# The Impact of Integration on Productivity and Welfare Distortions Under Monopolistic Competition

Swati Dhingra
CEP, LSE and Princeton University

John Morrow
Centre for Economic Performance, LSE

This Draft: November 27, 2011

**Abstract**

A fundamental question in monopolistic competition theory is whether the market allocates resources optimally. This paper generalizes the Spence-Dixit-Stiglitz framework to heterogeneous firms, addressing when the market provides optimal quantities, variety and productivity. Under constant elasticity demand, each firm prices above its average cost, yet we show market allocations are first-best. When demand elasticities vary, market allocations are not optimal and reflect the distortions of imperfect competition. After determining the nature of market distortions, we investigate how integration may serve as a remedy to imperfect competition. Both market distortions and the impact of integration depend on two demand side elasticities, and we suggest richer demand structures to pin down these elasticities. We also show that integration eliminates distortions, provided the post-integration market is sufficiently large.

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

---

*Acknowledgments.* We thank Bob Staiger for continued guidance, Katheryn Russ for detailed discussion and 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, Mathieu Parenti, Nina Pavcnik, Steve Redding, Andres Rodriguez-Clare, Jacques Thisse, John Van Reenen and Mian Zhu for insightful comments. This paper has benefited from helpful comments of participants at AEA, DIME-ISGEP, the Philadelphia Fed, Princeton University and Wisconsin-Madison. Preliminary draft circulated as "When is Selection on Firm Productivity a Gain from Trade?" in 2006. Swati thanks the IES Princeton for their hospitality during work on this paper. 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 and the impact of market integration on firm survival and markups (Bernard et al. 2007, Feenstra 2006). The introduction of firm heterogeneity in monopolistic competition models has provided new insights into how industries continually reallocate resources. A fundamental question within this setting is whether the market allocates resources optimally. Symmetric firm models explain when market allocations are optimal by explaining the tradeoff 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. This paper answers this question for a general demand structure, which allows rich interrelationships between markups, productivity and efficiency.

We focus on three key aspects of market distortions. First, we characterize which demand systems among the general class considered by Dixit and Stiglitz are efficient in open economies, and discuss the nature of distortions induced by imperfect competition. Second, we study how economic integration affects the ability of markets to reallocate resources optimally. For instance, the distortions of imperfect competition may be mitigated with entry of foreign firms, implying that trade liberalization provides opportunities to correct market failure. The impact of integration depends crucially on demand system characteristics, which we discuss in detail. Third, we examine whether large markets will push outcomes towards what we define as the monopolistically competitive limit, which eliminates distortions. This enables us to understand differences in the ability of small and large markets to allocate resources efficiently.

Differences in productivity across firms alter optimal allocation decisions in a fundamental way. In symmetric firm models, marginal cost pricing and average cost pricing serve as heuristics for first-best and second-best resource allocations. In heterogeneous firm models, inducing each firm to price at its marginal cost or average cost will not maximize welfare because neither scheme takes into account sunk entry costs or the effect of heterogeneity on input costs. Thus, different levels of production maximize welfare than what marginal or average cost pricing imply. It could be optimal to allocate 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. A better understanding of optimal resource allocations can inform policy, especially in regard to international integration.

Starting with constant elasticity of substitution (CES) demand, we show that the closed market equilibrium is first-best despite the existence of positive profits. This optimality result seems surprising, based on the logic of marginal or average cost pricing which is designed to return producer surplus back to consumers. With productivity differences, the market requires prices above average cost 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. Therefore, the market implements the first-best allocation and laissez-faire industrial policy is optimal under CES demand.

How broadly does this efficiency result hold? We generalize the demand structure to the variable elasticity of substitution (VES) form of Dixit and Stiglitz which provides a rich setting for a wide range of market outcomes (Spence 1976; Vives 2001; Zhelobodko et al. 2011).[1] Since optimality is unique to CES demand, any other VES demand generates distortions. Due to imperfect competition, the market maximizes aggregate real revenue instead of maximizing welfare. The nature of distortions can be determined by two demand side elasticities, the inverse demand elasticity and the elasticity of utility. The misalignment of these elasticities pins down the bias in market allocations.

As a heterogeneous cost environment presents a potentially severe information problem for policy, one potential tool for efficiency improvements is to increase competition through international integration. Melitz (2003a) shows that under CES demand, integration increases average productivity, resulting in welfare gains from trade. We show that this outcome is efficient. In the presence of fixed export costs, the firms a planner would close down in the open economy are exactly those that would not survive in the market. However, a planner 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.[2]

Under VES demand, firms respond to import competition by changing their markups, and productivity responds to market size (even in the absence of trade frictions). For this demand class, the impact of integration can be summarized by demand-side elasticities. For instance, changes in market productivity depend on the inverse demand elasticity, while changes in optimal productivity depend on the elasticity of utility. A comparison of these elasticities determines how integration affects the gap between market productivity and optimal productivity. When these elasticities are aligned, distortions will eventually dissipate. However, when the elasticities are misaligned, integration can exacerbate productivity distortions.

To capture the role of integration as a policy tool, we define the monopolistically competitive limit in which firm heterogeneity persists, but market size is so large that quantity sold from a firm to each worker is negligible. In this limit, VES demand operates much like CES demand, and is therefore socially optimal. Thus, constant markups implied by CES demand might approximate richer demand structures in large economies. However, the monopolistically competitive limit may

[1] 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 has an economically significant impact on productivity distributions across markets (Campbell and Hopenhayn 2005, Syverson 2004a, Syverson 2004b). Variable elasticity demand enables us to capture these rich interrelationships.

[2] Melitz assumes equally sized countries trading with each other. Market allocations are efficient even with asymmetric countries. But in the presence of trade frictions, asymmetry of country sizes introduces distributional concerns which we do not address in this paper.

require a market size which is unattainable even in fully integrated world markets. Then estimation of richer demand structures becomes imperative in understanding how integration impacts market distortions. Consequently, our last results provide suggestions about how to assess distortions empirically.

The paper is organized as follows. Section 2 relates this paper to previous work and Section 3 recaps trade models with firm heterogeneity. Section 4 presents efficiency results in a closed economy. Section 5 introduces international trade and contrasts the efficiency of CES demand with inefficiency of VES demand, also deriving a monopolistically competitive limit which shows how integration can eliminate distortions. Section 6 further analyzes the impact of integration on distortions. Section 7 gathers together some theoretical implications useful for designing empirical strategies and Section 8 concludes.

# 2 Related Work

Our paper is related to work on welfare gains in industrial organization and international economics. The tradeoff between variety and quantity occupies a prominent place in the industrial organization literature (e.g., Economides 1989, Mankiw and Whinston 1986). We contribute to this literature by studying the effects of firm heterogeneity and international trade. The analysis is motivated by efficiency properties which have been studied at length in symmetric firm models of monopolistic competition.[3] Recently, Bilbiie et al. (2006) show the market equilibrium with symmetric firms is socially optimal if and only if preferences are CES. We generalize the result to heterogeneous firms and show that efficiency is unrelated to the productivity distribution of firms. To the best of our knowledge, this is the first paper to show market outcomes in Melitz are first best.[4]

To highlight the potential scope of market imperfections, we generalize the well known CES demand structure to VES demand. In contemporaneous work, Zhelobodko et al. (2011) 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. Unlike Zhelobodko et al., our focus is on market efficiency.

We also study the limiting behavior of a VES economy. A large literature examines whether

[3] Spence (1976); Dixit and Stiglitz (1977); Bilbiie et al. (2006); Behrens and Murata (2009).

[4] We consider this to be the proof of a folk theorem. The idea of efficiency in Melitz has been "in the air." Within the heterogeneous firm literature, Baldwin and Robert-Nicoud (2008) and Feenstra and Kee (2008) discuss certain efficiency properties of the Melitz economy. In their working paper, Atkeson and Burstein (2010) consider a first order approximation and numerical exercises to show 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 et al. (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.

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). This literature considers the limiting behavior as the number of firms tends to infinity. Instead, we examine a monopolistically competitive model with a continuum of firms so there are infinitely many firms even in an economy with finite market size. After establishing the equivalence of increased international trade and increased market size, we study the limiting behavior in terms of the model primitive of market size becoming large.

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 (2003b) shows that opening to trade raises welfare through reallocation of resources towards high productivity firms. Considering 48 countries exporting to the US in 1980-2000, Feenstra and Kee (2008) estimate that rise in export variety accounts for an average 3.3 per cent rise in productivity and GDP for the exporting country. In recent influential work, Arkolakis et al. (forthcoming) show that the mapping between trade data and welfare is the same across several old and new trade models with different production structures. This equivalence holds for models which permit welfare to be summarized by import shares and trade elasticities (that can be derived from gravity equations). Unlike Arkolakis et al., we vary the demand side of new trade models and focus on the optimality of market outcomes. Once the Spence-Dixit-Stiglitz demand framework is considered, welfare inferences from import shares require additional information about demand and become more structural in nature. A large body of empirical studies use firm level production data to examine whether trade liberalization induces exit of low productivity firms and increases sales of high productivity firms.[5] Our results characterize when observed productivity gains reflect a narrowing of the distortionary gap between market and optimal productivity. Therefore, our work is more in line with Tybout (2003) and Katayama et al. (2009) who point to the limitations of the empirical literature in mapping observed productivity gains to welfare and optimal policies.

Our results speak directly to the mixed findings about trade liberalization and productivity in the empirical literature. Following trade liberalization, some countries show a reallocation towards high productivity firms while others showing a reallocation towards low productivity firms.[6] Tybout (2003) proposes that these mixed findings could mean that the selection effects emphasized by Melitz are not robust, or that firm size is a poor proxy for productivity. We address the first issue by examining the robustness of selection effects to general demand specifications. We show that differences in inverse demand elasticities induce different patterns of firm selection, reconciling the

[5]For a detailed survey of the literature, the reader is referred to Tybout (2003). While productivity estimation is fraught with difficulties in measuring technical efficiency, we focus on the relationship between productivity and welfare as exposited in the heterogeneous firm literature.

[6]Interpreting differences in firm size as productivities, Tybout (2003) notes that it was the high productivity firms that lost market share in Chile and Colombia while it was the low productivity firms that suffered a decline in Morocco.

mixed evidence for productivity changes across heterogeneous firms. The second issue of productivity measurement has been addressed in several studies. Instead of addressing measurement, we focus on how VES demand can better explain observed patterns. We show how observed markups and physical productivity vary with market size under general demand specifications. Our findings reiterate the importance of disentangling changes in markups and productivity to understand the sources of welfare gains from trade.

# 3 Trade Models with Heterogeneous Firms

Trade models with heterogeneous firms differ from earlier trade 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. 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 in the home and foreign countries. Let $M_e$ denote the mass of entering varieties and $q(c)$ denote 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:

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

Here $u$ denotes utility from an individual variety and $\int u(q)dG$ denotes utility from a unit bundle of differentiated varieties. In a Melitz economy, preferences take the special CES form with $u(q) = q^\rho$.[7] 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 with $u' > 0$ and $u'' < 0$.

[7] 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.

3. $u$ guarantees each monopolist's FOC uniquely determines their optimal quantity.[8]

4. The elasticity of marginal utility $\mu(q) \equiv |qu''(q)/u'(q)|$ is less than one.

For each good indexed by $c$, VES preferences induce an inverse demand $p(q(c)) = u'(q(c))/\delta$ where $\delta$ is a consumer's budget multiplier. As $u$ is strictly increasing and concave, for any fixed price vector the consumer's maximization problem is concave. The necessary condition which determines the inverse demand is sufficient, and has a solution provided inada conditions on $u$.[9] Multiplying both sides of the inverse demand by $q(c)$ and aggregating over all $c$, the budget multiplier is $\delta = M_e \int_0^{c_a} u'(q(c)) \cdot q(c) dG$.

## 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 $g$.[10]

After entry, should a firm produce for the domestic market it faces a cost function $\text{TC}(q(c)) \equiv cq(c) + f$ 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 profit of the firm from domestic sales 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 $\mu(q) \equiv |qu''(q)/u'(q)| = |d \ln p(q)/d \ln q|$.

