Contents lists available at ScienceDirect

Journal of International Economics

journal homepage: www.elsevier.com/locate/jie

Full Length Articles

Increasing marginal costs, firm heterogeneity, and the gains from “deep” international trade agreements

Jeffrey H. Bergstrand [a,d,*], Stephen R. Cray [b], Antoine Gervais [c]

[a] Department of Finance, Department of Economics, Kellogg Institute for International Studies, University of Notre Dame, Notre Dame, IN 46556, USA
[b] Department of Finance, University of Notre Dame, Notre Dame, IN 46556, USA
[c] Department of Economics, Université de Sherbrooke, Sherbrooke, Québec J1K 2R1, Canada
[d] CESifo, Munich, Germany

ARTICLE INFO

Article history:
Received 2 August 2021
Received in revised form 23 May 2023
Accepted 24 May 2023
Available online 30 May 2023

Dataset link: https://data.mendeley.com/datasets/6h6ktwjhmv/3

JEL classification numbers:
F1

Keywords:
International trade
Deep trade agreements
Melitz models
Increasing marginal costs
Gravity equations

ABSTRACT

Two parameters are central to modern quantitative models of trade flows: the elasticity of substitution in consumption (σ) and the inverse index of heterogeneity of firms' productivities (θ). However, structural parameter estimation using the seminal Feenstra econometric methodology focuses on estimates of only σ and a bilateral export-supply elasticity (labeled γ). Separately, modern trade agreements are increasingly “deep,” meaning they reduce fixed trade costs alongside variable trade costs. First, in the spirit of Arkolakis (2010), we extend the Melitz model of trade to allow for increasing marginal market-penetration costs in an empirically tractable manner to help understand the relative impacts on trade, extensive margins, intensive margins, and welfare of reducing fixed trade costs and variable trade costs. Second, we provide a microeconomic foundation for estimating all three parameters using the Feenstra methodology alongside a gravity equation. Third, we demonstrate the importance of increasing marginal costs for shallow and deep trade-agreement liberalizations using two counterfactual exercises.

1. Introduction

Central to the post-2000 modern quantitative models of international trade are two parameters. The first – and arguably most visible – is the elasticity of substitution in consumption among differentiated products, σ. This parameter is key in the seminal theoretical foundation for the gravity equation with Armington preferences in Anderson (1979), monopolistic competition model of intra-industry trade with Dixit-Stiglitz preferences in Krugman (1980), analysis of optimal tariffs in Broda et al. (2008) and Ossa (2016), and a vast array of applied computable general equilibrium (CGE) models used for trade-policy analyses, cf., United States International Trade Commission (2019). The second parameter, which surfaced over the last 20 years, is a (inverse) measure of heterogeneity of firms' productivities, which we denote θ. Motivated by theoretical models of Eaton and Kortum (2002) and Melitz (2003), θ is the key parameter in modern quantitative trade models with heterogeneous firms for cap-

[*] Corresponding author.
E-mail addresses: bergstrand.1@nd.edu (J.H. Bergstrand), Antoine.Gervais2@USherbrooke.ca (A. Gervais).

https://doi.org/10.1016/j.jinteco.2023.103774
0022-1996/© 2023 Elsevier B.V. All rights reserved.

turing the infamous “trade elasticity” (i.e., elasticity of bilateral trade with respect to *ad valorem* bilateral variable trade costs), one of two sufficient statistics to measure welfare effects of trade liberalizations in a broad set of quantitative trade models (cf., Arkolakis et al. (2012), henceforth, ACR).

A common assumption to these quantitative trade models is constant marginal costs. By contrast, the most widely respected structural method for estimating σ – introduced by Feenstra (1994) and further developed by Broda and Weinstein (2006) (henceforth, F/BW) and Broda et al. (2008) – assumes bilateral export supply prices are positive functions of the level of exports to foreign markets, which suggests increasing marginal costs of exporting to each destination market. We will refer to the parameter that governs the bilateral export supply elasticity as γ. Although σ and θ currently play central roles in trade theory and calibration exercises of new quantitative trade models, γ has been largely ignored. Moreover, the bilateral export supply elasticity has typically been incorporated in these econometric analyses in an ad hoc manner. For instance, in Feenstra (1994), Broda and Weinstein (2006), and Soderbery (2015, 2018), positively-sloped bilateral export supply curves were simply assumed. More recently, Feenstra et al. (2018) extend the method of Feenstra (1994) allowing firm heterogeneity based upon a standard Melitz model with constant marginal costs, but still introduce an equation that “plays the role of a supply curve” (p. 140).

Separately, modern international trade agreements – such as free trade agreements (FTAs) – are increasingly “deep,” meaning that, beyond the typical reductions in *ad valorem* tariff rates found in “shallow” agreements, they also reduce *fixed* trade costs. The World Bank has recently compiled a large data set on deep trade agreements' (DTAs) provisions. The database, summarized comprehensively in Hofmann et al. (2017), documents the extensive growth in DTAs over the past twenty years. A notable economic difference concerning these deep provisions is that they relate to regulatory convergences and administrative liberalizations that are unrelated to the quantity of goods exported and are more readily interpreted as reducing fixed trade costs. For instance, the most popular non-tariff measures included in modern trade agreements are customs administration (often referred to as trade facilitation measures), competition policy, sanitary and photosanitary (SPS) regulations, and technical barriers to trade (TBT) regulations.

