogeneity.

6.1 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. See online Appendix and Krugman (1979). □

Proposition 8 implies that the biases in market distortions detailed in Section 5 persist in integrated markets. Resource allocation in an integrated VES 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 trade, then there are gains from trade. 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 in the integrated economy. When firms are symmetric in c, trade is beneficial as long as variety does not fall after trade ($M_e \geq M_e^{\text{aut}}$) and average unit cost of the autarky bundle is lower after trade ($C(w,q) \cdot q^{\text{aut}} \leq C(w,q^{\text{aut}}) \cdot q^{\text{aut}}$). In terms of primitives, we find trade is beneficial when preferences are aligned. This is true for any cost distribution, but requires a regularity condition for decreasing private markups (see Appendix). It is therefore reasonable to expect small increases in market size to improve welfare. However, a more ambitious question is: can we ever expect trade to eliminate distortions? As acknowledged by Spence, "perfectly general propositions are hard to come by" and the nature of distortions can be highly dependent on parameter magnitudes.[21] To make progress, we follow Stiglitz (1986) and study market and optimal outcomes as market size becomes arbitrarily large. This allows us to examine when international trade enables markets to eventually mitigate distortions.

6.2 Efficiency in Large Markets

Looking at efficiency in large markets explains whether integrating with world markets can enable 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]

The large economy concept is similar in spirit to the idea of a competitive limit. As the size of the integrated market grows large, the number of entrants grows large. However, when firms are heterogeneous, simply knowing there are a large number of entrants does not explain the distribution of productivity, prices and quantity. At least three salient outcomes can occur. One outcome is that competitive pressures might weed out all firms but the most productive. This occurs for instance when marginal revenue is bounded, as when u is quadratic or CARA (constant absolute risk aversion). 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 with heterogeneous firms, 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.

[21] Even under symmetric firms and a specific utility function, Behrens and Murata (2012) argue it is difficult to show analytically that welfare in the market versus the optimum rises monotonically with L.

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

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 ($Lq(c)$) produced by a firm as follows:

Proposition 9. *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. See Appendix. □

Proposition 9 shows that when markups are interior and the cost cutoff converges, one of three things must happen. 1) Only the lowest cost firms remain ($\lim_{L\to\infty} c_d^{\text{mkt}} = 0$) and prices go to zero (akin to perfect competition), while the lowest cost firms produce infinite total quantities ($\lim_{L\to\infty} Lq\left(c_d^{\text{mkt}}\right) = \infty$). 2) Post-entry, all firms produce independent of cost ($\lim_{L\to\infty} c_d^{\text{mkt}} = \infty$) while prices become unbounded and the total quantities produced become negligible ($\lim_{L\to\infty} Lq\left(c_d^{\text{mkt}}\right) = 0$), akin to a "rentier" case where firms produce little after fixed costs are incurred. 3) The cost cutoff converges to a positive finite level ($\lim_{L\to\infty} c_d^{\text{mkt}} \in (0,\infty)$), 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 9 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). To obtain this result, we introduce one last regularity condition.

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 monopolistically competitive 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 monopolistically competitive 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]

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

Proposition 10. *Under the above assumptions, as market size L approaches infinity the market approaches the monopolistically competitive limit. This limit has the following characteristics:*

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

Proof. See Appendix. □

Proposition 10 shows that integration with large markets can push economies based on VES demand to the monopolistically competitive 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 market distortions until they are 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).[25] While the monopolistically competitive 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 VES 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.

6.2.1 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

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

[25] Stiglitz (1986) notes that the CES model violates the assumptions of the competitive limit of the monopolistically competitive economy derived by Hart (1985) who assumes markups are completely wiped out in the limit.

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 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.[26]

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

Proof. See 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.[27]

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 and provides a net welfare gain from trade. 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.

Modeling trade between equally sized countries makes the role of trade frictions extremely 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

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

[27] However, terms of trade externalities may exist and lead to a breakdown of laissez faire policies. Demidova and Rodriguez-Clare (2009) incorporate terms of trade considerations and provide domestic policies to obtain the first-best allocation in an open Melitz economy with Pareto cost draws. Chor (2009) also considers when policy intervention is appropriate in a heterogeneous firm model with multinationals and a homogeneous goods sector.

with cost distribution $G(c) = (c/c_{\max})^k$ and 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.

7 Conclusion

Is firm size optimal? Are there too few small or large businesses? To understand such questions, this paper examines the efficiency of market allocations when firms vary in productivity and markups. Considering the Spence-Dixit-Stiglitz framework with heterogeneous firms, the efficiency of CES demand is valid even with heterogeneous firms and trade frictions. Firms earn positive profits and charge prices higher than their average costs, yet market allocations are efficient. The market selects the right firms and optimally allocates resources across firms. These findings crucially depend on CES preferences which are necessary for market efficiency.

Generalizing to variable elasticities of substitution, firms charge heterogeneous markups which affect the trade-off between quantity, variety and productivity. Unlike symmetric firm models, the nature of market distortions depends on the elasticity of inverse demand and the elasticity of utility. Under CES demand, these two elasticities are constant and provide strong efficiency properties, but miss out on meaningful trade-offs. When these elasticities vary, introducing firm heterogeneity provides new insights into the biases in market allocations across firms. We characterize the nature of market distortions by demand-side elasticities, which reveals likely targets for policy. While the modeling framework we consider provides a theoretical starting point, 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 have characterized international integration as a key policy tool to realize these potential gains. Integration introduces foreign competition and may provide opportunities to eliminate monopolistic distortions. While integrating with small markets provides potential gains, these need not arise in imperfect markets. As Dixit and Norman (1988) put it, this may seem like a "sad note" on which to end. But we find that integration with large markets holds out the possibility of approaching the monopolistically competitive 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.

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

Bagwell, K. and R. W. Staiger, "Delocation and trade agreements in imperfectly competitive markets," *NBER Working Paper*, 2009.

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.

Berge, Claude and Karreman, *Topological spaces, including a treatment of multi-valued functions, vector spaces and convexity*, New York: Macmillan, 1963.

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., "Relative Thinking and Markups," *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. 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.

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.

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.

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,” *Harvard University, mimeo*, 2009.

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, all 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: $G(x) \equiv \int_0^x g(c)dc$, $R(x) \equiv \int_0^x c^{\rho/(\rho-1)} g(c)dc$.

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

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

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

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

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

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

where $h(c,x) \equiv xcg(c)$ and $v(c,x) \equiv x^\rho g(c)$. One may show that $V(q) - \lambda H(q)$ is strictly concave $\forall\lambda$.[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 nothing but $\rho q^*(c)^{\rho-1}g(c) = \lambda cg(c)$ 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 appropriate choice of λ since for fixed λ we have

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

so choosing λ as $\lambda^* \equiv \rho\left(L/M_e - f_e - fG(c_d)\right)^{\rho-1}/L^{\rho-1}R(c_d)^{\rho-1}$ will make q^* a solution. In

[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.