When the economy opens to trade, firms incur an iceberg transport cost $\tau \geq 1$ and a fixed cost $f_x \geq 0$ in order to export to other countries. As a result, firms face a cost function $\text{TC}(q_x(c)) \equiv \tau c q_x(c) + f_x$ and a demand function $p(q_x(c))$ for sales to the export market. Profit from foreign sales is $\pi_x(c) \equiv \max_{q_x(c)}[p(q_x(c)) - \tau c]q_x(c)L - f_x$ and the optimal $q_x$ choice is given by a similar $MR = MC$ condition.

[8] Sufficient conditions for this are $2u'' + u'''q < 0$ or that $u$ is the integral of a strictly decreasing and concave function.

[9] Utility functions not satisfying inada conditions are permissible but may require parametric restrictions to ensure existence.

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

### 3.3 Market equilibrium

Profit maximization implies that firms produce for the domestic and/or export markets if they can earn non-negative profits from sales in the domestic and/or export markets, respectively. We denote the cutoff cost level of firms that are indifferent between producing and exiting from the domestic market as $c_a$ in autarky and $c_d$ in the open economy. The cutoff cost level for firms indifferent between exporting and not producing for the export market is denoted by $c_x$. Formally, let $\iota = a, d, x$ denote autarky and the domestic and export markets of the open home economy respectively. Each $c_\iota$ is fixed by the Zero Profit Condition (ZPC),

$$\pi_\iota(c_\iota) = 0 \qquad \text{for } \iota = a, d, x. \qquad \text{(ZPC)}$$

Since firms with cost draws higher than the cutoff level do not produce, the mass of domestic producers ($M_\iota$) supplying to market $\iota$ is $M_\iota = M_e G(c_\iota)$.

In summary, each firm faces a two stage problem: in the second stage it maximizes profits from domestic and export sales 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, let $\Pi(c)$ denote the total expected profit from sales in all markets for a firm with cost draw $c$, then ex ante average $\Pi$ 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 for closed and open economies.

## 4 Efficiency in the Closed Economy

Having described an economy consisting of heterogeneous imperfectly competitive firms, we now examine the optimality of market outcomes in the closed economy. Outside of cases in which imperfect competition leads to competitive outcomes with zero profits, one would generally expect the coexistence of positive markups and positive profits to indicate inefficiency through loss of consumer surplus. Nonetheless, this Section shows that CES demand combined with the Melitz production framework exhibits positive markups and profits for surviving firms, yet it is allocationally efficient. However, we also show that the usual relationship between imperfect competition and welfare, that private incentives are not aligned with optimal production patterns, is true for all VES demand structures except CES.

## 4.1 Welfare under isoelastic demand

In a closed economy, a social planner maximizes individual welfare $U$ as given in Equation (1).[11] The social planner is unconstrained and chooses the mass of entrants, quantities and which firms of various productivities produce. At the optimum, zero quantities will be chosen for varieties above a cost threshold $c_a$. Therefore, all optimal allocative decisions can be summarized by quantity $q(c)$, potential variety $M_e$ and productivity $c_a$. Our approach for arriving at the optimal allocation is to think of optimal quantities $\hat{q}(c)$ as being determined implicitly by $c_a$ and $M_e$ so that per capita welfare can be written as

$$U = M_e \int_0^{c_a} u(\hat{q}(c))dG. \tag{2}$$

After solving for each $\hat{q}$ conditional on $c_a$ and $M_e$, Equation (2) can be maximized in $c_a$ and $M_e$. Proposition 1 shows the market provides the first-best quantity, variety and productivity.

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

*Proof.* See Appendix. □

The proof of Proposition 1 differs from standard symmetric firm monopolistic competition results because optimal quantity is a nontrivial function of unit cost, variety and cutoff productivity. As the proof is involved, we relegate details to the Appendix and discuss the rationale for optimality below.

In homogeneous firm models, we know that 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, homogeneous firms choose price equal to average cost so the profit exactly finances the fixed cost of production. Each firm therefore internalizes the effect of higher variety on consumer surplus, resulting in an efficient market equilibrium (Grossman and Helpman 1993, Bilbiie et al. 2006, Baldwin and Robert-Nicoud 2008).

With heterogeneous firms, markups continue to be constant, ensuring that market prices across firms are proportionate to unit costs. But, average cost pricing is too low to compensate firms for an efficient allocation, because it will not cover ex ante entry costs. The market ensures that surviving firms internalize the losses faced by exiting firms, losses which are determined by aggregate economic demand that depends on $q(c)$, $c_a$ and $M_e$. 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 $c_a$ and $M_e$ are at optimal levels that fix $p(c_a)$, thereby fixing absolute prices to optimal levels.

[11] Free entry implies zero expected profits so the planner's focus is on consumer surplus.

The way in which CES preferences cause firms to optimally internalize aggregate economic conditions can be made clear by defining the elasticity of utility $\varepsilon(q) \equiv qu'(q)/u(q)$ and the social markup $1-\varepsilon(q)$. We term $1-\varepsilon(q)$ the social markup because at the optimal allocation, it denotes the utility from consumption of a variety net of its resource cost. At the optimal allocation, there is a multiplier $\lambda$ which encapsulates the shadow cost of utility and ensures $u'(q(c)) = \lambda c$. Therefore, the social markup is

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

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

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

In contrast, the incentives that firms face in the market are

$$\max_{q(c)} L\mu(q(c))pq - f = \max \frac{\mu(q)}{1-\mu(q)}cq - f.$$

Since $\varepsilon$ and $\mu$ depend only on the primitive $u(q)$, we can examine which preferences would make firms choose optimal quantities. Clearly, if $\mu(q)/(1-\mu(q))$ is proportional to $(1-\varepsilon(q))/\varepsilon(q)$, firms will choose optimal quantities $q$ when they produce, but the set of producers might be smaller or larger than optimal, depending on which firms can make enough profits to clear the fixed cost $f$. For the market to also select the optimal range of productivity, $\mu(q)/(1-\mu(q))$ must not only be proportional to $(1-\varepsilon(q))/\varepsilon(q)$, but in fact be the same. Examining CES demand, we see precisely that $\mu(q)/(1-\mu(q)) = (1-\varepsilon(q))/\varepsilon(q)$ for all $q$. Thus, CES demand incentivizes exactly the right firms to produce, in addition to producing optimal quantities.

A direct implication of Proposition 1 is that laissez faire industrial policy is optimal under CES demand. This 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 under CES demand for differentiated goods. Dixit and Stiglitz consider two sectors (a homogeneous goods sector and a differentiated goods sector) and assume a general utility function to aggregate across these goods. With a general utility function, the elasticity of substitution between the homogeneous and differentiated goods is not constant, leading to inefficient market allocations. In the next subsection, we examine the role of elasticities in greater detail. In keeping with Melitz, we consider a single sector to develop results for market efficiency in terms of elasticities.

## 4.2 Welfare beyond isoelastic demand

Efficiency of the market equilibrium in a Melitz economy is tied to CES demand. To highlight the role of CES demand, we consider the general class of variable elasticity of substitution (VES) demand studied by Dixit and Stiglitz (1977) as specified in Equation (1). With regrad to efficiency, 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.[12]

**Proposition 2.** *Under VES demand, a necessary condition for the market equilibrium to be socially optimal is that u is CES.*[13]

*Proof.* Proof available upon request. □

Under general VES demand, market allocations are not socially optimal. Market allocations 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 in the closed economy.*

*Proof.* See Appendix. □

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

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

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

For CES demand, $u(q) = q^\rho$ while $u'(q)q = \rho q^\rho$ implying revenue maximization is perfectly aligned with welfare maximization. Outside of CES, quantities produced by firms are too low

[12] VES utility is additively separable and therefore does not include the quadratic utility of Melitz and Ottaviano (2008) and the translog utility of Feenstra (2003). However, Zhelobodko et al. (2011) show VES demand captures the qualitative features of market outcomes obtained under these forms of non-additive utility.

[13] For completeness, we note that constant elasticities of demand are necessary but not sufficient for optimality of market allocations. We extend the CES demand of Melitz to CES-Benassy preferences $U(M_e, c_a, q) \equiv \nu(M_e) \int_0^{c_d} q(c)^\rho g(c) dc$. In this example, $u$ is CES but varieties and the unit bundle are valued differently through $\nu(M_e)$. Market allocations under CES-Benassy preferences are the same as with CES preferences of Melitz. However, firms do not fully internalize consumers' taste for variety, leading to suboptimal levels of quantity, variety and productivity. Following Benassy (1996), Bilbiie et al. (2006) and Alessandria and Choi (2007), when $\nu(M_e) = M_e^{\rho(\nu_B+1)}$, these preferences disentangle "taste for variety" $\nu_B$ 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 ($\nu_B = (1-\rho)/\rho$).

or too high and in general equilibrium, this implies the average productivity of operating firms is also too low or too high. Market quantity, variety and productivity reflect distortions of imperfect competition, and therefore, increased competition through opening markets to trade might improve allocations. This leads us to an examination of the impact of trade on market distortions.

# 5 Efficiency in an Open Economy

Motivated by empirical studies of firm heterogeneity, Melitz (2003b) shows that reallocation of resources towards high productivity firms provides a new source of gains from trade. In this Section, we examine how international trade affects market and optimal allocations in a Melitz economy. We start by showing that CES demand continues to induce socially optimal allocations in an open economy. Under VES demand, market allocations are suboptimal so we examine when market expansion from trade eventually mitigates the distortions of imperfect competition while preserving firm heterogeneity.

## 5.1 Welfare under isoelastic demand

Trade provides productivity gains by reallocating resources towards low cost firms. One might therefore expect artificially selecting low cost firms to produce would improve welfare in autarky. In fact, this is not the case. Proposition 1 shows that the autarkic market equilibrium is efficient. This implies that the open economy productivity level is undesirable in autarky. It reduces entry and generates too little variety. However, as Proposition 4 below shows, the productivity level selected in an open economy is efficient. Thus trade itself makes reduced entry and reduced variety of home firms efficient.

**Proposition 4.** *Every market equilibrium of identical open Melitz economies is socially optimal.*

*Proof.* See Appendix. □

Why is the higher productivity level of the open economy inefficient in autarky? Proposition 4 implies that market selection of firms is optimal if an increase in size can only be attained at a cost of exogenous frictions $(\tau, f_x)$. Compared to a frictionless world, trade frictions reduce the potential welfare gains from trade. The market minimizes the losses from frictions by weeding out the right firms. It bids up resource prices and eliminates low productivity firms. Conditional on trade costs, market selection of firms is optimal and provides a net welfare gain from trade.

Proposition 4 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. Market allocations maximize social welfare under equal Pareto weights across the equally sized countries.[14]

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 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. This shows the market is implicitly favoring certain consumers, so that firm selection patterns reflect distributional outcomes in addition to cost competitiveness. The cross-country distribution of welfare gains is important but beyond the focus of this study. In what follows, we wish to study efficiency rather than distribution so we model the stylized case of frictionless trade and consider more general demand structures which can explain a greater range of trade effects.

## 5.2 Welfare beyond isoelastic demand

We examine the impact of opening to international trade on a VES economy with no trade costs. As discussed above, the absence of trade costs allows us to abstract from distributional issues. In particular, 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 $L = L_1 + ... + L_n$. Thus, opening to trade is equivalent to an increase in market size, echoing Krugman (1979). This result is summarized as Proposition 5.

**Proposition 5.** *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.* Available upon request, see also Krugman (1979). □

Proposition 5 allows us to think about increased trade as an increase in market size $L$ of a closed economy. An increase in market size has the identical effect of increased competition (except now instead of new foreign competition, there is new domestic competition) which will impact efficiency through altering market distortions. We turn to efficiency properties of the open

[14] 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.

VES economy, and investigate how far increased competition from trade can go towards improving market outcomes.