Recent empirical work using gravity equations indicates economically and statistically significant effects of indexes of DTAs' provisions on trade flows, cf., Kohl et al. (2016), Baier and Regmi (2022), Breinlich et al. (2022) and Fontagne et al. (2022). By contrast, there has been a dearth in numerical analyses of variable versus fixed bilateral trade costs in either standard CGE models (such as GTAP) or in the new quantitative trade models. Zhai (2008) is one of the earliest – and rare – studies to introduce a standard Melitz model (with constant marginal costs) into a global CGE model of world trade and to contrast the trade and welfare effects of a 5% variable trade-cost reduction relative to a 50% fixed trade-cost reduction.[1] In Zhai (2008), it would take a *29%* reduction in bilateral fixed trade costs to achieve the equivalent gain in welfare as a 4% reduction in *ad valorem* variable trade costs (a ratio of 7.25:1). More recently, however, Arkolakis et al. (2021) extend the canonical Melitz model of trade to allow multiproduct firms facing constant marginal costs in core-product production, but allowing increasing marginal market-penetration costs and increasing marginal costs in non-core products. Among several findings, one counterfactual implies that it would take a 13% reduction in fixed trade costs with countries to generate the same welfare gain as a 4 percentage point reduction in tariff rates (or a ratio of 3.25:1). Such estimates suggest evaluating the role of increasing versus constant marginal costs to address the question: Why have countries increasingly pursued deep trade agreements?

Given these considerations, we now summarize our paper's contributions. Our first contribution, motivated by Arkolakis (2010), is to introduce increasing marginal costs (IMC) into the Melitz model via an empirically-tractable formulation of increasing marginal market-penetration costs. To get a sense of the impact of IMC on the trade elasticity, consider a simple Armington trade model. Fig. 1 illustrates the attenuation of the intensive margin elasticity in the presence of a positively-sloped bilateral export supply curve, consistent with IMC. In the standard case of constant marginal costs (CMC), a 1% increase in *ad valorem* variable trade costs, $\Delta \ln \tau_{ij} = \overline{AD}$, lowers bilateral imports from country i to country j (IM_{ij}) by $\Delta \ln IM_{ij} = (1-\sigma)\Delta \ln \tau_{ij} = \overline{AB}$, where σ is the elasticity of substitution in consumption. However, with IMC, the same 1% increase in *ad valorem* variable trade costs lowers bilateral imports by less, $\Delta \ln IM_{ij} = \overline{AC} < \overline{AB}$. Fig. 1 clearly illustrates that under CMC the trade elasticity is a function solely of the elasticity of substitution, whereas under IMC the trade elasticity also depends on an index of the shape of the supply curve.

Our extended model yields several analytical results. First, we derive a gravity equation similar to that in ACR except that the extensive margin elasticity and the trade elasticity with respect to (*ad valorem*) variable trade costs are magnified; yet, the variable trade-cost intensive margin elasticity is diminished, consistent with Fig. 1. An implication is that variable trade-cost liberalizations with IMC will have more firm entry and exit and more labor reallocations than under CMC. Second, the fixed trade-cost trade elasticity – which is a function of the variable trade-cost extensive margin elasticity relative to the variable trade-cost intensive margin elasticity – is magnified under IMC. Moreover, a further implication of IMC is that the fixed trade-cost trade elasticity is magnified relative to the variable trade-cost trade elasticity, which will be important in understanding the welfare-equivalent impacts of fixed trade-cost liberalizations relative to variable trade-cost liberalizations in deep FTAs. Third, allowing IMC diminishes the welfare effect of a given change in the domestic trade share (for a given θ). The intuition is that real wage gains from a trade liberalization can be traced to changes in average productivity. In the Melitz model, changes in average productivity are proportionate to changes in output of the zero-cutoff-profit (ZCP) productivity firm. In the CMC case, the latter are directly proportionate to productivity changes of the ZCP firm. However, with increasing marginal costs

[1] We will discuss Balistreri et al. (2011) and Dixon et al. (2016) below in section 6.

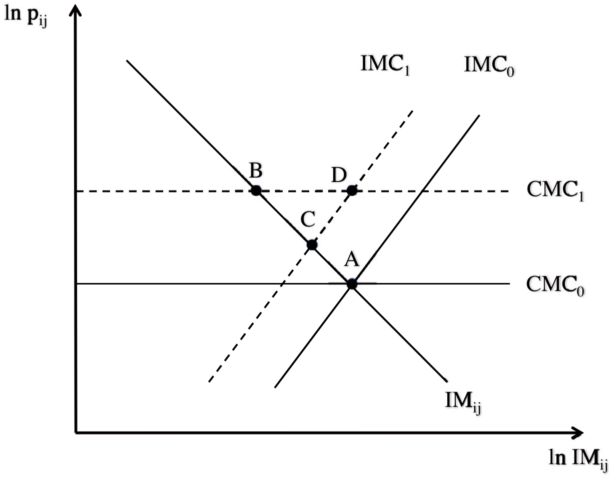

Fig. 1. Increasing Marginal Costs vs. Constant Marginal Costs.

($\gamma < \infty$), output of the ZCP firm rises less than proportionately to the change in the ZCP firm's productivity. The gains to average productivity are diminished at a rate of $1 + 1/\gamma$.

Our second contribution is to develop further the microeconomic foundation for the F/BW econometric approach to estimate σ and γ by accounting explicitly for firm heterogeneity. Unlike F/BW, our approach distinctly recognizes the importance of differences in the masses of exporting firms, which depend on the exporting country's labor-force size and the zero-cutoff-profit productivity threshold. In the context of the heterogeneous-firm models, one must account for both new import varieties from trade liberalizations as well as declining numbers of domestic varieties. The F/BW reduced-form estimating equation includes two variables and one interaction term. Our extension of the F/BW approach to account for firm heterogeneity motivates the inclusion of 6 additional variables, for a total of 8 variables (including the dependent variable) and 28 interaction terms. While we can show all 35 (right-hand-side) coefficients are functions of the three structural parameters only, the large number of nonlinear constraints precludes estimation of σ, γ, and θ simultaneously. Instead, we pursue a two-pronged approach, composed of two reduced-form estimating equations. In the first part of our estimation, we implement our extension of the F/BW reduced-form equation that allows us to estimate σ and γ while controlling explicitly for firm heterogeneity. In the second part, we use the gravity equation generated from our theoretical model to identify θ using the trade elasticity alongside the first part's estimate of γ. Our novel estimation approach yields median estimates across the distribution of industries of σ and γ of 6.45 and 6.00, respectively – approximately *35 and 50% larger*, respectively, than the comparable F/BW estimates ignoring firm heterogeneity. Moreover, our median estimate of θ from the second step is 8.50 – which is very close to Eaton and Kortum (2002)'s and Arkolakis (2010)'s preferred estimate of 8.28.