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. Solving this expression and taking logs shows that

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

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

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

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

A.2 Converse of the Folk Theorem

We now consider general consumer preferences of the form given by Equation (8).

$$U(M_e, c_d, q) \equiv v(M_e, c_d) \int_0^{c_d} u(q(c)) g(c) dc \tag{8}$$

where ν is positive and continuously differentiable, and u satisfies Definition 1.

Proposition. *Consider an economy with preferences as in Equation (8). The market equilibrium is socially optimal only if u is CES.*

Proof. Assume an equilibrium exists which is socially optimal with M_e and c_d fixed by that equilibrium. Also let $q^*(c)$ denote equilibrium quantities. If the equilibrium is efficient for these fixed M_e and c_d, the quantities $q_p(c)$ a policymaker would choose must be optimal. For convenience, define the functional $H(q)$ as in the above proof and let $U^*(q) \equiv U(M_e, c_d, q)$ be as in Equation (8). By Theorems 5.11 and 5.15 of Troutman, a necessary condition for q_p to be optimal is that either $\delta H(q_p;\xi) = 0 \; \forall \xi \in \mathscr{C}^1[0, c_d]$ or $\exists \lambda$ s.t. $\delta U^*(q_p) = \lambda \delta H(q_p;\xi) = 0 \; \forall \xi \in \mathscr{C}^1[0, c_d]$. We will rule out the first and exploit an implication of the second.

Case 1: $\delta H(q_p;\xi) = 0 \; \forall \xi \in \mathscr{C}^1[0, c_d]$. $\forall \xi$ we have that

$$\delta H(q_p;\xi) = \int_0^{c_d} \xi(c) c g(c) dc = 0$$

which implies $cg(c)$ is identically zero on $[0, c_d]$ which is clearly not optimal.

Case 2: $\delta U^*(q_p) = \lambda \delta H(q_p;\xi) \; \forall \xi \in \mathscr{C}^1[0, c_d]$. For any fixed M_e and c_d and $\forall \xi$ we have that

$$\nu(M_e, c_d) \int_0^{c_d} \xi(c) u'(q_p(c)) g(c) dc = \lambda M_e \int_0^{c_d} \xi(c) c g(c) dc$$

so for $\lambda' \equiv \lambda M_e / \nu(M_e, c_d)$ we have $\int_0^{c_d} [u'(q_p(c)) - \lambda' c] g(c) \xi(c) dc = 0$ and since g is $\mathscr{C}^1$ and strictly positive, we conclude

$$u'(q_p(c)) = \lambda' c \tag{9}$$

Using similar reasoning, a monopolist with costs c picks $q_m(c)$ according to

$$\max_{q_m(c)} [D(q_m(c)) - c] q_m(c) = \max_{q_m(c)} [u'(q_m(c))/\delta - c] q_m(c) \tag{Market}$$

so long as the resulting profit covers f. By assumption, the FOC $[u'(q_m(c))/\delta - c] + u''(q_m(c)) q_m(c)/\delta = 0$ uniquely determines each monopolist's optimal quantity which must be $q^*(c)$ in equilibrium. We conclude that $q^*(c)$ is implicitly determined by the monopolist FOC as given in Equation (10).

$$u'(q^*(c)) + u''(q^*(c)) q^*(c) = \delta c \tag{10}$$

We now show $q^* = q_p$. Since $H(q_p) = H(q^*)$ and $H(q)$ is linear in q, any convex combination $q_\alpha \equiv \alpha q^* + (1-\alpha) q_p$ has $H(q_\alpha) = H(q_p) = H(q^*)$ and so is attainable. Since u is strictly concave, a standard concavity argument shows that the optimality of q_p and q^* implies $q_p = q_\alpha = q^* \quad \forall \alpha \in [0,1]$. Now comparing Equations (9) and (10) with the knowledge that $q^* = q_p$ and dividing the second by the first we see Equation (11) holds on $[0, c_d]$.

$$1 + u''(q_p(c))q_p(c)/u'(q_p(c)) = \delta/\lambda' \tag{11}$$

Equation (11) implies for some constant k_0 that for each $c \in [0, c_d]$ that

$$u''(q_p(c))q_p(c) \quad = \quad k_0 u'(q_p(c))$$

Equation (10) paired with $u'' < 0$ shows that $q(c)$ is strictly decreasing so we have that $q([0, c_d]) = [q(c_d), q(0)]$. Consequently, $\forall x \in [q(c_d), q(0)]$ we have that $u''(x)x = k_0 u'(x)$. Standard solution techniques imply that the unique continuously differentiable solution for u on $[0, c_d]$ is $u(x) = \alpha + \beta x^\gamma$ for constants α, β, γ, which is precisely the CES form up to an affine transformation. $\square$

A.3 VES Market Allocation

Proposition. *The market equilibrium, when unique, maximizes aggregate real revenue in the economy. Formally, the market allocation solves*

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

Proof. 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 following conditions 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:

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

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

A.4 Static Distortion Results

Lemma. *For sufficiently high fixed costs, the quantities produced by all firms are close to the maximum quantity produced ($q(0)$).*

Proof. To clarify, we wish to show for sufficiently high f, for any producing firm with cost $c \le c_d$, either $|q(0) - q(c)|$ is arbitrarily small in the case that $q(0)$ is finite, otherwise when $q(0) = \infty$, $q(c)$ grows large. For both of these cases, we need only consider the impact of f on $q(c_d)$ since our assumptions imply it is the lowest quantity produced and the quantity $q(0)$ is unaffected by f in the market or social optimum. Furthermore, both cases hold iff as $f \longrightarrow \infty$, $\delta c_d \longrightarrow 0$ because (considering the market case, similar to the optimum case) we have

$$u'(q(c_d))\left[1 - \mu(q(c_d))\right] = \delta c_d$$

and the LHS is marginal revenue which is decreasing in quantity. Since δ is equal to revenue per capita which by above is maximized by the market, δ is decreasing in f. Direct comparative statics also show that c_d is decreasing in f. Therefore if either δ or $c_d \longrightarrow 0$ we are done and WLOG both δ and c_d are bounded away from 0 (at least on a subsequence, which monotonicity forces to be true on the whole sequence). In particular, $d\delta/df = -\delta G(c_d)M_e/L$ and since $\delta \ge 0$, necessarily $d\delta/df \longrightarrow 0$ which implies $M_e \longrightarrow 0$. Finally, $\delta = M_e \int_0^{c_d} u'(q(c))\, q(c) dG$ and as δ is bounded away from zero and $M_e \longrightarrow 0$, we conclude $\int_0^{c_d} u'(q(c))\, q(c) dG \longrightarrow \infty$. Noting that $u'(q(c))\, q(c) = \varepsilon(q(c)) \cdot u(q(c))$, since c_d is bounded away from zero and G is a probability distribution, $\int_0^{c_d} u'(q(c))\, q(c) dG \longrightarrow \infty$ implies $q(c) \longrightarrow q(0)$ for $c \in [0, \kappa]$ for some $\kappa > 0$. However this contradicts δc bounded away from zero as $u'(q(c))\left[1 - \mu(q(c))\right] = \delta c$. We conclude at least one of δ or $c_d \longrightarrow 0$, giving the result. □