### 5.2.1 Market Efficiency under VES Demand

Having established that opening to trade is equivalent to an increase in the size of a VES economy, we can follow the same reasoning as in the closed VES economy to infer that market allocations in an open economy are suboptimal. Marginal revenues do not correspond to marginal utilities so relative allocations in the market are not aligned with efficient allocations. When market allocations are suboptimal, 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.

This raises the question of when integration mitigates or exacerbates 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. To make progress, we follow Stiglitz (1986) and first 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. We emphasize productivity distortions as they are new to models with monopolistic competition and later we also examine how small increases in market size affect productivity distortions.

### 5.2.2 Market Efficiency in Large Markets

We consider the effects of trade when world markets are so large that the quantity supplied from each firm to each worker becomes negligible. For example, consider a small economy which integrates with the rest of the world. Looking at efficiency in large markets explains when integrating with world markets enables a small economy to overcome its market distortions. From a theoretical perspective we will 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.[15]

The large economy concept is similar in spirit to the idea of a competitive limit, although with heterogeneous firms 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) as in Behrens and Murata (2009). 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:

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

some, but not all, firms produce. Accordingly, we consider large economies which satisfy the following assumption.

**Assumption.** *In the large economy, both the market and the social planner select a non-degenerate cost distribution in which some but not all entrants produce.*

In considering general VES demand, we maintain the previous regularity conditions for a market equilibrium and add a few more conditions about the behavior of $u(q)$ when $q$ is close to zero in order to aid the analysis.

**Assumption.** *VES preferences exhibit the following properties for the large economy:*

1. *Quantity ratios distinguish price ratios for small q:*

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

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

3. *For small quantities, the inverse demand elasticity and elasticity of utility are monotone. Formally, for all sufficiently small q,* $\mu'(q) \neq 0$ *and* $\varepsilon'(q) \neq 0$.

Assumption 1 is a regularity condition which guarantees production levels across firms can be distinguished if the firms charge distinct prices. In particular, this implies $\lim_{q \to 0} u'(q) = \infty$. Following Dixit and Stiglitz, Assumptions 2 and 3 imply that market prices converge to non-zero constants as market size grows and $\lim_{q \longrightarrow 0} 1 - \varepsilon(q) = \lim_{q \longrightarrow 0} \mu(q)$ so the added utility provided per labor unit in the social optimum converges to a non-zero constant as in Solow (1998). An example of a utility function satisfying Assumptions 2 and 3 is $u(q) = q^{\rho} - bq$ for $\rho \in (0,1)$ and $b > 0$.[16] The main economic content of these assumptions is that they determine when the economy converges to a monopolistically competitive limit, rather than a competitive 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 $\mu$. Under the competitive limit, firms are left with no market power and $\mu$ drops to zero. Similarly, the social markup $(1-\varepsilon)$ does not drop to zero in the monopolistically competitive limit and each variety contributes at a positive rate to utility even at low levels of quantity. Combining these assumptions ensures the large economy goes to the monopolistically competitive limit, summarized as Proposition 6.

[16] Kuhn and Vives (1999) find that $\lim_{q \longrightarrow 0} 1 - \varepsilon(q) = \lim_{q \longrightarrow 0} \mu(q)$ under relatively mild assumptions. Let $\nu = |V'(q)q/V(q)|$ where $V(q) = u(q)/q$ in our notation. Then their assumption is that there exist positive constants $\rho$ and $\kappa$ such that $\lim_{q \to 0}[\nu(q) - \nu(0)]/q^{\rho} = \kappa$ so $\nu(q) - \nu(0)$ has an asymptotic expansion at $q = 0$ with a leading term of the constant elasticity type. Examples of such utility functions include $u(q) = q^{\rho} e^{\kappa q}$ for $\rho \in (0,1)$ and $\kappa < (1-\rho)/(1+\rho)\rho$, $u(q) = q^{\kappa}(1-q)$ for $\kappa \in (0,1)$ and $u(q) = q(1-q^{\kappa})$ for $\kappa > 0$.

**Proposition 6.** *Under the large economy 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 $q(c)/q(c_d)$ 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 6 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. Intuitively, we can explain Proposition 6 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 $c$. 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 and utility provided by the market approaches the first-best level.

It is somewhat remarkable that the large market outcome is socially optimal. Firms charge positive markups but they exactly recover both average costs and *ex ante* entry costs. Therefore, market allocations are first-best despite positive markups. Such persistence of imperfection in 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).[17] 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 markets are small, variable markups are crucial in understanding the productivity gap between the market and the social optimum. In the next Section, we examine the impact of trade on the productivity gap in a small variable elasticity economy.

[17] 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.

# 6 Distortions and the Impact of Integration

With variable markups, opening to trade can have positive or negative effects on productivity. Trade expands market size and has different effects on profitability across firms. With constant markups, firm profitability is unaffected by market expansion so trade affects productivity only through trade frictions and not market expansion.[18] Here we show that this is specific to CES demand. With VES demand, trade affects productivity even in the absence of trade frictions. We first show that the inverse demand elasticity determines when the market selects more (or less) productive firms after trade. The optimal productivity change depends on the elasticity of utility. As discussed above, market allocations in a VES economy are not optimal, so it is unclear if the direction of productivity changes is optimal. Accordingly, we compare the market and optimal outcomes to examine the distortions induced by imperfect competition.

## 6.1 Productivity changes in the market

The markup rate in a VES economy is $\mu(q) \equiv |qu''(q)/u'(q)|$. When $\mu'(q) > 0$, markups are positively correlated with quantity and when $\mu'(q) < 0$, markups are negatively correlated with quantity. Here we show that $\mu' > 0$ implies that productivity rises following market expansion, and $\mu' < 0$ implies productivity falls after market expansion. For CES demand, markups are constant ($\mu' = 0$) and so are uncorrelated with any variable. This distinction between $\mu' > 0$ and $\mu' < 0$ is brought out by the richer VES demand structure and, as we now spell out, the sign of $\mu'$ determines the impact of trade on productivity.

In a VES economy, inverse demand is $p(q(c)) = u'(q(c))/\delta$ where $\delta$ is a consumer's budget multiplier. The multiplier $\delta$ is an aggregate demand shifter that increases with market size. This can be seen from the free entry condition (FE), $\int_0^{c_a} [(p(c)-c)q(c)L - f]\, dG = f_e$. Differentiating FE with respect to $L$ and applying the envelope theorem (and noting $\pi(c_a) = f$), we have

$$\int_0^{c_a} [(p(c)-c)q(c) + L\partial p(c)/\partial L \cdot q(c)]\, dG = 0.$$

The first term on the LHS above is the rise in profits from higher sales in a bigger market while the second term reflects the shift in the residual demand curve due to market expansion. We can solve for the change in aggregate demand conditions as $-\partial \ln p(c)/\partial \ln L = \int_0^{c_a} (p(c)-c)q(c)dG / \int_0^{c_a} p(c)q(c)dG$.

[18] Unlike productivity changes from trade frictions, productivity changes from market size effects of trade reflect an expansion in production possibilities of the economy. However, the interaction between trade frictions, selection and pro-competitive effects of trade can introduce new questions that we do not address here.

Using the fact that $p(c)-c=\mu(c)\cdot p(c)$, we have

$$\partial \ln p(c)/\partial \ln L=-\int_0^{c_a}\mu(c)p(c)q(c)dG/\int_0^{c_a}p(c)q(c)dG<0.$$

As market size expands, more firms enter so residual demand of each firm falls. The percentage fall is the average markup in the economy, and we now examine how this rise affects the ability of the cutoff firm to survive.

From the cutoff cost condition (ZPC), $(p(c_a)-c_a)q(c_a)L=f$. Differentiating with respect to to $L$ and applying the envelope theorem, we have

$$(p(c_a)-c_a)\,q(c_a)+L(\partial p(c_a)/\partial L-dc_a/dL)\,q(c_a)=0.$$

The first terms on the LHS is the rise in profit from higher sales in a bigger market and the second term is the drop in residual demand from market expansion. Re-writing for $d\ln c_a/d\ln L$ and $\partial \ln p/\partial \ln L$, we have

$$\partial \ln p(c_a)/\partial \ln L+(p(c_a)-c_a)/p(c_a)=(d\ln c_a/d\ln L)\cdot c_a/p(c_a). \tag{3}$$

Substituting for $\partial \ln p(c)/\partial \ln L$ and noting $\mu=(p-c)/p$ and $1-\mu=c/p$, we see that

$$(d\ln c_a/d\ln L)\,(1-\mu(c_a))=\mu(c_a)-\int_0^{c_a}\mu(c)\cdot p(c)q(c)dG/\int_0^{c_a}p(c)q(c)dG. \tag{4}$$

Since $\mu(c_a)\in(0,1)$, $d\ln c_a/d\ln L$ is the same sign as the RHS of Equation (4) which depends on the markup of the cutoff firm relative to the average markup in the economy. Our regularity conditions guarantee high cost firms produce lower quantities so $q(c)$ is decreasing in $c$ and sign $d\mu(c)/dc=$ sign $(\mu'(q)q'(c))=-$sign $\mu'(q)$. Therefore, the change in the inverse demand elasticity $\mu'(q)$ determines how market expansion from trade changes the cutoff cost level. When $\mu'(q)>0$, $d\ln c_a/d\ln L<0$ and when $\mu'(q)<0$, $d\ln c_a/d\ln L>0$.

The intuitive explanation is as follows. When $\mu'(q)>0$, high productivity firms sell more $q$ and charge higher markups. This is the case studied by Krugman (1979) and we refer to it as addictive preferences because firms are able to charge higher markups when they sell higher quantities. This means markups are negatively correlated with unit costs so less productive firms have both low $q$ and low markups. With market expansion, the rise in competition ($\delta$) squeezes prices and the less productive firms are the least able to cushion this price drop through markups. They cannot survive so the cutoff cost level drops.

Conversely, when $\mu'(q)<0$, firms selling higher quantities charge lower markups, so we term these preferences congestive. Following market expansion, rise in competition forces a lower

supply of per capita quantity $q$, but $\mu'(q) < 0$ implies a compensating increase in markups to cushion this competition. This increase in markups is strong enough to make even higher cost firms profitable and the cutoff cost level rises with market expansion. Low productivity firms sell the lowest $q(c)$ and have the highest markups to cushion against import competition. These low productivity "boutique" firms fare much better following increased competition. This is consistent with Holmes and Stevens (2010) who find small US plants were less impacted than large plants during the import surge from China.

Under CES preferences, the inverse demand elasticity is $\mu(q) = 1 - \rho$ implying $\mu'(q) = 0$. Constant inverse demand elasticity implies there should be no correlation between markups and unit costs. Therefore, the cutoff cost level is not affected by market size. The drop in residual demand from higher competition exactly counterbalances the higher sales to a bigger market so firm decisions are unaffected. The cost cutoff does not change with market size and only trade frictions induce low productivity firms to exit the market. Under VES demand, this is no longer true. Firms adjust their markups in response to market expansion and the ensuing firm entry, so profitability of the least productive firm is affected by market size. Following market expansion from trade, productivity changes depend on whether the markup of low productivity firms provides enough cushion to absorb the downward shift in demand.

We summarize the implications of addictive and congestive VES preferences in Proposition 7. We show that the market selects high productivity firms when preferences are addictive and includes more low productivity firms when preferences are congestive.

**Proposition 7.** *Increases in market size (L) change the market cost cutoff ($c_a$) as follows:*

1. *When preferences are addictive, defined as $\mu'(q) > 0$, the cutoff decreases with size.*
2. *When preferences are congestive, defined as $\mu'(q) < 0$, the cutoff increases with size.*

We have seen the correlation of markups with quantities characterizes productivity changes. It is unknown if these changes are optimal. For CES demand, this correlation is zero and productivity changes are optimal. VES demand allows non-zero correlation between markups and quantities so we have good reason to expect that changes in market size fundamentally alter the optimality of market allocations. We now address this issue by considering optimal productivity changes under VES demand.