Our third contribution is to illustrate the impact of recognizing increasing marginal costs on the estimated effects of DTAs in the world. Goldberg and Pavcnik (2016) emphasized that economists have not paid sufficient attention to the study of the effects of trade-policy changes other than *ad valorem* tariff-rate changes, and that a better understanding of the effects of reduced fixed trade costs on international trade and economic welfare is critical. In this spirit, we conduct two numerical analyses. In the first exercise, we show that – even under IMC – the welfare gains from trade for an economy can be captured by a function of an economy's current intra-national trade share and the trade elasticity. This result is fully consistent with the main conclusion in ACR that the trade elasticity (independent of its structural interpretation) and the intra-national trade share are sufficient statistics to measure the welfare effect of a change in bilateral variable or fixed trade costs (τ_{ij} or f_{ij}, respectively). However, in the presence of IMC, the trade elasticity is higher (in absolute terms) and consequently the welfare gains lower, owing to a "welfare diminution effect" attributable to diminishing marginal returns. In a second exercise, we examine the *relative* impacts of variable trade-cost changes and fixed trade-cost changes. We show that, for typical values of σ and θ, under CMC ($\gamma = \infty$) the degree of liberalization of fixed trade costs needed to generate an equivalent increase in welfare is very large relative to the degree of liberalization of variable trade costs, questioning the increasing effort toward deep trade agreements. By contrast, under increasing marginal costs ($\gamma < \infty$), the degree of liberalization of fixed trade costs needed to generate an equivalent increase in welfare is *dramatically reduced* relative to the degree of liberalization of variable trade costs, which helps explain the attractiveness of deep trade agreements. For instance, we show for the United States that, under CMC, fixed trade costs would have to be reduced by 57% to provide the same increase in welfare as a reduction in variable trade costs of 3%. By contrast, under the empirically supported assumption of IMC, it would take only a 14% reduction in fixed trade costs to increase U.S. welfare by the same 3% variable trade-cost reduction.

The remainder of this paper is as follows. In section 2, we introduce and solve our Melitz model allowing increasing marginal costs, asymmetric countries, and a Pareto distribution of productivities. In section 3, we solve for our gravity equation and trade elasticity, derive the variable- and fixed-trade-cost elasticities of extensive and (for variable trade costs) intensive margins, discuss welfare implications, and provide the intuition behind our "welfare diminution effect." In section 4, we discuss our econometric methodology, empirical specifications, and data sources. In section 5, we provide estimates of σ, γ, θ, and the variable- and fixed-trade-cost trade elasticities. In section 6, we provide numerical estimates of a counterfactual analysis of the impact of introducing increasing marginal costs on the welfare effects from trade and another counterfactual analysis demonstrating the importance of recognizing empirically-justified increasing marginal costs toward evaluating the quantitative welfare significance of liberalizations of fixed trade costs relative to those of variable trade costs, two components of (modern) deep trade agreements. In section 7, we offer some conclusions.

2. Theory

Our theoretical framework builds on the Melitz (2003) heterogeneous firms model. As in Chaney (2008) and Redding (2011), we allow for differences in countries' labor endowments and bilateral trade barriers and we assume a Pareto distribution for productivity draws. The Pareto distribution is particularly useful because it yields closed-form solutions that we can use to obtain clear theoretical predictions and to develop our novel econometric approach for the estimation. A key difference with the Melitz (2003) model is that our framework features an empirically tractable adaptation of the increasing marginal market-penetration cost aspect of Arkolakis (2010) to allow for the *possibility* of increasing marginal costs of providing output to any market. It seems reasonable to study the more general version of the model – especially one that motivates the econometrically tractable structural bilateral import demand and bilateral export supply functions in F/BW – and let the data determine the slope of the bilateral export supply curve, instead of imposing CMC *ex ante*.

2.1. Consumer behavior

Our modeling of consumer behavior is standard. We assume a world with $j = 1, 2, \ldots, N$ countries. In each country, there is a mass of consumers, L_j, each endowed with one unit of labor (or a composite input we call "labor"). The preferences of the representative consumer in country j are a constant-elasticity-of-substitution (CES) function of the consumption of a continuum of differentiated varieties:

$$U_j = \left[\sum_{i=1}^{N} \int_{\nu \in \Omega_{ij}} b_i^{\frac{1-\sigma}{\sigma}} c_{ij}(\nu)^{\frac{\sigma-1}{\sigma}} d\nu\right]^{\frac{\sigma}{\sigma-1}}, \tag{1}$$

where $c_{ij}(\nu)$ is the quantity consumed of variety ν from country i, Ω_{ij} is the (endogenous) mass of varieties produced in country i and available for consumption in country j, $b_i > 0$ is an exogenous (inverse) preference parameter for country i's varieties (c.f., Anderson and van Wincoop (2003)) and $\sigma > 1$ is the elasticity of substitution between varieties.

