The Elusive Pro-Competitive Effects of Trade*

Costas Arkolakis
Yale and NBER

Arnaud Costinot
MIT and NBER

Dave Donaldson
Stanford and NBER

Andrés Rodríguez-Clare
UC Berkeley and NBER

July 2017

Abstract

We study the gains from trade liberalization in models with monopolistic competition, firm-level heterogeneity, and variable markups. For a large class of demand functions used in the international macro and trade literature, we derive a parsimonious generalization of the welfare formula in Arkolakis et al. (2012). We then use both estimates from micro-level trade data and evidence regarding firm-level pass-through to quantify the implications of this new formula. Within the class of models that we consider, our main finding is that gains from trade liberalization predicted by models with variable markups are equal to, at best, and slightly lower than, at worst, those predicted by models with constant markups. In this sense, pro-competitive effects of trade are elusive.

*We thank numerous colleagues, discussants, and seminar participants for helpful comments. Fabian Eckert, Federico Esposito, Brian Greanay, Cory Smith, and Anthony Tokman provided superb research assistance.

1 Introduction

How large are the gains from trade liberalization? Does the fact that trade liberalization affects firm-level markups, as documented in many micro-level studies, make these gains larger or smaller?

There are no simple answers to these questions. On the one hand, gains from trade liberalization may be larger in the presence of variable markups if opening up to trade reduces distortions on the domestic market. In the words of Helpman and Krugman (1989): "The idea that international trade increases competition [...] goes back to Adam Smith, and it has long been one of the reasons that economists give for believing that the gains from trade and the costs from protection are larger than their own models seem to suggest." On the other hand, gains may be smaller if opening up to trade leads foreign firms to increase their markups. Again in the words of Helpman and Krugman (1989): "An occasionally popular argument about tariffs is that they will be largely absorbed through a decline in foreign markups rather than passed onto consumers—the foreigner pays the tariff." If so, when trade costs go down, foreigners get their money back.

The goal of this paper is to characterize and estimate the pro-competitive effects of trade, by which we mean the differential impact of trade liberalization on welfare when markups vary and when they do not. We do so in the context of a new class of gravity models featuring monopolistic competition, firm-level heterogeneity, and variable markups. Our main theoretical contribution is a simple formula that relates the welfare gains from trade liberalization in such environments to three sufficient statistics based on both macro and micro data. To quantify the importance of variable markups, we compare the gains predicted by this formula to those predicted by a gravity model that is also consistent with macro data but ignores micro data and counterfactually restricts markups to be constant across firms. By construction, the difference between these two numbers measures the pro-competitive effects of trade, holding fixed the aggregate responses of trade flows to changes in trade costs.

While our theoretical analysis does not impose any a priori restriction on the magnitude or sign of the pro-competitive effects of trade, our main empirical finding is that gains from trade liberalization predicted by models with variable markups are no greater than those predicted by models with constant markups. Because a decline in trade costs indirectly lowers the residual demand for domestic goods, the former class of models predicts that domestic markups go down after trade liberalization, which reduces distortions and increases welfare. Yet, this indirect effect is (weakly) dominated by the direct effect of a change in trade costs on foreign markups, which leads to (weakly) lower welfare gains from trade liberalization overall. In short, pro-competitive effects of trade are elusive.

The benefit of focusing on gravity models for quantifying the pro-competitive effects of trade is twofold. First, gravity models are very successful empirically and the workhorse models for quantitative work in the field; see e.g. Head and Mayer (2014) and Costinot and Rodriguez-Clare (2014). Second, welfare gains from trade liberalization in gravity models with monopolistic competition, CES utility, and constant markups take a very simple form. Arkolakis et al. (2012), ACR hereafter, have shown that these gains are pinned down by two statistics: (i) the share of expenditure on domestic goods, λ; and (ii) an elasticity of imports with respect to variable trade costs, ε, which we refer to as the trade elasticity. If a small change in variable trade costs raises trade openness in some country, $d \ln \lambda < 0$, then the associated welfare gain is given by

$$d \ln W = -d \ln \lambda / \varepsilon,$$

where $d \ln W$ is the equivalent variation associated with the shock expressed as a percentage of the income of the representative agent. We show that for a general demand system that encompasses prominent alternatives to CES utility and generates variable markups under monopolistic competition, the welfare effect of a small trade shock is given by

$$d \ln W = -\left(1 - \eta\right) d \ln \lambda / \varepsilon,$$

where η is a constant that summarizes the effects of various structural parameters, including the average elasticity of markups with respect to firm productivity. Thus the only endogenous variable that one needs to keep track of for welfare analysis remains the share of expenditure on domestic goods. The net welfare implications of changes in domestic and foreign markups boils down to a single new statistic, η, the sign of which determines whether or not there are pro-competitive effects of trade.

The potential drawback of focusing on gravity models is that the same functional form assumption that gives rise to a constant trade elasticity—namely, the assumption that firm-level productivity follows a Pareto distribution—also gives rise to a constant univariate distribution of markups. This does not imply that pro-competitive effects must be zero, as η could be strictly positive or negative within the class of models that we consider, but this does restrict the channels through which such effects may arise. In our analysis, whether preferences are homothetic or not plays a critical role. When they are, as in Feenstra (2003), we show that the extent to which lower trade costs get (incompletely) passed-through to domestic consumers exactly compensates the extent to which domestic misallocations get alleviated. In this case, $\eta = 0$ and gains from trade are identical to those in ACR. When preferences are non-homothetic, however, common alternatives to CES utility, such as those

considered in Krugman (1979), imply that the first (negative) force dominates the second (positive) force. In this case, $\eta > 0$ and gains from trade liberalization predicted by our new gravity models are strictly lower than those predicted by models with CES utility.

The value of η is ultimately an empirical matter. In the second part of our paper we discuss two simple empirical strategies to estimate η. Our first strategy focuses on a parsimonious generalization of CES utility under which the sign of η depends only on one new demand parameter. Using micro-level U.S. trade data to estimate this alternative demand system, we find that $\eta \simeq 0.06$. Our second strategy draws on a range of existing estimates from the markup and pass-through literatures. This second set of estimates again implies that $\eta \geq 0$ (because in the cross-section productive firms appear to charge lower mark-ups, and in the time-series firms appear to pass-through cost changes incompletely to consumers), with the actual value of η ranging from 0 to 0.14. The robust conclusion that emerges from both of these strategies is that the relevant notion of demand is likely to be in the region where demand elasticities fall with the level of consumption, and hence $\eta \geq 0$. Since $d \ln \lambda$ and ε are the same in the class of gravity models we consider as in gravity models with constant markups, this implies weakly lower gains from trade liberalization, though the quantitative implications of this feature for the gains from trade liberalization, relative to the CES benchmark, are no larger than 14%.

The previous conclusions rely on a number of restrictive assumptions. Within the class of models that we consider, there is one representative agent in each country, all goods are sold in the same monopolistically competitive industry, labor supply is perfectly inelastic and all labor markets are perfectly competitive. Thus, our analysis has little to say about how variable markups may affect the distributional consequences of trade, alleviate misallocations between oligopolistic sectors, or worsen labor market distortions. We come back to some of these general issues in our review of the literature. Our baseline analysis also abstracts from welfare gains from new varieties, because of our focus on small changes in variable trade costs, and from changes in the distribution of markups, because of our focus on Pareto distributions. The final part of our paper explores the sensitivity of our results to these two restrictions through a number of simulations. Although in such environments it is less straightforward to compare models with and without variable markups while holding fixed the aggregate responses of trade flows to changes in trade costs, our simulations provide little support to the idea that the pro-competitive effects of trade in our baseline analysis are special and unusually low.

Our findings are related to, and have implications for, a large number of theoretical and empirical papers in the international trade literature. Many authors have studied the empirical relationship between international trade and firm-level markups; see e.g. Levinsohn (1993), Harrison (1994), Krishna and Mitra (1998), Konings et al. (2001), Chen et al. (2009),

de Loecker and Warzynski (2012), and de Loecker et al. (2016). Methodologies, data sources, and conclusions vary, but a common feature of the aforementioned papers is their exclusive focus on domestic producers. A key message from our analysis is that focusing on domestic producers may provide a misleading picture of the pro-competitive effects of trade. Here we find that a decrease in trade costs reduces the markups of domestic producers. Yet, because it also increases the markups of foreign producers, gains from trade liberalization may actually be lower than those predicted by standard models with CES utility.

A recent empirical paper by Feenstra and Weinstein (2017) is closely related to our analysis. The authors estimate a translog demand system—which is one of the demand systems covered by our analysis—to measure the contribution of new varieties and variable markups on the change in the U.S. consumer price index between 1992 and 2005. Using the fact that markups should be proportional to sales under translog, they conclude that the contribution of these two margins is of the same order of magnitude as the contribution of new varieties estimated by Broda and Weinstein (2006) under the assumption of CES utility. Our theoretical results show that in a class of homothetic demand systems that includes but is not limited to the translog case the overall gains from a hypothetical decline in trade costs are exactly the same as under CES utility.

Despite the apparent similarity between the two previous conclusions, it should be clear that the two exercises are very different. First, Feenstra and Weinstein (2017) is a measurement exercise that uses a translog demand system to infer changes in particular components of the U.S. price index from observed changes in trade flows. The exercise is thus agnostic about the origins of changes in trade flows—whether it is driven by U.S. or foreign shocks—as well as their overall welfare implications. In contrast, our paper is a counterfactual exercise that focuses on the welfare effect of trade liberalization, which we model as a change in variable trade costs. Second, the reason why Feenstra and Weinstein (2017) conclude that the overall gains are the same with translog demand as under CES utility is because the gains from the change in markups that they measure are offset by lower gains from new varieties. In our baseline exercise, the latter effect is absent and negative pro-competitive effects of trade necessarily reflect welfare losses from changes in markups.[1]

Feenstra (2014) comes back to the importance of offsetting effects in an economy where consumers have quadratic mean of order r (QMOR) expenditure functions—a demand system also covered by our analysis—and productivity distributions are bounded Pareto—our baseline analysis assumes that they are Pareto, but unbounded, which guarantees a gravity equation. When analyzing the effect of an increase in country size, he concludes that

[1]The previous observation does not create a contradiction between our results and those in Feenstra and Weinstein (2017). Seen through the lens of our model, their empirical results merely imply that the shocks that lead to changes in markups and varieties must have included more than small changes in trade costs.

changes in markups lead to positive welfare gains, though the overall welfare changes are below those predicted by a model with constant markups. We return to this point when studying the quantitative implications of small changes in trade costs under alternative distributional assumptions in Section 6.3.

The general idea that gains from international trade may be higher or lower in the presence of domestic distortions is an old one in the field; see e.g. Bhagwati (1971). Chief among such distortions are departures from perfect competition. As Helpman and Krugman (1985) note, "Once increasing returns and imperfect competition are introduced, there are both extra sources of potential gains and risks that trade may actually be harmful." This is true both under oligopolistic competition, as in the pioneering work of Brander and Krugman (1983), and under monopolistic competition. A number of recent papers have revisited that idea, either analytically or quantitatively, using variations and extensions of models with firm-level heterogeneity and monopolistic competition, as in Epifani and Gancia (2011), Dhingra and Morrow (2016), and Mrazova and Neary (2016a), Bertrand competition, as in de Blas and Russ (2015) and Holmes et al. (2014), and Cournot competition, as in Edmond et al. (2015). In line with our analysis of small changes in trade costs under monopolistic competition, Edmond et al. (2015) find pro-competitive effects that are close to zero around the observed trade equilibrium, though pro-competitive effects are substantial near autarky.[2]

Our approach differs from these recent papers in three important ways. First, we focus on trade models with variable markups that satisfy the same macro-level restrictions as trade models with constant markups. Besides the empirical appeal of focusing on gravity models, this provides an ideal theoretical benchmark to study how departures from CES utility may affect the welfare gains from trade liberalization. Since the macro-level behavior of new trade models considered in this paper is exactly controlled for, new gains may only reflect new micro-level considerations. Second, we provide a theoretical framework in which the welfare implications of variable markups can be signed and quantified using only one new statistic, η. Hence counterfactual welfare analysis can still be conducted in a parsimonious manner. Third, we develop a new empirical strategy to estimate η and to compute the welfare gains from trade liberalization using micro-level trade data.

The rest of the paper is organized as follows. Section 2 describes our theoretical framework. Section 3 characterizes the trade equilibrium. Section 4 derives our new welfare formula. Section 5 presents our empirical estimates. Section 6 explores the robustness of our results. Section 7 offers some concluding remarks.

[2]Perhaps surprisingly, given this last observation, Edmond et al. (2015) also find that total welfare gains from trade remain well approximated by the ACR formula.

2 Theoretical Framework

Consider a world economy comprising $i = 1, ..., n$ countries, one factor of production, labor, and a continuum of differentiated goods $\omega \in \Omega$. All individuals are perfectly mobile across the production of different goods and are immobile across countries. L_i denotes the population and w_i denotes the wage in country i.

2.1 Consumers

The goal of our paper is to study the implications of trade models with monopolistic competition for the magnitude of the gains from trade in economies in which markups are variable. This requires departing from the assumption of CES utility. Three prominent alternatives in the international trade and international macro literature are: (i) additively separable, but non-CES utility functions, as in the pioneering work of Krugman (1979) and the more recent work of Behrens and Murata (2012), Behrens et al. (2014), Saure (2012), Simonovska (2015), Dhingra and Morrow (2016) and Zhelobodko et al. (2011); (ii) a symmetric translog expenditure function, as in Feenstra (2003), Bergin and Feenstra (2009), Feenstra and Weinstein (2017), Novy (2013), and Rodriguez-Lopez (2011), as well as its strict generalization to quadratic mean of order r (QMOR) expenditure functions, as in Feenstra (2014); and (iii) Kimball preferences, as in Kimball (1995) and Klenow and Willis (2016). In our baseline analysis, we study a general demand system for differentiated goods that encompasses all of them.[3]

All consumers have the same preferences and the same income, y, which derives from their wages and the profits of firms in their country (if any). If a consumer with income y faces a schedule of prices $\boldsymbol{p} \equiv \{p_\omega\}_{\omega \in \Omega}$, her Marshallian demand for any differentiated good ω is

$$q_\omega(\boldsymbol{p}, y) = Q(\boldsymbol{p}, y) D(p_\omega / P(\boldsymbol{p}, y)), \tag{1}$$

where $D(\cdot)$ is a strictly decreasing function and $Q(\boldsymbol{p}, y)$ and $P(\boldsymbol{p}, y)$ are two aggregate demand shifters, which firms will take as given in subsequent sections. Note that whereas $Q(\boldsymbol{p}, y)$ only affects the level of demand, $P(\boldsymbol{p}, y)$ affects both the level and elasticity of demand, which will have implications for firm-level markups. As discussed in Burstein and Gopinath (2014), equation (1) is a common feature of many models in the macroeconomic literature on international pricing.

[3] A trivial generalization of this demand system also nests the case of quadratic, but non-separable utility function, as in Ottaviano et al. (2002) and Melitz and Ottaviano (2008), when a homogenous "outside good" is introduced. We have discussed the additional considerations associated with the existence of an outside good in the June 2012 version of this paper. Details are available upon request.

To complete the description of our demand system, we assume that $Q(\boldsymbol{p}, y)$ and $P(\boldsymbol{p}, y)$ are jointly determined as the solution of the following system of two equations,

$$\int_{\omega\in\Omega} \left[H\left(p_\omega/P\right)\right]^{\beta} \left[p_\omega Q D\left(p_\omega/P\right)\right]^{1-\beta} d\omega = y^{1-\beta}, \tag{2}$$

$$Q^{1-\beta}\left[\int_{\omega\in\Omega} p_\omega Q D\left(p_\omega/P\right) d\omega\right]^{\beta} = y^{\beta}, \tag{3}$$

with $\beta \in \{0,1\}$ and $H(\cdot)$ strictly increasing and strictly concave. As shown in Appendix A.1, utility functions are additively separable if and only if $\beta = 0$, whereas QMOR expenditure functions and Kimball preferences imply $\beta = 1$.[4] When $\beta = 0$, equation (2) reduces to the consumer's budget constraint with $P(\boldsymbol{p}, y)$ equal to the inverse of the Lagrange multiplier associated with that constraint, whereas equation (3) merely implies that $Q(\boldsymbol{p}, y) = 1$. When $\beta = 1$, $P(\boldsymbol{p}, y)$ remains determined by equation (2), which becomes $\int_{\omega\in\Omega} H\left(p_\omega/P\right) d\omega = 1$, but the consumer's budget constraint is now captured by equation (3) with $Q(\boldsymbol{p}, y)$ set such that budget balance holds.

Three properties of the general demand system introduced above are worth emphasizing. First, the own-price elasticity $\partial \ln D\left(p_\omega/P(\boldsymbol{p}, y)\right)/\partial \ln p_\omega$ is allowed to vary with prices, which will generate variable markups under monopolistic competition. Second, other prices only affect the demand for good ω through their effect on the aggregate demand shifters, $Q(\boldsymbol{p}, y)$ and $P(\boldsymbol{p}, y)$.[5] Third, the demand parameter β controls whether preferences are homothetic or not. If $\beta = 1$, equations (2) and (3) imply that $P(\boldsymbol{p}, y)$ is independent of y and that $Q(\boldsymbol{p}, y)$ is proportional to y. Thus preferences are homothetic. Conversely, if $\beta = 0$, preferences are non-homothetic unless $D(\cdot)$ is iso-elastic, i.e. utility functions are CES.[6] The

[4] We do not know whether there are other primitive assumptions that satisfy equations (1)-(3). We note, however, that a slight generalization of equations (1)-(3) would also encompass the case of additively separable indirect utility functions, as in Bertoletti et al. (2017). Specifically, we could leave equation (1) unchanged and generalize equations (2) and (3) to

$$\int [H(p(\omega)/P)]^{(1-\alpha)\beta} [p_\omega Q D(p_\omega/P)]^{1-\beta+\alpha} = y^{1+\alpha-\beta},$$

$$P^{\alpha} Q^{(1-\alpha)(1-\beta)} [\int p_\omega Q D(p_\omega/P) d\omega]^{\beta-\alpha} = y^{\beta},$$

Our baseline analysis corresponds to $\alpha = 0$ and $\beta \in \{0,1\}$, whereas the case of additively separable indirect utility functions corresponds to $\alpha = 1$ and $\beta = 1$. Since the analysis of Section 3 does not depend on equations (2) and (3), such a generalization would leave the structure of the trade equilibrium unchanged. We briefly discuss how it would affect our welfare formula in Section 4.

[5] In this regard, our specification is more restrictive than the Almost Ideal Demand System (AIDS) of Deaton and Muellbauer (1980). Compared to AIDS, however, our specification does not impose any functional form restriction on $Q(\boldsymbol{p}, y)$ and $P(\boldsymbol{p}, y)$.

[6] The formal argument can be found in Appendix A.1. Intuitively, CES utility functions correspond to the knife-edge case in which β admits multiple values. CES utility functions can be thought either as a special case of additively separable utility functions—and derived under the assumption $\beta = 0$—or as a special case of

parameter β will influence the magnitude of general equilibrium effects and play a crucial role in our welfare analysis.

Compared to most papers in the existing trade literature, either theoretical or empirical, we do not impose any functional form restriction on $D(\cdot)$. The only restriction that we impose on $D(\cdot)$ in our theoretical analysis is that it features a choke price.[7]

A1. [Choke Price] *There exists $a \in \mathbb{R}$ such that for all $x \geq a$, $D(x) = 0$.*

Without loss of generality, we normalize a to one in the rest of our analysis so that the aggregate demand shifter $P(\boldsymbol{p}, y)$ is also equal to the choke price. In the absence of fixed costs of accessing domestic and foreign markets—which is the situation that we will focus on—Assumption A1 implies that the creation and destruction of "cut-off" goods have no first-order effects on welfare at the margin. Indeed, if there was some benefit from consuming these goods, they would have been consumed in strictly positive amounts.

Assumption A1 provides an instructive polar case. In models with CES utility, such as those studied in Arkolakis et al. (2012), there are welfare gains from new "cut-off" goods, but markups are fixed. In contrast, firm-level markups can vary in our baseline analysis but there are no welfare gains from new "cut-off" goods. While Assumption A1 rules out CES utility, it is also worth pointing out that A1 does not impose any restriction on the magnitude of the choke price. As it becomes arbitrarily large, one might expect the economies that we consider to start behaving like economies without a choke price. The demand system that we consider in our empirical analysis provides one such example.

For future derivations, it is convenient to write the demand function in a way that makes explicit the symmetry across goods as well as the way in which the aggregate demand shifters, $Q(\boldsymbol{p}, y)$ and $P(\boldsymbol{p}, y)$, affect the demand for all goods. Thus, we write $q_\omega(\boldsymbol{p}, y) \equiv q(p_\omega, Q(\boldsymbol{p}, y), P(\boldsymbol{p}, y))$, with

$$q(p_\omega, Q, P) = QD\left(p_\omega / P\right). \tag{4}$$

2.2 Firms

Firms compete under monopolistic competition. Entry may be restricted or free. Under restricted entry, there is an exogenous measure of firms, $\bar{N}_i$, with the right to produce in each country i. Under free entry, there is a large number of ex ante identical firms that have

QMOR expenditure functions or Kimball preferences—and derived under the assumption $\beta = 1$.

[7] Throughout our welfare analysis, we also implicitly restrict ourselves to cases where there exist preferences that rationalize the Marshallian demand function described by equations (1)-(3). Since such a function necessarily satisfies homogeneity of degree zero and Walras' law, this is equivalent to restricting the Slutsky matrix to be symmetric and negative semidefinite. When $\beta = 0$, the assumption that $D(\cdot)$ is decreasing is sufficient for the previous restriction to hold.

the option of hiring $F_i > 0$ units of labor to enter the industry. Firms then endogenously enter up to the point at which aggregate profits net of the fixed entry costs, $w_i F_i$, are zero. We let N_i denote the measure of firms in country i.

Upon entry, production of any differentiated good is subject to constant returns to scale. For a firm with productivity z in country i, the constant cost of delivering one unit of the variety associated with that firm to country j is given by $w_i \tau_{ij}/z$, where $\tau_{ij} \geq 1$ is an iceberg trade cost. We assume that only international trade is subject to frictions, $\tau_{ii} = 1$. As mentioned earlier, there are no fixed costs of accessing domestic and foreign markets. Thus, the selection of firms across markets is driven entirely by the existence of a choke price, as in Melitz and Ottaviano (2008). Throughout our analysis, we assume that good markets are perfectly segmented across countries and that parallel trade is prohibited so that firms charge the optimal monopoly price in each market.

As in Melitz (2003), firm-level productivity z is the realization of a random variable drawn independently across firms from a distribution G_i. We assume that G_i is Pareto with the same shape parameter $\theta > 0$ around the world.