### 6.1.1 Optimal productivity changes

In a parallel fashion to the market productivity changes, we consider two types of preferences based on the sign of $\varepsilon'(q)$ and show that they have different implications for how trade affects

optimal productivity.[19] The argument is similar to the one for the impact of trade on market productivity above. Recalling the definition of the elasticity of utility $\varepsilon(q) = qu'(q)/u(q)$ and leaving calculations to the Appendix, we arrive at the following comparative static:

$$d\ln c_a/d\ln L = \varepsilon(c_a)^{-1}\left[(1-\varepsilon(c_a)) - \int_0^{c_a}(1-\varepsilon(c))\cdot u(q(c))dG/\int_0^{c_a}u(q(c))dG\right]. \quad (5)$$

The change in the cutoff cost level depends on the social markup at the cutoff $(1-\varepsilon(q))$ relative to the average social markup for all varieties. The sign of the RHS of Equation (5) can be determined and interpreted in a manner similar to the change in market productivity.

With Equation (5), we can use the sign of $(1-\varepsilon(q))'$ to determine how market expansion from trade affects the cutoff cost level. If the social markup rises with quantity $(1-\varepsilon(q))' > 0$, social markups are higher at higher levels of quantity. This implies a negative correlation between social markups $1-\varepsilon$ and unit costs $c$ as high cost varieties are produced in lower quantities. For all $c$, $1-\varepsilon(q(c)) > 1-\varepsilon(q(c_a))$ and the cost cutoff falls with market expansion. We refer to this case as *socially addictive* preferences because the social benefit from a variety rises more than proportionately with consumption.

Conversely, when $(1-\varepsilon(q))' < 0$, social markups are lower at higher levels of quantity. As a consumer has more of a variety, social markup from the variety rises less than proportionately so we refer to this as *socially congestive* preferences. These preferences imply a positive correlation between social markups and unit costs so $1-\varepsilon(q(c)) < 1-\varepsilon(q(c_a))$ and the cutoff cost level rises with market expansion. The optimal productivity level falls after trade because the boutique varieties which are consumed in small quantities provide relatively higher utility.

Under CES preferences, the elasticity of utility is $1-\varepsilon(q) = 1-\rho$ implying $(1-\varepsilon(q))' = 0$. Constant elasticity of utility implies the optimal cost cutoff does not change with market size. This result explains why it is not optimal to select high productivity firms in the absence of trade frictions in an open Melitz economy.

In summary, the elasticity of utility characterizes optimal productivity changes after trade, as summarized in Proposition 8. With socially addictive preferences, low productivity varieties have no cushion against the rise in social costs because they provide lower social markups. These varieties are closed down and the optimal cost cutoff falls. With socially addictive preferences, low productivity varieties have a cushion against the rise in social costs because they provide higher social markups. Consequently, low productivity varieties are retained and the optimal cost cutoff rises.

**Proposition 8.** *Increases in market size (L) change the optimal cost cutoff ($c_a$) as follows:*

[19] Kuhn and Vives (1999) and Vives (2001) also use the elasticity of utility to study excess entry in a second-best world with symmetric firms.

1. *When preferences are socially addictive, defined as* $(1-\varepsilon(q))' > 0$*, the cutoff decreases with size.*

2. *When preferences are socially congestive, defined as* $(1-\varepsilon(q))' < 0$*, the cutoff increases with size.*

*Proof.* See Appendix. □

We have so far determined the direction of market productivity changes and optimal productivity changes. We conclude the Section by discussing how these results determine the impact of trade on the difference in productivity levels between the market and optimal outcomes.

## 6.2 Market Distortions and Integration

Here we compare the market and optimal outcomes to understand the nature of distortions in a VES economy. Table 1 summarizes the bias in market allocations. A detailed exposition is under construction and detailed proofs are in the Appendix.

Table 1: Distortions by Demand Characteristics

| | $(1-\varepsilon)' < 0$ | $(1-\varepsilon)' > 0$ |
|---|---|---|
| $\mu' > 0$ | Quantities Too High:<br>$q^{\text{mkt}}(c) > q^{\text{opt}}(c)$<br><br>Productivity Too High: $c_d^{\text{mkt}} < c_d^{\text{opt}}$<br><br>Entry Ambiguous: (VC vs FC) | Quantities Low-Cost Skewed:<br>$q^{\text{mkt}}(c) > q^{\text{opt}}(c)$ for $c < c^*$<br>$q^{\text{mkt}}(c) < q^{\text{opt}}(c)$ for $c > c^*$<br><br>Productivity Too Low: $c_d^{\text{mkt}} > c_d^{\text{opt}}$<br><br>Entry Too Low: $M_e^{\text{mkt}} < M_e^{\text{opt}}$ |
| $\mu' < 0$ | Quantities High-Cost Skewed:<br>$q^{\text{mkt}}(c) < q^{\text{opt}}(c)$ for $c < c^*$<br>$q^{\text{mkt}}(c) > q^{\text{opt}}(c)$ for $c > c^*$<br><br>Productivity Too High: $c_d^{\text{mkt}} < c_d^{\text{opt}}$<br><br>Entry Too High: $M_e^{\text{mkt}} > M_e^{\text{opt}}$ | Quantities Too Low:<br>$q^{\text{mkt}}(c) < q^{\text{opt}}(c)$<br><br>Productivity Too Low: $c_d^{\text{mkt}} > c_d^{\text{opt}}$<br><br>Entry Ambiguous (VC vs FC) |

To understand how integration affects the gap between market and optimal allocations, Table 2 summarizes the impact of integration on distortions by demand characteristics. Depending on the inverse demand elasticity and the elasticity of utility, productivity distortions may be mitigated or exacerbated. When the elasticities are aligned, increased integration would eventually reduce distortions in a VES economy, as discussed earlier.

Table 2: Impact of Integration on Distortions

| | $(1-\varepsilon)' < 0$ | $(1-\varepsilon)' > 0$ |
|---|---|---|
| $\mu' > 0$ | Productivity Diverges: $c_d^{\text{mkt}} \downarrow$ & $c_d^{\text{opt}} \uparrow$<br>$c_d^{\text{mkt}} < c_d^{\text{opt}}$ so Productivity Distortions<br><br>Integration + Private-Social Misalignment:<br>Productivity Distortions Magnified<br>Scope for complementary policy | Productivity Co-moves: $c_d^{\text{mkt}} \downarrow$ & $c_d^{\text{opt}} \downarrow$<br>$c_d^{\text{mkt}} > c_d^{\text{opt}}$ so Possible Correction<br><br>Integration + Private-Social Alignment:<br>Monopolistically Competitive Limit<br>Market and Optimum Coincide |
| $\mu' < 0$ | Productivity Co-moves: $c_d^{\text{mkt}} \uparrow$ & $c_d^{\text{opt}} \uparrow$<br>$c_d^{\text{mkt}} < c_d^{\text{opt}}$ so Possible Correction<br><br>Integration + Private-Social Alignment:<br>Monopolistically Competitive Limit<br>Market and Optimum Coincide | Productivity Diverges: $c_d^{\text{mkt}} \uparrow$ & $c_d^{\text{opt}} \downarrow$<br>$c_d^{\text{mkt}} > c_d^{\text{opt}}$ so Productivity Distortions<br><br>Integration + Private-Social Misalignment:<br>Productivity Distortions Magnified<br>Scope for complementary policy |

Since the theoretical implications of this paper depend heavily on the nature of demand, in particular through $\mu(q)$ and $\varepsilon(q)$, we now turn to how the theory can inform empirical work.

# 7 Theoretical Insights for Empirical Strategies

This paper has so far illustrated that the underlying demand structure can have very different implications for welfare and productivity changes following market expansion. The nature of these demand systems across industries, countries and markets is an empirical question beyond the scope of this paper. However, as we have shown, there are observable differences that can distinguish different systems within the VES class. This Section recaps these differences for empirical work and discusses some theoretical considerations we hope might be useful when designing estimation strategies.

Propositions 7 and 8 identify at least four combinations with distinct relationships between market and optimal productivity. Are preferences congestive or addictive? Spence suggests preferences are socially congestive while Dixit and Stiglitz consider socially addictive preferences, so *a priori* there is lack of a consensus (Vives 2001). Empirical work to guide this analysis is limited (Weyl and Fabinger 2009). Therefore, the answer to this question must rely on further investigation and we discuss theoretical predictions that can guide this analysis. We begin with how observable implications of VES preferences can distinguish between different types of market demand. We then move on to distinguishing social welfare properties which enables policy inference. The section concludes by contrasting productivity and welfare changes in small versus large markets.

## 7.1 Directly observable features of VES demand

Addictive and congestive preferences can be distinguished in at least three ways. First, the productivity threshold for market survival can be directly observed from firm production data. Since addictive preferences imply selection of high productivity firms after a rise in market size (Proposition 7), increases in the productivity threshold are consistent with addictive preferences. Conversely, decreases in the productivity threshold are consistent with congestive preferences (and inconsistent with addictive preferences). We state this relationship as Remark 1.

*Remark* 1. Following a rise in market size, productivity increases are consistent with addictive preferences, while productivity decreases are consistent with congestive preferences.

Second, addictive preferences by definition mean that markups increase with quantity ($\mu'(q) > 0$). Congestive preferences mean that markups decrease with quantity ($\mu'(q) < 0$). Therefore, a direct approach to distinguishing addictive and congestive preferences is markup estimation using firm pricing and production data. In particular, addictive preferences imply that markups and quantities are positively correlated (Remark 2). Since the theory implies that quantity falls as unit cost increases, addictive preferences also imply markups and unit costs are negatively correlated (Remark 3). The opposite correlations hold for congestive preferences.

*Remark* 2. Addictive preferences imply $\text{Cov}(\mu, q) > 0$ and $\text{Cov}(\mu, c) < 0$.

*Remark* 3. Congestive preferences imply $\text{Cov}(\mu, q) < 0$ and $\text{Cov}(\mu, c) > 0$.

Third, if sufficient data is available, the markup function $\mu(q)$ can be estimated semi-parametrically to allow $\mu'(q)$ to vary in sign. Having obtained an estimate $\hat{\mu}(q)$, one can use the VES demand structure to examine the relation between markups and productivity directly. One strength of this approach is that recovering $\hat{\mu}(q)$ would allow recovery of the indirectly observable elasticity of utility $\varepsilon(q)$. In fact, $\varepsilon(q)$ and $\mu(q)$ are interrelated through the expression[20]

$$\ln \varepsilon(q)/q = \int_0^q -(\mu(t)/t)\,dt - \ln\left[\int_0^q \exp\left\{\int_0^s -(\mu(t)/t)\,dt\right\} ds\right]. \tag{6}$$

Equation (6) shows that using data on observed markups to semi-parametrically recover $\mu(q)/q$ will allow recovery of $\varepsilon(q)/q$. This fixes the demand system up to the consumer's budget multiplier $\delta$ and identifies rich productivity and welfare interrelationships as detailed in the above theory.

[20] This equation follows from the observation that $\ln u'(q) = \int_0^q -(\mu(t)/t)\,dt + \kappa$ for some constant $\kappa$ and by definition $u(0) = 0$. The change in $\varepsilon$ is $\varepsilon' = \varepsilon[1 - \varepsilon - \mu]/q$ which can be recovered from $\mu$ and $\varepsilon$.

## 7.2 Indirectly observable features of VES demand

Distinguishing socially addictive and socially congestive preferences is more challenging. Utility, and therefore elasticity of utility $\varepsilon = u'(q) \cdot q/u(q)$, is not directly observable from standard firm pricing and production data. For instance, we know from the theory that socially addictive preferences imply it is optimal to select higher productivity firms after a rise in market size (Proposition 8), but it is hard to say when such selection would 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. (forthcoming).[21] Consequently, for standard firm level datasets, policy inferences from productivity gains require more structure on demand.

To determine $\varepsilon(q)$ empirically, we propose modeling VES preferences in a way which nests all combinations of addictive, congestive, socially addictive and socially congestive preferences. This suggests using flexible demand systems that leave determination of these four possibilities up to the data. However, a preliminary question is whether any single demand system can generate all four possibilities. The answer is affirmative for the parametric specification of Equation (7):

$$u(q) = aq^{\rho} + bq^{\gamma}. \tag{7}$$

The VES form of Equation (7) allows all sign combinations of $\varepsilon'(q)$ and $\mu'(q)$ (shown in the Appendix). This parametric approach has lower data requirements than using firm pricing and production data semi-parametrically as suggested above.