The representative consumer maximizes utility subject to the standard income constraint, such that the optimal aggregate demand function for each variety is given by:

$$c_{ij}(\nu) = E_j P_j^{\sigma-1} b_i^{1-\sigma} p_{ij}^c(\nu)^{-\sigma}, \tag{2}$$

where E_j denotes aggregate expenditures in country j, $p_{ij}^c(\nu)$ is the price of a unit of variety ν from country i facing the consumer in country j, and P_j defined as:

$$P_j = \left[\sum_{i=1}^{N} \int_{\nu \in \Omega_{ij}} b_i^{1-\sigma} p_{ij}^c(\nu)^{1-\sigma} d\nu\right]^{\frac{1}{1-\sigma}} \tag{3}$$

is the price index dual to the consumption index $C_j \equiv U_j$. Because consumers have no taste for leisure, they always supply their unit of labor to the market at the prevailing wage rate, w_j. Hence, the equilibrium labor supply is L_j.

2.2. Cost function

The F/BW approach assumes upward-sloping bilateral export supply curves to estimate bilateral import demand elasticities (σ) for various industries. Starting with Feenstra (1994), the bilateral supply curve for j's imports from country i was specified as $p_{ij} = q_{ij}^{\omega} \xi_{ij}$, where q_{ij} is the quantity produced in i and exported to j and ξ_{ij} was assumed to be a random (technology) factor (cf., his Eq. (8)).[2] More than twenty years later, Soderbery (2018) also assumed the same positively sloped export supply curve stating "an upward sloping constant elasticity (bilateral) export supply curve of this nature was championed by Feenstra (1994), and has become standard with Broda and Weinstein (2006) and Broda et al. (2008) for structurally estimating (bilateral) import demand and export supply elasticities. Additionally, recent deviations from Feenstra (1994) by Feenstra and Weinstein (2017)

[2] We have modified his notation to be consistent with that of the current paper.

and Hottman et al. (2016) model a tighter link between exporter cost functions and export supply, but effectively assume that (bilateral) export supply is isoelastic and upward sloping" (p. 47). The "tighter link" that Soderbery (2018) refers to is Feenstra and Weinstein (2017) specifying that "marginal costs from each exporting country" to an importing country are an exponential function $mc_{ij} = \omega_{ij0} q_{ij}^{\omega}$, where ω_{ij0} is an undefined term.[3]

We note two issues in the extant literature. First, because the studies cited focus on demand considerations, they do not provide a micro-foundation for the supply side of the model. Second, these studies ignore the heterogeneity in firms' productivities that is now well documented. In this paper, we address both of these issues by extending the Melitz trade model to allow for fixed and variable marketing costs in the spirit of Arkolakis (2010), adapted to an empirically tractable framework.[4] The core idea put forward in Arkolakis (2010) is that "firms reach individual consumers rather than the market in its entirety" (p. 1152). Arkolakis (2010) introduced a variable cost component to the fixed export (marketing) component that yielded that only the most productive firms would enter a foreign market (selling to the "first consumer"), but to reach additional consumers (i.e., marginal market penetration) the firm faced "increasing *marginal* penetration costs" (p. 1151; italics added). The model in Arkolakis (2010) provides a rich extension of the Melitz model that matches empirical regularities in the data, such as the observation in Eaton et al. (2011) that the typical destination market for exporters has a large number of smaller firms. Specifically, Eaton et al. (2011) note that the size distribution of exporters within each destination market exhibits a Pareto distribution for relatively larger exporters, but they also note *deviations* from the Pareto distribution for a large proportion of French exporters in each market selling small amounts. Moreover, the rationale for introducing marginal marketing costs is supported empirically. As noted in Arkolakis (2010), Bagwell (2007) reviewed the literature on the economics of advertising and notes that most studies found that advertising's effectiveness is subject to diminishing returns.[5]

For our purposes, the specific features of the model in Arkolakis (2010) are constraining. First, by introducing variable marketing costs via the export fixed cost term, the model in Arkolakis (2010) yields a pricing function where price is a function of *constant* marginal costs, independent of destination output and consequently inconsistent with the typical F/BW positively sloped supply curve. Second, as Anderson (2011) pointed out, the marketing element in Arkolakis (2010) effectively has a "fixed-cost component and a variable-cost component *subject to diminishing returns*" (p. 140; italics added). Third, one of the benefits of Arkolakis embedding the variable cost marketing component inside export fixed costs is that – for his calibrations – he avoids having to specify "as many [export] fixed costs as destinations" (p. 1164). However, as Anderson (2011) noted, the introduction of numerous additional parameters is useful for his simulations, but "is not econometrically tractable" (p. 140). Consequently, we introduce in our model a simple explicit variable marketing cost in the production function, similar in spirit to iceberg transport costs, that captures increasing marginal market-penetration costs in an econometrically tractable manner consistent with the F/BW approach.

Let m_{ij} denote an *ad valorem* factor representing the additional output that must be produced by firms in country i to cover variable marketing costs of "marginal market penetration" from selling in country j. Like iceberg trade costs, variable marketing costs are a function of the quantity sold within the destination market, $m_{ij}\left(q_{ij}\right)$. However, unlike iceberg trade costs, variable marketing costs are an exponential function $m_{ij}\left(q_{ij}\right) = q_{ij}^{1/\gamma}$ where $0 < \gamma < \infty$ (and hence $0 < 1/\gamma < 1$), capturing that previous empirical studies noted above suggest that marketing expenditures exhibit diminishing returns to reach more consumers *within* market j.[6] Having defined all the components of costs, we can now introduce the cost function. Production uses only one input, labor. The labor required by a country-i firm with productivity φ to produce q_{ij} units of output for sale to country j is given by:

$$l_{ij}(\varphi) = \frac{1}{A_i}\left(f_{ij} + \frac{m_{ij}\left(q_{ij}\right) q_{ij}}{\varphi} \right) = \frac{1}{A_i}\left(f_{ij} + \frac{q_{ij}(\varphi)^{1+\frac{1}{\gamma}}}{\varphi} \right) \tag{4}$$

where $A_i > 0$ is incorporated as an exogenous parameter which captures the productivity of workers in the entire country.[7] As implied by Eq. (4), the fixed costs component (f_{ij}) is common across firms for a given origin-destination pair, whereas marginal costs vary across firms for two reasons.[8] First, as conventional to a Melitz model, more productive firms (i.e., with higher φ) need fewer

[3] Once again, we have modified notation in Feenstra and Weinstein (2017) to be consistent with that of the current paper; see that paper's equation (23) on page 1059. Also, Fajgelbaum et al. (2020) asume the same bilateral increasing marginal cost function.