Proposition. *When $(1-\varepsilon)'$ and μ' have different signs, $q^{\text{mkt}}(c)$ 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 $(1-\varepsilon)'$ and μ' have the same sign 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 $c > c^*$.*

Proof. This result relies heavily on the following relationship which we first prove:

$$\overline{\sigma} \equiv \sup_{c \le c_d^{\text{mkt}}} \varepsilon\left(q^{\text{mkt}}(c)\right) > \delta/\lambda > \inf_{c \le c_d^{\text{opt}}} \varepsilon\left(q^{\text{opt}}(c)\right) \equiv \underline{\sigma}. \tag{12}$$

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

$$\delta/\overline{\sigma} = M_e^{\text{mkt}} \int_0^{c_d^{\text{mkt}}} \left(\varepsilon\left(q^{\text{mkt}}(c)\right)/\overline{\sigma}\right) u\left(q^{\text{mkt}}(c)\right) dG < M_e^{\text{mkt}} \int_0^{c_d^{\text{mkt}}} u\left(q^{\text{mkt}}(c)\right) dG \tag{13}$$

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 (12).

Now note that

$$\left[u''\left(q^{\text{mkt}}(c)\right) q^{\text{mkt}}(c) + u'\left(q^{\text{mkt}}(c)\right)\right]/\delta = c, \qquad u'\left(q^{\text{opt}}(c)\right)/\lambda = c. \tag{14}$$

And it follows from Equations (14) we have

$$\left[1 - \mu\left(q^{\text{mkt}}(c)\right)\right] \cdot u'\left(q^{\text{mkt}}(c)\right)/u'\left(q^{\text{opt}}(c)\right) = \delta/\lambda. \tag{15}$$

Suppose $\mu' > 0 > (1-\varepsilon)'$, and it is sufficient to show $\inf\limits_{c \le c_d^{\text{mkt}}} 1 - \mu\left(q^{\text{mkt}}(c)\right) \ge \overline{\sigma}$, since then Equations (12) and (15) 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 $\lim\limits_{c\to 0} q^{\text{mkt}}(c) = \infty = \lim\limits_{c \to 0} q^{\text{opt}}(c)$,

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

Similarly, if $\mu' < 0 < (1-\varepsilon)'$ one may show that $\sup\limits_{c \le c_d^{\text{mkt}}} 1 - \mu\left(q^{\text{mkt}}(c)\right) \le \underline{\sigma}$, implying from Equations (12) and (15) that $q^{\text{mkt}}(c) < q^{\text{opt}}(c)$.

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

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

It follows from Equation (15) 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 (15), 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 $c \neq c^*$. □

Proposition. *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: $c_d^{\text{mkt}} < c_d^{\text{opt}}$.*

Proof. 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:

$$\max_{M_e, c_d, q(c)} LM_e \int_0^{c_d} v_\alpha(q(c))\,dG \text{ subject to } L \geq M_e \left(\int_0^{c_d} Lcq(c) + f\,dG + F_e \right). \tag{16}$$

Let the Lagrange multiplier associated with each α in Equation (16) be written as $\beta(\alpha)$. By appealing to the envelope theorem and differentiating Equation (16) in M_e we have $\beta(\alpha) = M_e \int_0^{c_d} v_\alpha(q(c))\,dG$ and that $d\beta/d\alpha = M_e \int_0^{c_d} w(q(c))\,dG = M_e \int_0^{c_d} u(q(c))\,[\varepsilon(q) - 1]\,dG < 0$. The conditions 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

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

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

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

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

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

Cancellation and expansion again show this is equivalent to

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

Finally, this expression can be rewritten $\varepsilon(q(c_d)) \gtrless \int_0^{c_d} \varepsilon(q(c))\,u(q(c))\,dG / \int_0^{c_d} u(q(c))\,dG$ and since $q(c)$ is strictly decreasing in c, we see $dc_d/d\alpha \gtrless 0$ when $\varepsilon' \lessgtr 0$. Note that Equation (16) 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 $\varepsilon' > 0$. □

Proposition. *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* $(1-\varepsilon)', \mu' > 0$ *and* $\mu'(q)q/\mu \leq 1$, *the market has too little entry:* $M_e^{\text{mkt}} < M_e^{\text{opt}}$.

Proof. 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

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

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

$$B_v'(c) = q(c)/\varepsilon_v(q(c)) + c\,(dq(c)/dc)\left[1-\varepsilon_v'(q(c))\,q(c)/\varepsilon_v(q(c))\right]/\varepsilon_v(q(c)). \tag{17}$$

Equation (17) can be considerably simplified using two relationships. The first is

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

The second is that manipulating the necessary conditions shows that $dq(c)/dc = -(q(c)/c)\cdot(1/\mu_v(q(c)))$. Substituting these relationships into Equation (17) yields

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

Now consider that 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 in integrands)

$$1/M_e^{\text{opt}} - 1/M_e^{\text{mkt}} = \int_0^{c_d^{\text{opt}}} cq^{\text{opt}}/\varepsilon\left(q^{\text{opt}}\right)dG(c) - \int_0^{c_d^{\text{mkt}}} cq^{\text{mkt}}/\left[1-\mu\left(q^{\text{mkt}}\right)\right]dG \tag{18}$$

and similarly (suppressing the c arguments):

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

Now assume $\varepsilon' < 0 < \mu'$, so by above $c_d^{\text{mkt}} > c_d^{\text{opt}}$ and for the result, from Equation (18) 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, $\mu'(q)q/\mu \leq 1$ implies $q/\mu(q)$ is increasing 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 $c > c^*$. Put together 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 same argument goes through since clearly $\mu'(q)q/\mu(q) \leq 1$. □

A.5 Welfare Gains from Trade

The sufficient condition for welfare 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 as $d\ln\bar{\varepsilon}/d\ln L = (u_d/\bar{\varepsilon}\int udG)(\varepsilon_d - \bar{\varepsilon})c_d g(c_d)(d\ln c_d/d\ln L) + \int(\varepsilon' u + u'\varepsilon - u'\bar{\varepsilon})(d\ln q/d\ln L)dG/\int \varepsilon u dG$. An increase in market size raises the marginal utility of income at the rate of average markups ($d\ln\delta/d\ln L = \int \mu pqdG/\int pqdG \equiv \tilde{\mu}$. Combining $d\ln\delta/d\ln L$ and $d\ln\bar{\varepsilon}/d\ln L$, change in welfare is

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

When preferences are aligned, the first term in square brackets is positive because private markups μ and social markups $(1-\varepsilon)$ 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 as long as preferences are aligned, given regularity conditions in the following Lemma:

Lemma. *Trade increases welfare when preferences are aligned, provided* $(\mu q)'' \leq 0$ *whenever* $\mu' < 0$.