Another approach to recovering $\varepsilon(q)$ is to directly use price and quantity data. As $\varepsilon(q) = u'(q)q/u(q)$, we can use $u(q) = \int u'(q)dq$ with the initial condition $u(0) = 0$ to infer $u(q)$. Multiplying and dividing $\varepsilon(q)$ by $\delta$, we have

$$\varepsilon(q) = (u'(q)/\delta)q/\int (u'(q)/\delta)dq = p(q)q/\int p(q)dq.$$

which can be obtained using firm price and quantity data. To recover the area under the demand curve, $\int p(q)dq$, we need to account for the fact that the observed price-quantity distribution reflects the cost distribution $G(c)$ and not the uniform quantity distribution over which the demand curve should be integrated. For instance, at the mode of the cost distribution, we will observe more price-quantity pairs but these observations over-represent the demand curve at the mode, so these observations need to be appropriately weighted when constructing a sample analog of the integral $\int p(q)dq$.

[21] This is shown graphically in the Appendix for the VES system of Equation (7) below.

One approach to recovering $\int p(q)dq$, based on an ordered sample $\{(p_i, q_i)\}$, is to approximate

$$\int_0^{q_k} p(q)dq \approx \sum_{q_i \leq q_k} p_i \cdot (q_{i+1} - q_{i-1})/2$$

which weights each observed price by the length of the "quantity interval" $[(q_i + q_{i-1})/2, (q_{i+1} + q_i)/2]$ over which the demand curve is being integrated. Then the sample analog of the elasticity of utility is

$$\hat{\varepsilon}(q_k) = p_k q_k / \left[ \sum_{q_i \leq q_k} p_i \cdot (q_{i+1} - q_{i-1})/2 \right]. \tag{8}$$

This equation provides a first pass at recovering the elasticity of utility from firm level data without recourse to semi-parametric or non-linear methods.

Once the elasticity of utility has been recovered, we can determine the results as being socially addictive or socially congestive, and thereby compare actual productivity changes with optimal changes along the lines of Proposition 8. We summarize two distinguishing characteristics, parallel to how addictive and congestive preferences can be distinguished, in Remarks 4-5.

*Remark* 4. Socially addictive preferences imply $\text{Cov}(1-\varepsilon, q) > 0$ and $\text{Cov}(1-\varepsilon, c) < 0$.

*Remark* 5. Socially congestive preferences imply $\text{Cov}(1-\varepsilon, q) < 0$ and $\text{Cov}(1-\varepsilon, c) > 0$.

A careful empirical approach such as those just suggested can address the magnitude of the productivity gap and identify the impact of integration on productivity distortions. We now consider the role of integration with large markets.

## 7.3 The Role of Market Size

In small markets, differences from a CES approximation are likely to be more pronounced and the relationship between market size, markups and firm productivity can be addressed by the more detailed VES demand.

In large markets, Proposition 6 shows that in the monopolisitically competitive limit, per capita consumption of each variety is negligible so the variable elasticities do not play a role in market and optimal outcomes. A theoretical insight that may prove useful in determining how large markets need to be is the idea that markups should tend to align across firms in large markets. Although firms continue to charge positive markups, these markups become more uniform. Therefore decreased dispersion of markups following integration is consistent with positive steps towards the monopolistically competitive limit. We summarize this as Remark 6.

*Remark* 6. The monopolistically competitive limit is consistent with positive markups which become more uniform with increased market size.

Another consequence of Proposition 6 is that the distribution of firm productivity is stationary in the limit. Thus, to explain sizeable productivity shifts from trade, the most promising modeling choice is VES demand in small markets far from the monopolistically competitive limit. Since market allocations are not optimal in the presence of variable elasticities, this also highlights the importance of estimating welfare and evaluating potential policies, conditional on both demand and the productivity distribution of the economy. We leave these avenues to further research.

# 8 Summary and Conclusion

This paper examines when market forces induce an optimal tradeoff between quantity, variety and productivity. Generalizing the Spence-Dixit-Stiglitz framework to heterogeneous firms, the optimality of CES demand is valid even with heterogeneous firms, even though positive profits obtain in equilibrium. Each firm charges a price higher than its average cost and expected profits exactly compensate for ex ante expected losses of non-producers.

These findings crucially depend on CES preferences which are necessary for market optimality. Generalizing to preferences with variable elasticities of substitution, productivity varies with market size, but suboptimally. 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 may miss out on meaningful trade-offs. VES demand offers the potential to explain a wider variety of productivity and welfare changes. Careful work using flexible VES demand offers a method to quantify productivity-welfare relationships present in imperfectly competitive markets to inform policy.

Considering the distortions present when firms charge variable markups also highlights the special role of integration as a policy tool to reduce distortions. Sufficiently large markets hold out the possibility of approaching the monopolistically competitive limit which induces constant markups and therefore (as this paper has shown) an optimal outcome. Even though integration can cause market and social objectives to perfectly align, "How Large is Large?" is an open question. Further empirical work might quantify these relationships and thereby exhibit the scope of integration as a tool to overcome market distortions.

# References

ALESSANDRIA, G. AND H. CHOI (2007): "Do Sunk Costs of Exporting Matter for Net Export Dynamics?" *The Quarterly Journal of Economics*, 122, 289–336.

ARKOLAKIS, C., A. COSTINOT, AND A. RODRIGUEZ-CLARE (forthcoming): "New trade models, same old gains?" *American Economic Review*.

ATKESON, A. AND BURSTEIN (2010): "Innovation, Firm Dynamics, and international Trade," *Journal of political economy*, 118, 433–484.

BAGWELL, K. AND R. W. STAIGER (2009): "Delocation and trade agreements in imperfectly competitive markets," *NBER Working Paper*.

BALDWIN, R. E. AND F. ROBERT-NICOUD (2008): "Trade and growth with heterogeneous firms," *Journal of International Economics*, 74, 21–34.

BAUMOL, W. J. AND D. F. BRADFORD (1970): "Optimal Departures From Marginal Cost Pricing," *The American Economic Review*, 60, 265–283.

BEHRENS, K. AND Y. MURATA (2009): "Trade, Competition, and Efficiency," *Cahier de recherche/Working Paper*, 9, 40.

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

BERGE, C. AND KARREMAN (1963): *Topological spaces, including a treatment of multi-valued functions, vector spaces and convexity*, New York: Macmillan.

BERNARD, A. B., J. B. JENSEN, S. J. REDDING, AND P. K. SCHOTT (2007): "Firms in International Trade," *The Journal of Economic Perspectives*, 21, 105–130.

BILBIIE, F. O., F. GHIRONI, AND M. J. MELITZ (2006): "Monopoly power and endogenous variety in dynamic stochastic general equilibrium: distortions and remedies," *manuscript, University of Oxford, Boston College, and Princeton University*.

CAMPBELL, J. R. AND H. A. HOPENHAYN (2005): "Market Size Matters," *Journal of Industrial Economics*, 53, 1–25.

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

DEMIDOVA, S. AND A. RODRIGUEZ-CLARE (2009): "Trade policy under firm-level heterogeneity in a small economy," *Journal of International Economics*, 78, 100–112.

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

ECONOMIDES, N. (1989): "Symmetric equilibrium existence and optimality in differentiated product markets," *Journal of Economic Theory*, 47, 178–194.

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

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

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

GROSSMAN, G. M. AND E. HELPMAN (1993): *Innovation and Growth in the Global Economy*, MIT Press.

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

HELPMAN, E., O. ITSKHOKI, AND S. J. REDDING (2011): "Trade and Labor Market Outcomes," *NBER Working Paper*.

HOLMES, T. J. AND J. J. STEVENS (2010): "An alternative theory of the plant size distribution with an application to trade," *NBER Working Paper*.

KATAYAMA, H., S. LU, AND J. R. TYBOUT (2009): "Firm-level productivity studies: illusions and a solution," *International Journal of Industrial Organization*, 27, 403–413.

KHAN, M. A. AND Y. SUN (2002): "Non-cooperative games with many players," *Handbook of Game Theory with Economic Applications*, 3, 1761–1808.

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

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

KUHN, K. U. AND X. VIVES (1999): "Excess entry, vertical integration, and welfare," *The Rand Journal of Economics*, 30, 575–603.

MANKIW, N. G. AND M. D. WHINSTON (1986): "Free entry and social inefficiency," *The RAND Journal of Economics*, 48–58.

MELITZ, M. J. (2003a): "The Impact of Trade on Intra-Industry Reallocations and Aggregate Industry Productivity," *Econometrica*, 71, 1695–1725.

——— (2003b): "The Impact of Trade on Intra-Industry Reallocations and Aggregate Industry Productivity," *Econometrica*, 71, 1695–1725.

MELITZ, M. J. AND G. I. P. OTTAVIANO (2008): "Market Size, Trade, and Productivity," *Review of Economic Studies*, 75, 295–316.

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

SOLOW, R. M. (1998): *Monopolistic competition and macroeconomic theory*, Cambridge University Press.

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

STIGLITZ, J. E. (1986): "Towards a more general theory of monopolistic competition," *Prices, competition and equilibrium*, 22.

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

——— (2004b): "Product Substitutability and Productivity Dispersion," *Review of Economics and Statistics*, 86, 534–550.

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

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

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

WEYL, E. G. AND M. FABINGER (2009): "Pass-through as an Economic Tool," *Harvard University, mimeo*.

ZHELOBODKO, E., S. KOKOVIN, M. PARENTI, AND J. F. THISSE (2011): "Monopolistic competition in general equilibrium: Beyond the CES," *PSE Working Papers*.

# A Appendix: Proofs

## A.1 Social welfare

To assess the optimality of market allocations resulting from international trade, we need to clarify the planner'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 planner 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

$$W\left(\{Q_{ji}\}\right) \equiv \int_{k \text{ is a worker}} \min_{i,j} \left\{U(Q_{ji})/\omega_{ji}\right\} dk \tag{9}$$

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 $\omega_{JI}$ 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

$$W = \sum_{I \text{ is a country}} L_I \min_{I,J} \{U(Q_{JI})/\omega_{JI}\} = \min_{I,J} \{U(Q_{JI})/\omega_{JI}\} \cdot \sum_I L_I. \tag{10}$$

From Equation (10), dividing both sides by the world population shows any socially optimal allocation maximizes per capita welfare, using appropriate Pareto weights for each country pairing $(J,I)$.[22] 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 (10).

[22] 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.

## A.2 A Folk Theorem

In this context we need to define the Social Planner's 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 SP is allowed to choose from. A sufficiently rich class for our purposes are $q(c)$ which are positive and continuously differentiable on some closed interval and zero otherwise. This follows from the basic principle that a SP 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 closed Melitz economy is socially optimal.