A2. [Pareto] *For all* $z \geq b_i$, $G_i(z) = 1 - (b_i/z)^{\theta}$, *with* $\theta > 0$.

While by far the most common distributional assumption in models of monopolistic competition with firm-level heterogeneity—even when utility functions are not CES, see e.g. Melitz and Ottaviano (2008), Behrens et al. (2014), Simonovska (2015), and Rodriguez-Lopez (2011)—Assumption A2 is obviously a strong restriction on the supply-side of our economy. So it is worth pausing to discuss its main implications.

As we will demonstrate below, the main benefit of Assumption A2 is that trade flows will satisfy the same gravity equation as in models with CES utility. This will allow us to calibrate our model and conduct counterfactual analysis in the exact same way as in ACR. Accordingly, we will be able to ask and answer the following question: Conditional on being consistent with the same macro data, do models featuring variable markups predict different welfare gains from trade liberalization? In our view, this is a theoretically clean way to compare the welfare predictions of different trade models.

Given the generality of the demand system considered in Section 2.1, it should be clear that Assumption A2 is no less appealing on empirical grounds than under the assumption of CES utility. As documented by Axtell (2001) and Eaton et al. (2011), among others, Pareto distributions provide a reasonable approximation for the right tail of the observed distribution of firm sales. Since Pareto distributions of firm sales can be generated from a model of monopolistic competition with CES utility and Pareto distributions of firm-level productivity, the previous facts are often given as evidence in favor of Assumption A2. Although demand functions derived from CES utility do not satisfy A1, one can construct general-

izations of CES demands that satisfy A1, behave like CES demands for the right tail of the distribution of firm sales, and provide a better fit for the left tail. We come back to this point in our empirical application.

Perhaps the main concern regarding Assumption A2 is that it may be too much of a straight jacket, i.e., that we may be assuming through functional form assumptions whether gains from trade liberalization predicted by models with variable markups will be larger, smaller, or the same. As Proposition 1 will formally demonstrate, this is not so. Although Assumption A2 has strong implications for the univariate distribution of firm-level markups—as we will see, it is unaffected by changes in trade costs—this knife-edge feature does not preclude the existence of variable markups to increase or decrease—in theory—the welfare gains from trade liberalization. As we discuss in Section 4.2, what matters for welfare is not the univariate distribution of markups, but the bivariate distribution of markups and employment, which is free to vary in our model. In Section 6, we further discuss the sensitivity of our results to departures from Assumption A2.

3 Trade Equilibrium

In this section we characterize the trade equilibrium for arbitrary values of trade costs. We proceed in two steps. We first study how the demand system introduced in Section 2 shapes firm-level variables. We then describe how firm-level decisions aggregate up to determine bilateral trade flows and the measure of firms active in each market.

3.1 Firm-level Variables

Consider the optimization problem of a firm producing good ω in country i and selling it in a certain destination j. To simplify notation, and without risk of confusion, we drop indices for now and denote by $c \equiv w_i \tau_{ij}/z$ the constant marginal cost of serving the market for a particular firm and by Q and P the two aggregate shifters of demand in the destination country, respectively. Under monopolistic competition with segmented good markets, the firm chooses its market-specific price p in order to maximize profits in each market,

$$\pi(c, Q, P) = \max_p \{(p - c)\, q(p, Q, P)\},$$

taking Q and P as given. The associated first-order condition is

$$(p - c)/p = -1/(\partial \ln q(p, Q, P)/\partial \ln p),$$

which states that monopoly markups are inversely related to the elasticity of demand.

Firm-level markups. We use $m \equiv p/c$ as our measure of firm-level markups. Combining the previous expression with equation (4), we can express m as the implicit solution of

$$m = \varepsilon_D(m/v)/(\varepsilon_D(m/v) - 1), \tag{5}$$

where $\varepsilon_D(x) \equiv -\partial \ln D(x)/\partial \ln x$ measures the elasticity of demand and $v \equiv P/c$ can be thought of as a market-specific measure of the efficiency of the firm relative to other firms participating in that market, as summarized by P. Equation (5) implies that the aggregate demand shifter P is a sufficient statistic for all indirect effects that may lead a firm to change its price in a particular market.

We assume that for any $v > 0$, there exists a unique $m \equiv \mu(v)$ that solves equation (5). Assuming that $\varepsilon_D' > 0$ is a sufficient, but not necessary condition for existence and uniqueness. The properties of the markup function $\mu(v)$ derive from the properties of $D(\cdot)$. Since $\lim_{x\to 1} D(x) = 0$ by Assumption A1, we must also have $\lim_{x\to 1} \varepsilon_D(x) = \infty$, which implies $\mu(1) = 1$. Thus, the choke price in a market is equal to the marginal cost of the least efficient firm active in that market. Whether markups are monotonically increasing in productivity depends on the monotonicity of ε_D. As is well-known and demonstrated in Appendix A.2, more efficient firms charge higher markups, $\mu' > 0$, if and only if demand functions are log-concave in log-prices, $\varepsilon_D' > 0$.[8]

Firm-level sales and profits. In any given market, the price charged by a firm with marginal cost c and relative efficiency v is given by $p(c, v) = c\mu(v)$. Given this pricing rule, the total sales faced by a firm with marginal cost c and relative efficiency v in a market with aggregate demand shifter Q and population L, are equal to

$$x(c, v, Q, L) \equiv LQc\mu(v)D(\mu(v)/v). \tag{6}$$

In turn, the profits of a firm with marginal cost c and relative efficiency v selling in a market with aggregate shifter Q and population L are given by

$$\pi(c, v, Q, L) \equiv ((\mu(v) - 1)/\mu(v))\, x(c, v, Q, L). \tag{7}$$

The relationship between profits and sales is the same as in models of monopolistic compe-

[8] Mrazova and Neary (2016b) refer to the condition $\varepsilon_D' > 0$ as the "subconvexity" of the (inverse) demand function. Although Assumption A1 requires demand functions to be log-concave in log-prices locally around the choke price, we wish to emphasize that it does not require them to be log-concave in log-prices away from that neighborhood. Accordingly, our theoretical analysis encompasses environments where, on average, the elasticity of the markup is negative. As we will demonstrate in Section 4, such environments may have very distinct implications for the welfare gains from trade liberalization.

tition with CES utility, except that markups are now allowed to vary across firms.

3.2 Aggregate Variables

Aggregate sales, profits, and income. Let X_{ij} denote the total sales by firms from country i in country j. Only firms with marginal cost $c \leq P_j$ sell in country j. Thus there exists a productivity cut-off $z^*_{ij} \equiv w_i \tau_{ij} / P_j$ such that a firm from country i sells in country j if and only if its productivity $z \geq z^*_{ij}$. Accordingly, we can express the bilateral trade flows between the two countries as

$$X_{ij} = N_i \int_{z^*_{ij}}^{\infty} x(w_i \tau_{ij}/z, z/z^*_{ij}, Q_j, L_j) dG_i(z).$$

Combining this expression with equation (6) and using our Pareto assumption A2, we get, after simplifications,

$$X_{ij} = \chi N_i b_i^{\theta} \left(w_i \tau_{ij}\right)^{-\theta} L_j Q_j \left(P_j\right)^{1+\theta}. \tag{8}$$

where $\chi \equiv \theta \int_1^{\infty} \left(\mu(v)/v\right) D(\mu(v)/v) v^{-\theta-1} dv > 0$ is a constant that affects overall sales.[9]

Let Π_{ij} denote aggregate profits by firms from country i in country j gross of fixed entry costs. This is given by

$$\Pi_{ij} = N_i \int_{z^*_{ij}}^{\infty} \pi(w_i \tau_{ij}/z, z/z^*_{ij}, Q_j, L_j) dG_i(z).$$

Using equations (6) and (7), and again invoking Assumption A2, we get

$$\Pi_{ij} = \pi N_i b_i^{\theta} \left(w_i \tau_{ij}\right)^{-\theta} L_j Q_j \left(P_j\right)^{1+\theta}, \tag{9}$$

where $\pi \equiv \theta \int_1^{\infty} \left(\mu(v) - 1\right) D(\mu(v)/v) v^{-\theta-2} dv > 0$ is a constant that affects overall profits. For future reference, note that Equations (8) and (9) imply that aggregate profits are a constant share of aggregate sales,

$$\Pi_{ij} = \zeta X_{ij}, \tag{10}$$

where $\zeta \equiv (\pi/\chi) \in (0,1)$. Finally, let $Y_j \equiv y_j L_j$ denote aggregate income country j. It is equal to the sum of wages and profits, which must add up to the total sales of firms from

[9] Equation (8) implicitly assumes that the lower-bound of the Pareto distribution b_i is small enough so that the firm with minimum productivity b_i always prefers to stay out of the market, $b_i < z^*_{ij}$. This implies that the "extensive" margin of trade is active for all country pairs, which is the empirically relevant case. It also implicitly assumes that the behavior of the distribution of firm-level productivity and demand in the upper-tail is such that χ is finite. Given specific functional form assumptions on D, the associated restrictions on θ can be made explicit; see e.g. Feenstra (2014) for the case of QMOR expenditure functions.

country j,

$$Y_j = \sum_i X_{ji}. \tag{11}$$

Measures of firms and wages. The measure of firms in each country is such that

$$N_i = \begin{cases} \bar{N}_i, & \text{if entry is restricted,} \\ \sum_j \Pi_{ij}/(w_i F_i), & \text{if entry is free.} \end{cases}$$

Wages are such that labor supply equals labor demand,

$$w_i L_i = \begin{cases} \sum_j X_{ij} - \sum_j \Pi_{ij}, & \text{if entry is restricted,} \\ \sum_j X_{ij} - \sum_j \Pi_{ij} + w_i F_i N_i, & \text{if entry is free.} \end{cases}$$

Together with equation (10), the two previous expressions imply

$$N_i = \begin{cases} \bar{N}_i, & \text{if entry is restricted,} \\ \zeta(L_i/F_i), & \text{if entry is free;} \end{cases} \tag{12}$$

$$w_i L_i = \begin{cases} (1-\zeta)(\sum_j X_{ij}), & \text{if entry is restricted,} \\ \sum_j X_{ij}, & \text{if entry is free.} \end{cases} \tag{13}$$

Regardless of whether entry is free or restricted, equations (12) and (13) imply that the measure of firms N_i is invariant to changes in trade costs and that the total wage bill, $w_i L_i$, is proportional to total sales, $\sum_j X_{ij}$.

Summary. A trade equilibrium corresponds to price schedules, $(\boldsymbol{p}_1, ..., \boldsymbol{p}_n)$, measures of firms, $(N_1, ..., N_n)$, and wages, $(w_1, ..., w_n)$, such that (i) prices set in country j by firms with productivity z located in country i maximize their profits:

$$p_{ij}(z) = (w_i \tau_{ij}/z)\mu\left(P_j z/w_i \tau_{ij}\right) \tag{14}$$

if $z \geq w_i \tau_{ij}/P_j$ and $p_{ij}(z) \geq w_i \tau_{ij}/z$ otherwise; (ii) measures of entrants are given by equation (12); and (iii) wages are consistent with labor market clearing, equation (13), with aggregate demand shifters, Q_j and P_j, determined by equations (2) and (3), aggregate sales X_{ij} determined by equation (8), and aggregate income Y_j determined by equation (11). Note that under the previous equilibrium conditions, trade is necessarily balanced: since the Marshallian demand in Section 2.1 must satisfy the budget constraint of the representative con-

sumer, $y_j = \sum_i X_{ij}/L_j$, equation (11) immediately implies

$$\sum_i X_{ji} = \sum_i X_{ij}. \tag{15}$$

3.3 Discussion

In spite of the fact that the pricing behavior of firms, as summarized by equation (14), is very different in the present environment than in trade models with CES utility, bilateral trade flows still satisfy a gravity equation. Indeed, by equations (8), (11), and (15), we have

$$X_{ij} = \frac{N_i b_i^\theta \left(w_i \tau_{ij}\right)^{-\theta} Y_j}{\sum_k N_k b_k^\theta \left(w_k \tau_{kj}\right)^{-\theta}}. \tag{16}$$

Together, equations (10), (15), and (16) imply that the macro-level restrictions imposed in ACR still hold in this environment. As shown in Appendix A.3, it follows that once calibrated to match the trade elasticity θ and the observed trade flows $\{X_{ij}\}$, the models with variable markups considered in this paper must have the same macro-level predictions, i.e., the same counterfactual predictions about wages and bilateral trade flows in response to changes in variable trade costs, as gravity models with CES utility, such as Krugman (1980), Eaton and Kortum (2002), Anderson and Van Wincoop (2003), and Eaton et al. (2011). Yet, as we will see, differences in the behavior of firms at the micro-level open up the possibility of new welfare implications.[10]

Before we turn to our welfare analysis, it is worth emphasizing again that there will be no gains from new varieties associated with small changes in trade costs in the present environment. Such gains must derive from either a change in the measure of entrants, N_i, or from changes in the productivity cut-offs, z_{ij}^*. Here, aggregate profits are a constant share of aggregate revenues, which rules out the former changes, and there are no fixed costs of accessing domestic and foreign markets, which rules out welfare effects from the latter changes. Thus our focus in this paper is squarely on the welfare implications of variable markups at the firm-level.

[10] Whereas the positive predictions of ACR for wages and trade flows only depend on three macro-level restrictions, R1, R2, and R3', their normative predictions also rely on restrictions about preferences, technology, and market structure, including the assumption of CES utility. This is the critical assumption that we have relaxed in this paper.

4 Welfare Analysis

In this section we explore the pro-competitive effects of trade, or lack thereof, in the economic environment described in Sections 2 and 3. We focus on a small change in trade costs from $\tau \equiv \{\tau_{ij}\}$ to $\tau' \equiv \{\tau_{ij} + d\tau_{ij}\}$. ACR show that under monopolistic competition with Pareto distributions of firm-level productivity and CES utility, the equivalent variation associated with such a change—namely, the percentage change in income that would be equivalent to the change in trade costs in terms of its welfare impact—is given by

$$d \ln W_j = -d \ln \lambda_{jj} / \theta,$$

where, like in the present paper, θ is the shape parameter of the Pareto distribution and $d \ln \lambda_{jj}$ is the change in the share of domestic expenditure on domestic goods caused by the change from τ to τ'. Since $\theta > 0$, the equivalent variation $d \ln W_j$ is positive if a change in trade costs leads to more trade, $d \ln \lambda_{jj} < 0$. We now investigate how going from CES utility to the demand system described in equation (1) affects the above formula.

4.1 A New Formula

Without loss of generality, we use labor in country j as our numeraire so that $w_j = 1$ before and after the change in trade costs. Under both restricted and free entry, income per capita in country j is proportional to the wage w_j. Thus, the percentage change in income, $d \ln W_j$, equivalent to the change in trade costs from τ to τ' can be computed as the negative of the percentage change in the expenditure function, $d \ln e_j$, of a representative consumer in country j. This is what we focus on next.[11]

By Shephard's lemma, we know that $de_j / dp_{\omega,j} = q(p_{\omega,j}, Q_j, P_j) \equiv q_{\omega,j}$ for all $\omega \in \Omega$. Since all price changes associated with a move from τ to τ' are infinitesimal,[12] we can express

[11] Since we have not restricted preferences to be (quasi-) homothetic, it should be clear that the assumption of a representative agent in each country is stronger than usual. In Section 2, we have not only assumed that all individuals share the same preferences, but also that they have the same endowments, as in Krugman (1979). Absent this assumption, the aggregate welfare gains from trade liberalization could still be computed by summing up equivalent variations across individuals, or more generally, by specifying a social welfare function; Galle et al. (2014) and Antras et al. (2016) provide an example of such an approach. Given our interest in variable markups rather than the distributional consequences of trade liberalization, however, we view the economies with representative agents that we consider as a useful benchmark.

[12] In principle, price changes may not be infinitesimal because of the creation of "new" goods or the destruction of "old" ones. This may happen for two reasons: (i) a change in the number of entrants N or (ii) a change in the productivity cut-off z^*. Since the number of entrants is independent of trade costs, as argued above, (i) is never an issue. Since the price of goods at the productivity cut-off is equal to the choke price, (ii) is never an issue either. This would not be true under CES utility functions and fixed exporting costs. In this case, changes in productivity cut-offs are associated with non-infinitesimal changes in prices since goods at the margin go from a finite (selling) price to an (infinite) reservation price, or vice versa. We come back to this point in detail

the associated change in expenditure as

$$de_j = \sum_i \int_{\omega \in \Omega_{ij}} q_{\omega,j} dp_{\omega,j} d\omega,$$

where Ω_{ij} is the set of goods produced in country i and exported to country j and $dp_{\omega,j}$ is the change in the price of good ω in country j caused by the move from τ to τ'. The previous expression can be rearranged in logs as

$$d \ln e_j = \sum_i \int_{\omega \in \Omega_{ij}} \lambda_{\omega,j} d \ln p_{\omega,j} d\omega, \tag{17}$$

where $\lambda_{\omega,j} \equiv p_{\omega,j} q_{\omega,j} / e_j$ is the share of expenditure on good ω in country j in the initial equilibrium. Using equation (14) and the fact that firms from country i only sell in country j if $z \geq z_{ij}^*$, we obtain

$$d \ln e_j = \sum_i \int_{z_{ij}^*}^{\infty} \lambda_{ij}(z) \left(d \ln c_{ij} + d \ln m_{ij}(z) \right) dG_i(z), \tag{18}$$

where

$$\lambda_{ij}(z) \equiv \frac{N_i x(w_i \tau_{ij}/z, z/z_{ij}^*, Q_j, L_j)}{\sum_k \int_{z_{kj}^*}^{\infty} N_k x(w_k \tau_{kj}/z, z/z_{kj}^*, Q_j, L_j) dG_k(z)}$$

denotes the share of expenditure in country j on goods produced by firms from country i with productivity z, $c_{ij} \equiv w_i \tau_{ij}$, and $m_{ij}(z) \equiv \mu(z/z_{ij}^*)$. Equation (18) states that the percentage change in expenditure is equal to a weighted sum of the percentage change in prices, with the percentage changes in prices themselves being the sum of the percentage change in marginal costs, $d \ln c_{ij}$, and markups, $d \ln m_{ij}(z)$.

Let $\lambda_{ij} \equiv X_{ij}/E_j$ denote the total share of expenditure on goods from country i in country j and let $\rho_{ij} \equiv \int_{z_{ij}^*}^{\infty} \rho(z/z_{ij}^*) \frac{\lambda_{ij}(z)}{\lambda_{ij}} dG_i(z) dz$ denote the weighted average of the markup elasticities, $\rho(v) \equiv d \ln \mu(v) / d \ln v$**.** Using this notation, we can simplify equation (18) into

$$d \ln e_j = \sum_i \lambda_{ij} \left(d \ln c_{ij} - \rho_{ij} d \ln z_{ij}^* \right). \tag{19}$$

Using Assumption A2, as well as the definition of $\lambda_{ij}(z)$, one can show that the markup elasticity, like the trade elasticity, must be common across countries (i.e., $\rho_{ij} = \rho$ for all i, j)

in Section 6.

and given by the constant

$$\rho \equiv \int_1^{\infty} \frac{d \ln \mu(v)}{d \ln v} \frac{(\mu(v)/v) D(\mu(v)/v) v^{-\theta-1}}{\int_1^{\infty} (\mu(v')/v') D(\mu(v')/v') (v')^{-\theta-1} dv'} dv. \tag{20}$$

Finally, using the fact that the productivity cut-off satisfies $z_{ij}^* = c_{ij}/P_j$, we can rearrange the expression above as

$$d \ln e_j = \underbrace{\sum_i \lambda_{ij} d \ln c_{ij}}_{\text{Change in marginal costs}} + \underbrace{(-\rho) \sum_i \lambda_{ij} d \ln c_{ij}}_{\text{Direct markup effect}} + \underbrace{\rho d \ln P_j}_{\text{Indirect markup effect}}. \tag{21}$$

To fix ideas, consider a "good" trade shock, $\sum_i \lambda_{ij} d \ln c_{ij} < 0$. If markups were constant, $\rho = 0$, the only effect of such a shock would be given by the first term on the RHS of (21). Here, the fact that firms adjust their markups in response to a trade shock leads to two additional terms. The second term on the RHS of (21) is a direct effect. Ceteris paribus, a decrease in trade costs makes exporting firms relatively more productive, which leads to changes in markups, by equation (5). If $\rho > 0$, we see that the direct effect of markups tends to *lower* gains from trade liberalization. The reason is simple. There is incomplete pass-through of changes in marginal costs from foreign exporters to domestic consumers. Firms that become more productive because of lower trade costs tend to raise their markups ($\rho > 0$), leading to lower welfare gains ($-\rho \sum_i \lambda_{ij} d \ln c_{ij} > 0$). The third term on the RHS of (21) is an indirect effect. It captures the change in markups caused by changes in the aggregate demand shifter, P_j. If trade liberalization leads to a decline in P_j, reflecting a more intense level of competition, then $\rho > 0$ implies a decline in domestic and foreign markups and *higher* gains from trade liberalization. If $\rho < 0$, the sign of the direct and indirect markup effects are reversed.

Based on the previous discussion, whether or not there are pro-competitive effects of trade liberalization, in the sense of larger welfare gains than in models with constant markups, depends on a horse race between the direct and indirect markup effects. In order to compare these two effects, we need to compare the change in marginal costs, $\sum_i \lambda_{ij} d \ln c_{ij}$, to the change in the aggregate demand shifter, $d \ln P_j$. We can do so by using equations (2) and (3). Depending on whether entry is restricted or free, income per capita is either equal to w_j or $w_j/(1-\zeta)$. Given our choice of numeraire and Assumption A2, we therefore have

$$\kappa Q_j^{1-\beta} P_j^{\theta+1-\beta} \left(\sum_i N_i b_i^{\theta} c_{ij}^{-\theta} \right) = (1-\zeta)^{(1-\phi)(\beta-1)}, \tag{22}$$

$$\chi^{\beta} Q_j P_j^{\beta(1+\theta)} \left(\sum_i N_i b_i^{\theta} c_{ij}^{-\theta} \right)^{\beta} = (1-\zeta)^{-(1-\phi)\beta}, \tag{23}$$

with $\kappa \equiv \theta \int_1^\infty \left[H\left(\mu(v)/v\right)\right]^\beta \left[\left(\mu\left(v\right)/v\right) D(\mu(v)/v)\right]^{1-\beta} v^{-1-\theta} dv$ and ϕ is a dummy variable equal to 1 if entry is free and zero if it is restricted. For $\beta \in \{0,1\}$, equations (22) and (23) imply $P_j = \left(\kappa(1-\zeta)^{(\phi-1)(\beta-1)} \sum_i N_i b_i^\theta c_{ij}^{-\theta}\right)^{-1/(\theta+1-\beta)}$. Taking logs and totally differentiating, we therefore have

$$d \ln P_j = \left(\theta/(\theta+1-\beta)\right) \sum_i \lambda_{ij} d \ln c_{ij}. \tag{24}$$

Since $\theta > 0$ and $\beta \leq 1$, we see that a "good" trade shock, $\sum_i \lambda_{ij} d \ln c_{ij} < 0$, is necessarily accompanied by a decline in the aggregate demand shifter, $d \ln P_j < 0$, as hinted to in the previous paragraph. As we can also see from equation (24), the ranking of the direct and indirect markup effects is pinned down by the preference parameter β. Namely, the indirect markup effect is larger if preferences are homothetic ($\beta = 1$) than if they are not ($\beta = 0$).