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

$$1 + \int \frac{1-\mu-\bar{\varepsilon}}{1-\mu+\mu'q/\mu}\frac{1-\mu}{\mu}\frac{\varepsilon u}{\bar{\varepsilon}\int udG}dG = \int\left[1-\bar{\varepsilon}+\mu'q/(1-\mu)\right]\gamma dG \geq 0. \tag{19}$$

This clearly holds for $\mu' \geq 0$, and for the other case where preferences are aligned, we have $\mu' < 0 < \varepsilon'$. Expanding Equation (19) 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 $\int [\overline{\mu}-\mu]\,\gamma dG$. This sufficient condition is equivalent to

$$\int \mu \frac{u}{\int udG} dG \geq \int \mu\eta \frac{u}{\int udG} dG \tag{20}$$

where $\eta(c) \equiv \gamma(c)\cdot\left(\int udG/u\right)/\int\gamma$. Since $\int \eta dG = 1$ and $d\mu/dc > 0$, it follows that if $d\eta/dc < 0$, then Equation (20) holds by stochastic dominance. As $d\eta/dc < 0$ iff $d\eta/dq > 0$, consider that

$$\begin{aligned}\operatorname{sign}\{d\eta/dq\} &= \operatorname{sign}\left\{d\ln\left(\mu+\mu' q/(1-\mu)\right)^{-1}\left(\varepsilon/\bar{\varepsilon}\int\gamma\right)/d\ln q\right\}\\ &= \operatorname{sign}\left\{-\left(\mu'' q + 2\mu'\right)q/(1-\mu) + \left(\varepsilon' q/\varepsilon - \mu' q/(1-\mu)\right)\left(\mu+\mu' q/(1-\mu)\right)\right\}.\end{aligned} \tag{21}$$

The additional hypothesis that $(\mu q)'' \leq 0$ guarantees that each term in Equation (21) is positive, so $d\eta/dq > 0$ and we conclude Equation (20) holds, giving the result. $\square$

A.6 Results Regarding the Impact of Large Markets

Lemma. *As market size becomes large:*

1. *Under the market, revenue is increasing in market size and goes to infinity.*

2. *Under the optimum, utility per capita is increasing in market size and goes to infinity.*

3. *Market entry goes to infinity.*

Proof. From above, the market allocation solves

$$\max_{M_e, c_d, q(c)} LM_e \int_0^{c_d} u'(q(c))\,q(c)dG \text{ subject to } L \geq M_e\left(\int_0^{c_d} Lcq(c) + fdG + F_e\right).$$

Let $R(L) \equiv M_e \int_0^{c_d} u'(q(c))\,q(c)dG$ be the revenue per capita under the market allocation. Fix L and let $\{q(c), c_d, M_e\}$ denote the market allocation with L resources. Consider an increased resource level $\tilde{L} > L$ with allocation $\left\{\tilde{q}(c), \tilde{c}_d, \tilde{M}_e\right\} \equiv \left\{\left(L/\tilde{L}\right)\cdot q(c), c_d, \left(\tilde{L}/L\right)\cdot M_e\right\}$ which direct inspection shows is feasible. This allocation generates revenue per capita of

$$\tilde{M}_e \int_0^{\tilde{c}_d} u'(\tilde{q}(c))\,q(c)dG = M_e \int_0^{c_d} u'\left(\left(L/\tilde{L}\right)\cdot q(c)\right)q(c)dG \leq R\left(\tilde{L}\right).$$

Since u is concave, it follows that $R\left(\tilde{L}\right) > R(L)$. Since $\tilde{q}(c) = \left(L/\tilde{L}\right)\cdot q(c) \longrightarrow 0$ for all $c > 0$ and $\lim_{q\longrightarrow 0} u'(q) = \infty$, revenue per capita goes to infinity as $\tilde{L} \longrightarrow \infty$. A similar argument holds for the social optimum.

First note that $q(c)$ is fixed by $u'(q(c))[1-\mu(q(c))] = \delta c$, and $\delta \longrightarrow \infty$ and $\mu(q(c))$ is bounded, it must be that $u'(q(c)) \longrightarrow \infty$ for $c > 0$. This requires $q(c) \longrightarrow 0$ for $c > 0$. Since revenue $u'(q(c))q(c)$ is equal to $\varepsilon(q(c))u(q(c))$ and ε is bounded, revenue also goes to zero for each $c > 0$. Revenue is also decreasing in δ for every c, so we can bound revenue with a function $B(c)$. In particular, for any fixed market size $\tilde{L}$ and implied allocation $\left\{\tilde{q}(c), \tilde{c}_d, \tilde{M}_e\right\}$, for $L \geq \tilde{L}$:

$$u'(q(c))q(c)\mathbf{1}_{[0,c_d]}(c) \leq u'(\tilde{q}(c))\tilde{q}(c)\mathbf{1}_{[0,\tilde{c}_d]}(c) + u'(\tilde{q}(\tilde{c}_d))\tilde{q}(\tilde{c}_d)\mathbf{1}_{[\tilde{c}_d,\infty]}(c) \equiv B(c) \tag{22}$$

where we appeal to the fact that $q(c)$ is decreasing in c for any market size. Since for any L, $\int_0^{c_d} u'(q(c))q(c)dG = \delta/M_e$, it is clear that $\int_0^{\infty} B(c)dG = \int_0^{\tilde{c}_d} u'(\tilde{q}(c))\tilde{q}(c)dG + u'(\tilde{q}(c_d))\tilde{q}(c_d) < \infty$. Since $u'(q(c))q(c)$ converges pointwise to zero for $c > 0$, we conclude

$$\lim_{L \longrightarrow \infty} \int_0^{c_d} u'(q(c))q(c)dG = \int_0^{c_d} \lim_{L \longrightarrow \infty} u'(q(c))q(c)dG = 0$$

by dominated convergence. Therefore $\lim_{L \longrightarrow \infty} \delta/M_e = 0$ which with $\delta \longrightarrow \infty$ shows $M_e \longrightarrow \infty$. The optimal allocation case is similar. □

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. For any costs $c_L < c_H$, $q(c_H)$ is in the choice set of a firm with costs c_L and therefore

$$\pi(c_L) \geq (p(c_H) - c_L)q(c_H)L - f = \pi(c_H) + (c_H - c_L)q(c_H)L. \tag{23}$$

Furthermore, for every $\tilde{c} > 0$, we argue that $\pi(\tilde{c})$ is bounded. For $\underline{c} \equiv \tilde{c}/2$, $\pi(\tilde{c}) \leq \pi(\underline{c})$ while $\pi(\underline{c})$ is bounded since $\lim\limits_{L\to\infty} \int_0^{c_d} \pi(c)dG = F_e$ and $\limsup\limits_{L\to\infty} \pi(\underline{c}) = \infty$ would imply $\limsup\limits_{L\to\infty} \int_0^{c_d} \pi(c)dG = \infty$. It follows from Equation (23) that $Lq(c)$ is bounded. Substituting ϖ for π leads to similar arguments for the social optimum. □

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

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

Similarly, under the optimal allocation:

1. $\lim\limits_{L\to\infty} c_d^{\text{opt}} = \infty$ *iff* $\lim\limits_{L\to\infty} u \circ q\left(c_d^{\text{opt}}\right)/\lambda q\left(c_d^{\text{opt}}\right) = \infty$ *iff* $\lim\limits_{L\to\infty} Lq\left(c_d^{\text{opt}}\right) = 0$.