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

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

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 SP 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 $\forall \lambda$.[23] Now for fixed $(M_e, c_d)$, consider the problem of finding $q^*$ given by

$$\max_{q \in \mathscr{Q}_{[0,c_d]}} V(q) \text{ subject to } H(q) = L/M_e - f_e - fG(c_d). \qquad (11)$$

Following Troutman (1996), if some $q^*$ maximizes $V(q) - \lambda H(q)$ on $\mathscr{Q}_{[0,c_d]}$ for some $\lambda$ and satisfies the constraint then it is a solution to Equation (11). For any $\lambda$, a sufficient condition for some

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

$q^*$ to be a global maximum on $\mathscr{Q}_{[0,c_d]}$ is

$$D_2 v(c,q^*(c)) \quad = \quad \lambda D_2 h(c,q^*(c)). \tag{12}$$

This follows because (12) 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 $\delta$ denotes the Gateaux derivative in the direction of $\xi$) and $q^*$ is a global max since $V(q)-\lambda H(q)$ is strictly concave. The condition (12) is nothing but $\rho q^*(c)^{\rho-1} g(c) = \lambda c g(c)$ which implies $q^*(c) = (\lambda c/\rho)^{1/(\rho-1)}$.[24] 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 $\lambda$ since for fixed $\lambda$ we have

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

so choosing $\lambda$ 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 summary, for each $(M_e, c_d)$ a globally optimal $q^*$ satisfying the resource constraint is

$$q^*(c) \quad = \quad c^{1/(\rho-1)} \left(L/M_e - f_e - fG(c_d)\right) / LR(c_d) \tag{13}$$

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

$$W(M_e, c_d) \equiv M_e L \int_0^{c_d} q^*(c)^\rho g(c) dc = M_e L^{1-\rho} [L/M_e - f_e - fG(c_d)]^\rho R(c_d)^{1-\rho}. \tag{14}$$

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 (14), 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

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

using $R'(c_d) = c_d^{\rho/(\rho-1)} g(c_d)$ yields

$$B'(c_d) = \left\{ c_d^{\rho/(\rho-1)} - R(c_d) f / \left[ f_e + f G(c_d) \right] \right\} / g(c_d) R(c_d). \tag{15}$$

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 (15), $B'(c_d) = 0$ 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.3 Melitz Open Economy

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

*Proof.* Following the discussion of social welfare in the text, 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 $\left\{ \omega_{ji} \right\}$. 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 maximizing

$$W^1 \equiv \min_i \left\{ U(Q_{1i})/\omega_{1i} \right\}. \tag{16}$$

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 $\left\{ Q_{1i}^* \right\}$ 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 (16) 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 $\left\{ Q_{1i}^* \right\}$ we may set $\omega_{1i} \equiv U(Q_{1i}^*)$ so that $\left\{ Q_{1i}^* \right\}$ maximizes Equation (16).

$\mathscr{W}^1$ 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 $C_f\left(\overline{M},\overline{c}\right) \equiv \overline{M}\left(f_e + G(\overline{c})f\right)$ which is incurred from fixed costs at home. Next define "variable" costs at home $C_1\left(Q_{11}\right)$ and abroad $C_j\left(Q_{1j}\right)$ by

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

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

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 $\left\{\mathscr{L}_j\right\}$, with $C_f + \sum \mathscr{L}_j = L$, that constrain $C_j \leq \mathscr{L}_j$. We may then maximize each $U(Q_{1j})$ subject to the constraint $C_j \leq \mathscr{L}_j$ separately and we may assume WLOG that each $\mathscr{L}_j > 0$.[25] 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

$$q_{11}^*(c) = c^{1/(\rho-1)}\mathscr{L}_1 / M_e^{11} L R(c_d^{11}), \tag{17}$$

$$q_{1j}^*(c) = c^{1/(\rho-1)}[\mathscr{L}_j / M_e^{1j} - f_x G(c_d^{1j})] / L R(c_d^{1j})\tau. \tag{18}$$

Having found the optimal quantities of Equations (17-18) 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 continuous correspondence, so by the Theorem of the Maximum $S(\overline{M},\overline{c})$ is continuous on $C_f^{-1}\left([0,L]\right)$ (Berge and Karreman, 1963). Since $C_f$ is continuous, $C_f^{-1}\left([0,L]\right)$ 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 (17-18) yield respectively

$$U(Q_{11}) = R(c_d^{11})^{1-\rho} L^{1-\rho} M_e^{11} \left(\mathscr{L}_1 / M_e^{11}\right)^{\rho}, \tag{19}$$

$$U(Q_{1j}) = R(c_d^{1j})^{1-\rho} L^{1-\rho} M_e^{1j} \left(\mathscr{L}_j / M_e^{1j} - f_x G(c_d^{1j})\right)^{\rho} \tau^{-\rho}. \tag{20}$$

Equation (19) is increasing in both $M_e^{11}$ and $c_d^{11}$ so it follows that at any optimum, $M_e^{11*} = \overline{M}$ and

[25] 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 $> 0$.

$c_d^{11*} = \bar{c}$. Equation (20) 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 $c_d^{1j*} < \bar{c}$ 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 $c_d^{1j*} < \bar{c}$ implies $M_e^{1j*} = \overline{M}$ and $M_e^{1j*} < \overline{M}$ implies $c_d^{1j*} = \bar{c}$. 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 $c_d^{1j*} = \bar{c}$ cannot be optimal. Therefore we conclude that $M_e^{1j*} = \overline{M}$ and $c_d^{1j*} < \bar{c}$. In particular, $c_d^{1j*}$ must solve the implicit equation

$$\rho R(c_d^{1j*})/\left(c_d^{1j*}\right)^{\rho/(\rho-1)} + (1-\rho)G(c_d^{1j*}) = (1-\rho)\mathscr{L}_j/\overline{M}f_x \tag{21}$$

derived from the first order necessary condition.

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

$$\mathscr{W}^1 = \left(\overline{M}L\right)^{1-\rho}\left\{R(\bar{c})^{1-\rho}\mathscr{L}_1^{\rho} + \tau^{-\rho}\sum_{j>1} R(c_d^{1j})^{1-\rho}\left(\mathscr{L}_j - \overline{M}f_x G(c_d^{1j})\right)^{\rho}\right\}. \tag{22}$$

Now consider maximizing $\mathscr{W}^1$ as given in Equation (22) over $\overline{M}, \bar{c}, \mathscr{L}_j, c_d^{1j}$ with $c_d^{1j}$ unconstrained by $\bar{c}$ for $j > 1$. Using a standard Lagrangian approach, the candidate solution from the necessary conditions implies $c_d^{1j*} = (f_x/f)^{(\rho-1)/\rho}\bar{c}/\tau$ and since it is assumed $(f/f_x)^{(1-\rho)/\rho} < \tau$ for trade in a market equilibrium in the Melitz framework, $c_d^{1j*} < \bar{c}$. The candidate solution with $c_d^{1j}$ unconstrained also yields Equation (21) so the unconstrained candidate solution coincides with the solution including the omitted constraints $c_d^{1j*} < \bar{c}$. 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.4 Results Characterizing Market Distortions and Integration

**Lemma.** *Suppose* $(1-\varepsilon(q))' \neq 0$ *and* $\mu'(q) \neq 0$. *At extreme quantities, social and private markups align as follows:*

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

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

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

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

**Lemma.** *If $c_d$ converges to $c_d^{\infty}$ for the market, then profits ($\pi(c)$) and total quantities ($Lq(c)$) are bounded for all $c>0$.*
*Similarly, if $c_d$ converges to $c_d^{\infty}$ for the social optimal, then social profits ($\varpi(c)\equiv(1-\varepsilon(c))/\varepsilon(c)\cdot cq(c)L-f$) and total quantities ($Lq(c)$) are bounded for all $c>0$.*

*Proof.* First note that for any costs $c_L<c_H$, $q(c_H)$ is in the choice set of the firm with costs $c_L$ and therefore

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

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 $\lim_{L\longrightarrow\infty}\int_0^{c_d}\pi(c)dG=F_e$ and $\limsup\pi(\underline{c})=\infty$ would imply $\limsup\int_0^{c_d}\pi(c)dG=\infty$. Substituting $\varpi$ for $\pi$ leads to similar arguments for the social optimum. □

**Proposition.** *Integration will select a few highly productive firms for the market and social optimum under the following conditions:*

1. *If $\mu'\neq 0$ and $\lim_{q\longrightarrow 0}\mu\left(q\right)=0$ then $c_d^{Mkt}\longrightarrow 0$ as $L\longrightarrow\infty$.*
2. *If $(1-\varepsilon)'\neq 0$ and $\lim_{q\longrightarrow 0}1-\varepsilon\left(q\right)=0$ then $c_d^{SP}\longrightarrow 0$ as $L\longrightarrow\infty$.*

*Proof.* Considering the market case, since $\mu(q)\in[0,1]$, clearly $\mu'>0$ which implies $c_d^{\text{Mkt}}$ decreases monotonically in $L$. By above, for each fixed $c>0$, $Lq(c)$ and $\pi(c)$ are bounded in $L$. Since $\pi(c)=L\mu\left(q(c)\right)/\left(1-\mu\left(q(c)\right)\right)\cdot cq(c)$, and $\mu\left(q(c)\right)/\left(1-\mu\left(q(c)\right)\right)\longrightarrow 0$ while $Lq(c)$ is bounded, $\pi(c)\longrightarrow 0$ as $L\longrightarrow\infty$ so $c_d^{\infty}<c$ for each $c>0$, implying $c_d^{\text{Mkt}}\longrightarrow 0$. The case for the social optimum is similar. □

**Lemma.** *Maintain the previous assumptions. If $\lim_{q\longrightarrow 0}\mu(q)\in(0,1)$ and $\mu'\gtrless 0$ then for the market: Each price $p(c)$ converges in $(0,\infty)$ for $c>0$, $Lq(c_d)$ converges in $(0,\infty)$, and*

$$\lim_{L\longrightarrow\infty}u'\circ q(c)/u'\circ q(c_d)=c/c_d.$$

*Similarly, if* $\lim_{q\longrightarrow 0} 1-\varepsilon(q) \in (0,1)$ *and* $(1-\varepsilon)' \lesseqgtr 0$ *then at the social optimum: Each* $u\circ q(c)/\lambda q(c)$ *converges in* $(0,\infty)$ *for* $c>0$*,* $Lq(c_d)$ *converges in* $(0,\infty)$*, and*

$$\lim_{L\longrightarrow\infty} \left(u\circ q(c)/q(c)\right)/\left(u\circ q(c_d)/q(c_d)\right) = c/c_d.$$

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

**Lemma.** *Maintain the previous assumptions. Assume* $\kappa\neq\tilde{\kappa}$ *implies* $\lim_{q\longrightarrow 0} u'(\kappa q)/u'(q) \neq \lim_{q\longrightarrow 0} u'(\tilde{\kappa} q)/u'(q)$*. 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. $q(c)/q(c_d)\geq 1$ has at least one limit point and if it has two limit points, say $a$ and $b$ with $a<b$, we can choose subsequences $\left(q(c)/q(c_d)\right)_{a_n}\longrightarrow a$ and $\left(q(c)/q(c_d)\right)_{b_n}\longrightarrow b$. There also exist distinct $\kappa$ and $\tilde{\kappa}$ in $(a,b)$ so that eventually

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

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, $\lim_{q\longrightarrow 0} u'(\kappa q)/u'(q) > \lim_{q\longrightarrow 0} u'(\tilde{\kappa} q)/u'(q)$ but

$$\lim_{n\longrightarrow\infty}\left(u'\circ q(c)/u'\circ q(c_d)\right)_{a_n} = c/c_d = \left(u'\circ q(c)/u'\circ q(c_d)\right)_{b_n}$$

contradicting $a<b$.