Plugging equation (24) into equation (21), we finally get

$$d \ln e_j = \left(1 - \rho\left((1-\beta)/(1-\beta+\theta)\right)\right) \sum_i \lambda_{ij} d \ln c_{ij}. \tag{25}$$

As in ACR, by differentiating the gravity equation (16), one can show that $\sum_i \lambda_{ij} d \ln c_{ij}$ is equal to $d \ln \lambda_{jj}/\theta$. Combining this observation with equation (25), we obtain

$$d \ln e_j = \left(1 - \rho\left((1-\beta)/(1-\beta+\theta)\right)\right) d \ln \lambda_{jj}/\theta. \tag{26}$$

Given free entry and our choice of numeraire, we have already argued that $d \ln W_j = -d \ln e_j$. Thus, the main theoretical result of our paper can be stated as follows.

Proposition 1 *Suppose that Assumptions A1 and A2 hold. Then the equivalent variation associated with a small trade shock in country j is given by*

$$d \ln W_j = -\left(1-\eta\right) d \ln \lambda_{jj}/\theta, \text{ with } \eta \equiv \rho\left((1-\beta)/(1-\beta+\theta)\right).$$

Although markups are allowed to vary at the firm-level, we see that welfare analysis can still be conducted using only a few sufficient statistics. In particular, like in ACR, the share of expenditure on domestic goods, λ_{jj}, is the only endogenous variable whose changes need to be observed in order to evaluate the welfare consequences of changes in trade costs.

Compared to ACR, however, Proposition 1 highlights the potential importance of micro-level data. In spite of the fact that the models analyzed in this paper satisfy the same macro-level restrictions as in ACR, different predictions at the micro-level—namely the variation in markups across firms—lead to different welfare conclusions. Since bilateral trade flows satisfy the gravity equation (16) and the measure of entrants is independent of trade costs, the value of $d \ln \lambda_{jj}/\theta$ caused by a given trade shock is exactly the same as in ACR. Yet, welfare changes are no longer pinned down by $d \ln \lambda_{jj}/\theta$, but depend on an extra statistic, η. Here,

welfare changes depend both on the expenditure-weighted sum of marginal cost changes, which are captured by the original ACR formula, $-d \ln \lambda_{jj}/\theta$, as well as the expenditure-weighted sum of markup changes, which are captured by the extra term, $\eta d \ln \lambda_{jj}/\theta$.[13] According to Proposition 1, if $\eta < 0$, then an increase in trade openness, $d \ln \lambda_{jj} < 0$, must be accompanied by a negative expenditure-weighted sum of markup changes, which raises the gains from trade liberalization. Conversely, if $\eta > 0$, the change in markups must lead to smaller welfare gains.[14]

The sign of η, in turn, depends on two considerations. First, is the preference parameter β equal to zero or one? This determines the relative importance of the direct and indirect markup effects. Second, is the average markup elasticity ρ positive or negative? This determines which of the direct and indirect markup effects is welfare enhancing. While the answer to these questions is ultimately an empirical matter, which we deal with in Section 5, a number of theoretical issues are worth clarifying at this point.

4.2 Discussion

In Section 2, we have mentioned three special cases of our general demand system: (i) additively separable utility functions, which imply $\beta = 0$; (ii) QMOR expenditure functions, which imply $\beta = 1$; and (iii) Kimball preferences, which also imply $\beta = 1$. In cases (ii) and (iii), Proposition 1 implies that gains from trade liberalization are exactly the same as those predicted by the models with constant markups considered in ACR. In case (i), whether $\eta > 0$ or < 0 depends on the sign of the (average) markup elasticity, ρ. Since the pioneering work of Krugman (1979), the most common assumption in the literature is that the demand elasticity is decreasing with the level consumption, and hence increasing with the level of prices, $\varepsilon_D' > 0$, which implies $\rho > 0$.[15] Under this assumption, $\eta > 0$, the gains from trade liberalization predicted by models with variable markups are *lower* than those predicted by

[13] Formally, the above analysis establishes that

$$\sum_i \int_{z_{ij}^*}^{\infty} \lambda_{ij}(z)\, d \ln m_{ij}(z)\, dG_i(z) = \eta d \ln \lambda_{jj}/\theta.$$

[14] Profit maximization, however, requires $\eta \leq 1$. To see this, note that the profits of firms, $(p-c)\, q(p, Q, P)$, are supermodular in (c, p). Thus by Milgrom and Shannon's (1994) Monotonicity Theorem, firms with lower costs must have lower prices. This requires $d \ln \mu(v)/d \ln v \leq 1$ for all v, and, in turn, $\rho \leq 1$. It follows that $\eta = \rho\left((1-\beta)/(1-\beta+\theta)\right)$ is also less than one. Economically speaking, variable markups can lower the welfare gains from trade liberalization, because pass-through is incomplete, but they cannot turn gains into losses, because pass-through cannot be negative.

[15] In the words of Krugman (1979), "this seems to be necessary if this model is to yield reasonable results, and I make the assumption without apology." As Mrazova and Neary (2016b) note, this condition is sometimes called "Marshall's Second Law of Demand," as Marshall (1920) argued it was the normal case. Zhelobodko et al. (2011) and Dhingra and Morrow (2016) offer recent exceptions that study the predictions of monopolistically competitive models when $\varepsilon_D' > 0$.

models with constant markups. In other words, under the most common alternatives to CES utility, the existence of variable markups at the firm-level in the class of gravity models that we consider (weakly) dampens rather than magnifies the gains from trade liberalization.[16]

What are the economic forces behind lower gains from trade liberalization under $\eta > 0$? As we formally establish in Appendix A.4, a strong implication of Assumption A2 is that if markups are an increasing function of firm-level productivity—as they would be under standard alternatives to CES utility—then the univariate distribution of markups is independent of the level of trade costs. This reflects the countervailing effects of a change in trade costs on markups. On the one hand, a decline in trade costs, τ_{ij}, leads current exporters from country i to increase their markups in country j. On the other hand, it leads less efficient firms from country i to start exporting to j, and such firms charge lower markups. When firm-level productivity is distributed Pareto, the second effect exactly offsets the first one so that the markup distribution is not affected. Yet the entry of the less efficient firms is irrelevant from a welfare standpoint, which explains why the invariance of the markup distribution does not preclude changes in markups to affect the welfare gains from trade liberalization. In our analysis, welfare changes depend on the expenditure weighted sum of markup changes, which may be positive or negative. This is reflected in the fact that η could be positive or negative in Proposition 1.

The economic forces behind our welfare results echo the two quotes from Helpman and Krugman (1989) given in the Introduction. First, the existence of variable markups affects how trade cost shocks get passed through from foreign firms to domestic consumers. This is reflected in $(-\rho)\sum_{i\neq j}\lambda_{ij}\left(d\ln c_{ij} - d\ln P_j\right)$ in equation (21), which captures both the direct and indirect effects on foreign markups. Second, the existence of variable markups implies that changes in trade costs also affect the degree of misallocation in the economy. This is reflected in $\rho\lambda_{jj}d\ln P_j$ in equation (21), which captures the indirect effect on domestic markups. While domestic markups per se are a transfer from consumers to producers, it is a matter of simple algebra to check that under Assumption A2, changes in domestic markups, $\rho\lambda_{jj}d\ln P_j$, are proportional to the negative of the covariance between firm-level markups on the domestic market and changes in firm-level employment shares for that market; see Appendix A.4. Thus whenever domestic markups go down on average, workers get reallocated towards firms with higher markups. Since their goods are under-supplied in the initial equilibrium, this increases welfare above and beyond what a model with constant markups

[16] If one generalizes equations (2) and (3) to allow for additively separable indirect utility functions, as discussed in footnote 4, then the correction term η generalizes to

$$\eta \equiv \rho\left((1-\beta+\nu\theta)/(1-\beta+\theta)\right),$$

with the case of additively separable indirect utility functions corresponding to $\nu = 1$ and $\beta = 1$. In this case, we see that $\eta = \rho$. Hence, if $\rho > 0$, gains from trade liberalization must also be lower.

would have predicted.[17]

The connection between pro-competitive effects of trade and misallocations is perhaps best illustrated in the context of a symmetric world economy. In such an environment, there are no general equilibrium effects; welfare in each country is equal to world welfare; and the ACR formula reduces to $-d \ln \lambda_{jj}/\theta = -(1-\lambda_{jj}) d \ln \tau$. This corresponds to the first-best welfare change, i.e., the one that would be associated with a small change in trade cost if the world economy was efficient. Accordingly, the pro-competitive effects of trade, defined as the difference between the welfare change predicted by Proposition 1 and the ACR formula, here $\eta \left(1-\lambda_{jj}\right) d \ln \tau$, simply measure the extent to which trade integration reduces misallocation, i.e., the welfare gap between the distorted and efficient economies.[18]

At this point, it should therefore be clear that our theoretical analysis is perfectly consistent with a scenario in which after trade liberalization: (i) the least efficient domestic firms exit; (ii) domestic firms that stay in the industry reduce their markups; and yet (iii) welfare gains from trade liberalization are lower than those predicted by a simple trade model with constant markups and no firm heterogeneity like Krugman (1980). The underlying economics are simple: the exit of the least efficient firms has no first-order welfare effects; the decrease in domestic markups raises welfare by reducing distortions on the domestic market; but the welfare consequences of trade liberalization also depend on changes in foreign markups, which tend to push welfare in the opposite direction.

The role of non-homotheticity in preferences. Since Assumption A2 rules out changes in the distribution of markups, our welfare analysis gives a central role to non-homotheticity in preferences. In general, reallocations of workers between firms with different markups may arise because of changes in the relative markups charged by these firms. Here, non-homotheticity is the only source of such reallocations.

A corollary of Proposition 1 is that if preferences are homothetic, which corresponds to $\beta = 1$ and hence $\eta = 0$, the direct and indirect markup effects exactly compensate one another, implying that welfare changes are equal to those predicted by models with constant markups considered in ACR. Intuitively, a good trade shock in an open economy is like a positive income shock in a closed economy. If preferences are homothetic, such a shock does

[17] The fact that changes in the degree of misallocation should be picked up by the covariance between markups and changes in factor share is not specific to the particular model that we consider; see Basu and Fernald (2002) for a general discussion.

[18] The previous comparison implicitly holds fixed the level of openness, $1-\lambda_{jj}$, in the distorted and efficient economies. Instead, as done in Edmond et al. (2015), one could imagine holding fixed the initial level of trade costs, τ. Under this alternative approach, even if all markups were to remain constant in response to a trade shock, one would conclude that there are positive pro-competitive effects of trade, i.e., a positive difference between the welfare change in the distorted and efficient economies, provided that the former exhibits a higher level of openness, $(1-\lambda_{jj}^{distorted}) > (1-\lambda_{jj}^{planner})$. According to our definition, pro-competitive effects of trade only arise if the expenditure-weighted sum of markup changes is non-zero.

not affect how domestic consumers allocate their expenditures across goods and, in turn, has no additional welfare effects even if the economy is distorted. In contrast, if preferences are non-homothetic, a positive income shock may additionally lower welfare in a distorted economy if it triggers a reallocation towards goods that have lower markups. This is what happens if $\rho > 0$ and $\beta = 0$.[19]

Under the assumption that preferences are homothetic, it is worth noting that the equivalence between models with variable and constant markups extends beyond small changes in trade costs. Homotheticity in preferences implies that consumers that are subject to an income shock equivalent to the trade shock still consume goods in the exact same proportions as consumers that are not. In order to compute the equivalent variation associated with an arbitrary change in trade costs from τ to τ', we can therefore integrate the expression given in Proposition 1 between the initial and final equilibria. Formally, if Assumptions A1 and A2 hold and $\beta = 1$, then the equivalent variation associated with any trade shock in country j is given by

$$\hat{W}_j = \left(\hat{\lambda}_{jj}\right)^{-1/\theta},$$

where $\hat{\lambda}_{jj} \equiv \lambda'_{jj}/\lambda_{jj}$ denotes the proportional change in the share of expenditure on domestic goods caused by the trade shock. This is the exact same expression for large welfare changes as in ACR.

Although the set of models with homothetic preferences considered in this paper is rich enough to rationalize any cross-sectional distribution of markups—by appropriately choosing the demand function $D(\cdot)$ that enters equation (5)—any model within that set would predict the same welfare gains from trade liberalization as in ACR, regardless of whether trade shocks are small or not.

Relationship to Krugman (1979). While the demand system described in equation (1) nests the case of additively separable utility functions considered in Krugman (1979), our analysis differs from his in three dimensions. First, we impose the existence of a choke price. Second, we assume that firms are heterogeneous in their productivity. Third, we focus on changes in iceberg trade costs, whereas he focuses on changes in market size. The last two differences

[19] Our restriction to non-homothetic demand functions that are additively separable is crucial for establishing a simple relationship between the sign of the markup elasticity and the sign of the welfare adjustment. Additive separability implies that the (absolute value of the) income and price elasticities are both larger for goods consumed in low quantities and, in turn, that the consumption of goods with lower markups must expand in response to a positive income shock. Formally, one can check that $d \ln P_j / d \ln w_j > 0$, and hence the covariance between firm-level markups and log-changes in firm-level employment shares caused by the positive income shock is negative. More generally, demand functions that are both non-homothetic and non-additively separable, as in Comin et al. (2015), could lead to richer predictions. Finally, we note that under the assumption that preferences are additively separable, the only way to approach homotheticity is to go from a finite choke price, which can always be normalized to one, to an infinite one. Hence, we cannot smoothly approach the homothetic (CES) case.

have strong implications for the nature of distortions in the class of models that we analyze compared to his.

In models of monopolistic competition with homogeneous firms and no trade costs, the level of the markups may change with the size of the market, but they are always common across goods in a given equilibrium. Thus markups are not a source of misallocation across producing firms. The only distortion in the economy is that there may be too many or too few goods produced in equilibrium, or equivalently, that *all* producing firms may be producing too little or too much. Ceteris paribus, the pro-competitive effects in Krugman (1979) are therefore positive if an increase in country size raises output per firm, and firms were producing too little before market integration, or it lowers output, and they were producing too much. The formal argument can be found in Appendix A.4.[20]

In contrast, because of Assumption A2, the measure of entrants in our model is independent of changes in trade costs, as discussed in Section 3.2. The only distortion in the models that we consider is that markups vary across goods from the same country.[21] Our focus here is on the existence of variable markups at the firm-level and whether, conditional on the same observed macro data, models that feature such markups should lead us to conclude that welfare gains from trade liberalization are larger than previously thought.

How would alternative market structures affect the pro-competitive effects of trade? Many popular models in international trade, from Krugman (1980) to Melitz (2003), feature monopolistic competition with CES utility, thereby leading to constant markups. In this paper, we have chosen to introduce variable markups by maintaining monopolistic competition, but departing from CES utility in a flexible way. One could have instead maintained CES utility and depart from monopolistic competition by assuming Bertrand or Cournot competition, as in Bernard et al. (2003), Atkeson and Burstein (2008), and Edmond et al. (2015).

Although a general analysis of the pro-competitive effects under oligopolistic competition is beyond the scope of our paper, we briefly discuss the potential channels through which the introduction of oligopolistic competition may or may not affect our results. At the firm-level, we know that if CES utility is maintained, then markups under Bertrand and Cournot competition can still be expressed as a function of the ratio of the firm-level price and an aggregate price index, as discussed in Burstein and Gopinath (2014). It follows that the first part of our welfare analysis in Section 4.1, leading to equation (19), would remain

[20] As also noted in Appendix A.4, a full analysis of the pro-competitive effects of trade in Krugman (1979) would also require to take a stand on which macro moments to hold constant when comparing models with and without variable markups. Since Krugman (1979) does not satisfy a gravity equation away from the CES case, this is less straightforward than in the class of models that we analyze.

[21] Under Assumption A2, the distribution of markups in a given destination is also the same across all source countries. Thus all markup distortions are "within" rather than "between" distortions; see Appendix A.4 for details.

unchanged. The second part, however, would change. Aggregating across a finite rather than a continuum of firms makes our analysis potentially more complex. Since we can no longer invoke the law of large numbers, Assumption A2 is no longer sufficient to derive a gravity equation, as in Eaton et al. (2013). Similarly, there is no guarantee that the markup elasticity, like the trade elasticity, would be common across countries (i.e., $\rho_{ij} = \rho$ for all i, j).[22]

The previous observations notwithstanding, we note that Bernard et al. (2003) offer one example of an oligopoly model with a continuum of sectors that generates variable markups at the micro level, a gravity equation at the macro level, a fixed univariate distribution of markups, and the same welfare implications as the (homothetic) monopolistically competitive models that we consider. For the interested reader, Neary (2016) provides further results on the welfare gains from trade in an economy that includes both Cournot competition and non-CES utility.

4.3 Multi-Sector Extension

In our baseline analysis, we have focused on a single monopolistically competitive sector. This is a useful theoretical benchmark, but one that imposes implausibly strong restrictions on the pattern of substitution across goods. In practice, we do not expect the elasticity of substitution between goods from the same sector, say cotton and non-cotton t-shirts, to be equal to the elasticity of substitution between goods from different sectors, say t-shirts and motor vehicles. Before moving to our empirical analysis, we therefore describe how our theoretical analysis can be extended to accommodate multiple sectors and a flexible pattern of substitution across those.

Compared to Section 2, we focus on an economy comprising multiple sectors, indexed by k, and a continuum of goods within each sector, indexed by ω. We assume that consumers have weakly separable preferences so that consumption on goods in sector k, $\boldsymbol{q}^k \equiv \{q_\omega^k\}_{\omega \in \Omega^k}$, only depends on the schedule of prices, $\boldsymbol{p}^k \equiv \{p_\omega^k\}_{\omega \in \Omega^k}$, and the expenditure per capita, y^k, in that sector. We do not impose any restriction on the structure of preferences across sectors. All other assumptions are the same as in our baseline model. In particular, the Marshallian demand for any differentiated good ω in sector k is

$$q_\omega^k(\boldsymbol{p}^k, y^k) = Q^k\left(\boldsymbol{p}^k, y^k\right) D^k\left(p_\omega^k / P^k(\boldsymbol{p}^k, y^k)\right), \tag{27}$$

[22] Starting from equation (19), and using $z_{ij}^* = c_{ij}/P_j$, we have

$$d\ln e_j = \textstyle\sum_i \lambda_{ij} d\ln c_{ij} - \sum_i \lambda_{ij}\rho_{ij} d\ln c_{ij} + \sum_i \lambda_{ij}\rho_{ij} d\ln P_j,$$

instead of equation 21.

where $Q^k\left(\boldsymbol{p}^k, y^k\right)$ and $P^k\left(\boldsymbol{p}^k, y^k\right)$ are sector-level demand shifters determined as the solution of the following system of equations,

$$\int_{\omega\in\Omega^k}\left[H^k\left(p_\omega^k/P^k\right)\right]^{\beta^k}\left[p_\omega^k Q^k D^k\left(p_\omega^k/P^k\right)\right]^{1-\beta^k}d\omega = \left(y^k\right)^{1-\beta^k}, \tag{28}$$

$$\left(Q^k\right)^{1-\beta^k}\left[\int_{\omega\in\Omega^k} p_\omega^k Q^k D^k\left(p_\omega^k/P^k\right)d\omega\right]^{\beta^k} = \left(y^k\right)^{\beta^k}. \tag{29}$$

Consider first the case of restricted entry. Let η^k and ζ^k denote the sector-level counterparts of η and ζ defined in previous sections, and let $s^k \equiv y^k/\sum_{k'} y^{k'}$ denote the sector-level expenditure shares. In Appendix A.5, we establish the following multi-sector generalization of Proposition 1.

Proposition 2 *Suppose that Assumptions A1 and A2 hold sector by sector, entry is restricted in all sectors, and $\eta^k = \eta$ and $\zeta^k = \zeta$ for all k. Then the equivalent variation associated with a small trade shock in country j is given by*

$$d\ln W_j = -(1-\eta)\sum_k s_j^k d\ln\lambda_{jj}^k/\theta^k.$$

When studying monopolistically competitive models with multiple sectors, restricted entry, and constant markups, Arkolakis et al. (2012) found that $d\ln W_j = -\sum_k s_j^k d\ln\lambda_{jj}^k/\theta^k$. Hence, like in the one-sector case analyzed in Section 4.1, the welfare implications of variable markups reduce to one extra statistic, η, the sign of which determines whether pro-competitive effects of trade are positive or negative.

It is worth noting that Proposition 2 requires both η^k and ζ^k to be constant across sectors. In general, letting $\eta_j \equiv \sum_k s_j^k \eta^k$, we have

$$d\ln W_j = -\sum_k\left(1-\eta^k\right)s_j^k d\ln\lambda_{jj}^k/\theta^k - \sum_k \eta^k ds_j^k + \left(1-\eta_j\right)d\ln(\sum_k L_j^k/(1-\zeta^k)). \tag{30}$$

The second term on the right-hand side shows that reallocation of expenditure towards sectors with a lower η^k leads to additional welfare gains, while the third term shows that the same happens with reallocation of employment towards sectors with a higher average markup (i.e., higher $1/\left(1-\zeta^k\right)$). These cross-sector effects are ruled by the assumption that $\eta^k = \eta$ and $\zeta^k = \zeta$ for all k, implying that the focus of Proposition 2 is on within- rather than between-sector distortions.[23]

[23] Epifani and Gancia (2011) provide empirical evidence of the dispersion of markups between sectors and study its implication for the welfare consequences of international trade.

Now consider the case of free entry. Even under the assumption that total labor supply is inelastic, trade shocks may now lead to changes in sector-level employment and, in turn, the measure of firms, N_i^k. We already know from the work of Arkolakis et al. (2012) that such considerations matter for welfare. When studying monopolistically competitive models with multiple sectors, constant markups, but free rather than restricted entry, they find that the equivalent variation associated with a small trade shock in country j becomes $-\sum_k s_j^k (d \ln \lambda_{jj}^k - d \ln L_j^k)/\theta^k$, where L_j^k denotes employment in sector k. The relevant question for our purposes is the extent to which the introduction of variable markups within each sector affects the previous formula.