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:

$$u'(q(c_d))/\delta - f/\left[Lq(c_d)\cdot\mu\circ q(c_d)/(1-\mu\circ q(c_d))\right] = c_d, \tag{24}$$

$$Lc_d q(c_d)\cdot\mu\circ q(c_d)/(1-\mu\circ q(c_d)) = f. \tag{25}$$

First, if $\lim_{L\to\infty} Lq(c_d) = 0$, Equation (25) 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 $\mu\circ q(c_d)/(1-\mu\circ q(c_d))$ is bounded. It follows from Equation (25) that $Lc_d q(c_d)$ is bounded, so from Equation (24), $Lq(c_d)\cdot u'(q(c_d))/\delta$ 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 (25), $c_d\longrightarrow 0$. Now assume $c_d\longrightarrow 0$ so from Equation (25), $Lq(c_d)\cdot\mu\circ q(c_d)/(1-\mu\circ q(c_d))\longrightarrow\infty$ which implies with Equation (24) that $u'(q(c_d))/\delta\longrightarrow 0$. Finally, if $u'(q(c_d))/\delta\longrightarrow 0$, Equation (24) 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, $u\circ q(c)/\lambda q(c)$ converges in $(0,\infty)$ for $c>0$, $Lq(c_d)$ converges in $(0,\infty)$.*

Proof. Since $q(c)\longrightarrow 0$ for all $c>0$, $\lim_{q\to 0}\mu(q)\in(0,1)$ shows $\lim_{L\to\infty} p(c)$ aligns with constant markups and thus converges for all $c>0$. In particular, $p(c_d)$ converges and $L(p(c_d)-c_d)q(c_d) = f$ so it follows $Lq(c_d)$ converges. Similar arguments hold for the social optimum. □

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

Proof. Fix any $c>0$ and first note that for both the market and social planner, $q(c)/q(c_d) = Lq(c)/Lq(c_d)$ and both $Lq(c)$ and $Lq(c_d)$ are bounded, so $q(c)/q(c_d)$ is bounded.

Now consider the market. $q(c)/q(c_d)\ge 1$ has at least one limit point and if it has two limit points, say a and b with $a<b$, there exist subsequences $(q(c)/q(c_d))_{a_n}\to a$ and $(q(c)/q(c_d))_{b_n}\to$

b. There also exist distinct κ and $\tilde{\kappa}$ in (a,b) so that eventually

$$(q(c))_{a_n} < \kappa q(c_d)_{a_n} < \tilde{\kappa} q(c_d)_{b_n} < (q(c))_{b_n}.$$

With $u'' < 0$ this implies

$$\begin{aligned}\left(u'(q(c))/u'(q(c_d))\right)_{a_n} &> \left(u'(\kappa q(c_d))/u'(q(c_d))\right)_{a_n} > \left(u'(\tilde{\kappa} q(c_d))/u'(q(c_d))\right)_{b_n} \\ &> \left(u'(q(c))/u'(q(c_d))\right)_{b_n}.\end{aligned}$$

By assumption, $\lim_{q\to 0} u'(\kappa q)/u'(q) > \lim_{q\to 0} u'(\tilde{\kappa} q)/u'(q)$ but since $q(c) \longrightarrow 0$,

$$\begin{aligned}\lim_{n\longrightarrow\infty} \left(u' \circ q(c)/u' \circ q(c_d)\right)_{a_n} &= \lim_{n\longrightarrow\infty} \left([1-\mu \circ q(c)]\, c/[1-\mu \circ q(c_d)]\, c_d\right)_{a_n} = c/c_d \\ &= \lim_{n\longrightarrow\infty} \left(u' \circ q(c)/u' \circ q(c_d)\right)_{b_n}\end{aligned}$$

where we have used the fact that $\lim_{q\to 0} \mu(q) \in (0,1)$, however by assumption this contradicts $a < b$.

For the social optimum, we could repeat this argument (substituting $\varepsilon \neq 0$ for $u'' < 0$ where appropriate) so long as

$$\kappa \neq \tilde{\kappa} \text{ implies } \lim_{q\longrightarrow 0} (u(\kappa q)/\kappa q)\,/\,(u(q)/q) \neq \lim_{q\longrightarrow 0} (u(\tilde{\kappa} q)/\kappa q)\,/\,(u(q)/q). \tag{26}$$

Since $\lim_{q\to 0} u'(q) = \infty$ and $\lim_{q\to 0} \varepsilon \in (0,\infty)$ it follows that $\lim_{q\to 0} u(q)/q = \infty$. By L'Hospital's rule, $\lim_{q\to 0} (u(\kappa q)/\kappa q)\,/\,(u(q)/q) = \lim_{q\to 0} u'(\kappa q)/u'(q)$ for all κ so the condition (26) in holds because $\kappa \neq \tilde{\kappa}$ implies $\lim_{q\to 0} u'(\kappa q)/u'(q) \neq \lim_{q\to 0} u'(\tilde{\kappa} q)/u'(q)$. □

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. By assumption, $\lim_{q\to 0} \varepsilon(q) > 0$. Expanding this limit via L'Hospital's rule shows

$$\begin{aligned}\lim_{q\longrightarrow 0} \varepsilon(q) = \lim_{q\longrightarrow 0} q/\left(u(q)/u'(q)\right) \qquad &= \lim_{q\longrightarrow 0} 1/\lim_{q\longrightarrow 0} \left(1-u(q)u''(q)/(u'(q))^2\right) \\ = 1/\lim_{q\longrightarrow 0} \left(1+\mu(q)/\varepsilon(q)\right) &= \lim_{q\longrightarrow 0} \varepsilon(q)/\lim_{q\longrightarrow 0} \left(\varepsilon(q)+\mu(q)\right)\end{aligned}$$

which gives the first part of the result. Identical steps for $q \longrightarrow \infty$ give the second part. □

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)

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

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 $\lim_{q\to 0} f_q'(z) = -\mu^\infty\Upsilon^{-1}(z)/z$. 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 $q \longrightarrow 0$. It follows (Rudin 1964, Thm 7.17) that

$$\lim_{q \longrightarrow 0} f_q'(z) = d \lim_{q \longrightarrow 0} f_q(z)/dz = -\mu^\infty \Upsilon^{-1}(z)/z = d\Upsilon^{-1}(z)/dz. \tag{27}$$