[4] For simplicity here, we assume a single industry as in Melitz (2003). As common to the literature, we could instead have multiple industries with Cobb-Douglas preferences. Nevertheless, our estimation method recognizes that the structural parameters vary across industries.

[5] Arkolakis (2010) also notes several other studies supporting that advertising expenditures are subject to diminishing returns, cf., Simonovska and Waugh (1980), Saunders (1987), Sutton (1991), and Jones (1995).

[6] See Gervais (2015), equation (2), and Flach and Unger (2022), equation (5), for a similar formulation in the context of models with quality differentiation.

[7] As standard to this literature, for the domestic market, the fixed costs f_{ii} capture the costs of setting up a production facility, as well as advertising and domestic distribution costs. For foreign markets ($i \neq j$), the fixed costs f_{ij} represent only the additional fixed costs of selling to the foreign market (such costs associated with advertising, distribution, and conforming to foreign regulations).

[8] In our model, we follow Bernard et al. (2011) in assuming that, for export fixed costs, domestic labor is employed. However, it is straightforward to consider instead the cases where labor in the foreign market is used as in Redding (2011) or labor from both countries is used as in ACR's equation (23). Naturally, this would have the associated implications for our results as discussed in ACR.

workers to produce a given level of firm output.[9] Second, marginal costs are a function of destination output such that, all else equal, larger firms face higher marginal costs to reach more consumers in a market. The parameter γ determines the marginal cost elasticity of output. For any value of $\gamma \in (0, \infty)$, marginal costs are increasing. When γ goes to infinity, we obtain the constant marginal cost function in most workhorse trade models.[10]

As common in this literature, sales to foreign consumers are subject to iceberg trade costs. Firms in country i must ship $\tau_{ij} \geq 1$ units of output to sell one unit in destination j. As typical, we assume $\tau_{ij} > 1$ for all $i \neq j$ and $\tau_{ii} = 1$ for all i. As in Feenstra (2010), we let $p_{ij}(\varphi)$ and $q_{ij}(\varphi)$ denote the factory gate price and quantity shipped. Since a firm in country i producing for and selling to market j incurs *ad valorem* iceberg costs τ_{ij}, only $c_{ij} = q_{ij}/\tau_{ij}$ arrives at destination j. Moreover, drawing upon section 2.1, it follows that, for consumers in j, the unit price will be $p_{ij}^c = \tau_{ij} p_{ij}$.

2.3. Firm behavior

Firms make two decisions for each potential market (including the domestic market). First, they must decide whether or not to enter the market. Second, for each market they enter, they must choose the sale price of a unit of output (or, equivalently, the quantity of output to sell). We look at each decision, beginning with the pricing one.

Firm profits in each market are given by revenues less labor costs:

$$\pi_{ij}(\varphi) = r_{ij}(\varphi) - w_i l_{ij}(\varphi) = p_{ij}(\varphi) q_{ij}(\varphi) - \frac{w_i}{A_i}\left[f_{ij} + \frac{q_{ij}(\varphi)^{\frac{1+\gamma}{\gamma}}}{\varphi}\right], \tag{5}$$

where the second equality uses cost function (4). Because each firm produces only one of a continuum of varieties, its pricing decision has no impact on the price index in the destination market (P_j). In other words, the structure of the model eliminates strategic interactions between firms. Firm profit maximization yields the following optimal (factory-gate) pricing rule[11]:

$$p_{ij}(\varphi) = \left(\frac{1+\gamma}{\gamma}\right)\left(\frac{\sigma}{\sigma-1}\right)\frac{w_i q_{ij}(\varphi)^{\frac{1}{\gamma}}}{A_i \varphi} \tag{6}$$

We note that all country-i firms with productivity φ will charge the same price in destination j, such that the price of varieties can be identified by an origin country and a firm productivity, $p_j(\nu) = p_{ij}(\varphi)$.

Pricing rule (6) differs from standard Melitz models in two respects. First, the markup is no longer a function of only the elasticity of substitution (σ), but also depends on the inverse marginal cost elasticity of output (γ). As a result, conditional on the distribution of firm productivities, prices will be higher by a factor of $1 + 1/\gamma$ under IMC. Second, prices are an increasing function of quantity; this provides a rationale for the upward-sloping bilateral export supply functions in F/BW. We note that, when γ goes to infinity, the first term of the pricing rule converges to 1 and quantity vanishes from the equation such that we obtain the CMC pricing rule typical to a standard Melitz model and most workhorse trade models.

Next, we consider the decision to enter a market or not. As a first step, we compute firm profits. We can use pricing rule (6) to express firm profits, defined in Eq. (5), as:

$$\pi_{ij}(\varphi) = \left(\frac{\sigma+\gamma}{1+\gamma}\right)\frac{r_{ij}(\varphi)}{\sigma} - \frac{w_i}{A_i} f_{ij} \tag{7}$$

where $r_{ij}(\varphi)$ is the firm's optimal revenue. This result is analogous to a standard Melitz model with the exception of the first term $(\sigma+\gamma)/(1+\gamma)$, which exceeds unity because $\sigma > 1$. Our model implies that profits are higher when marginal costs are increasing in output (i.e., $1/\gamma > 0$). Again, when γ goes to infinity, the benchmark result obtains.