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

$$\kappa\neq\tilde{\kappa} \text{ implies } \lim_{q\longrightarrow 0}\left(u(\kappa q)/\kappa q\right)/\left(u(q)/q\right) \neq \lim_{q\longrightarrow 0}\left(u(\tilde{\kappa} q)/\kappa q\right)/\left(u(q)/q\right). \tag{24}$$

Since $\lim_{q\longrightarrow 0} u'(q)=\infty$ and $\lim_{q\longrightarrow 0}\varepsilon\in(0,\infty)$ it follows that $\lim_{q\longrightarrow 0} u(q)/q=\infty$. By L'Hospital's rule, $\lim_{q\longrightarrow 0}\left(u(\kappa q)/\kappa q\right)/\left(u(q)/q\right) = \lim_{q\longrightarrow 0} u'(\kappa q)/u'(q)$ for all $\kappa$ so the condition (24) in holds because $\kappa\neq\tilde{\kappa}$ implies $\lim_{q\longrightarrow 0} u'(\kappa q)/u'(q) \neq \lim_{q\longrightarrow 0} u'(\tilde{\kappa} q)/u'(q)$. □

**Lemma.** *Maintain the previous assumptions. It is also true that as market size grows large*

- $q(c)/q(c_d) \longrightarrow (c/c_d)^{-1/\alpha}$ *with* $\alpha = \lim_{q \longrightarrow 0} \mu(q)$.
- *The cost cutoffs for the social optimum and market converge to the same value.*
- *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 $\Upsilon$ is weakly decreasing, and the results above show $\Upsilon$ is one to one, so it is strictly decreasing. Define $f_q(z) \equiv u'(zq)/u'(q)$ so $\lim_{q \longrightarrow 0} f_q(z) = \Upsilon^{-1}(z)$ 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 $\lim_{q \longrightarrow 0} \mu(zq) = \mu^\infty \in (0,1)$ and $\lim_{q \longrightarrow 0} u'(zq)/zu'(q) = \Upsilon^{-1}(z)/z$, we know $\lim_{q \longrightarrow 0} f_q'(z) = -\mu^\infty \Upsilon^{-1}(z)/z$. On any strictly positive closed interval $I$, $\mu$ and $u'(zq)/zu'(q)$ are monotone in $z$ so $f_q'(z)$ converges uniformly on $I$ as $q \longrightarrow 0$. It follows (Rudin's Principles, Thm 7.17) that

$$\lim_{q \longrightarrow 0} f_q'(z) = d \lim_{q \longrightarrow 0} f_q(z)/dz = -\mu^\infty \Upsilon^{-1}(z)/z = d\Upsilon^{-1}(z)/dz. \tag{25}$$

We conclude that $\Upsilon^{-1}(z)$ is differentiable and thus continuous, and given the form deduced in (25), $\Upsilon^{-1}(z)$ is continuously differentiable. Since $d\Upsilon^{-1}(z)/dz = 1/\Upsilon' \circ \Upsilon^{-1}(z)$, composing both sides with $\Upsilon(z)$ and using Equation (25) we have $\Upsilon'(z) = -\Upsilon(z)/\mu^\infty z$. Therefore $\Upsilon$ is CES, in particular $\Upsilon(z) = z^{-1/\mu^\infty}$.

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

$$q^{\text{so}}(c)/q^{\text{so}}(c_d^{\text{so}}) \longrightarrow (c/c_\infty^{\text{so}})^{-1/\alpha} \text{ and } q^{\text{mkt}}(c)/q^{\text{mkt}}(c_d^{\text{mkt}}) \longrightarrow (c/c_\infty^{\text{mkt}})^{-1/\alpha}. \tag{26}$$

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

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

It necessarily follows that

$$\lim_{L\longrightarrow\infty} L\int_0^{c_d^{\text{mkt}}} \mu\circ q^{\text{mkt}}(c)/\left(1-\mu\circ q^{\text{mkt}}(c)\right)\cdot cq^{\text{mkt}}(c)dG - fG(c_d^{\text{mkt}}) =$$
$$\lim_{L\longrightarrow\infty} L\int_0^{c_d^{\text{so}}} \left(1-\varepsilon\circ q^{\text{so}}(c)\right)/\varepsilon\circ q^{\text{so}}(c)\cdot cq^{\text{so}}(c)dG - fG(c_d^{\text{so}}). \tag{27}$$

Using Equation (26), we see that $Lq^{\text{so}}(c)$ and $Lq^{\text{mkt}}(c)$ converge uniformly on any strictly positive closed interval. Combined with the fact that $\lim_{q\longrightarrow 0}\mu(q) = \lim_{q\longrightarrow 0} 1-\varepsilon(q)$, we see from Equation (27) the limits of the $\mu/(1-\mu)$ and $(1-\varepsilon)/\varepsilon$ terms are equal and factor out of Equation (27), leaving

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

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

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

so that finally evaluating the limits, we have

$$\int_0^{c_\infty^{\text{mkt}}} \left[(c/c_\infty^{\text{mkt}})^{1-1/\alpha} - 1\right]dG = \int_0^{c_\infty^{\text{so}}} \left[(c/c_\infty^{\text{so}})^{1-1/\alpha} - 1\right]dG. \tag{28}$$

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 $c_\infty^{\text{so}}$, namely $c_\infty^{\text{so}} = c_\infty^{\text{mkt}}$ such that Equation (28) holds. Checking the conditions for $L/M_e$ show they coincide between the market and social optimum as well. □

**Proposition.** *When $(1-\varepsilon)'$ and $\mu'$ have different signs, $q^{Mkt}(c)$ and $q^{SP}(c)$ never cross, specifically:*

1. *If $\mu' > 0 > (1-\varepsilon)'$, market quantities are too high: $q^{Mkt}(c) > q^{SP}(c)$.*
2. *If $\mu' < 0 < (1-\varepsilon)'$, market quantities are too low: $q^{Mkt}(c) < q^{SP}(c)$.*

*In contrast, when $(1-\varepsilon)'$ and $\mu'$ have the same sign, $q^{Mkt}(c)$ and $q^{SP}(c)$ always have a unique crossing at some $c^*$ (perhaps beyond both market and optimal cost cutoffs). Specifically:*

1. *If $\mu' > 0$ and $(1-\varepsilon)' > 0$, $q^{Mkt}(c) > q^{SP}(c)$ for $c < c^*$ and $q^{Mkt}(c) < q^{SP}(c)$ for $c > c^*$.*
2. *If $\mu' < 0$ and $(1-\varepsilon)' < 0$, $q^{Mkt}(c) < q^{SP}(c)$ for $c < c^*$ and $q^{Mkt}(c) > q^{SP}(c)$ for $c > c^*$.*

*Proof.* First, note that the limiting assumptions on $\mu$ and $\varepsilon$ combined with inada conditions on $u$ guarantee that both $q^{\text{Mkt}}(c)$ and $q^{\text{SP}}(c)$ range from 0 to $\infty$ as $c$ varies. We will also use the fact that

$$\left[u''\left(q^{\text{Mkt}}(c)\right)q^{\text{Mkt}}(c) + u'\left(q^{\text{Mkt}}(c)\right)\right]/\delta = c, \qquad u'\left(q^{\text{SP}}(c)\right)/\lambda = c. \tag{29}$$

From Equations (29) we have

$$\left[1-\mu\left(q^{\text{Mkt}}(c)\right)\right]\cdot u'\left(q^{\text{Mkt}}(c)\right)/u'\left(q^{\text{SP}}(c)\right) = \delta/\lambda. \tag{30}$$

Suppose $\mu' > 0 > (1-\varepsilon)'$. From above, this implies $\lim_{q\longrightarrow\infty} 1-\varepsilon(q) \in (0,1)$ and therefore $\lim_{q\longrightarrow\infty} 1-\varepsilon(q) = \lim_{q\longrightarrow\infty}\mu(q)$. For the result, it is sufficient to show that $\inf_q 1-\mu(q) > \delta/\lambda$, since then Equation (30) shows that $u'\left(q^{\text{Mkt}}(c)\right) < u'\left(q^{\text{SP}}(c)\right)$ which implies $q^{\text{Mkt}}(c) > q^{\text{SP}}(c)$. To see that $\inf_q 1-\mu(q) > \delta/\lambda$, first note that

$$\inf_q 1-\mu(q) = \lim_{q\longrightarrow\infty} 1-\mu(q) = \lim_{q\longrightarrow\infty}\varepsilon(q) = \sup_q \varepsilon(q)$$

and also that $\delta = M_e^{\text{Mkt}}\int_0^{c_d^{\text{Mkt}}} u'\left(q^{\text{Mkt}}(c)\right)q^{\text{Mkt}}(c)dG$ and $\lambda = M_e^{\text{SP}}\int_0^{c_d^{\text{SP}}} u\left(q^{\text{SP}}(c)\right)dG$. Thus from Equation (30), $\inf_q 1-\mu(q) = \sup_q\varepsilon(q) > \delta/\lambda$ so long as

$$M_e^{\text{SP}}\int_0^{c_d^{\text{SP}}} u\left(q^{\text{SP}}(c)\right)dG > M_e^{\text{Mkt}}\int_0^{c_d^{\text{Mkt}}} u'\left(q^{\text{Mkt}}(c)\right)q^{\text{Mkt}}(c)dG/\left(\sup_q\varepsilon(q)\right). \tag{31}$$

Equation (31) must in fact hold since

$$\begin{aligned}\delta/\left(\sup_q\varepsilon(q)\right) &= M_e^{\text{Mkt}}\int_0^{c_d^{\text{Mkt}}}\left(\varepsilon\left(q^{\text{Mkt}}(c)\right)/\sup_q\varepsilon(q)\right)u\left(q^{\text{Mkt}}(c)\right)dG \\ &< M_e^{\text{Mkt}}\int_0^{c_d^{\text{Mkt}}} u\left(q^{\text{Mkt}}(c)\right)dG\end{aligned} \tag{32}$$

and Equation (32) must be less than or equal to $M_e^{\text{SP}}\int_0^{c_d^{\text{SP}}} u'\left(q^{\text{SP}}(c)\right)q^{\text{SP}}(c)dG$ because this expression is per capita welfare which is maximized at the social optimium. For $\mu' < 0 < (1-\varepsilon)'$,

a similar argument holds exploiting the relationship

$$\sup_q 1-\mu(q) = \lim_{q\longrightarrow\infty} 1-\mu(q) = \lim_{q\longrightarrow\infty} \varepsilon(q) = \inf_q \varepsilon(q)$$

and using the fact that the market maximizes marginal revenue per capita, $M_e^{\text{Mkt}} \int_0^{c_d^{\text{Mkt}}} u'\left(q^{\text{Mkt}}(c)\right) q^{\text{Mkt}}(c) dG$ as shown above.

Now consider the cases when $\mu'$ and $\varepsilon'$ have different signs, and 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 cases where $\lim_{q\longrightarrow z} \varepsilon(q) = 0$ for $z \in \{0,\infty\}$ require the regularity condition $\lim_{q\longrightarrow z} d\ln u'/d\ln q = 0$.) 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 (30) that for some $c^*$, $\left[1-\mu\left(q^{\text{Mkt}}(c^*)\right)\right] = \delta/\lambda$ and therefore $u'\left(q^{\text{Mkt}}(c^*)\right) = u'\left(q^{\text{SP}}(c^*)\right)$ so $q^{\text{Mkt}}(c^*) = q^{\text{SP}}(c^*)$. Furthermore, $q^{\text{Mkt}}(c)$ is strictly decreasing in $c$ so with $\mu' \neq 0$, $c^*$ is unique. Returning to Equation (30), using the fact that $q^{\text{Mkt}}(c)$ is strictly decreasing in $c$ also shows the relative magnitudes of $q^{\text{Mkt}}(c)$ and $q^{\text{SP}}(c)$ for $c \neq c^*$. □

**Proposition.** *In general, market productivity is too low or high, as follows:*

1. *If* $(1-\varepsilon)' > 0$*, market productivity is too low:* $c_d^{Mkt} > c_d^{SP}$.
2. *If* $(1-\varepsilon)' < 0$*, market productivity tis too high:* $c_d^{Mkt} < c_d^{SP}$.