We conclude that $\Upsilon^{-1}(z)$ is differentiable and thus continuous, and given the form deduced in (27), $\Upsilon^{-1}(z)$ is continuously differentiable. Since $d\Upsilon^{-1}(z)/dz = 1/\Upsilon' \circ \Upsilon^{-1}(z)$, composing both sides with $\Upsilon(z)$ and using Equation (27) 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 $c > 0$:

$$q^{\text{opt}}(c)/q^{\text{opt}}\left(c_d^{\text{opt}}\right) \longrightarrow \left(c/c_\infty^{\text{opt}}\right)^{-1/\alpha} \text{ and } q^{\text{mkt}}(c)/q^{\text{mkt}}\left(c_d^{\text{mkt}}\right) \longrightarrow \left(c/c_\infty^{\text{mkt}}\right)^{-1/\alpha}. \tag{28}$$

Now consider the parallel conditions involving F_e for the market and social optimum, $\int_0^{c_d^{\text{mkt}}} \pi(c)dG = F_e = \int_0^{c_d^{\text{opt}}} \varpi(c)dG$. Expanding these we see that

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

It necessarily follows that

$$\lim_{L\longrightarrow\infty} L\int_0^{c_d^{\text{mkt}}} \mu\circ q^{\text{mkt}}(c)/\left(1-\mu\circ q^{\text{mkt}}(c)\right)\cdot cq^{\text{mkt}}(c)dG - fG(c_d^{\text{mkt}}) =$$
$$\lim_{L\longrightarrow\infty} L\int_0^{c_d^{\text{opt}}} \left(1-\varepsilon\circ q^{\text{opt}}(c)\right)/\varepsilon\circ q^{\text{opt}}(c)\cdot cq^{\text{opt}}(c)dG - fG(c_d^{\text{opt}}). \tag{29}$$

Using Equation (28), 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 (29) the limits of the $\mu/(1-\mu)$ and $(1-\varepsilon)/\varepsilon$ terms are equal and factor out of Equation (29), leaving

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

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

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

so that finally evaluating the limits, we have

$$\int_0^{c_\infty^{\text{mkt}}} \left[(c/c_\infty^{\text{mkt}})^{1-1/\alpha} - 1\right] dG = \int_0^{c_\infty^{\text{opt}}} \left[(c/c_\infty^{\text{opt}})^{1-1/\alpha} - 1\right] dG. \tag{30}$$

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

A.7 Melitz Open Economy

A.7.1 Social Welfare

To assess the optimality of market allocations resulting from international trade, we need to clarify the policymaker's objective function over different international pairings between producers and consumers. This is because every linkage between a producer in country j and a consumer in country i may encounter trade frictions distinct from one another, and a policymaker will factor the costs of each linkage in their decisions. We define social welfare W over allocations of goods

$\{Q_{ji}\}$ produced in j and sold in country i to a worker k as

$$W\left(\{Q_{ji}\}\right) \equiv \int_{k \text{ is a worker}} \min_{i,j} \left\{U(Q_{ji})/\omega_{ji}\right\} dk \tag{31}$$

where U is each worker's utility and $\omega_{ji} > 0$ is the Pareto weight for country i's consumption of goods from j.

In our setting, workers are treated identically by producers within each country. Accordingly, we constrain the social planner to provide the same allocation to all workers within a country. We identify each worker i with her country I and a country-wide Pareto weight ω_{JI} which weights utility from goods produced in J. Each country has a mass L_I of workers, which allows us to aggregate within each country and write social welfare as

$$W = \sum_{I \text{ is a country}} L_I \min_{I,J} \{U(Q_{JI})/\omega_{JI}\} = \min_{I,J} \{U(Q_{JI})/\omega_{JI}\} \cdot \sum_I L_I. \tag{32}$$

From Equation (32), dividing both sides by the world population shows any socially optimal allocation maximizes per capita welfare, using appropriate Pareto weights for each country pairing (J,I).[30] For any Pareto efficient allocation $\{Q^*_{JI}\}$, defining weights so that $\omega_{JI}/\omega_{J'I'} = U(Q^*_{JI})/U(Q^*_{J'I'})$ shows $\{Q^*_{JI}\}$ must maximize W (otherwise a Pareto improvement is possible). Since every Pareto efficient allocation corresponds to some set of weights $\{\omega_{ji}\}$, ranging over all admissible weights $\{\omega_{JI}\}$ sweeps out the Pareto frontier of allocations in which there is a representative worker for each country. Thus, any market allocation can be evaluated for Pareto efficiency in the usual way using Equation (32).

A.7.2 CES Efficiency

Proposition. *Every market equilibrium of identical open Melitz economies is socially optimal.*

Proof. Following the discussion of social welfare under trade, we will show that the market allocation is Pareto efficient. Concretely, the products that j produces and are consumed by i are a triple $Q_{ji} = \left(M_e^{ji}, c_d^{ji}, q_{ji}\right)$ which provides welfare of $U\left(Q_{ji}\right) \equiv M_e^{ji} L_i \int_0^{c_d^{ji}} \left(q_{ji}(c)\right)^{\rho} g(c)dc$. As laid out in the definition of social welfare, these j and i are representative, and the optimal allocation is one that maximizes $W \equiv \min_{i,j} \left\{U(Q_{ji})/\omega_{ji}\right\}$ for some Pareto weights $\{\omega_{ji}\}$. Since labor is not mobile and resources are symmetric ($L_j = L$ for all j), one can maximize W by considering the goods produced by each country j separately. Accordingly, fix $j = 1$ so maximizing W amounts to

[30] Our specification of social welfare is consistent with the trade agreement literature. Bagwell and Staiger (2009) focus on equal weights as home and foreign labor are directly comparable in their model due to the presence of an outside homogeneous good.

maximizing

$$W^1 \equiv \min_i \{U(Q_{1i})/\omega_{1i}\}. \tag{33}$$

Since U is increasing (if every element of a product vector Q' is strictly greater than a product vector Q then $U(Q') > U(Q)$) it is easy to see that any $\{Q^*_{1i}\}$ that maximizes W^1 is characterized exactly by simultaneously being on the Pareto frontier while $U(Q_{1i})/U(Q_{1j}) = \omega_{1i}/\omega_{1j}$. Since Equation (33) is difficult to deal with directly, we will now maximize an additive social welfare function $\mathscr{W}^1 \equiv U(Q_{11}) + \sum_{j>1} U(Q_{1j})$. This is because any allocation which maximizes $\mathscr{W}^1$ must be Pareto efficient, as any Pareto improvement increases $\mathscr{W}^1$. Since the Pareto weights are free, at any maximum $\{Q^*_{1i}\}$ we may set $\omega_{1i} \equiv U(Q^*_{1i})$ so that $\{Q^*_{1i}\}$ maximizes Equation (33).