We can combine the zero-cutoff-profit (ZCP) condition $\pi_{ij}\left(\varphi_{ij}^*\right) = 0$, the optimal pricing Eq. (6), and profits Eq. (7) to solve for the output and the productivity of the ZCP firm as follows:

$$q_{ij}\left(\varphi_{ij}^*\right) = \left[\frac{\gamma}{\sigma+\gamma}(\sigma-1) f_{ij} \varphi_{ij}^*\right]^{\frac{\gamma}{1+\gamma}}, \tag{8}$$

[9] We model higher productivity as producing a symmetric variety at lower marginal cost. However, higher productivity may also be thought of as producing a higher quality variety at equal cost. As noted in Melitz (2003), given the form of product differentiation, the modeling of either type of productivity difference is isomorphic.

[10] The cost function assumed here allows closed-form analytical solutions in a world with asymmetrically-sized countries and asymmetric bilateral trade costs. It is also feasible to follow instead Vannoorenberghe (2012) in a special case of symmetric country sizes and bilateral trade costs where marginal costs are simply increasing in total firm output; hence, Vannoorenberghe (2012) was the first to introduce increasing marginal costs in total firm output in a Melitz framework. We solve this case in Online Appendix C, noting that – with a large number of countries – the trade, extensive-margin, and intensive-margin elasticities are identical.

[11] Detailed derivations are available in sections 1 and 2 of Online Appendix A.

where φ^*_{ij} is the productivity level of the ZCP firm and:

$$\left(\varphi^*_{ij}\right)^{-\theta} = \left[\frac{\left(\frac{1+\gamma}{\gamma}\frac{\sigma}{\sigma-1}\frac{w_i}{A_i}\right)^{\sigma}}{b_i^{1-\sigma}E_jP_j^{\sigma-1}}\right]^{\frac{-\theta}{\frac{\gamma}{1+\gamma}(\sigma-1)}}\left[\frac{\gamma}{\sigma+\gamma}(\sigma-1)f_{ij}\right]^{\frac{-\theta\frac{1+\gamma}{\gamma}}{\frac{1+\gamma}{\sigma+\gamma}(\sigma-1)}}\tau_{ij}^{-\theta\frac{1+\gamma}{\gamma}} \tag{9}$$

Because $\gamma/(\sigma+\gamma)$ and $\gamma/(1+\gamma)$ in Eq. (8) are both positive and smaller than one, for a given φ^*_{ij} the level of output $q_{ij}\left(\varphi^*_{ij}\right)$ is smaller than in the CMC case. Eq. (9) provides an explicit link between *ad valorem* variable trade costs (τ_{ij}) and a country-pair's export cutoff productivity (φ^*_{ij}).[12] Under CMC (i.e., $\gamma=\infty$), these two variables are proportionate. However, under IMC, a 1% change in τ_{ij} has a more-than-proportionate effect on φ^*_{ij}. We will show later that this implies the trade elasticity is larger under IMC relative to CMC. Finally, we note that when $\gamma\rightarrow\infty$, Eqs. (8) and (9) simplify to the standard result in the benchmark CMC case.

Revenue is increasing in firm productivity, so that profits are also increasing in firm productivity. As a result, firms in country i with productivity above the productivity cutoff φ^*_{ij} will enter market j, while those with productivity below the cutoff will not. Furthermore, Eq. (9) implies that the ratio of export and domestic cutoff productivities is:

$$\frac{\varphi^*_{ij}}{\varphi^*_{ii}} = \left(\frac{E_iP_i^{\sigma-1}f_{ij}^{\frac{1+\gamma}{\sigma+\gamma}}}{E_jP_j^{\sigma-1}f_{ii}^{\frac{1+\gamma}{\sigma+\gamma}}}\right)^{\frac{1}{\sigma-1}\left(\frac{1+\gamma}{\gamma}\right)}\tau_{ij}^{\frac{1+\gamma}{\gamma}} \equiv \Gamma_{ij} \quad \Rightarrow \quad \varphi^*_{ij} = \Gamma_{ij}\varphi^*_{ii} \tag{10}$$

As in Bernard et al. (2011), we assume that $\Gamma_{ij}>1, \forall i \neq j$ (see page 1284). In that case, only the most productive firms export, while intermediate productivity firms serve only the domestic market and the low productivity firms exit. The assumption that there are no "pure exporters" is consistent with the empirical literature on firms in international trade.[13]

2.4. Trade flows

We can now characterize equilibrium aggregate trade flows.[14] Imposing the labor-market-clearing condition and assuming a Pareto distribution for firms' productivities, we can solve for the mass of incumbent firms in each country i that sell to each destination j:

$$M_{ij} = \left(\frac{\gamma}{1+\gamma}\right)\left(\frac{\sigma-1}{\sigma}\right)\frac{A_iL_i}{\delta\theta f^e}\left(\varphi^*_{ij}\right)^{-\theta}. \tag{11}$$

In the case of $\gamma=\infty$, M_{ij} simplifies to the respective term in a standard Melitz model with Pareto distribution. Next, using pricing rule (6) and mass of firms Eq. (11), we can express bilateral trade flows as:

$$X_{ij} \equiv M_{ij}\int_{\varphi^*_{ij}}^{\infty} r_{ij}(\varphi)\mu_{ij}(\varphi)d\varphi = \left[\frac{\frac{\gamma}{\sigma+\gamma}(\sigma-1)}{\theta-\frac{\gamma}{\sigma+\gamma}(\sigma-1)}\right]\frac{w_iL_if_{ij}}{\delta f^e}\left(\varphi^*_{ij}\right)^{-\theta} \tag{12}$$