In Appendix A.5, we show that for the three types of demand systems discussed in Section 2.1—additively separable preferences, QMOR expenditure functions, and Kimball preferences—if Assumptions A1 and A2 hold sector by sector, entry is free in all sectors, and $\eta^k = \eta$ for all k then the welfare formula in Proposition 1 becomes

$$d \ln W_j = -(1-\eta) \sum_k s_j^k \left(d \ln \lambda_{jj}^k - d \ln L_j^k \right) / \theta^k. \tag{31}$$

In short, for arbitrary preferences across sectors and regardless of whether entry is restricted or free, our theoretical analysis points towards η as a sufficient statistic for the measurement of the pro-competitive effects of trade.[24] We now describe a procedure to estimate η in Proposition 1.

5 Empirical Estimates

As presented in Proposition 1, the direction and magnitude of pro-competitive effects of trade hinges on the value η. The purpose of this section is to discuss empirical evidence that speaks to this value.

Recall that η is defined as the product of two terms: $(1-\beta)/(1-\beta+\theta)$ and ρ (the sales-weighted integral over firms' elasticities of mark-ups with respect to marginal costs, as in equation 20). As discussed above, in the homothetic case for which $\beta = 1$ we then have $\eta = 0$, and hence no pro-competitive effects, irrespective of the value taken by other parameters. By contrast, in the non-homothetic case ($\beta = 0$) the value of η hinges on the product of $1/(1+\theta)$ and ρ. In our model, θ is equal to the elasticity of aggregate trade flows with respect to trade costs. We therefore use $\theta = 5$, which is in line with recent estimates of this

[24] We conjecture that the result in (31) remains valid for any demand system satisfying (27)-(29), but this is not something that we have been able to prove in general. Note that if η^k varies across sectors, then equation 31 features an extra term, $-\sum_k \eta^k d s_j^k$, exactly as in equation 30. This captures the first-order welfare effects associated with reallocation of expenditures across sectors; see Appendix A.5.

"trade elasticity" parameter—e.g. Eaton et al. (2011), Simonovska and Waugh (2014), and Costinot et al. (2012)—and is equal to the median estimate in the meta-analysis of gravity-based estimates in Head and Mayer (2014).

This logic implies that η lies between zero (for homothetic demand) and $\rho/6$ (for non-homothetic demand). To complete these bounds we therefore require an estimate of ρ, so we turn now to two different strategies for estimating this parameter. The first involves estimating $D(\cdot)$ directly and using these estimates to evaluate ρ; this method has the advantage of also providing an estimate of the demand-side primitives, beyond their implications for ρ, that are needed for some of our extensions in Section 6. The second strategy for estimating ρ draws on estimates of firm-level pass-through of costs into prices; this has the advantage of focusing on the true spirit of ρ, given that it is defined as a particular pass-through elasticity. Ultimately we see these strategies as complementary—and the fact that they point to the same broad conclusion, despite drawing on different forms of empirical variation, is reassuring.

5.1 Estimating ρ from Demand

We follow a large literature that uses detailed data on bilateral U.S. merchandise imports within narrowly defined product codes to estimate a representative U.S. consumer's demand parameters; see e.g. Broda and Weinstein (2006) and Feenstra and Weinstein (2017). This section contains a short summary of our procedure and results; for details, see Appendix B.1.

We focus on the the case of additively separable preferences in the "Pollak family"; see Pollak, 1971. This corresponds to

$$D(p_\omega/P) = (p_\omega/P)^{1/\gamma} - \alpha.$$

This nests the CES case (when $\alpha = 0$) but also allows for the possibility of either $\rho > 0$ (which occurs when $\alpha > 0$) or $\rho < 0$ (which occurs when $\alpha < 0$). [25,26]

The best available data are at the 10-digit HS level, annually from 1989-2009. We assume that a variety ω in the model corresponds to a particular 10-digit HS product, indexed by g, from a particular exporting country, indexed by i, and that a sector k in the model cor-

[25] Simple algebra reveals that $\varepsilon_D' > 0$ if and only if $\alpha > 0$. From Appendix A.2, it follows that $\mu'(v) > 0$, and hence $\rho > 0$, if and only if $\alpha > 0$. See Mrazova and Neary (2016a) for a more general analysis of the implications of separable preferences in the Pollak family.

[26] We note also that whereas the sign of α is critical for the value of ρ, its absolute value is not. Specifically, in the region of the parameter space where Assumption A1 holds, i.e. $\alpha > 0$, the value of ρ is independent of α. A change in α is isomorphic to a change in the number of efficiency units of labor, which has no effect on the markup elasticity or any of our results.

responds to a 4-digit HS category.[27] In our baseline analysis we let the demand shifter P_t^k vary across sectors and over time, but restrict the demand parameters α and γ to be common across all sectors. In practice, we estimate the inverse demand relation given by

$$\Delta_t \Delta_{gi} \ln p_{git}^k = \gamma \Delta_t \Delta_{gi} \ln(q_{git}^k + \alpha) + \Delta_t \Delta_{gi} \ln \epsilon_{git}^k, \tag{32}$$

where the notation Δ_t refers to mean-differencing over time and Δ_{gi} to mean-differencing over product-country gi observations within a sector-year kt. The error term $\Delta_t \Delta_{gi} \epsilon_{git}^k$ could arise from unobserved demand differences, measurement error, or product quality differences that are not removed by our double-differencing procedure.[28] Because of standard endogeneity concerns, we use the (log of one plus the) relative value of tariff duties charged as an instrumental variable when estimating equation (32).

Our non-linear IV estimate (along with 95% confidence intervals, block-bootstrapped at the exporting country level) is $\widehat{\gamma} = -0.347$ $[-0.373, -0.312]$ and $\widehat{\alpha} = 3.053$ $[0.633, 9.940]$.[29] Notably, the α estimate has a 95% confidence interval that excludes zero, so the CES case is rejected at standard levels of significance.

These parameter estimates (along with $\theta = 5$) imply, using equations (5) and (20), that $\hat{\rho} = 0.36$ and $\hat{\eta} = \hat{\rho}/6 = 0.06$. Thus, micro-level trade data lead us to conclude (following Proposition 1) that gains from trade liberalization are 6% lower than what we would have predicted by assuming (wrongly) that markups are constant across firms.

5.2 Estimates of ρ from Pass-Through

One potential concern about the previous empirical strategy is that the source of variation used to estimate ρ, and hence η, relies too much on the particular structure of the model. Economically speaking, ρ measures how, on average, changes in marginal costs map into changes in markups. Under monopolistic competition, ρ can be inferred by using information about the shape of demand and the distribution of firm-level sales. But one may imagine instead measuring the elasticity of markups with respect to productivity directly. We now discuss estimates of ρ, and hence η, based on evidence from the existing literature

[27] The resulting dataset has 13,746 unique products, 242 unique exporters, 1387 unique sectors, and ultimately 3,563,993 observations for the estimation of equation (32).

[28] This differencing removes the empirical analog of the unobserved P_t^k as well as any unobserved shifters of product-country gi demand (such as quality or units differences) that are constant over time.

[29] By way of comparison, under the CES restriction of $\alpha = 0$ the IV estimate (and standard error clustered at the exporter level) is $\widehat{\gamma} = -0.206$ (0.036). This corresponds to an elasticity of substitution equal to $1/\widehat{\gamma} = -4.854$, in line with typical estimates of the CES demand parameter in international trade settings, which suggests that our tariff-based instrumental variable is generating similar exogenous variation in trade costs to that which is typically exploited by other researchers. The first-stage F-statistic is 27.28, which implies that finite-sample IV bias is likely to be small.

on the response of markups to changes in marginal costs.

Cross-sectional evidence. Given the static nature of our model, we view ρ as a long-run elasticity. A natural way to estimate such an elasticity is to analyze how markups vary with productivity in a cross-section of firms. The recent empirical work of de Loecker and Warzynski (2012) and de Loecker et al. (2016) provides state-of-the-art estimates of markups and productivity. In a cross-section of Slovenian manufacturing firms, de Loecker and Warzynski (2012) estimate an elasticity of markups to productivity equal to 0.3. Ignoring heterogeneity in markup elasticities across firms, this alternative estimation strategy would immediately lead to $\hat{\rho} = 0.3$ and hence $\hat{\eta} = \hat{\rho}/6 = 0.05$, which is close to the 6% downward adjustment computed above using our demand estimates. de Loecker et al. (2016) use a similar methodology to estimate marginal costs for Indian manufacturing firms. When estimating a cross-sectional regression of (log) prices on (log) marginal cost, they find a "pass-through" coefficient of 0.35. For a given firm in our model, the pass-through coefficient is equal to one minus the markup elasticity. This alternative estimation strategy would lead to $\hat{\rho} = 0.65$ and gains from trade liberalization that are up to 11% lower. We are not aware of similar cross-sectional estimates for all U.S. manufacturing firms, though we note that the positive correlation between TFPR and TFPQ in Foster et al. (2008)—obtained for a small number of industries with information on physical productivity—also points towards $\hat{\rho} > 0$, which, through the lens of our theoretical analysis, again implies weakly lower gains from trade liberalization in the presence of variable markups.

Time-series evidence. Alternatively, one can estimate ρ by studying how marginal cost shocks, such as those caused by changes in exchange rates, tariffs, or energy prices, get passed through to changes in prices over time.

There is a large literature in international macro on exchange rate pass-through. Burstein and Gopinath (2014) offers a review of existing empirical evidence. In the case of the United States, they document long-run pass-through rates using aggregate price indices that range from 0.14 to 0.51. Ignoring again heterogeneity in pass-through rates across firms, this corresponds to $\hat{\rho}$ between 0.49 and 0.86 and hence downward adjustments to the gains from trade liberalization that range from 8% to 14%.

Two recent papers by Berman et al. (2012) and Amiti et al. (2014) document heterogeneity in firm-level pass-through across French and Belgian exporters, respectively. While pass-through rates are nearly complete for small firms, they find pass-through rates of around 0.25 and 0.50 for large firms, respectively.[30] This implies that $\hat{\rho}$ must be below 0.5 (in the

[30] We note that this finding is inconsistent with the estimates of demand from Section 5.1. The Pollak family is flexible enough to generate both incomplete pass-through, $\rho > 0$, and pass-through rates that are lower for larger firms. The previous pattern, however, requires $\alpha > 0$ and $\gamma < -1$. At our estimated parameters, $\widehat{\alpha} = 3.053$ and $\widehat{\gamma} = -0.347$, pass-through is incomplete, but higher for larger firms. We come back briefly to

case of large French exporters) and 0.75 (in the case of large Belgian exporters). This leads to welfare gains that are lower by no more than 12.5%.

For reasons outside of our static model, the pass-through rate of exchange rate shocks, that we have so far discussed here, and the pass-through rate of trade cost shocks, that we need to evaluate ρ in our new welfare formula, may be very different in practice, perhaps because the former shocks are much more volatile than the latter. Having estimated marginal costs for the same Indian firm at different points in time, de Loecker et al. (2016) also run a panel regression of (log) price on (log) marginal cost with firm fixed effects. This yields a pass-through coefficient of 0.2, which would imply $\hat{\rho} = 0.8$ and a downward adjustment no greater than 13%, in the same range as those inferred from exchange pass-through. In terms of U.S. manufacturing firms, Ganapati et al. (2016) focus on the same subset of sectors as in Foster et al. (2008) and study the response of firm-level prices to energy cost shocks. They estimate an average pass-through rate of 0.3, in line with other estimates discussed above.[31]

To summarize, both our empirical strategy, based on the estimation of demand, and alternative empirical strategies, based on cross-section and time-series evidence on the response of markups to changes in marginal cost, point towards markup elasticities implying gains from trade liberalization that are between 5% to 14% lower than those under constant markups.

6 Sensitivity Analysis

We have designed our baseline analysis with two objectives in mind: (i) generate the same aggregate predictions across models with and without variable markups; and (ii) abstract from welfare gains from new varieties. While this provides a clear benchmark to study the welfare implications of variable markups, conditions (i) and (ii) rely on strong assumptions. The goal of this final section is to relax these assumptions and explore the robustness of our earlier conclusions. Namely, we allow for changes in trade costs that are not infinitesimal, for distributions of productivity that are not Pareto, and for fixed marketing costs that are not zero.

6.1 Calibrated Economy

To analyze welfare changes in these more general environments, we rely on numerical simulations. We focus on a world economy comprising two symmetric countries. We set country

this point in Section 6.

[31] Ganapati et al. (2016) also document variation in pass-through rates across sectors, a possibility that we are abstracting from in this paper, as already discussed in Section 4.3.

Parameter	Value	Target/Choice Calibration
Panel A: Demand (Sections 6.2-6.4)		
α	1	Baseline estimate (Table 1, Panel B, normalized)
γ	-0.35	Baseline estimate (Table 1, Panel B)
Panel B: Pareto productivity distribution (Sections 6.2 and 6.4)		
θ	5	Trade elasticity (Head and Mayer (2014))
τ	1.56	Exports/Output = 9.9% (World Input Output Tables, 2002)
Panel C: Lognormal productivity distribution (Section 6.3)		
τ	1.66	Targets for all the three parameters:
μ_l	-2.56	(i) trade elasticity = 5, (ii) exports/output = 9.9%,
σ_l	0.49	and (iii) share of firms exporting = 18% (BJRS, 2007)
Panel D: Bounded Pareto productivity distribution (Section 6.3)		
τ	1.69	Targets for all the three parameters:
θ	2.95	(i) trade elasticity = 5, (ii) exports/output = 9.9%,
$\bar{z}_i^u$	0.42	and (iii) share of firms exporting = 18% (BJRS, 2007)

Table 1: Calibration procedure. Procedure for model parameter calibration discussed in Section 6. BJRS (2007) refers to Bernard et al. (2007).

size to $L = 1$ and fixed entry costs to $F = 1$. This affects welfare levels in the initial equilibrium—by affecting the number of firms—but not the welfare changes that we are interested in. In all simulations, we use the demand system estimated in Section 5.1 under the same normalization of the choke price as in Sections 2 through 4; this corresponds to $\alpha = 1$ and $\gamma = -0.35$. Finally, we set trade costs and parameters of the firm-level productivity distributions to match the trade elasticity, the U.S. exports to (gross) output ratio for U.S. manufacturing firms in 2002, and, in the case of lognormal and bounded Pareto distributions, the share of U.S. manufacturing firms exporting in 2002 reported by Bernard et al. (2007). The values of all calibrated parameters and the targets can be found in Table 1. For the baseline calibration with Pareto distribution this calibration implies a choice of the Pareto elasticity of $\theta = 5$.

Before turning to our counterfactual exercises, we briefly discuss the positive implications of our calibrated model. In the previous literature, a number of models with CES demand have been constructed to match salient features of firm-level data, including the distribution of exporting sales and the difference in measured productivity between exporters and non-exporters. Since our demand estimates have lead us to depart from CES, it is natural to ask how well our calibrated model performs along these two dimensions.

Figure 1 depicts the distribution of total firm revenue —normalized by mean sales— for

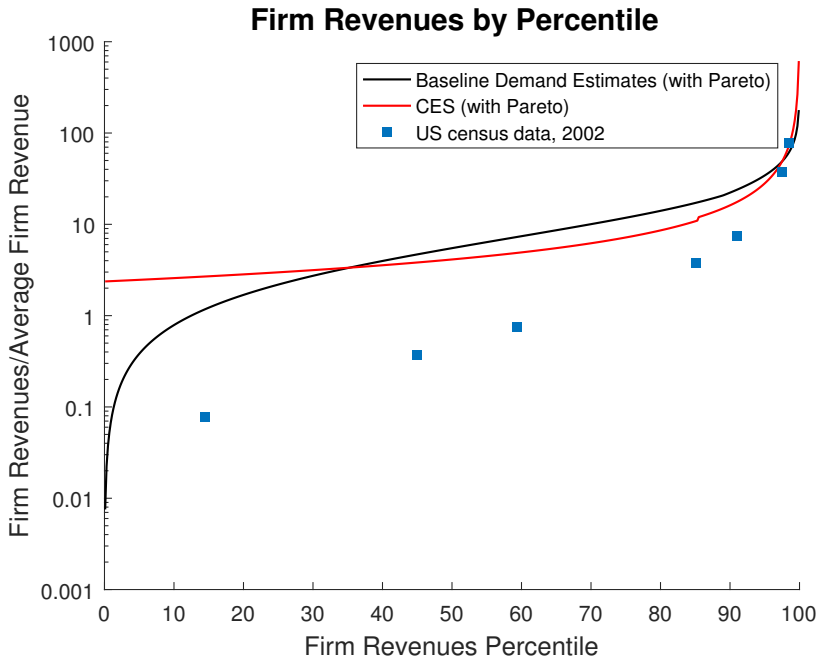

Figure 1: Distribution of Firm Sales. Source: US Census 2002.

US manufacturing firms in 2002 across different percentiles using census data obtained from the small business administration.[32] The predictions of our calibrated model are plotted on the same figure (black line). For comparison, we also plot the predictions of the model with CES demand instead (red line).[33] In both cases, we use the same Pareto elasticity ($\theta = 5$). The two models do well in matching the observed distribution of sales for the largest firms. Intuitively, our estimated demand function asymptotically resembles a CES function. So, given Pareto distributions of productivity, both models predict a Pareto distribution of sales in the right tail. On the left tail, however, our calibrated model comes closer to the observed distribution than the model with CES demand.[34]

Another important feature of firm-level data is the difference between the measured productivity of exporters and non-exporters. Bernard et al. (2003) report that the relative advantage of US exporters to non-exporters in log-productivity is 33% overall and 15% within the same sector. As in the model of Bernard et al. (2003), measured productivity in our model corresponds to the sum of revenues divided by the sum of labor payments, $\sum_j r_{ij}(z) / (\sum_j w_i \tau_{ij} q_{ij}(z) / z)$. For domestic firms this ratio is equal to their markup while for

[32] See Arkolakis (2016) for additional information about the U.S. census data.

[33] In the CES case, we use the estimates of demand in Panel A of Table 1, $\alpha = 0$ and $\gamma = -0.2$. This implies an elasticity of substitution equal to 5.

[34] One can improve the fit of the CES model by introducing demand shocks and fixed marketing costs, as in Eaton et al. (2011) and Arkolakis (2010). The fit of our model for the firm-level distribution of sales is as good as the fit of these richer models.

exporters it is a weighted average of their domestic and foreign markups. At the calibrated parameters, we find that that the exporter's productivity advantage is 13%, very close to the 15% observed within sectors in the data. Absent any fixed cost of production, of course, the same model with CES demand would predict no variation in markups and hence no variation in measured productivity across firms.

Finally, we can compare the implications of our calibrated model for pass-through to the estimates presented in Section 5.2. When running a regression of log domestic price on log marginal cost and a constant using data generated from our model, we find a pass-through coefficient equal to 0.61. This is somewhat higher than the estimates obtained by de Loecker et al. (2016) (discussed above) of 0.35 and 0.2. We note also that our model predicts near complete pass-through for the largest firms given the generalized CES demand. While this feature of our calibrated model helps us match the right tail of the distribution of sales under the assumption that firm-level productivity is Pareto distributed, this is inconsistent with the empirical findings of Berman et al. (2012) and Amiti et al. (2014) on exchange-rate pass-through, as also discussed in Section 5.2.[35]

6.2 Large Changes in Trade Costs

For our first series of numerical exercises, we maintain the exact same assumptions as in our baseline analysis, but consider large changes in trade costs. Namely, we let symmetric iceberg trade costs, τ, vary from twenty percent below to twenty percent above the calibrated value, $\tau = 1.56$.

To understand why large changes may affect our earlier conclusions, let us return to the expenditure minimization problem in country j. Under the restrictions imposed on demand in Section 5, one can check that the expenditure function is given by

$$e_j = \min_{\{q_{ij}(z)\}} \sum_i N_i \int_{z_{ij}^*} p_{ij}(z)\, q_{ij}(z)\, dG_i(z) \tag{33}$$
$$\sum_i N_i \int_{z_{ij}^*} u_{ij}\left(q_{ij}(z)\right) dG_i(z)\, dz \geq \bar{u},$$

[35] In principle, one could construct the demand function $D(\cdot)$ to match (exactly) the relationship between firm-level productivity and pass-through and then, given $D(\cdot)$, construct the distribution of productivity to match (exactly) the distribution of sales. Compared to our baseline estimates of the pro-competitive effects of trade with Pareto distributions and Pollak demand (Section 5.1), neither the estimates without Pollak, but with Pareto (Section 5.2), or with Pollak, but without Pareto (Section 6.3), looked very different. Although we have not explored simultaneous departures from Pareto and Pollak, we have no reason to believe that they would lead to significantly larger pro-competitive effects of trade.

with $u_{ij}(q) = (q+\alpha)^{1+\gamma}$. The Envelope Theorem then implies that

$$d\ln e_j = \sum_i \frac{N_i \int_{z_{ij}^*} \left[p_{ij}(z)\, q_{ij}(z,\bar{u}) - \xi u_{ij}\left(q_{ij}(z,\bar{u})\right)\right] \lambda dG_i(z)\, dz}{e_j} d\ln N_i \tag{34}$$
$$- \sum_i \frac{N_i \left[p_{ij}\left(z_{ij}^*\right) q_{ij}\left(z_{ij}^*,\bar{u}\right) - \xi u_{ij}\left(q_{ij}\left(z_{ij}^*,\bar{u}\right)\right)\right] g_i\left(q_{ij}\left(z_{ij}^*\right)\right)}{e_j} dz_{ij}^*$$
$$+ \sum_i \frac{N_i \int_{z_{ij}^*} \left[p_{ij}(z)\, q_{ij}(z)\, d\ln p_{ij}(z)\right] dG_i(z)}{e_j},$$

where ξ is the Lagrange multiplier associated with the utility constraint and $q_{ij}(z,\bar{u})$ is the compensated (Hicksian) demand. The first term in equation (34) corresponds to the total surplus associated with a change in the measure of varieties from country i, which must be equal to zero if entry is restricted; the second term corresponds to the surplus associated with cut-off varieties; and the third term measures the effects of changes in the prices of existing varieties, either through changes in marginal costs or markups. This last term is the only one that is non-zero in our baseline analysis.

When productivity distributions are Pareto, the number of entrants is fixed even under free entry. So, the first term must always be equal to zero, regardless of whether changes in trade costs are large or small. Away from the initial equilibrium, however, the second term may not be. Although the consumer in the decentralized equilibrium would never consume the cut-off variety, the consumer whose utility has been held at some constant level $\bar{u}$ may very well choose to do so. Put differently, non-homotheticities imply that gains and losses from cut-off varieties, which the formula in Proposition 1 ignores, may no longer be zero as one goes from small to large changes in trade costs.