$\mathscr{W}^1$ must be maximized subject to a joint cost function $C(\{Q_{1i}\})$ we now detail. For brevity define the two "max" terms $\overline{M} \equiv \max_j \{M_e^{1j}\}$ and $\overline{c} \equiv \max_j \{c_d^{1j}\}$ and the "fixed" cost function $C_f(\overline{M}, \overline{c}) \equiv \overline{M}(f_e + G(\overline{c})f)$ which is incurred from fixed costs at home. Next define "variable" costs at home $C_1(Q_{11})$ and abroad $C_j(Q_{1j})$ by

$$C_1 \equiv M_e^{11} L \int_0^{c_d^{11}} c q_{11}(c) g(c) dc \quad \text{and} \quad C_j \equiv M_e^{1j} \int_0^{c_d^{1j}} \left(L\tau c q_{1j}(c) + f_x\right) g(c) dc$$

where $\tau = \tau_{ji}$ denotes the symmetric transport cost. Then total costs are given by $C(\{Q_{1i}\}) = C_f(\overline{M}, \overline{c}) + C_1(Q_{11}) + \sum_{j>1} C_j(Q_{1j})$.

Now fix $\left\{M_e^{1j}\right\}$ and $\left\{c_d^{1j}\right\}$ which fixes C_f. Also fix some allocation of labor across variable costs, say $\{\mathscr{L}_j\}$, with $C_f + \sum \mathscr{L}_j = L$, that constrain $C_j \le \mathscr{L}_j$. We may then maximize each $U(Q_{1j})$ subject to the constraint $C_j \le \mathscr{L}_j$ separately and we may assume WLOG that each $\mathscr{L}_j > 0$.[31] As in the argument for the closed economy, sufficient conditions for maximization with $\left\{M_e^{1j}\right\}$ and $\left\{c_d^{1j}\right\}$ fixed are

$$q^*_{11}(c) = c^{1/(\rho-1)} \mathscr{L}_1 / M_e^{11} L R(c_d^{11}), \tag{34}$$
$$q^*_{1j}(c) = c^{1/(\rho-1)} [\mathscr{L}_j / M_e^{1j} - f_x G(c_d^{1j})] / L R(c_d^{1j}) \tau. \tag{35}$$

Having found the optimal quantities of Equations (34-35) in terms of finite dimensional variables, we now prove existence of an optimal allocation. Note that for any fixed pair $(\overline{M}, \overline{c})$, the remaining choice variables are restricted to a compact set $K(\overline{M}, \overline{c})$ so that continuity of the objective function (by defining $U(Q_{1j}) = 0$ when $\mathscr{L}_j = 0$) guarantees existence of a solution and we denote the value of $\mathscr{W}^1$ at the maximum by $S(\overline{M}, \overline{c})$. In fact, $K(\overline{M}, \overline{c})$ can be shown to be a con-

[31] If $\mathscr{L}_j = 0$ for all j then autarkic allocations are optimal, and as shown above the optimal autarkic allocation coincides with the market. Any set of exogenous parameters which result in trade imply welfare beyond autarky, so if countries trade in the market equilibrium, $\mathscr{L}_j = 0$ for all j cannot be optimal. Inada type conditions on $U(Q_{1j})$ imply that if it is optimal to have at least one $\mathscr{L}_j > 0$ then all $\mathscr{L}_j$ are > 0.

tinuous correspondence, so by the Theorem of the Maximum $S(\overline{M},\overline{c})$ is continuous on $C_f^{-1}([0,L])$ (Berge and Karreman, 1963). Since C_f is continuous, $C_f^{-1}([0,L])$ is compact and therefore a global max of $S(\overline{M},\overline{c})$ exists. Therefore there is an allocation that maximizes $\mathscr{W}^1$ which we now proceed to characterize.

Now evaluating welfare at the quantities of Equations (34-35) yield respectively

$$U(Q_{11}) = R(c_d^{11})^{1-\rho} L^{1-\rho} M_e^{11} \left(\mathscr{L}_1/M_e^{11}\right)^{\rho}, \tag{36}$$

$$U(Q_{1j}) = R(c_d^{1j})^{1-\rho} L^{1-\rho} M_e^{1j} \left(\mathscr{L}_j/M_e^{1j} - f_x G(c_d^{1j})\right)^{\rho} \tau^{-\rho}. \tag{37}$$

Equation (36) is increasing in both M_e^{11} and c_d^{11} so it follows that at any optimum, $M_e^{11*} = \overline{M}$ and $c_d^{11*} = \overline{c}$. Equation (37) is first increasing in M_e^{1j}, attains a critical point at $(1-\rho)\mathscr{L}_j/f_x G(c_d^{1j})$ and is then decreasing, so at any optimum $M_e^{1j*} = \min\left\{(1-\rho)\mathscr{L}_j/f_x G(c_d^{1j}), \overline{M}\right\}$. If $c_d^{1j*} < \overline{c}$ then the first order necessary condition implies

$$M_e^{1j} = (1-\rho)\mathscr{L}_j/f_x \left(\rho R(c_d^{1j})/\left(c_d^{1j}\right)^{\rho/(\rho-1)} + (1-\rho)G(c_d^{1j})\right) < (1-\rho)\mathscr{L}_j/f_x G(c_d^{1j})$$

so $c_d^{1j*} < \overline{c}$ implies $M_e^{1j*} = \overline{M}$ and $M_e^{1j*} < \overline{M}$ implies $c_d^{1j*} = \overline{c}$. Ruling out the latter case, $M_e^{1j*} < \overline{M}$ implies $U(Q_{1j}) = \tau^{-\rho} L^{1-\rho} (1-\rho)^{1-\rho} \rho^{\rho} \mathscr{L}_j f_x^{\rho-1} \left(R(c_d^{1j})/G(c_d^{1j})\right)^{1-\rho}$ which is decreasing in c_d^{1j} so $c_d^{1j*} = \overline{c}$ cannot be optimal. Therefore we conclude that $M_e^{1j*} = \overline{M}$ and $c_d^{1j*} < \overline{c}$. In particular, c_d^{1j*} must solve the implicit equation

$$\rho R(c_d^{1j*})/\left(c_d^{1j*}\right)^{\rho/(\rho-1)} + (1-\rho)G(c_d^{1j*}) = (1-\rho)\mathscr{L}_j/\overline{M} f_x \tag{38}$$

derived from the first order necessary condition.