*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 $\alpha$) defined as:

$$\max_{M_e, c_d, q(c)} LM_e \int_0^{c_d} v_\alpha(q(c))\, dG \text{ subject to } L \geq M_e \left( \int_0^{c_d} Lcq(c) + f dG + F_e \right). \tag{33}$$

Let the Lagrange multiplier associated with each $\alpha$ in Equation (33) be written as $\beta(\alpha)$. By appealing to the envelope theorem twice in succession, differentiating Equation (33) in $L$ 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

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

so cancellation and rearrangement, using the expressions for $\beta$, $d\beta/d\alpha$ 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 then 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 (33) shows $\alpha = 0$ corresponds to the social optimum while $\alpha = 1$ corresponds to the market equilbrium. It follows that when $\varepsilon' < 0$ that $dc_d/d\alpha > 0$ so we have $c_d^{\text{Mkt}} > c_d^{\text{SP}}$ and vice versa for $\varepsilon' > 0$. □

**Proposition.** *In general, the market over or under produces varieties, as follows:*

1. *If $(1-\varepsilon)', \mu' < 0$, the market has too much entry: $M_e^{Mkt} > M_e^{SP}$.*
2. *If $(1-\varepsilon)', \mu' > 0$, the market has too little entry: $M_e^{Mkt} < M_e^{SP}$, provided $\mu'(q)q/\mu \leq 1$.*

*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

$$B'_v(c) = q(c)/\varepsilon_v(q(c)) + c(dq(c)/dc)\left[1 - \varepsilon'_v(q(c))q(c)/\varepsilon_v(q(c))\right]/\varepsilon_v(q(c)). \tag{34}$$

Equation (34) 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 (34) yields

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

Now consider that the social planner problem corresponds to $v(q)=u(q)$ while the market allocation is generated by maximizing $v(q)=u'(q)q$ so that (supressing the $c$ argument to $q$ in integrands)

$$1/M_e^{\mathrm{SP}} - 1/M_e^{\mathrm{Mkt}} = \int_0^{c_d^{\mathrm{SP}}} cq^{\mathrm{SP}}/\varepsilon\left(q^{\mathrm{SP}}\right)dG(c) - \int_0^{c_d^{\mathrm{Mkt}}} cq^{\mathrm{Mkt}}/\left[1-\mu\left(q^{\mathrm{Mkt}}\right)\right]dG(c) \tag{35}$$

and similarly (supressing the $c$ arguments):

$$B_u = cq^{\mathrm{SP}}/\varepsilon\left(q^{\mathrm{SP}}\right), \qquad B_u' = -q^{\mathrm{SP}}/\mu\left(q^{\mathrm{SP}}\right),$$
$$B_{u'q} = cq^{\mathrm{Mkt}}/\left[1-\mu\left(q^{\mathrm{Mkt}}\right)\right], \; B_{u'q}' = -q^{\mathrm{Mkt}}/\left[\mu\left(q^{\mathrm{Mkt}}\right)+\mu'\left(q^{\mathrm{Mkt}}\right)q^{\mathrm{Mkt}}/\left(1-\mu\left(q^{\mathrm{Mkt}}\right)\right)\right].$$

Now assume $\varepsilon' < 0 < \mu'$, so by above $c_d^{\mathrm{Mkt}} > c_d^{\mathrm{SP}}$ and for the result, from Equation (35) it is sufficient to show that $\int_0^{c_d^{\mathrm{SP}}} B_u(c) - B_{u'q}(c)dG(c) \le 0$. From above, there is also a $c^*$ such that $q^{\mathrm{Mkt}}(c) > q^{\mathrm{SP}}(c)$ for $c < c^*$ and $q^{\mathrm{Mkt}}(c) < q^{\mathrm{SP}}(c)$ for $c > c^*$. For $c < c^*$, $B_u(c) - B_{u'q}(c) < 0$ as $q^{\mathrm{Mkt}}(c) > q^{\mathrm{SP}}(c)$ and $\varepsilon' < 0$ implies

$$cq^{\mathrm{Mkt}}/\left[1-\mu\left(q^{\mathrm{Mkt}}\right)\right] > cq^{\mathrm{SP}}/\left[1-\mu\left(q^{\mathrm{SP}}\right)\right] > cq^{\mathrm{SP}}/\varepsilon\left(q^{\mathrm{SP}}\right).$$

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

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

Finally, $\mu'(q)q/\mu \le 1$ implies $q/\mu(q)$ is increasing in $q$. With $q^{\mathrm{Mkt}}(c) < q^{\mathrm{SP}}(c)$ for $c > c^*$, this implies $\left(B_u(c) - B_{u'q}(c)\right)' \le 0$ so $B_u(c) \le B_{u'q}(c)$ for $c > c^*$. Put together with above, $\int_0^{c_d^{\mathrm{SP}}} B_u(c) - B_{u'q}(c)dG(c) \le 0$ giving the result. For the case $\varepsilon' > 0 > \mu'$, the same argument goes through since clearly $\mu'(q)\,q/\mu(q) \le 1$. □

## A.5 VES Utility

**Proposition.** Increases in market size ($L$) change the optimal cost cutoff ($c_a$) as follows: When $(1-\varepsilon(q))' > 0$, the cost cutoff decreases with size. When $(1-\varepsilon(q))' < 0$, the cost cutoff increases with size.

*Proof.* Let the normalized resource constraint $R$ be defined as

$$R \equiv 1 - M_e \left( \int_0^{c_a} cq(c)dG + G(c_a)f/L + f_e/L \right).$$

The social planner maximizes $M_e \int_0^{c_a} u(q(c))dG + \lambda R$ where $\lambda$ is the shadow value of an extra unit of resources. The optimality conditions for the three outcomes of quantity, mass of varieties and cost cutoff determine the optimal allocations along with the resource constraint $R = 0$.

Optimal quantity equates the marginal social benefit to the marginal social cost implying $u'(q(c)) = \lambda c$. The FOC for optimal $M_e$ with the binding resource constraint implies $\int_0^{c_a} u(q(c))dG = \lambda (1-R)/M_e = \lambda/M_e$. The FOC for the optimal cost cutoff shows that the welfare contribution of the marginal variety is equal to its per capita shadow cost, $u(q(c_a)) = \lambda (c_a q(c_a) + f/L)$.

Differentiating the cost cutoff equation wrt to $L$ shows

$$\left(u'(q(c_a)) - \lambda c_a\right)(dq(c_a)/dL) - (c_a q(c_a) + f/L)(d\lambda/dL) - \lambda q(c_a)(dc_a/dL) + \lambda f/L^2 = 0.$$

Substituting for $u'(q(c)) = \lambda c$ and multiplying through by $L/\lambda$, we have

$$(c_a q(c_a) + f/L)(d\ln\lambda/d\ln L) + c_a q(c_a)(d\ln c_a/d\ln L) = f/L. \tag{36}$$

Equation (36) shows $dc_a/dL$ is tied to $d\lambda/dL$. Changes in the cost cutoff depend on how the shadow value of labor changes with market size, namely $d\lambda/dL$. Differentiating the $M_e$ FOC wrt to $L$ and rearranging shows

$$d\ln\lambda/d\ln L = d\ln M_e/d\ln L + LM_e \int_0^{c_a} c\,(dq(c)/dL)\,dG + LM_e\,(c_a q(c_a) + f/L)\,g(c_a)\,(dc_a/dL). \tag{37}$$

The binding resource constraint shows $0 = d(1-R)/dL$ and substituting for Equation (37) implies $d\ln\lambda/d\ln L - M_e(G(c_a)f/L + f_e/L) = 0$. The shadow value of labor rises with market size and the percentage rise in $\lambda$ reflects the better amortization of fixed and sunk costs in bigger markets. Using the expression for $R = 0$, we have $d\ln\lambda/d\ln L = 1 - M_e \int_0^{c_a} cq(c)dG$. Substituting this into

Equation (36) and rearranging gives

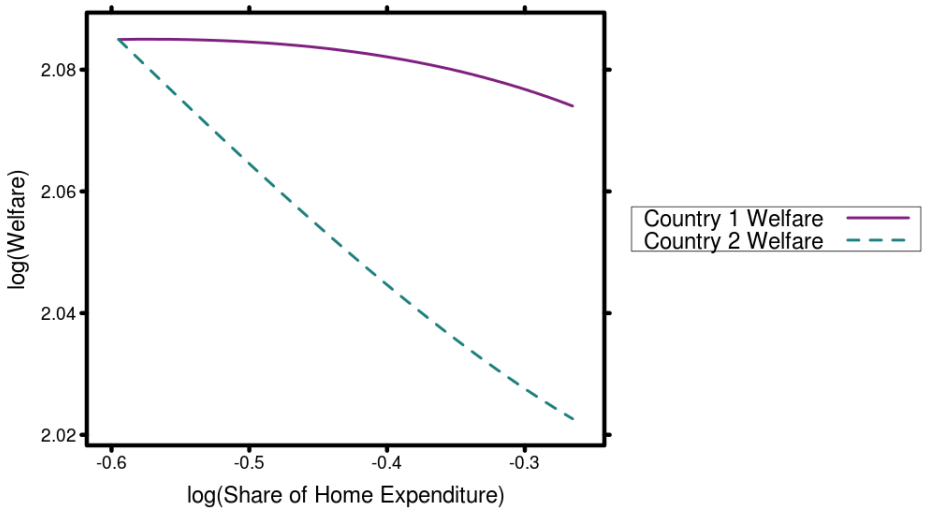
$$\begin{aligned} d\ln c_a/d\ln L &= [c_a q(c_a)/(c_a q(c_a)+f/L)]^{-1}\left[M_e\int_0^{c_a} cq(c)dG - c_a q(c_a)/(c_a q(c_a)+f/L)\right] \\ &= [c_a q(c_a)/(c_a q(c_a)+f/L)]^{-1}\int_0^{c_a} u'(q(c))\cdot q(c)dG/\int_0^{c_a} u(q(c))dG - 1 \end{aligned}$$

where the second line follows from $u'(q(c)) = \lambda c$ and the $M_e$ FOC. Substituting for the elasticity of utility $\varepsilon \equiv qu'(q)/u(q)$, we have Equation (5). □

### A.5.1 VES Specific Utility

The VES demand system implied by $u(q) = aq^{\rho} + bq^{\gamma}$ can generate all four combinations of congestive-addictive preferences as we now briefly discuss. First, note that $\varepsilon'(q) = ab(\rho-\gamma)^2 q^{\rho-\gamma-1}/(aq^{\rho-\gamma}+b)$ and $\mu'(q) = -ab\rho\gamma(\rho-\gamma)^2 q^{\rho-\gamma-1}/(a\rho q^{\rho-\gamma}+b\gamma)$. 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. (forthcoming), as shown in Figure 1 for $u(q) = q^{1/2} + q^{1/4}$.

Figure 1: Welfare and Share of Home Expenditure as Home Tariff Increases



# B Online Appendix

## B.1 VES Market Allocation

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

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

*Proof.* Consider a social planner 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 $LM_e \int_0^{c_d} u'(q(c))\,q(c)dG$, where $\delta$ denotes the Lagrange multiplier:

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

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

## B.2 Trade and Market Size

**Proposition.** In the absence of trade costs, trade between countries with identical VES demand and with 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 mutlipliers are equal in each economy $\delta = \delta^*$. We will show that this allocation constitutes a market equilibrium so opening to trade is equivalent to an increase in market size from $L$ to $L+L^*$.

The $MR = MC$ condition implies a home firm chooses $p(c)[1-\mu(q(c))] = c$ in the home

market and $p_x(c)[1-\mu(q_x(c))] = c$ in the foreign market. A foreign firm chooses $p^*(c)[1-\mu(q^*(c))] = c$ in the foreign market and $p_x^*(c)[1-\mu(q_x^*(c))] = c$ in the home market. When $\delta = \delta^*$, quantity allocations are identical, $q(c) = q_x^*(c) = q^*(c) = q_x(c)$.

This implies cost cutoffs are also the same across countries. The cost cutoff condition for home firms is $(p(c_a) - c_a)\, q(c_a)L + (p_x(c_a) - c_a)\, q_x(c_a)L^* = f$. Substituting for optimal $q^*$ and $q_x^*$ in the analogous foreign cost cutoff condition implies $c_a = c_a^*$. From the resource constraint, this fixes the relationship between entry across countries as $L/M_e = \int_0^{c_a}[cq(c) + cq_x(c) + f]dG + f_e = L^*/M_e^*$.

We show that these allocations can be supported by a terms of trade vector $(e, e^*)$ that is consistent with equal budget multipliers. The home terms of trade $(e)$ determines the total profits received by home firms in home labor units $(\pi + e\pi_x)$. To determine $e$, we use the consumer budget constraint which gives

$$e \equiv \frac{M_e^* \int_0^{c_a} p_x^* q_x^* L dG}{M_e \int p_x q_x L^* dG} = \frac{M_e^* L}{M_e L^*} = 1.$$

Similarly, the implied foreign terms of trade is $e^* = 1$. We conclude that $e = e^* = 1$ supports a market equilibrium with $\delta = \delta^*$. □