We use the goods-market-clearing condition, $R_i=E_i$, to express trade flows as a gravity equation. Substituting Eq. (12) into expenditure function $E_j=\sum_{k=1}^{N}X_{kj}$, using the solution for the productivity threshold in Eq. (9), and solving yields the following gravity equation:

$$X_{ij} = \left[\frac{A_i^{\theta\left(\frac{1+\gamma}{\gamma}\right)\left(\frac{\sigma}{\sigma-1}\right)}L_iw_i^{1-\theta\left(\frac{1+\gamma}{\gamma}\right)\left(\frac{\sigma}{\sigma-1}\right)}b_i^{-\theta\left(\frac{1+\gamma}{\gamma}\right)}\tau_{ij}^{-\theta\left(\frac{1+\gamma}{\gamma}\right)}f_{ij}^{1-\frac{\theta\left(\frac{1+\gamma}{\gamma}\right)}{\frac{1+\gamma}{\sigma+\gamma}(\sigma-1)}}}{\sum_{k=1}^{N}A_k^{\theta\left(\frac{1+\gamma}{\gamma}\right)\left(\frac{\sigma}{\sigma-1}\right)}L_kw_k^{1-\theta\left(\frac{1+\gamma}{\gamma}\right)\left(\frac{\sigma}{\sigma-1}\right)}b_k^{-\theta\left(\frac{1+\gamma}{\gamma}\right)}\tau_{kj}^{-\theta\left(\frac{1+\gamma}{\gamma}\right)}f_{kj}^{1-\frac{\theta\left(\frac{1+\gamma}{\gamma}\right)}{\frac{1+\gamma}{\sigma+\gamma}(\sigma-1)}}}\right]w_jL_j \tag{13}$$

We note that when $\gamma\rightarrow\infty$ the benchmark result obtains.[15,16]

[12] Detailed derivations are available in section 3 of Online Appendix A.

[13] The findings in Lu (2010) to the contrary are explained in Dai et al. (2016) as processing trade.

[14] Derivation details are provided in sections 4–8 of Online Appendix A.

[15] Note that the wage-rate elasticity is equivalent to that in Bernard et al. (2011) if one assumes $\gamma=\infty$, as we have followed their assumption of export fixed costs using the exporter's (i's) labor. By contrast, Redding (2011) assumes export fixed costs use the importer's (j's) labor. ACR's equation (23) allows either of those two cases; our setting is analogous to ACR in their case of $\mu=1$. In the case of $\gamma=\infty$ and $\mu=1$, our wage-rate elasticity is equivalent mathematically to ACR's.

[16] As shown in section 11 of Online Appendix A, equation (13) and the associated variable- and fixed-trade-cost trade elasticities are consistent also with a "structural gravity" representation that is common in the literature. As a result, the method developed in Head and Mayer (2014) to estimate the general equilibrium trade impacts (GETI) of changes in trade barriers remains applicable for us.

2.5. General equilibrium

In section 9 of Online Appendix A, we develop the dynamic aspect of the model and show that it is possible to define a set of free entry conditions that depend only on parameters and the productivity cutoffs. These conditions serve to identify equilibrium values for the productivity thresholds. As explained in section 10 of Online Appendix A, we can determine the general equilibrium using the recursive structure of the model as in Bernard et al. (2011).

3. Implications

In this section, we provide several theoretical implications from the model. In section 3.1, we derive novel *ad valorem* variable trade-cost and fixed trade-cost trade elasticities under IMC. With IMC, the variable trade-cost trade elasticity changes *relative* to the fixed trade-cost trade elasticity (relative to CMC), which has implications for estimating the relative welfare benefits of fixed trade-cost liberalizations relative to variable trade-cost liberalizations within deep trade agreements. In section 3.2, we show that under IMC the welfare effect of a change in trade costs is still measured by the change in the domestic trade share raised to the (negative of the) inverse of the (variable trade-cost) trade elasticity, as in ACR. However, the welfare effect is diminished for a given domestic trade share; we explain the source of this "welfare diminution effect."

3.1. Trade elasticities

As shown in section 12 of Online Appendix A, the (positively defined) *ad valorem* variable trade-cost trade elasticity (ε_τ) is:

$$\varepsilon_\tau \equiv -\frac{\partial X_{ij}}{\partial \tau_{ij}}\frac{\tau_{ij}}{X_{ij}} = -\left[\underbrace{-\theta\left(\frac{1+\gamma}{\gamma}\right)}_{\text{extensive}} + \underbrace{\frac{1+\gamma}{\sigma+\gamma}(1-\sigma)}_{\text{intensive}} + \underbrace{\frac{1+\gamma}{\sigma+\gamma}(\sigma-1)}_{\text{compositional}}\right] = \theta\left(\frac{1+\gamma}{\gamma}\right) \tag{14}$$

Following Head and Mayer (2014), we decompose this trade elasticity into extensive margin, intensive margin, and compositional margin components.[17] The extensive- and intensive-margin components have the usual interpretations. The extensive margin elasticity is caused by changes in the mass of firms serving each market. The intensive margin elasticity is caused by changes in firm-level exports.[18] The compositional-margin elasticity is caused by the fact that new entrants or exitors do not have the same productivity as the existing exporters. This margin is a function of the difference between the average shipment of the incumbent firms (X_{ij}/M_{ij}) and that of the marginal firm. All three components converge to the benchmark Melitz model values as $\gamma \to \infty$.

In line with previous results for Melitz models, the trade elasticity is determined entirely by the extensive margin elasticity. At the intensive margin, lower *ad valorem* trade costs increase exports of a given firm to a given country, which raise average exports per firm. At the compositional margin, lower *ad valorem* trade costs induce low productivity firms to enter the export market, which lowers average exports per firm. With a Pareto productivity distribution, the intensive margin and compositional margin elasticities offset one another exactly.