With these results in hand, $\mathscr{W}^1$ reduces to

$$\mathscr{W}^1 = \left(\overline{M}L\right)^{1-\rho} \left\{ R(\overline{c})^{1-\rho} \mathscr{L}_1^{\rho} + \tau^{-\rho} \sum_{j>1} R(c_d^{1j})^{1-\rho} \left(\mathscr{L}_j - \overline{M} f_x G(c_d^{1j})\right)^{\rho} \right\}. \tag{39}$$

Now consider maximizing $\mathscr{W}^1$ as given in Equation (39) over $\overline{M}, \overline{c}, \mathscr{L}_j, c_d^{1j}$ with c_d^{1j} unconstrained by $\overline{c}$ for $j > 1$. Using a standard Lagrangian approach, the candidate solution from the necessary conditions implies $c_d^{1j*} = (f_x/f)^{(\rho-1)/\rho} \overline{c}/\tau$ and since it is assumed $(f/f_x)^{(1-\rho)/\rho} < \tau$ for trade in a market equilibrium in the Melitz framework, $c_d^{1j*} < \overline{c}$. The candidate solution with c_d^{1j} unconstrained also yields Equation (38) so the unconstrained candidate solution coincides with the solution including the omitted constraints $c_d^{1j*} < \overline{c}$. We conclude the necessary conditions embodied in the candidate solution are also necessary to maximize $\mathscr{W}^1$ with constraints. Since these necessary conditions are exactly those which fix the unique market allocation, the market allocation maximizes $\mathscr{W}^1$. □

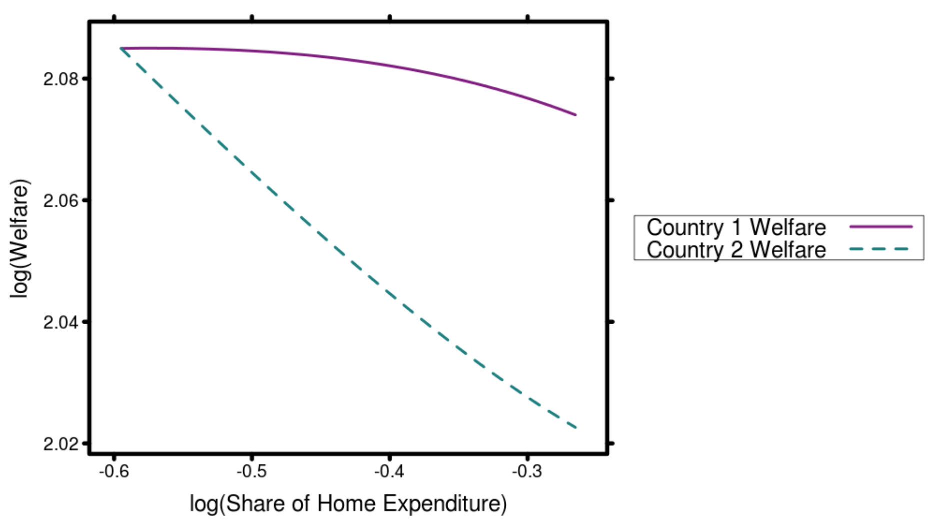
A.8 VES Specific Utility

The VES demand system implied by $u(q) = aq^{\rho} + bq^{\gamma}$ can generate all four combinations of increasing and decreasing, private and social markups as we now briefly discuss. First, note that

$$\varepsilon'(q) = ab(\rho-\gamma)^2 q^{\rho-\gamma-1} / \left(aq^{\rho-\gamma}+b\right)^2,$$
$$\mu'(q) = -ab\rho\gamma(\rho-\gamma)^2 q^{\rho-\gamma-1} / \left(a\rho q^{\rho-\gamma}+b\gamma\right).$$

For $\rho = \gamma$, $\varepsilon'(q) = \mu'(q) = 0$ and we are in a CES economy. For $\rho \neq \gamma$, sign $\varepsilon'(q) =$ sign ab and sign $\mu'(q) =$ sign $-ab \cdot \rho\gamma$, exhibiting all four combinations for appropriate parameter values. In addition, this demand system does not exhibit the log-linear relationship between welfare and share of expenditure on home goods discussed in Arkolakis et al. (2012a), as shown in Figure 2 for $u(q) = q^{1/2} + q^{1/4}$.

Figure 2: Welfare and Share of Home Expenditure as Home Tariff Increases



B Online Appendix

B.1 Trade and Market Size

Proposition. *In the absence of trade costs, 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. Consider a home country of size L opening to trade with a foreign country of size L^*. Suppose the consumer's budget multipliers are equal in each country so $\delta = \delta^*$ and that the terms of trade are unity. We will show that the implied allocation can be supported by a set of prices and therefore constitutes a market equilibrium. The implied quantity allocation, productivity level and per capita entry are the same across home and foreign consumers, so opening to trade is equivalent to an increase in market size from L to $L + L^*$.

Let e denote the home terms of trade, so

$$e \equiv M_e^* \int_0^{c_d^*} p_x^* q_x^* L dG / M_e \int_0^{c_d} p_x q_x L^* dG$$

and by assumption $e = e^* = 1$. Then the $MR = MC$ condition implies a home firm chooses $p(c)[1 - \mu(q(c))] = c$ in the home market and $e \cdot p_x(c)[1 - \mu(q_x(c))] = c$ in the foreign market. A foreign firm chooses $e^* \cdot p^*(c)[1 - \mu(q^*(c))] = c$ in the foreign market and $p_x^*(c)[1 - \mu(q_x^*(c))] = c$ in the home market. When $\delta = \delta^*$ and $e = e^* = 1$, quantity allocations and prices are identical, i.e. $q(c) = q_x^*(c) = q^*(c) = q_x(c)$ and $p(c) = p_x^*(c) = p^*(c) = p_x(c)$.

This implies cost cutoffs are also the same across countries. The cost cutoff condition for home firms is $\pi + e\pi_x = (p(c_d) - c_d)\, q(c_d) L + e\,(p_x(c_d) - c_d)\, q_x(c_d) L^* = f$. Substituting for optimal q^* and q_x^* in the analogous foreign cost cutoff condition implies $c_d = c_d^*$. From the resource constraint, this fixes the relationship between entry across countries as $L/M_e = \int_0^{c_d} [cq(c) + cq_x(c) + f] dG + f_e = L^*/M_e^*$. Thus, $\delta = \delta^*$ and $e = e^* = 1$ completely determines the behavior of firms. What remains is to check that $\delta = \delta^*$ and $e = e^* = 1$ is consistent with the consumer's problem and the balance of trade at these prices and quantities consistent with firm behavior.

For the consumer's problem, we require at home that $1 = M_e \int_0^{c_d} pqdG + M_e^* \int_0^{c_d^*} p_x^* q_x^* dG$, which from $L/M_e = L^*/M_e^*$ is equivalent to

$$L/M_e = L \int_0^{c_d} pqdG + L^* \int_0^{c_d^*} p_x^* q_x^* dG = L \int_0^{c_d} pqdG + L/M_e - L \int_0^{c_d} p_x q_x dG.$$

Therefore to show the consumer's problem is consistent, it is sufficient to show expenditure on home goods is equal to expenditure on exported goods ($\int_0^{c_d} pqdG = \int_0^{c_d} p_x q_x dG$), which indeed holds by the above equalities of prices and quantities. To show the balance of trade is consistent, we use the consumer budget constraint which gives

$$e = M_e^* \int_0^{c_d^*} p_x^* q_x^* L dG / M_e \int_0^{c_d} p_x q_x L^* dG = M_e^* L / M_e L^* = 1.$$

Similarly, the implied foreign terms of trade is $e^* = 1$. Thus $\delta = \delta^*$ and $e = e^* = 1$ generate an allocation consistent with monopolistic competition and price system consistent with consumer maximization and free trade. $\square$