To assess the importance of these considerations, we compute the equivalent variation associated with an arbitrary change in trade costs given by the expenditure function in (33). We refer to this number, expressed as a fraction of the country's initial income, as the exact welfare change.[36] We then compare this number to the welfare change that one would obtain by integrating the welfare formula in Proposition 1, i.e. $(\lambda'_{jj}/\lambda_{jj})^{-\frac{1-\eta}{\theta}} - 1$, with λ'_{jj} the share of expenditure on domestic goods in the equilibrium with the new trade costs, as computed in Appendix A.3. Figure 2 plots the exact welfare change (bold line) and the welfare change obtained using our new formula with $\eta = 0.06$ (dotted line) as a function of iceberg trade costs, τ. For completeness, we also report the welfare change one would

[36] Formally, the exact welfare change in country j is computed as $e\left(\mathbf{p}_j, u'_j\right)/w_j - 1$, with $\mathbf{p}_j$ and w_j the schedule of good prices and the wage in the initial equilibrium, respectively, and u'_j the utility level in the counterfactual equilibrium.

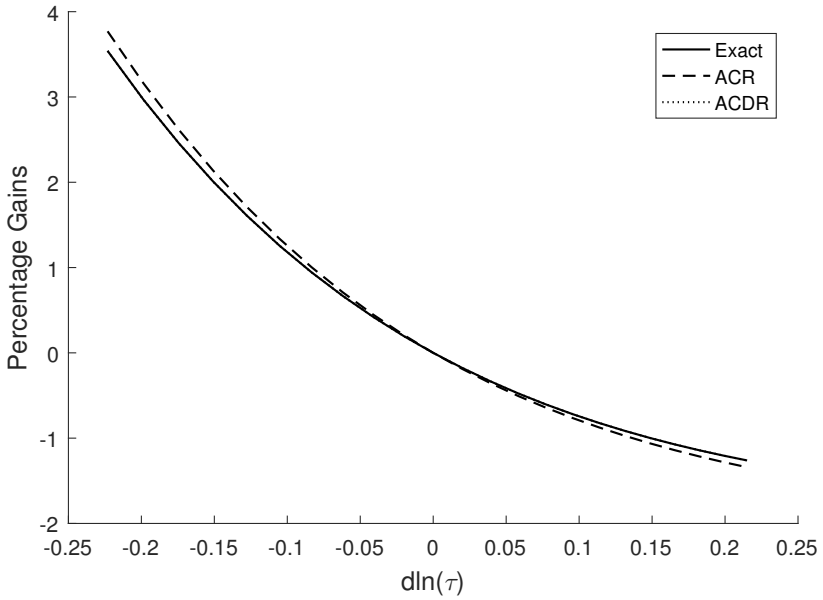

Figure 2: Welfare Gains Relative to Baseline ($\tau = 1.56$), Pareto

obtain by using ACR's welfare formula, i.e. with $\eta = 0$ (dashed line). The bold and dotted curves almost coincide, and so it is not possible to tell them apart in the figure. This implies that our formula in Proposition 1 which holds exactly for small changes in trade costs, also provides an accurate approximation to the case of large changes. In this numerical example, the impact of cut-off varieties on the welfare implications of trade liberalization is minor.

6.3 Alternative Productivity Distributions

In our baseline analysis, we have assumed that the distribution of firm-level productivity is Pareto. This implies a gravity equation, which facilitates comparisons with earlier work, but it also implies strong restrictions on the univariate distribution of mark-ups and the share of aggregate profits (gross of fixed entry costs) in aggregate revenue. Namely, both must be invariant to changes in trade costs. Though one should not expect this prediction to hold away from the Pareto case, it is not a priori obvious how departing from this benchmark case should affect aggregate welfare. Intuitively, one would expect changes in the univariate distribution of markups and the share of aggregate profits to be related and have opposite welfare effects. Take, for instance, an economy where trade liberalization lowers all markups by 10%, a situation that Assumption A2 rules out. Such a decrease would be accompanied by a 10% decrease in the prices faced by consumers, but also a decrease in the share of aggregate profits in aggregate revenue and, under free entry, a decrease in the measure of entrants that may very well offset the benefits from lower prices. If entry is restricted, the

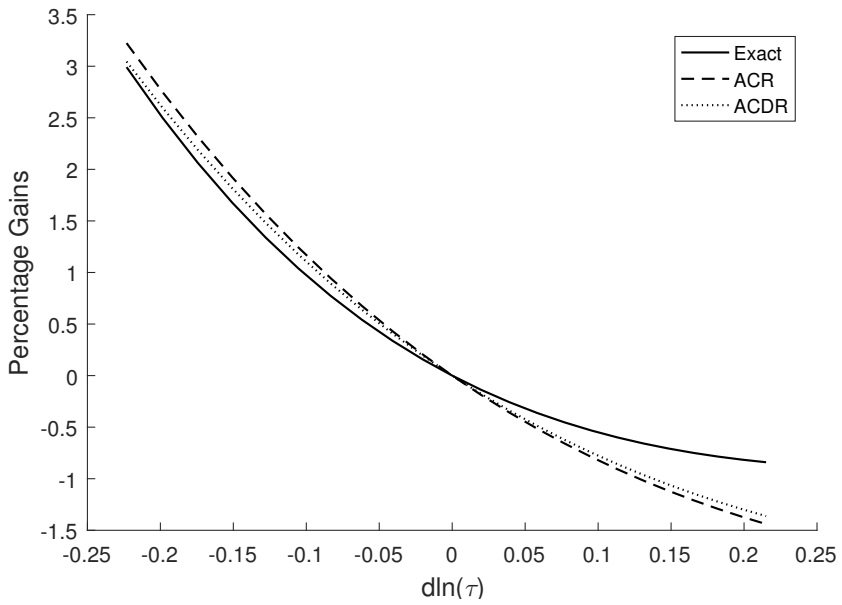

Figure 3: Welfare Gains Relative to Baseline ($\tau = 1.66$), Log-normal

situation is even starker. Namely, under the maintained assumption of a representative agent with perfectly inelastic labor supply, uniform changes in markups cannot have any effect on misallocation and welfare: any decrease in the consumer's expenditure function must be exactly compensated by a decrease in income.

The CES case with free entry nicely illustrates the potential importance of offsetting effects when studying aggregate welfare changes. Away from Pareto, we know that changes in trade costs not only affect the share of expenditure on domestic goods, but also the number of entrants in a given country. Yet, because the allocation is efficient under CES, we know from the work of Atkeson and Burstein (2010) that

$$d \ln e_j = (1 - \lambda_{jj}) d \ln \tau. \tag{35}$$

In a two-country symmetric economy, the formal definition of the trade elasticity in ACR reduces to $\varepsilon = d \ln((1 - \lambda_{jj})/\lambda_{jj})/d \ln \tau$. Using this definition and changing variable in the previous equation, one therefore gets

$$d \ln e_j = d \ln \lambda_{jj}/\varepsilon.$$

In this CES example, the local version of the ACR formula always holds, regardless of distributional assumptions and regardless of whether the number of entrants changes.

Without CES, and hence without efficiency, the situation is more subtle. To explore how our welfare results are affected by departures from Pareto under our estimated demand

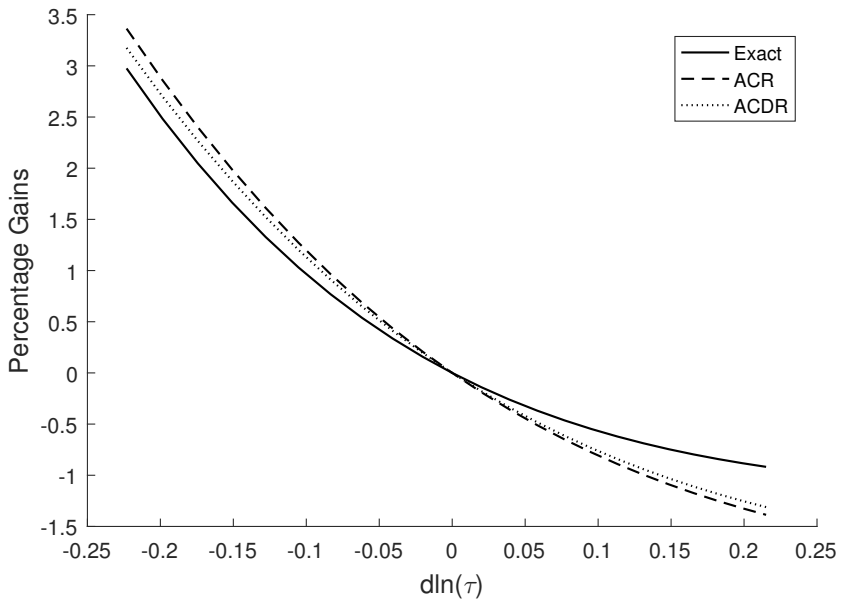

Figure 4: Welfare Gains Relative to Baseline ($\tau = 1.69$), Truncated Pareto

system, we focus on the two alternatives that have recently received attention in the literature: (i) log-normal distributions with mean μ_l and standard deviation σ_l, as in Head et al. (2014); and (ii) bounded Pareto distributions with shape parameter θ and upper-bound $\bar{z}_i^u$, as in Feenstra (2014). The calibrated values of these parameters are reported in Table 1. As discussed above and shown in Table 1, we set these parameters, together with the baseline iceberg trade cost, to target the following three moments: the U.S. manufacturing exports to output ratio, the trade elasticity, and the share of U.S. manufacturing firms that export. Since the trade elasticity is no longer constant, we target its value for a 1% change in trade costs around the calibration point using the formal definition in ACR, applied to the case of two symmetric countries: $\varepsilon = d\ln((1-\lambda_{jj})/\lambda_{jj})/d\ln\tau$.[37]

We then follow the same procedure as in Section 6.2. We compute the exact welfare change using the expenditure function in (33)—with the distribution G_i being either log-normal or bounded Pareto—and we compare those to the welfare change that one would obtain by integrating our new welfare formula or the ACR formula. These results are reported in Figures 3 and 4 for the case of free entry; the results under restricted entry are very similar.[38] In both cases, we see that our formulae over-estimate both the gains from trade

[37]For these alternative productivity distributions, we obtain predictions for the distribution of exporting sales and for the productivity advantage of exporters relative to non-exporters that are similar to those in the Pareto case. Results are available upon request.

[38]When integrating our new formula and the ACR formula, we let the trade elasticity and the average markup elasticity vary as variable trade costs change from their initial to their counterfactual values.

liberalization and the losses from trade protection.[39]

For our purposes, the important take-away from Figures 3 and 4 is that they provide little support to the idea that the welfare gains in the Pareto case are special and unusually low, perhaps because the univariate distribution of markups is fixed. Whether there exist other distributions that could lead to significantly larger gains remains an open question, but under these two alternative distributional assumptions, gains from trade are lower, not larger.[40]

6.4 Fixed Marketing Costs

For our last series of simulations, we introduce fixed marketing costs in our model. Such costs are potentially interesting from a welfare standpoint since they imply that creation and destruction of cut-off varieties may have first-order welfare effects, i.e. the second term in equation (34) is no longer zero, even for small changes in trade costs.

The economic environment is the same as in Section 2, except for the fact that after receiving their random productivity draws, firms must incur a fixed marketing cost, $w_j f_j$, in order to sell in market j. Fixed costs do not affect firm-level markups, which remain a function of relative efficiency alone, but they do affect firm-level profits. Without risk of confusion, let us drop the country indices as we did in Section 3.1. For a firm with marginal cost c and efficiency v, profits are now given by

$$\pi(c, v, Q, L) \equiv ((\mu(v) - 1)/\mu(v))\, x\,(c, v, Q, L) - wf,$$

with firm-level sales, $x(c, v, Q, L)$, still given by (6). Accordingly, a firm will enter a given market if and only if $v \geq v^*$, with v^* implicitly defined by

$$(\mu(v^*) - 1)D(\mu(v^*)/v^*) = (wfv^*)/(QLP). \tag{36}$$

[39]The interpretation of these numerical results is less straightforward than before. As we go from Pareto distributions to other distributions, we not only change the extent of firm-level distortions, but also the aggregate predictions of the model. Although we still target the same trade elasticity in the initial equilibrium, it now varies with the level of the trade of costs, a point emphasized by Head et al. (2014) and Melitz and Redding (2015) in the CES case. More precisely, the trade elasticity increases in absolute value with the level of trade costs, as documented in Appendix **??**. The new welfare numbers therefore reflect different behavior both at the macro and micro levels.

[40]When looking at the effects of trade liberalization under a truncated Pareto distribution, Feenstra (2014) concludes that there are positive pro-competitive effects of trade. A critical difference between his conclusion and ours comes from the definition of pro-competitive effects. According to Feenstra (2014), pro-competitive effects measure the welfare impact of variable markups through their effects on consumer prices, but not on aggregate profits and entry. Under such a definition, a uniform decrease in markups, with no effect on misallocation, would always be counted as a positive pro-competitive effect. We prefer to focus on the aggregate welfare implications of variable markups, independently of the particular channel through which they operate.

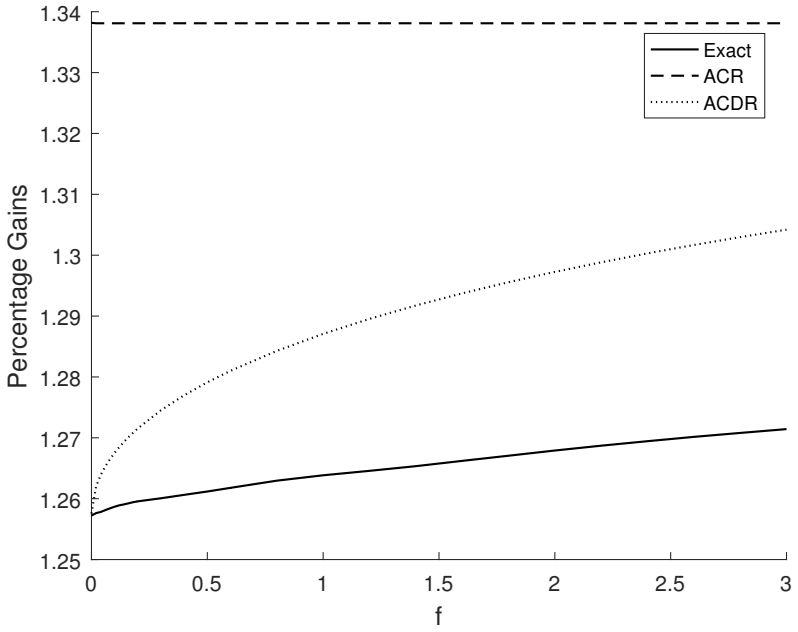

Figure 5: Welfare Gains and Fixed Costs ($\tau = 1.56$ to 1.40)

When $f = 0$, equation (36) implies $v^* = 1$. So, firms only enter a market if their marginal cost, c, is below the reservation price, P. When $f > 0$, marginal costs must be strictly below P for firms to break even. The gravity equation (16) is the same as before. The labor market clearing condition must be modified in order to take into account the resources associated with fixed marketing costs.

To quantify the importance of fixed marketing costs, we focus on the case with free entry but the results for restricted entry are again very similar. We consider a 10% decrease in trade costs from the calibrated value, $\tau = 1.56$ to a counterfactual value, $\tau = 1.40$. We then vary the fixed marketing cost from $f = 0$ to $f = 3$. Figure 5 reports the exact welfare changes together with the predictions that one would obtain by integrating our new welfare formula ($\eta = 0.06$) or the ACR formula ($\eta = 0$). Exact welfare changes are always bounded from above by our two formulas. As fixed costs increase, we see that both the exact welfare changes and our new formula converge towards the ACR formula. This is intuitive. As fixed marketing costs increase, only the most productive firms select into a market. These firms operate in parts of the demand curve that are very close to CES. Hence, markups are close to constant across firms and welfare changes are well-approximated by the ACR formula.[41]

[41] In numerical simulations we have focused on the case with $\beta = 0$. In the homothetic case, $\beta = 1$, one can check that although the efficiency cut-off, v^*, is no longer equal to one, it remains unaffected by trade costs. Accordingly, the distribution of markups and the number of entrants remain constant. Thus, whether fixed marketing costs are zero or not, gains from trade liberalization are given by the ACR formula.

7 Concluding Remarks

We have studied the gains from trade liberalization in models with monopolistic competition, firm-level heterogeneity, and variable markups. Under standard restrictions on consumers' demand and the distribution of firms' productivity, we have developed a generalized version of the ACR formula that highlights how micro- and macro-level considerations jointly shape the welfare gains from trade.

In the case of homothetic preferences, we have shown that this generalized version necessarily reduces to the ACR formula, hence pro-competitive effects must be zero. In the non-homothetic case, we have used a range of empirical estimates (based on both micro-level trade data and firm-level pass-through) to quantify our new formula. Our main finding here is that (rightly) taking into account variable markups leads to gains from trade liberalization that are up to 14% lower than those that one would have predicted by (wrongly) assuming constant markups. In this sense, pro-competitive effects remain elusive.

Our theoretical and empirical results only apply to a particular class of models. Monopolistic competition plays a central role in the field of international trade, but it is not the only market structure under which variable markups may arise. Likewise, gravity models have become the workhorse model for quantitative work in the field, but they rely on very strong functional restrictions that may be at odds with the data; see e.g. Adao et al. (2017). Hence, it goes without saying that the main lesson from our analysis is not that pro-competitive effects must, everywhere and always, be small. In our view, there are two robust messages that emerge from our analysis.

First, domestic and foreign markups are likely to respond very differently to trade liberalization. Whereas changes in domestic markups only reflect shifts in aggregate demand at the sector level, changes in foreign markups also reflect the direct effect of changes in trade costs. Because of this asymmetry, it is perfectly possible for domestic and foreign markups to move in opposite directions, as Helpman and Krugman (1989) stress and as our analysis illustrates. If one is interested in the aggregate implications of variable markups, this suggests caution when extrapolating from evidence on the behavior of domestic producers alone.

Second, information about the cross-sectional or time variation in markups alone is unlikely to be sufficient for evaluating the pro-competitive effects of trade. In the present paper, the average elasticity of markups matters, but so do non-homotheticities in demand. Intuitively, whether trade liberalization is likely to alleviate or worsen underlying misallocations does not only depend on the distribution of markups in the economy. It also depends on whether in response to a "good" income shock, such as the one created by trade liberalization, consumers spend more or less on goods with higher markups. The often imposed

assumption of homothetic preferences may not be innocuous in this context.

References

Adao, Rodrigo, Arnaud Costinot, and Dave Donaldson, "Nonparametric Counterfactual Predictions in Neoclassical Models of International Trade," *American Economic Review*, 2017, *107* (3), 633–689.

Amiti, Mary, Oleg Itskhoki, and Jozek Konings, "Importers, Exporters, and Exchange Rate Disconnect," *American Economic Review*, 2014, *104* (7), 1942–1978.

Anderson, James E. and Eric Van Wincoop, "Gravity with Gravitas: A Solution to the Border Puzzle," *American Economic Review*, 2003, *93* (1), 170–192.

Antras, Pol, Alonse de Gortari, and Oleg Itskhoki, "Globalization, Inequality and Welfare," *mimeo Harvard University*, 2016.

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

_ , "A Unified Theory of Firm Selection and Growth," *Quarterly Journal of Economics*, 2016, *131* (1), 89–155.

_ **, Arnaud Costinot, and Andres Rodríguez-Clare**, "New Trade Models, Same Old Gains?," *American Economic Review*, 2012, *102* (1), 94–130.

Atkeson, Andrew and Ariel Burstein, "Pricing to Market, Trade Costs, and International Relative Prices," *American Economic Review*, 2008, *98* (5), 1998–2031.

_ **and** _ , "Innovation, Firm Dynamics, and International Trade," *Journal of Political Economy*, 2010, *118* (3), 433–489.

Axtell, Rob L., "Zipf Distribution of U.S. Firm Sizes," *Science*, 2001, *293* (5536), 1818–1820.

Baldwin, Richard and James Harrigan, "Zeros, Quality and Space: Trade Theory and Trade Evidence," *American Economic Journal: Microeconomics*, 2011, *3* (2), 60–88.

Basu, Susanto and John G. Fernald, "Aggregate Productivity and Aggregate Technology," *European Economic Review*, 2002, *46*, 963–991.

Behrens, Kristian and Yasusada Murata, "Globalization and Individual Gains from Trade," *Journal of Monetary Economics*, 2012, *59* (8), 703–720.

_ **, Giordano Mion, Yasusada Murata, and Jens Sudekum**, "Trade, Wages, and Productivity," *International Economic Review*, 2014, *55* (4), 1305–1348.

Bergin, Paul R. and Robert Feenstra, "Pass-Through of Exchange Rates and Competition Between Floaters and Fixers," *Journal of Money Credit and Banking*, 2009, *41* (1), 35–70.

Bergson, A., "Real Income, Expenditure Proportionality, and Frisch's "New Methods of Measuring Marginal Utility"," *Review of Economic Studies*, 1936, *4*, 33–52.

Berman, Nicolas, Philippe Martin, and Thierry Mayer, "How Do Different Exporters React to Exchange Rate Changes?," *Quarterly Journal of Economics*, 2012, *127*, 437–492.

Bernard, Andew B., Jonathan Eaton, J. Bradford Jensen, and Samuel Kortum, "Plants and Productivity in International Trade," *American Economic Review*, 2003, *93* (4), 1268–1290.

Bernard, Andrew, Bradford Jensen, Steve Redding, and Peter J. Schott, "FIrms in International Trade," *Journal of Economic Perspectives*, 2007, *21* (3), 105–130.

Bertoletti, Paolo, Federico Etro, and Ina Simonovska, "International Trade with Indirect Additivity," *forhcoming, American Economic Journal: Microeconomics*, 2017.

Bhagwati, Jagdish N., *The Generalized Theory of Distortions and Welfare*, Vol. Trade, Balance of Payments, and Growth: Papers in International Economics in Honor of Charles P. Kindleberger, Amsterdam: North-Holland, 1971.

Blackorby, Charles, Daniel Primont, and R. Robert Russell, *Duality, Separability, and Functional Structure: Theory and Economic Applications*, North Holland, New York, 1978.

Brander, James and Paul Krugman, "A 'Recirpocal Dumping' Model of International Trade," *Journal of International Economics*, 1983, *15*, 313–321.

Broda, Cristian and David Weinstein, "Globalization and the Gains from Variety," *Quarterly Journal of Economics*, 2006, *121* (2), 541–585.

Burstein, Ariel and Gita Gopinath, "International prices and exchange rates," in Gita Gopinath, Elhanan Helpman, and Kenneth Rogoff, eds., *Handbook of International Economics*, Vol. 4, North Holland, 2014, chapter 7.

Chen, Natalie, Jean Imbs, and Andrew Scott, "The Dynamics of Trade and Competition," *Journal of International Economics*, 2009, *77* (1), 50–62.

Comin, D., D. Lashkari, and M. Mestieri, "Structural Change and Long-Run Income and Price Effects," *mimeo*, 2015.

Costinot, Arnaud and Andres Rodriguez-Clare, "Trade Theory with Numbers: Quantifying the Consequences of Globalization," in Gita Gopinath, Elhanan Helpman, and Kenneth Rogoff, eds., *Handbook of International Economics*, Vol. 4, North Holland, 2014, chapter 4.

_ , Dave Donaldson, and Ivana Komunjer, "What Goods Do Countries Trade? A Quantitative Exploration of Ricardo's Ideas," *Review of Economic Studies*, 2012, *79* (2), 581–608.

de Blas, Beatriz and Katheryn Niles Russ, "Understanding Markups in the Open Economy," *American Economic Journal: Macroeconomics*, 2015, *7* (2), 157–180.

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

_ , Pinelopi Goldberg, Amit Khandelwal, and Nina Pavcnik, "Prices, Markups and Trade Reform," *Econometrica*, 2016, *84* (2), 445–510.

Deaton, Angus and John Muellbauer, "An Almost Ideal Demand System," *American Economic Review*, 1980, *70* (3), 312–326.

Dhingra, Swati and John Morrow, "Monopolistic Competition and Optimum Product Diversity under Firm Heterogeneity," *mimeo, LSE*, 2016.

Eaton, Jonathan and Samuel Kortum, "Technology, Geography and Trade," *Econometrica*, 2002, *70* (5), 1741–1779.

_ , _ , and Francis Kramarz, "An Anatomy of International Trade: Evidence from French Firms," *Econometrica*, 2011, *79* (5), 1453–1498.

_ , _ , and Sebastian Sotelo, "International Trade: Linking Micro and Macro," in Daron Acemoglu, Manuel Arellano, and Eddie Dekel, eds., *Advances in Economics and Econometrics, Tenth World Congress (Volume II)*, Cambridge University Press, 2013.

Edmond, Chris, Virgiliu Midrigan, and Daniel Xu, "Competition, Markups and the Gains from Trade," *American Economic Review*, 2015, *105* (10), 3183–3221.

Epifani, Paolo and Gino Gancia, "Trade, Markup Heterogeneity and Misallocations," *Journal of International Economics*, 2011, *83* (1), 1–13.

Feenstra, Robert, "Restoring the Product Variety and Pro-competitive Gains from Trade with Heterogeneous Firms and Bounded Productivity," *NBER working paper*, 2014, *19833.*

Feenstra, Robert C., "A Homothetic Utility Function for Monopolistic Competition Models, Without Constant Price Elasticity," *Economics Letters*, 2003, *78* (1), 79–86.

_ **and David E. Weinstein**, "Globalization, Markups, and the U.S. Price Level," *Journal of Political Economy*, 2017, *125* (4).

Foster, Lucia, John Haltiwanger, and Chad Syverson, "Reallocation, Firm Turnover, and Efficiency: Selection on Productivity or Profitability?," *The American Economic Review*, 2008, *98* (1), 394–425.

Galle, Simon, Andres Rodríguez-Clare, and Moises Yi, "Slicing the Pie: Quantifying the Aggregate and Distributional Effects of Trade," 2014. Unpublished manuscript, UC Berkeley.

Ganapati, S., J. Shapiro, and R. Walker, "Energy prices, pass-through, and incidence in U.S. manufacturing," *mimeo Yale University*, 2016.

Harrison, Ann E., "Productivity, Imperfect Competition and Trade Reform. Theory and Evidence," *Journal of International Economics*, 1994, *36* (1-2), 53–73.

Head, Keith and Thierry Mayer, "Gravity Equations: Toolkit, Cookbook, Workhorse," in Gita Gopinath, Elhanan Helpman, and Kenneth Rogoff, eds., *Handbook of International Economics*, Vol. 4, North Holland, 2014, chapter 3.

_ **,** _ **, and Mathias Thoenig**, "Welfare and Trade Without Pareto," *American Economic Review Papers and Proceedings*, 2014, *104* (5), 310–314.

Helpman, Elhanan and Paul Krugman, *Market Structure and Foreign Trade: Increasing Returns, Imperfect Competition, and the International Economy*, Cambridge, Massachusetts: MIT Press, 1985.

_ **and Paul R. Krugman**, *Trade Policy and Market Structure*, Cambridge, Massachussetts: MIT Press, 1989.

Holmes, Thomas J., Wen-Tai Hsu, and Sanghoon Lee, "Allocative Efficiency, Mark-ups, and the Welfare Gains from Trade," *Journal of International Economics*, 2014, *92* (2), 195–206.

Kimball, Miles S, "The Quantitative Analytics of the Basic Neomonetarist Model," *Journal of Money, Credit and Banking*, 1995, pp. 1241–1277.

Klenow, Peter J and Jonathan L Willis, "Real rigidities and nominal price changes," *Economica*, 2016, *83*, 443–472.

Konings, Jozef, Patrick Van Cayseele, and Frederic Warzynski, "The Dynamics of Industrial Mark-Ups in Two Small Open Economies: Does National Competition Policy Matter?," *International Journal of Industrial Organization*, 2001, *19* (5), 841–859.

Krishna, Pravin and Devashish Mitra, "Trade Liberalization, Market Discipline and Productivity Growth: New Evidence from India," *Journal of Development Economics*, 1998, *56* (2), 447–462.

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

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

Levinsohn, James, "Testing the Imports-as-Market-Discipline Hypothesis," *Journal of International Economics*, 1993, *35* (1), 1–22.

Marshall, Alfred, *Principles of Economics, An Introductory Volume*, London: Mcmillan, eighth edition, 1920.

Melitz, Marc and Stephen Redding, "New Trade models, New Welfare Implications," *American Economic Review*, 2015, *105* (3), 1105–1146.

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

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

Milgrom, Paul and Chris Shannon, "Monotone Comparative Statics," *Econometrica*, 1994, *62* (1), 157–180.

Mrazova, Monika and Peter Neary, "Not so demanding: preference structure and firm behavior," *mimeo Oxford University*, 2016.

_ **and** _ , "Selection Effects with Heterogeneous Firms," *mimeo Oxford University*, 2016.

Neary, J. Peter, "International Trade in General Oligopolistic Equilibrium," *Review of International Economics*, 2016, *24* (4), 669–698.

Novy, Dennis, "International Trade Without CES: Estimating Translog Gravity," *Journal of International Economics*, 2013, *89* (2), 271–282.

Ottaviano, Gianmarco I.P., Takatoshi Tabuchi, and Jacques-François Thisse, "Agglomeration and Trade Revisited," *International Economic Review*, 2002, *43* (2), 409–436.

Pierce, Justin R and Peter K Schott, "Concording US harmonized system categories over time," Technical Report, National Bureau of Economic Research 2009.

Pollak, Robert A., "Additive Utility Functions and Linear Engel Curves," *Review of Economic Studies*, 1971, *38* (4), 401–414.

Rodriguez-Lopez, Jose-Antonio, "Prices and Exchange Rates: A Theory of Disconnect," *Review of Economic Studies*, 2011, *78* (3), 1135–1177.

Saure, Philip, "Bounded Love of Variety and Patterns of Trade," *Open Economies Review*, 2012, *23* (4), 645–674.

Simonovska, Ina, "Income Differences and Prices of Tradables: Insights from an Online Retailer," *Review of Economic Studies*, 2015, *82* (4), 1612–1656.

_ **and Michael Waugh**, "The Elasticity of Trade: Estimates and Evidence," *Journal of International Economics*, 2014, *92* (1), 34–50.

Zhelobodko, Evgeny, Sergey Kokovin, Mathieu Parenti, and Jacques-François Thisse, "Monopolistic Competition in General Equilibrium: Beyond the CES," *Econometrica*, 2011, *80* (6), 2765–2784.

Online Appendix for "The Elusive Pro-Competitive Effects of Trade"

Costas Arkolakis, Arnaud Costinot, Dave Donaldson, Andres Rodriguez-Clare

Abstract

This online appendix provides the proofs for various theoretical results (Section A) as well as additional information regarding the empirical estimation and the quantitative exercises in the main paper (Section B).

A Proofs

A.1 Section 2.1

Additively Separable Utility. We first establish that our demand system under $\beta = 0$ encompasses the case of additively separable utility functions considered in Krugman (1979). Using our notation, his model corresponds to a situation in which preferences are represented by a utility function, $U = \int_{\omega \in \Omega} u(q_\omega) d\omega$. The first-order conditions associated with utility maximization imply $u'(q_\omega) = \lambda p_\omega$, where λ is the Lagrangian multiplier associated with the budget constraint. Inverting the first-order conditions implies

$$q_\omega = u'^{-1}(\lambda p_\omega), \tag{A.1}$$

together with the budget constraint,

$$\int_{\omega \in \Omega} p_\omega q_\omega d\omega = y. \tag{A.2}$$

Under $\beta = 0$, equations (2) and (3) in the main text are equivalent to equation (A.2) and $Q = 1$, respectively. In turn, equation (1) in the main text and $Q = 1$ imply $q_\omega = D(p_\omega / P)$. Thus, setting $P \equiv 1/\lambda$ and $D(\cdot) \equiv u'^{-1}(\cdot)$, we see that if utility functions are additively separable, then the associated demand must satisfy equations (1)-(3) in the main text.

When $\beta = 0$, one can further show that the converse also holds. That is, if the demand system satisfies equations (1)-(3) in the main text, then the utility function of the representa-

tive agent must be additively separable. To see this, note that since $D(\cdot)$ is strictly decreasing, equation (1) in the main text implies

$$p_\omega = PD^{-1}(q_\omega).$$

From the first-order conditions associated with utility maximization we know that

$$dU/dq_\omega = \lambda p_\omega.$$

The two previous expressions imply that for any pair of goods, $\omega_1$and ω_2,

$$\frac{dU/dq_{\omega_1}}{dU/dq_{\omega_2}} = \frac{D^{-1}(q_{\omega_1})}{D^{-1}(q_{\omega_2})}.$$

Thus the Leontief-Sono condition for separability (Blackorby et al. (1978), p.53) is satisfied:

$$\frac{d}{dq_{\omega_3}}\left(\frac{dU/dq_{\omega_1}}{dU/dq_{\omega_2}}\right) = 0 \text{ for any } \omega_3 \neq \omega_1, \omega_2.$$

The fact that U is additively separable, up to a monotonic transformation, then follows from the Representation Theorem 4.8 in Blackorby et al. (1978), p. 136.

Kimball Preferences. We now show that our demand system under $\beta = 1$ encompasses the case of Kimball preferences. Under Kimball preferences, utility Q from consuming $\{q_\omega\}_{\omega\in\Omega}$ is implicitly given by

$$\int \Upsilon\left(\frac{q_\omega}{Q}\right) d\omega = 1, \tag{A.3}$$

for some function Υ that satisfies $\Upsilon' > 0$ and $\Upsilon'' < 0$. The utility maximization program of the consumer is to $\max_{Q,\{q_\omega\}} Q$ subject to equations (A.3) and (A.2). Let γ and λ denote the Lagrange multipliers associated with these two constraints. Manipulating the first-order conditions of this problem we get

$$q_\omega = Q\Upsilon'^{-1}\left(\frac{\lambda \int q_\omega \Upsilon'\left(\frac{q_\omega}{Q}\right) d\omega}{Q} p_\omega\right) \text{ for all } \omega. \tag{A.4}$$

The demand system under Kimball preferences is characterized by equations (A.2)-(A.4). Under $\beta = 1$, equations (2) and (3) in the main text are equivalent to $\int_{\omega\in\Omega} H(p_\omega/P)\, d\omega = 1$ and equation (A.2), respectively. Thus, setting $P \equiv Q/\left(\lambda \int q_\omega \Upsilon'\left(\frac{q_\omega}{Q}\right) d\omega\right)$, $D(\cdot) \equiv \Upsilon'^{-1}(\cdot)$, and $H(\cdot) \equiv \Upsilon(D(\cdot))$, our demand system with $\beta = 1$ replicates the demand system under Kimball preferences.

QMOR Expenditure. Finally, we show that our demand system under $\beta = 1$ also encompasses the demand system corresponding to QMOR expenditure functions in Feenstra (2014). The QMOR demand system entails $q_\omega = QD(p_\omega/P)$ with

$$D(x) \equiv \begin{cases} \varsigma x^{r-1}\left[1 - x^{-r/2}\right] & if \quad x \leq 1 \\ 0 & if \quad x > 1 \end{cases}, \tag{A.5}$$

where P acts as a choke price defined implicitly by

$$P = \left(\left(\frac{N}{N - \left(\widetilde{N} - \varsigma/\varrho \right)} \right)^{r/2} \int_{p_\omega \leq P} \frac{1}{N} p_\omega^{r/2} d\omega \right)^{2/r}, \tag{A.6}$$

and where Q is determined such that the budget constraint (A.2) is satisfied.[A.1] In the previous expressions, ς and ϱ are parameters, $\widetilde{N} \equiv \int_\Omega d\omega$ is the measure of all possible goods, $N \equiv \int_{p_\omega \leq P} d\omega$ is the measure of the set of goods with prices equal or below the choke price P. To proceed, note that equation (A.6) can be rearranged as

$$1 = \frac{1}{N - \left(\widetilde{N} - \varsigma/\varrho \right)} \int_{p_\omega \leq P} \left(\frac{p_\omega}{P} \right)^{r/2} d\omega. \tag{A.7}$$

To conclude, let us show that this is equivalent to equation (2) in the main text under $\beta = 1$ if one sets

$$H\left(\frac{p_\omega}{P} \right) \equiv \frac{1}{\varsigma \left(\varsigma/\varrho - \widetilde{N} \right)} \left(\frac{p_\omega}{P} \right)^{1-r/2} D\left(\frac{p_\omega}{P} \right).$$

Together with the definition of $D(\cdot)$ in equation (A.5), the previous definition implies

$$\int_\Omega H\left(\frac{p_\omega}{P} \right) d\omega = \frac{1}{\varsigma/\varrho - \widetilde{N}} \int_{p_\omega \leq P} \left[\left(\frac{p_\omega}{P} \right)^{r/2} - 1 \right] d\omega.$$

Thus, as argued above, $\int_\Omega H\left(\frac{p_\omega}{P} \right) d\omega = 1$ is equivalent to equation (A.7).[A.2]

[A.1] Equations (A.5) and (A.6) are the counterparts of equations (7) and (2) in Feenstra (2014), respectively.

[A.2] Since the translog expenditure system is a special case of QMOR expenditure functions, as shown in Feenstra (2014), this establishes that our demand system encompasses the translog case. But it is useful to show directly that our demand system leads to translog demand if we set $D(x) \equiv \zeta x^{-1} \ln x^{-1}$ for $x \leq 1$ and $D(x) = 0$ otherwise, with ζ some positive constant, and $H(x) \equiv xD(x)$. Equation (2) in the main text with $\beta = 1$ then implies $\int_{p_\omega \leq P} \zeta \ln(p_\omega/P)^{-1} d\omega = 1$, which is equivalent to

$$\ln P = \frac{1}{\zeta N} + \frac{1}{N} \int_{p_\omega \leq P} \ln p_\omega d\omega,$$

which is the condition that determines P in the translog demand; see equation (8) in Feenstra (2014). Equa-

Homothetic Preferences. In Section 2.1 we have also argued that if $D(\cdot)$ satisfies Assumption A1, then consumers have homothetic preferences if and only if $\beta = 1$. We now establish this result formally. Throughout this proof we will repeatedly use the fact that preferences are homothetic if and only if the income elasticity, $\partial \ln q_\omega(\boldsymbol{p}, y)/\partial \ln y$, is equal to one for all goods $\omega \in \Omega$.

Suppose first that $\beta = 1$. Then equation (2) in the main text implies $\int_{\omega\in\Omega} H(p_\omega/P)\, d\omega = 1$, so $P(\boldsymbol{p}, y)$ is independent of y. Differentiating equation (1) in the main text, we therefore get:

$$\frac{\partial \ln q_\omega(\boldsymbol{p}, y)}{\partial \ln y} = \frac{\partial \ln Q(\boldsymbol{p}, y)}{\partial \ln y}.$$

But Equation (3) in the main text implies $\frac{\partial \ln Q(\boldsymbol{p},y)}{\partial \ln y} = 1$, hence the income elasticity is equal to one for all goods $\omega \in \Omega$, so preferences are homothetic.

Now suppose that $\beta = 0$. As established above, this requires additively separable utility functions. From Bergson (1936), we also know that such functions are homothetic only if they are CES. Since Assumption A1 rules out the CES case, we conclude that preferences cannot be homothetic if $\beta = 0$.

A.2 Section 3.1

In Section 3.1 we have argued that more efficient firms charge higher markups, $\mu' > 0$, if and only if $\varepsilon_D' > 0$.

Suppose first that $\varepsilon_D' > 0$. Let $f(m, v) \equiv m - \frac{\varepsilon_D(m/v)}{\varepsilon_D(m/v)-1}$. Equation (5) in the main text entails $f(m, v) = 0$. Differentiating with respect to m and v, we obtain

$$\begin{aligned}
\frac{\partial f(m,v)}{\partial m} &= 1 + \frac{\varepsilon_D'(m/v)}{(\varepsilon_D(m/v)-1)^2}\frac{1}{v} > 0, \\
\frac{\partial f(m,v)}{\partial v} &= -\frac{\varepsilon_D'(m/v)}{(\varepsilon_D(m/v)-1)^2}\frac{m}{v^2} < 0,
\end{aligned}$$

where the two inequalities derive from $\varepsilon_D' > 0$. By the Implicit Function Theorem, equation (5) therefore implies $\mu'(v) = -\left(\partial f(m,v)/\partial v\right) / \left(\partial f(m,v)/\partial m\right) > 0$.

Now suppose that $\mu' > 0$. We proceed by contradiction. If $\varepsilon_D' \leq 0$, then $\mu'(v) = -\left(\partial f(m,v)/\partial v\right) / \left(\partial f(m,v)/\partial m\right) > 0$ implies

$$1 + \frac{\varepsilon_D'(m/v)}{(\varepsilon_D(m/v)-1)^2}\frac{1}{v} < 0.$$

tion (3) in the main text with $\beta = 1$ is just the budget constraint, which given equation (2) in the main text immediately implies $Q = w/P$.

Using the fact that $m = \frac{\varepsilon_D(m/v)}{\varepsilon_D(m/v)-1}$, this can be rearranged as

$$\varepsilon_D(m/v)\left(\varepsilon_D(m/v)-1\right)+(m/v)\varepsilon_D'(m/v)<0.$$

By definition, $\varepsilon_D(x) = -xD'(x)/D(x)$, which implies

$$\varepsilon_D'(x) = -\frac{D''(x)x}{D(x)} - \frac{D'(x)}{D(x)} + \frac{\left(D'(x)\right)^2 x}{\left(D(x)\right)^2}, \text{ for all } x.$$

Using this expression, we can rearrange the above inequality as

$$2\left(D'(m/v)\right)^2 - D(m/v)D''(m/v) < 0.$$

From the second-order condition of the firm's profit maximization problem, we know that

$$2(\partial q(p,Q,P)/\partial p) + (p-c)(\partial^2 q(p,Q,P)/\partial p^2) \leq 0$$

Together with the first-order condition, $(p-c)/p = -1/(\partial \ln q(p,Q,P)/\partial \ln p)$, this implies

$$2\left(\partial q(p,Q,P)/\partial p\right)^2 - q(p)(\partial^2 q(p,Q,P)/\partial p^2) \geq 0.$$

Using equation (4) in the main text, $m = p/c$, and $v = P/c$, we therefore have

$$2\left(D'\left(m/v\right)\right)^2 - D\left(m/v\right)D''\left(m/v\right) \geq 0,$$

a contradiction.

A.3 Section 3.3

In Section 3.3, we have argued that once models with variable markups considered in this paper are calibrated to match the trade elasticity θ and the observed trade flows $\{X_{ij}\}$, they must predict the exact same changes in wages and trade flows for any change in variable trade costs as gravity models with CES utility, such as Krugman (1980), Eaton and Kortum (2002), Anderson and Van Wincoop (2003), and Eaton et al. (2011). We now establish this result formally.

Relative to ACR, their restriction R1 follows from equation (15), R2 from equation (10),

and R3' from equation (16) in the main text. Combining these three conditions, we obtain

$$\begin{aligned} \lambda_{ij} &= \frac{N_i b_i^\theta \left(w_i \tau_{ij}\right)^{-\theta}}{\sum_k N_k b_k^\theta \left(w_k \tau_{kj}\right)^{-\theta}}, \\ w_i L_i &= \sum_j \lambda_{ij} w_j L_j, \end{aligned}$$

with N_i invariant to changes in trade costs, as established in equation (12) in the main text. These are the same equilibrium conditions as in gravity models with CES utility in ACR. To show that counterfactual changes in wages and trade flows only depend on trade flows and expenditures in the initial equilibrium as well as the value of the trade elasticity, we can use the same argument as in the proof of Proposition 2 in ACR. Consider a counterfactual change in variable trade costs from $\tau \equiv \{\tau_{ij}\}$ to $\tau' \equiv \left\{\tau'_{ij}\right\}$. Let $\hat{x} \equiv x'/x$ denote the change in any variable x between the initial and the counterfactual equilibrium. Since N_i is fixed for all i, one can show that $\{\hat{w}_i\}_{i \neq j}$ are implicitly given by the solution of

$$\hat{w}_i = \sum_{j'=1}^n \frac{\lambda_{ij'} \hat{w}_{j'} Y_{j'} \left(\hat{w}_i \hat{\tau}_{ij'}\right)^{-\theta}}{Y_i \sum_{i'=1}^n \lambda_{i'j'} \left(\hat{w}_{i'} \hat{\tau}_{i'j'}\right)^{-\theta}}. \tag{A.8}$$

where $\hat{w}_j = 1$ by choice of numeraire. Given changes in wages, $\{\hat{w}_i\}$, changes in expenditure shares are then given by

$$\hat{\lambda}_{ij} = \frac{\left(\hat{w}_i \hat{\tau}_{ij}\right)^{-\theta}}{\sum_{i'=1}^n \lambda_{i'j} \left(\hat{w}_{i'} \hat{\tau}_{i'j}\right)^{-\theta}}. \tag{A.9}$$

Equations (A.8) and (A.9) imply $\{\hat{w}_i\}$ and $\left\{\hat{\lambda}_{ij}\right\}$ only depend on the value of trade flows and expenditures in the initial equilibrium as well as the trade elasticity. Once changes in expenditure shares, $\left\{\hat{\lambda}_{ij}\right\}$, are known, changes in bilateral trade flows can be computed using the identity, $\hat{X}_{ij} = \hat{\lambda}_{ij} \hat{w}_j$. Thus the same observation applies to changes in bilateral trade flows, which concludes the argument.

A.4 Section 4.2

Invariance of Distribution of Markups. In Section 4.2, we have argued that if markups are an increasing function of firm-level productivity, then the univariate distribution of markups is independent of the level of trade costs. We now establish this result formally. Let $M_{ij}(m; \tau)$ denote the distribution of markups set by firms from country i in country j in a trade equilibrium if trade costs are equal to $\tau \equiv \{\tau_{ij}\}$. Since firm-level markups only depend on the

relative efficiency of firms, we can express

$$M_{ij}(m;\tau) = \Pr\left\{\mu\left(v\right) \leq m | v \geq 1\right\},$$

where the distribution of v depends, in principle, on the identity of both the exporting and the importing country. Recall that $v \equiv P/c$ and $c = c_{ij}/z$. Thus for a firm with productivity z located in i and selling in j, we have $v = P_j z/c_{ij} = z/z_{ij}^*$. Combining this observation with Bayes' rule, we can rearrange the expression above as

$$M_{ij}(\mu;\tau) = \frac{\Pr\left\{\mu(z/z_{ij}^*) \leq m, z_{ij}^* \leq z\right\}}{\Pr\left\{z_{ij}^* \leq z\right\}}.$$

Using Assumption A2 and the fact that $\mu\left(\cdot\right)$ is monotone, we can rearrange the previous expression as

$$M_{ij}(m;\tau) = \frac{\int_{z_{ij}^*}^{z_{ij}^*\mu^{-1}(m)} dG_i(z)}{\int_{z_{ij}^*}^{\infty} dG_i(z)} = 1 - \left(\mu^{-1}\left(m\right)\right)^{-\theta}.$$

Since the function $\mu\left(\cdot\right)$ is identical across countries and independent of τ, by equation (5) in the main text, this establishes that for any exporter i and any importer j, the distribution of markups $M_{ij}(\cdot;\tau)$ is independent of the identity of the exporter i, the identity of the importer j, and the level of trade costs τ. As a result, the overall distribution of markups in any country j is also invariant to changes in trade costs.

Domestic Markups and Misallocation. In Section 4.2, we have argued that changes in domestic markups, $\rho\lambda_{jj}d\ln P_j$, are proportional to the opposite of the covariance between firm-level markups on the domestic market and changes in firm-level employment shares for that market. We now establish this result formally.

Let us denote by $L_{jj}\left(z\right)$ the number of workers allocated by a firm with productivity z in country j to production of goods for market j. We must have

$$L_{jj}\left(z\right) = \tau_{jj} q_{jj}\left(z\right)/z,$$

where $q_{jj}\left(z\right)$ is such that

$$q_{jj}\left(z\right) = Q_j D\left(z_{jj}^*\mu(z/z_{jj}^*)/z\right).$$

Similarly, let us denote by $\sigma_{jj}\left(z\right) \equiv L_{jj}\left(z\right)/L_{jj}$ denote the employment share that goes to a

firm with productivity z. We have

$$\sigma_{jj}(z) = \frac{D\left(z_{jj}^{*}\mu(z/z_{jj}^{*})/z\right)/z}{\int_{z_{jj}^{*}}^{\infty} N_j D\left(z_{jj}^{*}\mu(z'/z_{jj}^{*})/z'\right)/z' dG_j(z')}.$$

Let us now compute the average of markups, $\bar{m}_{jj} \equiv \int_{z_{jj}^{*}}^{\infty} m_{jj}(z)\,\sigma_{jj}(z)\,N_j dG_j(z)$, for firms from country j selling in country j weighted by employment. We have:

$$\bar{m}_{jj} = \int_{z_{jj}^{*}}^{\infty} m_{jj}(z) \frac{D\left(z_{jj}^{*}m_{jj}(z)/z\right)/z}{\int_{z_{ij}^{*}}^{\infty} D\left(z_{jj}^{*}m_{jj}(z')/z'\right)/z' dG_j(z')} dG_j(z).$$

Under Assumption A2, we can rearrange the previous expression as

$$\bar{m}_{jj} = \int_{1}^{\infty} \mu(v) \frac{D(\mu(v)/v)\,v^{-\theta-2}dv}{\int_{1}^{\infty} D(\mu(v')/v')\,(v')^{-\theta-2}dv'}.$$

This implies

$$\frac{d\bar{m}_{jj}}{dz_{jj}^{*}} = \int_{z_{jj}^{*}}^{\infty} \frac{dm_{jj}(z)}{dz_{jj}^{*}} \sigma_{jj}(z) N_j dG_j(z) + \int_{z_{jj}^{*}}^{\infty} m_{jj}(z) \frac{d\sigma_{jj}(z)}{dz_{jj}^{*}} N_j dG_j(z) = 0,$$

where we have used the fact that $\sigma_{jj}\left(z_{jj}^{*}\right) = 0$. The first term can be rearranged as

$$\int_{z_{jj}^{*}}^{\infty} \frac{dm_{jj}(z)}{dz_{jj}^{*}} \sigma_{jj}(z) N_j dG_j(z) = -\frac{\rho\bar{m}_{jj}}{z_{jj}^{*}}.$$

By construction, $\int_{z_{jj}^{*}}^{\infty} \sigma_{jj}(z) N_j dG_j(z) = 1$. Using again $\sigma_{jj}\left(z_{jj}^{*}\right) = 0$, we therefore have $\int_{z_{jj}^{*}}^{\infty} \frac{d\sigma_{jj}(z)}{z_{jj}^{*}} N_j dG_j(z) = 0$. Thus the second term can be rearranged as

$$\int_{z_{jj}^{*}}^{\infty} m_{jj}(z) \frac{d\sigma_{jj}(z)}{dz_{jj}^{*}} N_j dG_j(z) = \int_{z_{jj}^{*}}^{\infty} \left(m_{jj}(z) - \bar{m}_{jj}\right) \left(\frac{d\sigma_{jj}(z)}{dz_{jj}^{*}} - 0\right) N_j dG_j(z),$$

Combining the three previous expressions we therefore get

$$\frac{\rho\bar{m}_{jj}}{z_{jj}^{*}} = \int_{z_{jj}^{*}}^{\infty} \left(m_{jj}(z) - \bar{m}_{jj}\right) \left(\frac{d\sigma_{jj}(z)}{dz_{jj}^{*}} - 0\right) N_j dG_j(z).$$

To conclude note that $z_{jj}^{*} = 1/P_j$, by our choice of numeraire. Thus the previous expression

implies

$$\rho\lambda_{jj}d\ln P_j = -\left(\frac{\lambda_{jj}}{\bar{m}_{jj}}\right)\left(\int_{z_{jj}^*}^{\infty}\left(m_{jj}\left(z\right)-\bar{m}_{jj}\right)\left(d\sigma_{jj}\left(z\right)-0\right)N_j dG_j\left(z\right)\right),$$

where the integral on the right-hand side is equal to the covariance between firm-level markups on the domestic market and changes in firm-level employment shares for that market.

Pro-Competitive Effects in Krugman (1979). In Section 4.2, we have argued that, ceteris paribus, the pro-competitive effects in Krugman (1979) are positive if an increase in country size raises output per firm, and firms were producing too little before market integration, or it lowers output, and they were producing too much. We now establish this result formally.

Consider a closed economy with a measure L of identical agents with additively separable preferences over a continuum of symmetric varieties,

$$U = Nu(q/L)$$

where N is the measure of available varieties and q is total output per variety. Let $c(q)$ denote the total labor cost of producing q units of a given variety. In Krugman (1979), $c(q) = f + q$ if $q > 0$ and zero otherwise. Let $\pi(q, L)$ denote the profit of a representative firm given total output, q, and market size, L. In Krugman (1979), $\pi(q, L) = \frac{\epsilon_D(q/L)}{\epsilon_D(q/L)-1}c'(q)q - c(q)$, where $\epsilon_D(q/L) \equiv -\frac{u'(q/L)}{(q/L)u''(q/L)}$ denotes the elasticity of demand faced by each firm as a function of consumption per capita, q/L.[A.3]

To study the welfare implications of an increase in market size, it is convenient to focus on the following constrained planning problem:

$$V(L, W) = \max_{N,q} Nu(Wq/L)$$

subject to

$$Nc(q) = L, \tag{A.10}$$

$$\pi(q, L) = 0 \tag{A.11}$$

Equations (A.10) and (A.11) correspond to the resource constraint and the free entry condition, respectively. By construction, (q, N) in the decentralized equilibrium is equal to the solution to the constrained planning problem for $W = 1$.

[A.3] By definition, we have $\epsilon_D(q/L) = \varepsilon_D(u'(q/L))$, where ε_D is the elasticity of demand as function of price used in the main text.

We are interested in computing the percentage change in income, $d \ln W$, equivalent to a percentage change in market size, $d \ln L$, i.e.,

$$d \ln W = \left(\frac{L}{W}\frac{dV/dL}{dV/dW}\right)_{W=1} d \ln L.$$

Let $q(L)$ denote the output level that solves equation (A.11). By the Envelope Theorem, we have

$$\frac{dV}{dL} = -\frac{NWq(L)u'(Wq(L)/L)}{L^2} + \lambda + q'(L)\left(\frac{NWu'(Wq(L)/L)}{L} - \lambda N c'(q(L))\right)$$
$$\frac{dV}{dW} = \frac{Nq(L)u'(Wq(L)/L)}{L},$$

where λ is the Lagrange multiplier associated with equation (A.10). This leads to

$$d \ln W = \left(-1 + \frac{\left(\lambda + q'(L)\left(\frac{NWu'(Wq(L)/L)}{L} - \lambda N c'(q(L))\right)\right) L^2}{Nq(L)u'(q(L)/L)}\right) d \ln L. \tag{A.12}$$

The first-order condition with respect to N, evaluated at $W = 1$, further implies $u(q/L) = \lambda c(q)$. Together with the resource constraint, we therefore have $\lambda = Nu(q(L)/L)/L$. Substituting for the Lagrange multiplier, λ, in equation (A.12), we therefore obtain, after simplifications,

$$d \ln W = \left(\frac{1 - \epsilon_u}{\epsilon_u} + \epsilon_q \vartheta\right) d \ln L, \tag{A.13}$$

where $\epsilon_u(x) \equiv \left(\frac{d \ln u}{d \ln x}\right)_{x=q(L)/L}$, $\epsilon_q \equiv \frac{d \ln q(L)}{d \ln L}$, and $\vartheta = \frac{u'(q(L)/L) - Nu(q(L)/L)c'(q(L))}{u'(q(L)/L)}$ captures the wedge between the marginal benefit of increasing output per variety, $(N/L)u'(q(L)/L)$, and its marginal cost, $\lambda c'(q(L)) = (N^2/L)u(q(L)/L)c'(q(L))$.

In the case with constant markups and CES utility considered by Krugman (1980), the decentralized equilibrium is efficient, $\vartheta = 0$. Thus gains from market integration only reflects gains from new varieties, as captured by $(1 - \epsilon_u)/\epsilon_u$.[A.4] Accordingly, we can express the pro-competitive effects from trade, defined as the differential impact of trade liberalization on welfare when markups vary and when they do not, as $\Delta(\frac{1-\epsilon_u}{\epsilon_u}) + \epsilon_q \vartheta$, where $\Delta(\frac{1-\epsilon_u}{\epsilon_u})$ denotes the difference between the welfare gains from new varieties in models with and without variable markups (a difference that depends, in general, on which moments one chooses to hold fix when comparing these models). For a given value of $\Delta(\frac{1-\epsilon_u}{\epsilon_u})$, the previous analysis establishes that welfare gains from market integration will be higher if an increase in market size raises output per firm, $\epsilon_q > 0$, and firms were producing too little

[A.4] In the CES case, one can also check that $\epsilon_u = \frac{\epsilon_D - 1}{\epsilon_D}$. Thus, the gains from market integration can be rearranged in a familiar way as $d \ln W = d \ln L/(\epsilon_D - 1)$.

before market integration, $\vartheta > 0$, or it lowers output,$\epsilon_q < 0$, and they were producing too much, $\vartheta < 0$.

A.5 Section 4.3

In the multi-sector case, and ignoring for now the country sub-index, the expenditure minimization problem of the representative consumer is given by

$$\begin{aligned} e(\boldsymbol{p}, U) \equiv \min_{\boldsymbol{q}} & \sum_k \int_{\Omega^k} p^k(\omega) q^k(\omega) d\omega \\ \text{s.t.} \quad & U(C^1(\boldsymbol{q}^1), ..., C^K(\boldsymbol{q}^K)) \geq U. \end{aligned}$$

Since preferences are weakly separable, the solution to the previous problem can be computed in two stages. At the lower stage, the optimal consumption of varieties within each sector solves

$$\begin{aligned} e^k(\boldsymbol{p^k}, C^k) \equiv \min_{\boldsymbol{q}^k} & \int_{\Omega^k} p^k(\omega) q^k(\omega) d\omega \\ \text{s.t.} \quad & C^k(\boldsymbol{q}^k) \geq C^k. \end{aligned}$$

At the upper stage, the optimal level of consumption between sectors solves

$$\begin{aligned} e(\boldsymbol{p}, U) \equiv \min_{C^1, ..., C^K} & \sum_k e^k(\boldsymbol{p^k}, C^k) \\ \text{s.t.} \quad & U(C^1, ..., C^K) \geq U. \end{aligned}$$

We are interested in $d \ln W = d \ln y - d \ln e$, with y being per-capita income. By Shephard's lemma, we know that a foreign shock implies that

$$d \ln e = \sum_k s^k d \ln e^k. \tag{A.15}$$

To compute $d \ln y$ and $d \ln e^k$, we consider separately the cases of restricted and free entry.

Restricted entry. Under restricted entry equation (17) in the main text remains valid at the sector level. So we can use the exact same approach as in the one-sector case to derive

$$d \ln e_j^k = \left(1 - \rho^k\right) \sum_i \lambda_{ij}^k d \ln c_{ij}^k + \rho^k d \ln P_j^k. \tag{A.16}$$

To compute $d \ln P_j^k$, we use the sector-level counterpart of equations (22)-(23) in the main

text, which imply

$$\begin{aligned}
\kappa^k \left(Q_j^k\right)^{1-\beta^k} \left(P_j^k\right)^{\theta^k+1-\beta^k} \left(\sum_i N_i^k \left(b_i^k\right)^{\theta^k} \left(c_{ij}^k\right)^{-\theta^k}\right) &= \left(y_j^k\right)^{1-\beta^k}, \\
\left(\chi^k\right)^{\beta^k} Q_j^k \left(P_j^k\right)^{\beta^k\left(1+\theta^k\right)} \left(\sum_i N_i^k \left(b_i^k\right)^{\theta^k} \left(c_{ij}^k\right)^{-\theta^k}\right)^{\beta^k} &= \left(y_j^k\right)^{\beta^k},
\end{aligned}$$

with

$$\begin{aligned}
\kappa^k &\equiv \theta^k \int_1^\infty \left[H^k\left(\mu^k(v)/v\right)\right]^{\beta^k} \left[\left(\mu^k(v)/v\right) D^k(\mu^k(v)/v)\right]^{1-\beta^k} v^{-1-\theta^k} dv, \\
\chi^k &\equiv \theta^k \int_1^\infty \left(\mu^k(v)/v\right) D^k(\mu^k(v)/v) v^{-\theta^k-1} dv.
\end{aligned}$$

From the two previous equations, we obtain

$$P_j^k = \left(\frac{\kappa^k \sum_i N_i^k \left(b_i^k\right)^{\theta^k} \left(c_{ij}^k\right)^{-\theta^k}}{\left(y_j^k\right)^{1-\beta^k}}\right)^{-1/\left(\theta^k+1-\beta^k\right)}, \tag{A.17}$$

and in turn, under restricted entry,

$$d\ln P_j^k = \frac{\theta^k}{\theta^k+1-\beta^k} \sum_i \lambda_{ij}^k d\ln c_{ij}^k + \frac{1-\beta^k}{\theta^k+1-\beta^k} d\ln y_j^k.$$

Together with equations (A.15) and (A.16), the previous expression yields

$$d\ln e_j = \sum_{i,k} s_j^k \lambda_{ij}^k \left(1-\eta^k\right) d\ln c_{ij}^k + \sum_k s_j^k \eta^k d\ln y_j^k,$$

with $\eta^k \equiv \rho^k\left((1-\beta^k)/(1-\beta^k+\theta^k)\right)$. Using the fact that $y_j^k = s_j^k y_j$, we can rearrange the second term on the right-hand side as

$$\sum_k s_j^k \eta^k \left(d\ln s_j^k + d\ln y_j\right) = \sum_k \eta^k d s_j^k + \eta_j d\ln y_j,$$

with $\eta_j \equiv \sum_k s_j^k \eta^k$. Since $d\ln W_j = d\ln y_j - d\ln e_j$, we get

$$d\ln W_j = \left(1-\eta_j\right) d\ln y_j - \sum_{i,k} \left(1-\eta^k\right) s_j^k \lambda_{ij}^k d\ln c_{ij}^k - \sum_k \eta^k d s_j^k. \tag{A.18}$$

Proceeding as in the one sector case, one can show that $\sum_i \lambda_{ij}^k d\ln c_{ij}^k$ is equal to $d\ln \lambda_{jj}^k/\theta^k$. To establish equation (30) in the main text, we therefore only need to solve for $d\ln y_j$. Under restricted entry, per-capita income in country j is given by $y_j = 1 + \sum_{i,k} \Pi_{ji}^k / L_j$, where we have set $w_j = 1$ by choice of numeraire. As in the one-sector case, sector-level profits are such that $\Pi_{ji}^k = \zeta^k X_{ji}^k$, with

$$\begin{aligned}
&\zeta^k \equiv \pi^k/\chi^k, \\
&\pi^k \equiv \theta^k \int_1^\infty \left(\mu^k(v) - 1\right) D^k(\mu^k(v)/v) v^{-\theta^k - 2} dv > 0.
\end{aligned}$$

As in the one-sector case, under restricted entry and with $w_j = 1$, sector-level employment is such that $L_j^k = (1-\zeta^k)(\sum_i X_{ji}^k)$. Combining the previous observations, we obtain

$$d\ln y_j = d\ln(\sum_k L_j^k/(1-\zeta^k)).$$

Plugging into (A.18), we obtain equation (30) in the main text. Proposition 2 derives from this expression and the joint observation that $\eta^k = \eta$ for all k implies $\sum_k \eta^k ds_j^k = \eta \sum_k ds_j^k = 0$ whereas $\zeta^k = \zeta$ for all k implies $d\ln y_j = d\ln L_j/(1-\zeta) = 0$.

Free Entry. Under free entry, equation (17) in the main text is no longer valid since we may have $d\ln N_i^k \neq 0$ for some i and k. To capture the welfare implications of the previous changes, we restrict ourselves to the three examples of demand functions discussed in Section 2.1: (i) additively separable utility functions; (ii) quadratic mean of order r (QMOR) expenditure functions; and (iii) Kimball preferences.

We first consider the case of additively separable utility functions and Kimball preferences. Under both cases, using Assumption A2, we can write the sector-level expenditure function as

$$\begin{aligned}
e_j^k = &\min_{\boldsymbol{q}_j^k} \sum_i \int_{b_i^k}^\infty p_{ij}^k(z) q_{ij}^k(z) \theta^k \left(b_i^k\right)^{\theta^k} N_i^k z^{-\theta^k - 1} dz \\
s.t. &\sum_i \int_{b_i^k}^\infty \Psi_j^k \left(q_{ij}^k(z) / \left(C_j^k\right)^{\beta^k}\right) \theta^k \left(b_i^k\right)^{\theta^k} N_i^k z^{-\theta^k - 1} dz \geq \left(C_j^k\right)^{1-\beta^k},
\end{aligned}$$

where $p_{ij}^k(z)$ is the price in country j of a variety with productivity z in sector k produced in country i and $q_{ij}^k(z)$ is the corresponding quantity. In the case of additively separable utility functions, we have $\beta_j^k = 0$ and the function Ψ_j^k is country j's sub-utility function u_j^k, while in the case of Kimball preferences we have $\beta_j^k = 1$ and the function Ψ_j^k is the sector-level counterpart of the function Y in Appendix A.1. Using the change of variable $\tilde{z} = N_i^k (b_i^k/z)^{\theta^k}$

and letting $\tilde{N}_i^k \equiv N_i^k \left(b_i^k\right)^{\theta^k}$, we now have

$$e_j^k = \min_{\boldsymbol{q}_j^k} \sum_i \int_0^{\tilde{N}_i^k \left(b_i^k\right)^{-\theta^k}} p_{ij}^k \left(\left(\tilde{N}_i^k / \tilde{z} \right)^{1/\theta^k} \right) q_{ij}^k \left(\left(\tilde{N}_i^k / \tilde{z} \right)^{1/\theta^k} \right) d\tilde{z}$$

$$s.t. \sum_i \int_0^{\tilde{N}_i^k \left(b_i^k\right)^{-\theta^k}} \Psi_j^k \left(q_{ij}^k \left(\left(\tilde{N}_i^k / \tilde{z} \right)^{1/\theta^k} \right) / \left(C_j^k \right)^{\beta^k} \right) d\tilde{z} \geq \left(C_j^k \right)^{1-\beta^k}.$$

Applying the Envelope Theorem and using the fact that demand is zero for the least productive firm, we get

$$d \ln e_j^k = \sum_i \int_0^{(\tilde{z}_{ij}^k)^*} \lambda_{ij}^k \left(\left(\tilde{N}_i^k / \tilde{z} \right)^{1/\theta^k} \right) d \ln p_{ij}^k \left(\left(\tilde{N}_i^k / \tilde{z} \right)^{1/\theta^k} \right) d\tilde{z}, \tag{A.19}$$

where $(\tilde{z}_{ij}^k)^* = \tilde{N}_i^k \left((z_{ij}^k)^* \right)^{-\theta^k}$ is the (rank) productivity cut-off; $\lambda_{ij}^k \left(\left(\tilde{N}_i^k / \tilde{z} \right)^{1/\theta^k} \right)$ is the expenditure share,

$$\lambda_{ij}^k \left(\left(\tilde{N}_i^k / \tilde{z} \right)^{1/\theta^k} \right) = x^k \left(c_{ij}^k \left(\tilde{N}_i^k / \tilde{z} \right)^{-1/\theta^k}, \left((\tilde{z}_{ij}^k)^* / \tilde{z} \right)^{1/\theta^k}, Q_j^k, L_j^k \right) / y_j^k;$$

and $d \ln p_{ij}^k \left(\left(\tilde{N}_i^k / \tilde{z} \right)^{1/\theta^k} \right)$ is the total derivative of the log price, including both the change in the price schedule conditional on productivity and the change in the normalized measure of entrants, $\tilde{N}_i^k$. To compute the latter, note that $p_{ij}^k(z) = (c_{ij}^k / z) \mu^k(z / z_{ij}^{k*})$, which implies

$$p_{ij}^k \left(\left(\tilde{N}_i^k / \tilde{z} \right)^{1/\theta^k} \right) = c_{ij}^k \left(\tilde{N}_i^k / \tilde{z} \right)^{-1/\theta^k} \mu^k \left(\left((\tilde{z}_{ij}^k)^* / \tilde{z} \right)^{1/\theta^k} \right),$$

and, in turn,

$$d \ln p_{ij}^k \left(\left(\tilde{N}_i^k / \tilde{z} \right)^{1/\theta^k} \right) = d \ln \left(c_{ij}^k \left(\tilde{N}_i^k \right)^{-1/\theta_k} \right) + \rho^k \left(\left((\tilde{z}_{ij}^k)^* / \tilde{z} \right)^{1/\theta^k} \right) d \ln \left((\tilde{z}_{ij}^k)^* \right)^{1/\theta^k},$$

with $\rho^k(z) \equiv d \ln \mu^k(z) / d \ln z$. Noting that $\int_0^{(\tilde{z}_{ij}^k)^*} \lambda_{ij}^k \left(\left(\tilde{N}_i^k / \tilde{z} \right)^{1/\theta^k} \right) d\tilde{z} = \lambda_{ij}^k$ and substituting into equation (A.19), we get

$$d \ln e_j^k = \sum_i \lambda_{ij}^k d \ln \left(c_{ij}^k \left(\tilde{N}_i^k \right)^{-1/\theta_k} \right) + \sum_i \rho^k \lambda_{ij}^k d \ln \left((\tilde{z}_{ij}^k)^* \right)^{1/\theta^k}, \tag{A.20}$$

with

$$\rho^k = \int_0^{(\tilde{z}_{ij}^k)^*} \rho^k \left(\left((\tilde{z}_{ij}^k)^* / \tilde{z} \right)^{1/\theta^k} \right) \frac{\lambda_{ij}^k \left(\left(\tilde{N}_i^k / \tilde{z} \right)^{1/\theta^k} \right)}{\lambda_{ij}^k} d\tilde{z}.$$

Note that in line with equation (20) in Section 4.1, a simple change of variable, $v = \left((\tilde{z}_{ij}^k)^* / \tilde{z} \right)^{1/\theta^k}$, implies

$$\rho^k = \int_1^\infty \frac{d \ln \mu^k(v)}{d \ln v} \frac{\left(\mu^k(v)/v \right) D^k(\mu^k(v)/v) v^{-1-\theta^k}}{\int_1^\infty \left(\mu^k(v')/v' \right) D^k \left(\mu^k(v')/v' \right) (v')^{-1-\theta^k} dv'} dv.$$

Since $(\tilde{z}_{ij}^k)^* = \tilde{N}_i^k \left(z_{ij}^{k*} \right)^{-\theta^k}$ and $z_{ij}^{k*} = c_{ij}^k / P_j^k$, equation (A.20) further implies

$$d \ln e_j^k = \sum_i \left(1 - \rho^k \right) \lambda_{ij}^k d \ln \left(c_{ij}^k \left(\tilde{N}_i^k \right)^{-1/\theta^k} \right) + \sum_i \rho^k \lambda_{ij}^k d \ln P_j^k.$$

To compute $d \ln P_j^k$, we can start from equation (A.17), which remains valid under free entry. Log-differentiation yields

$$d \ln P_j^k = \frac{\theta^k}{\theta^k + 1 - \beta^k} \sum_i \lambda_{ij}^k d \ln \left(c_{ij}^k \left(\tilde{N}_i^k \right)^{-1/\theta^k} \right) + \frac{1 - \beta^k}{\theta^k + 1 - \beta^k} d \ln y_j^k. \tag{A.21}$$

Combining the two previous expressions, we obtain

$$d \ln e_j^k = \sum_i \left(1 - \eta^k \right) \lambda_{ij}^k d \ln \left(c_{ij}^k \left(\tilde{N}_i^k \right)^{-1/\theta^k} \right) + \eta^k d \ln y_j^k. \tag{A.22}$$

Combined with equations (A.15) and (A.22), we then have

$$d \ln W_j = d \ln y_j - \sum_{i,k} s_j^k \left(1 - \eta^k \right) \lambda_{ij}^k d \ln \left(c_{ij}^k \left(\tilde{N}_i^k \right)^{-1/\theta^k} \right) - \sum_k s_j^k \eta^k d \ln y_j^k.$$

Under free entry, we know that $y_j = 1$, where we have again set $w_j = 1$ by choice of numeraire. This immediately implies $d \ln y_j = 0$. Given that $y_j^k = s_j^k y_j$, this further implies that $\sum_k s_j^k \eta^k d \ln y_j^k = \sum_k \eta^k d s_j^k$, and hence

$$d \ln W_j = - \sum_{i,k} s_j^k \left(1 - \eta^k \right) \lambda_{ij}^k d \ln \left(c_{ij}^k \left(\tilde{N}_i^k \right)^{-1/\theta^k} \right) - \sum_k \eta^k d s_j^k. \tag{A.23}$$

To conclude, note that sector-level trade flows still satisfy gravity,

$$\lambda_{ij}^k = \frac{\tilde{N}_i^k \left(c_{ij}^k\right)^{-\theta^k}}{\sum_l \tilde{N}_l^k \left(c_{lj}^k\right)^{-\theta^k}},$$

which implies

$$\sum_i \lambda_{ij}^k d \ln \left(\tilde{N}_i^k \left(c_{ij}^k\right)^{-\theta^k}\right) = d \ln \tilde{N}_j^k - d \ln \lambda_{jj}^k. \tag{A.24}$$

Combining this result with equation (A.23) and noting that $N_j^k = \zeta^k (L_j^k / F_j^k)$ implies $d \ln \tilde{N}_j^k = d \ln N_j^k = d \ln L_j^k$, we get

$$d \ln W_j = -\sum_k s_j^k \left(1 - \eta^k\right) \left(d \ln \lambda_{jj}^k - d \ln L_j^k\right) / \theta^k - \sum_k \eta^k d s_j^k.$$

If $\eta^k = \eta$ for all k, this simplifies into equation (31) in the main text.

Finally, consider the case of the QMOR expenditure functions analyzed by Feenstra (2014). Lemma 1 in Feenstra (2014) and the fact that the Herfindahl index is constant when productivity is distributed Pareto together imply that (in our notation) $d \ln e_j^k = d \ln P_j^k$. Combining this observation with equations (A.21) and (A.24), which remain valid in this case, and using the fact that $\beta^k = 1$ and $\eta = 0$ for this case, we again obtain equation (31) from the main text.

B Estimation and quantitative exercises

B.1 Section 5.1

This section describes a number of details behind the procedure used to estimate η from micro trade data that was described in Section 5.1.

From theory to data. We aim to estimate a parametric demand system that satisfies equations (1)-(3) in the main text. Our choice of a particular parameterization is motivated by parsimony, as well as the two following considerations. First, we want to nest the case of CES demand because of its prominence in prior work and because it provides a reference point in which markups will be constant under monopolistic competition. And second, we want to allow the average elasticity of markups—and hence η—to be positive or negative, so that data can speak to whether the existence of variable markups increases or decreases the gains from trade liberalization. In order to achieve these goals, we restrict attention to

additively separable preferences in the "Pollak family"; see Pollak (1971) and Mrazova and Neary (2016a). This implies the following parametric restriction on $D(\cdot)$:

$$D(p_\omega/P) = (p_\omega/P)^{1/\gamma} - \alpha,$$

where α and γ are the two structural parameters to be estimated.[B.1] In turn, the parameter β in equations (2) and (3) in the main text is equal to 0 if $\alpha \neq 0$ and to either 0 or 1 if $\alpha = 0$. Assumption A1 is only satisfied if $\alpha > 0$ but we do not impose this restriction on the estimation.

When $\alpha = 0$, the previous demand system reduces to the CES case, with elasticity of substitution given by $-1/\gamma$. In this case, trade liberalization has no effects on markups and $\eta = 0$. In contrast, when $\alpha > 0$, the demand elasticity is decreasing with the level of consumption, and hence increasing with the level of prices, $\varepsilon_D' > 0$, which implies $\rho > 0$ and $\eta = \rho/(1+\theta) > 0$. Finally, when $\alpha < 0$, the opposite happens, and hence $\varepsilon_D' < 0$ and $\rho < 0$.

Our estimation of this demand system draws on detailed data on bilateral U.S. merchandise imports within narrowly defined product codes to estimate the representative U.S. consumer's demand parameters. In particular, we use annual data (from 1989-2009) at the 10-digit HS level.[B.2] In mapping these data to our model we assume that a variety ω in the model corresponds to a particular 10-digit HS product, indexed by g, from a particular exporting country, indexed by i; that is, a "variety" ω in the model is a "product-country" pair gi in the data.[B.3] There are 13,746 unique products and 242 unique exporters. Because the demand system in equation (4) in the main text is intended to represent demand for varieties within a differentiated sector, we assume that a "sector", which we index by k, in the data is a level of product aggregation that is higher than the 10-digit level and in practice take this to be the 4-digit HS category (of which there are 1387) level . In what follows, we

[B.1] Simonovska (2015) uses the log-version of this demand system to analyze the relationship between income and prices across countries.

[B.2] We download this dataset from Peter Schott's homepage and use the concordances provided in Pierce and Schott (2009) to adjust for changes in 10-digit HS codes over this time period. The July 2015 version of this paper reported results from an earlier dataset spanning 1989-2005 only.

[B.3] While this practice is standard in the literature (e.g. Broda and Weinstein 2006), we note that the issue of "hidden varieties" is more problematic here than in the CES case. Under the assumption of CES demand, the fact that an unobserved number of firms from the same country may be producing a particular 10-digit HS product simply acts as an unobserved quality shifter. This is no longer true if $\alpha \neq 0$. We are unaware of a study that documents the extent of firm-level concentration at the country-HS10-digit level for US imports. But Feenstra and Weinstein (2017) estimate that for US imports in 1998 (the closest among their tabulated years to the mid-point of our sample) the trade-weighted average of the Herfindahl index within exporter-HS 4-digit product groups was 0.190. This would imply, for equally sized firms, about five firms per exporting country within each 4-digit industry. By comparison, on average there are approximately ten 10-digit HS products within each 4-digit group. There is therefore ample scope for the possibility that most exporter-HS10-digit product cells are served by only one firm.

let the price aggregator P_t^k vary across sectors and over time, but restrict (in our baseline analysis) the demand parameters α and γ to be common across all sectors.

We focus on the following empirical demand equation:

$$q_{git}^k = \left(\varepsilon_{git}^k p_{git}^k / P_t^k\right)^{1/\gamma} - \alpha, \tag{B.1}$$

where p_{git}^k is the price paid by U.S. importers when buying quantity q_{git}^k for a product g in sector k from an exporting country i in year t. The import data contain measures of total (that is, aggregated across all importers) expenditure, i.e., the empirical analogue of $q_{git}^k \times p_{git}^k$, and measures of total quantities purchased, which we take as our measure of q_{git}^k. To construct a measure of prices p_{git}^k we therefore simply use the ratio of expenditure to quantity. The variety-specific demand shifter, ε_{git}^k, captures the fact that physical units in the data may differ from the choice of units in Section 2, under which all varieties are implicitly assumed to enter utility in a symmetric fashion. Such differences in units of account can be interpreted as unobserved quality differences; see e.g. Baldwin and Harrigan (2011).

Estimation procedure. There are two key challenges involved in estimating equation (B.1): (i) the price aggregator P_t^k is unobserved and correlated with p_{git}^k; and (ii) the demand shifter ε_{git}^k is unobserved and correlated with p_{git}^k. We describe below, in turn, a procedure to estimate the demand parameters, α and γ, that overcomes these challenges.

First, consider the problem that the price aggregator P_t^k is unobserved and correlated with p_{git}^k. The key restriction imposed in equation (B.1), however, is that the demand for all varieties depends symmetrically on this aggregator; that is, the price aggregator does not vary across products g and exporters i within sector k. This suggests that identification of the demand parameters, α and γ, can be achieved through a differencing procedure designed to eliminate the unobserved and endogenous P_t^k term in equation (B.1). Specifically, inverting our demand function and taking logs, we have

$$\ln p_{git}^k = \gamma \ln(q_{git}^k + \alpha) - \ln P_t^k + \ln \varepsilon_{git}^k.$$

Taking differences with respect to one reference product-country within the same sector k, we then obtain

$$\Delta_{gi} \ln p_{git}^k = \gamma \Delta_{gi} \ln(q_{git}^k + \alpha) + \Delta_{gi} \ln \varepsilon_{git}^k, \tag{B.2}$$

where Δ_{gi} denotes the corresponding difference operator. While in principle the difference Δ_{gi} could be taken across any two product-country gi observations within a sector-year kt, we use the convention of mean differencing such that, for any variable Z, $\Delta_{gi} Z_{git}^k = Z_{git}^k -$

$\frac{1}{M_{kt}} \sum_{gi \in \mathcal{I}_{kt}} Z^k_{git}$ where $\mathcal{I}_{kt}$ is the set of product-country pairs gi in sector k and year t and M_{kt} is the number of observations in this set.

Second, consider the problem posed by the correlation between p^k_{git} and the unobserved demand-shifter, ε^k_{git}. We first follow the literature on demand system estimation using international trade data—e.g. Broda and Weinstein (2006) and Feenstra and Weinstein (2017)—and decompose this demand-shifter into two terms:

$$\ln \varepsilon^k_{git} = \ln \delta^k_{gi} + \ln \epsilon^k_{git}.$$

In this decomposition, the first term, $\ln \delta^k_{gi}$, reflects systematic differences in quality or units of account across products from different countries within a sector, whereas the second term, $\ln \epsilon^k_{gct}$, reflects idiosyncratic determinants of demand that are free to vary over time. To eliminate systematic unobserved differences in quality, we take a second difference of equation (*B*.2), now across time periods, to obtain

$$\Delta_t \Delta_{gi} \ln p^k_{git} = \gamma \Delta_t \Delta_{gi} \ln(q^k_{git} + \alpha) + \Delta_t \Delta_{gi} \ln \epsilon^k_{git}, \tag{B.3}$$

where Δ_t denotes the corresponding difference operator. Again, while the difference Δ_t could be taken across any two time periods we use mean differencing, as in Δ_{gi} defined above. While this double-differencing procedure will remove cross-sectional sources of bias due to unobserved quality shifters, endogeneity bias concerns due to potentially time-varying quality shifters (or measurement error in prices) remain. A natural solution is to use an instrumental variable (IV) approach, where here the instrument must be exogenous with respect to the error term $\Delta_t \Delta_{gc} \ln \epsilon^k_{git}$ and must be correlated with the endogenous variable, i.e. the double-demeaned quantity $\Delta_t \Delta_{gi} \ln(q^k_{git} + \alpha)$, for any value of α. In our model a natural candidate for such an instrument is trade costs. For this purpose we use the (log of one plus the) value of tariff duties charged, expressed as a percentage of import value, as a measure of trade costs; this variable is reported in the US 10-digit HS imports data. This procedure of using trade costs as exogenous demand shifters in an international trade setting is commonly employed in the empirical gravity literature; see e.g. Head and Mayer (2014).

Since the estimating equation (*B*.3) is linear in γ, but non-linear in α, we separate our estimation procedure into an inner-loop and an outer-loop. In the inner-loop, we take the value of α as given and compute $\hat{\gamma}(\alpha)$ as the IV estimator of γ with $\Delta_t \Delta_{gi} \ln(t^k_{git} + \alpha)$ the instrumental variable for $\Delta_t \Delta_{gi} \ln(q^k_{git} + \alpha)$, where t^k_{git} denotes the tariff rate charged by the United States on imports of product g in sector k from country i in year t. In the outer-loop, we then search for the value of α that minimizes the sum of the squared residuals across

	γ	α
Panel A: CES demand		
	-0.206^{***}	
	(0.036)	
Panel B: Generalized CES demand		
	-0.347^{***}	3.053^{***}
	$[-0.373, -0.312]$	$[0.633, 9.940]$

Table 2: Demand Estimates. Panel A reports IV estimates of equation (B.3) with $\alpha = 0$ and standard errors clustered at the exporter level. Panel B reports IV estimates of equation (B.3) without restrictions and with 95 percent confidence intervals from a block-bootstrap procedure, with blocks at the exporter level. The number of observations in both panels is 3,563,993. *** indicates p<0.05.

all linear IV regressions, and denote this value $\hat{\alpha}$.[B.4] Our estimator of γ is finally given by $\hat{\gamma} = \hat{\gamma}(\hat{\alpha})$.

Demand estimation and welfare implications. We begin by estimating the demand system in equation $(B.3)$ under the restriction that $\alpha = 0$. This reduces equation $(B.3)$ to the CES case, in which the estimating equation is linear. Our results are reported in Panel A of Table 2. In this restricted (CES) case, our IV estimate is $\widehat{\gamma} = -0.206$ with a standard error—clustered at the exporting country level to account for serial correlation over time and across products within exporters—that implies that the point estimate is statistically significantly different from zero at the 95% confidence level. This finding corresponds to an elasticity of substitution equal to $1/\widehat{\gamma} = -4.854$, which is in line with typical estimates of the CES demand parameter in international trade settings. This suggests that our particular instrumental variable, based on the reported value of tariff duties charged, isolates exogenous variation in trade costs that is similar to that used in the literature. Reassuringly, the F-statistic (again adjusted for clustering at the exporter level) on the instrumental variable in the first-stage is 27.28, implying that finite-sample bias due to a weak instrument is unlikely to be a first-order concern here.

We then estimate equation $(B.3)$ without any restriction on α—this corresponds to estimating unrestricted Pollak (rather than CES) demand. These results are reported in Panel B of Table 2. Our non-linear IV estimate of equation $(B.1)$ results in estimates of $\widehat{\gamma} = -0.347$ and $\widehat{\alpha} = 3.053$, with 95% confidence intervals, block-bootstrapped at the exporting country level, with 200 bootstrap replications, shown in parentheses in the table. Notably, this

[B.4] In practice we conduct a grid search over α subject to the restriction that $q^k_{git} + \alpha$ must be strictly positive for $\ln(q^k_{git} + \alpha)$ to be well-defined. Namely, we require α to be greater than minus the lowest value of q^k_{git} in our dataset, which is equal to 1 in all years. After first verifying with a coarse grid that the best-fitting value of α lies below 10, we consider a grid of 400 evenly-spaced values between -1 and 10.

estimate of α has a 95% confidence interval that excludes zero, suggesting that the departure from CES that is modeled in equation $(B.1)$ is a statistically significant feature of these data.[B.5] Furthermore, $\hat{\alpha}$ is positive. As argued above, this implies that η must be positive as well. So, regardless of the value of other structural parameters, Proposition 1 establishes that there cannot be any pro-competitive effect of trade in the sense that welfare gains from trade liberalization must be lower than those predicted by a model with constant markups.

As discussed in Section 5.1, the demand parameter estimates reported in Table 2, Panel B imply that $\hat{\rho} = 0.36$ and in turn $\hat{\eta} = \hat{\rho}/(1+\theta) = 0.06$.[B.6]

B.2 Section 6.3

All models that we consider are calibrated so that the trade elasticity for a 1% change in trade costs is equal to 5 in the initial equilibrium. Except when the distribution of productivity is Pareto, however, this elasticity will vary with the level of trade costs. Figure 6 plots the trade elasticity as a function of trade costs in the case of Pareto, log-normal and bounded Pareto distributions. In both the log-normal and bounded Pareto cases, we see that the trade elasticity increases, in absolute value, with the level of trade costs, as noted in Section 6.3.

[B.5] We have also explored this by HS "section", the coarsest level of disaggregation for which the HS system is designed. Across 22 such sections (two of which we do not include since they do not have the required tariff variation), the median estimates are $\widehat{\gamma} = -0.321$ $[-0.358, -0.210]$ and $\widehat{\alpha} = 0.898$ $[-0.999, 20]$, the 25th percentile estimates are $\widehat{\gamma} = -0.372$ $[-0.530, -0.211]$ and $\widehat{\alpha} = -0.729$ $[-0.999, -0.143]$, and the 75th percentile estimates are $\widehat{\gamma} = -0.200$ $[-0.326, -0.168]$ and $\widehat{\alpha} = 6.153$ $[1.490, 23.898]$. For two sections the estimates fail to reject the null hypothesis of $\gamma = 0$, whereas for six sections the estimates reject the null of $\alpha = 0$ (two of which have a point estimate in the $\alpha < 0$ region). Because of the imprecision of many of these estimates, and in line with the theoretical analysis of Section 4.3, we abstract from misallocations associated with heterogeneity in the values of α, γ, and, in turn, η across sectors.

[B.6] Since we focus on non-zero trade flows, one may be concerned that the previous estimates are subject to selection bias. To explore the potential importance of the previous concern, we have rerun our baseline estimation on a subsample that only includes bilateral trade flow observations at or above the 15th percentile value. We find (with 95% confidence intervals given in brackets) $\widehat{\gamma} = -0.287$ $[-0.304, -0.236]$ and $\widehat{\alpha} = 6.212$ $[1.305, 16]$. This implies that $\widehat{\eta} = 0.05$, only slightly lower than our baseline estimate.

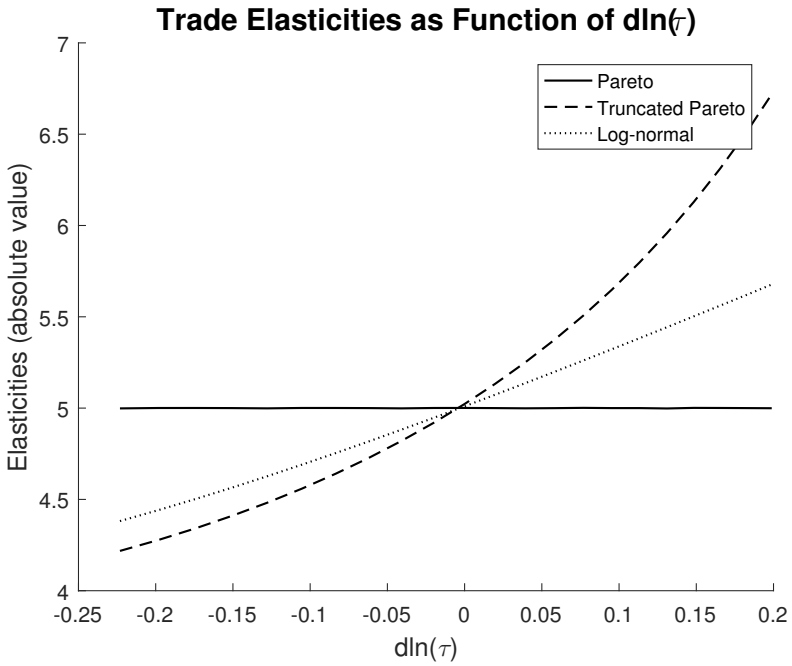

Figure 6: Trade elasticity