Under IMC, the elasticity of trade with respect to *ad valorem* trade costs, ε_τ, depends on θ, as in the benchmark, but is scaled up by the additional term $\frac{1+\gamma}{\gamma}$. Whenever $\gamma < \infty$, the trade elasticity is magnified relative to the benchmark ($\gamma \to \infty$). The intuition can be traced back to Eqs. (9) and (11). Eq. (9) reveals that, with IMC, a fall in τ_{ij} has a magnified effect of $\frac{1+\gamma}{\gamma}$ on lowering the country-pair's export cutoff productivity. In light of Eq. (11), this lower export productivity threshold makes it profitable for more firms to export from i to j and hence M_{ij} increases, enlarging the aggregate trade flow from i to j. Due to diminishing marginal returns, the trade elasticity is augmented and is now a nonlinear function of the two supply-side parameters, θ and γ.

As shown in section 13 of Online Appendix A, we can also decompose the (positively defined) elasticity of trade with respect to fixed trade costs (ε_f) into three margins:

$$\varepsilon_f \equiv -\frac{\partial X_{ij}}{\partial f_{ij}}\frac{f_{ij}}{X_{ij}} = -\left[\underbrace{-\frac{\theta\left(\frac{1+\gamma}{\gamma}\right)}{\frac{1+\gamma}{\sigma+\gamma}(\sigma-1)}}_{\text{extensive}} + \underbrace{0}_{\text{intensive}} + \underbrace{1}_{\text{compositional}}\right] = \frac{\theta\left(\frac{1+\gamma}{\gamma}\right)}{\frac{1+\gamma}{\sigma+\gamma}(\sigma-1)} - 1 \tag{15}$$

[17] We note that this decomposition nests other decompositions proposed in the literature. First, in the decomposition of Redding (2011), the intensive and compositional margins are lumped together and labeled as the "intensive margin." It also nests the decomposition proposed by Chaney (2008), which is obtained by taking the sum of the extensive and the compositional margins and calling it the "extensive margin."

[18] The intensive-margin elasticity here is consistent with that in a special case of Bergstrand (1985) with homogeneous firms. We address this in Online Appendix B.

All components converge to the benchmark values as $\gamma \to \infty$. The fixed trade-cost trade elasticity is also scaled up compared to the CMC case where $\varepsilon_f^{CMC} = \theta/(\sigma - 1) - 1$.[19] An explanation for the different elasticity under IMC also can be traced intuitively back to Eqs. (9) and (11). Using Eq. (9), with increasing marginal costs a fall in f_{ij} has a magnified effect on lowering the country-pair's export cutoff productivity relative to the case of CMC. In the IMC case, the scaling up of the numerator by $\frac{1+\gamma}{\gamma}$ and scaling down of the denominator of this elasticity by $\frac{1+\gamma}{\sigma+\gamma}$ augments the reduction in the country-pair's export productivity cutoff. Using Eq. (11), this lower export productivity threshold makes it profitable for more firms to export from i to j and hence M_{ij} increases, enlarging the aggregate trade flow from i to j.[20]

So far we have shown that, for given values of the structural parameters, the elasticities of trade are magnified under IMC. As a result, any trade-policy liberalization or transport-cost reduction that lowers bilateral *ad valorem* variable trade costs or fixed trade costs will have a *larger* impact on trade flows and consequently on the domestic expenditure share than in the CMC case. Moreover, Eqs. (14) and (15) reveal not only that IMC increases both elasticities in absolute terms, but also the fixed trade-cost trade elasticity increases *relative* to the variable trade-cost trade elasticity. To understand why fixed trade-cost reductions have a relatively larger effect on trade than variable trade-cost reductions with IMC, consider Eqs. (8) and (9). The variable trade-cost trade elasticity in a Melitz model is determined by extensive margin effects solely; consistent with these equations, lower τ_{ij} increases trade exclusively by increasing the mass of firms exporting from i to j (as under Pareto, the intensive margin effect is offset perfectly by the compositional margin effect). Due to IMC, the trade elasticity scales up θ by $\frac{1+\gamma}{\gamma}$ due to diminishing marginal returns, cf., Eq. (9). By contrast, the fixed trade-cost trade elasticity is determined by the *ratio* of the extensive margin elasticity to the intensive margin elasticity. Recall, under CMC, reductions in τ_{ij} change φ_{ij}^* proportionately; however, reductions in f_{ij} change φ_{ij}^* less than proportionately, i.e., φ_{ij}^* is proportionate to $f_{ij}^{1/(\sigma-1)}$ in Eq. (9) (when $\gamma = \infty$). The introduction of IMC causes both the variable trade-cost trade elasticity to increase from θ to $\theta\frac{1+\gamma}{\gamma}$, but also the intensive margin effect to decline from $\sigma - 1$ to $\frac{1+\gamma}{\sigma+\gamma}(\sigma - 1)$. This is confirmed in Eq. (9). As we will show later, this result is important for evaluating the relative trade and welfare benefits of "shallow" trade agreements (that only lower variable trade costs) with those of "deep" trade agreements (that also reduce fixed trade costs).

3.2. Welfare

In this section, we show two results related to welfare effects under IMC relative to CMC. First, we show that under IMC that the two sufficient statistics to measure the welfare effects of international trade-cost shocks remain the share of domestic expenditure on domestic output and the trade elasticity as in ACR. Second, because for a set of parameter values the trade elasticity is magnified under IMC relative to that under CMC, the predicted welfare gains from trade are smaller.

First, in section 14 of Online Appendix A, we show that the change in welfare of a given "foreign" shock (to τ_{ij} or f_{ij}) that leaves unchanged country j's labor endowment, L_j, as well as the costs to serve its own market (τ_{jj} and f_{jj}) can be expressed as:

$$\hat{W}_j = \hat{\lambda}_{jj}^{-1\big/\left[\theta\left(1+\frac{1}{\gamma}\right)\right]} = \hat{\lambda}_{jj}^{-1/\varepsilon_\tau}, \tag{16}$$

where $\hat{\lambda}_{jj} \equiv \lambda_{jj}^{'}/\lambda_{jj}$ is the (gross) change in the share of domestic expenditure (where $\lambda_{jj} = X_{jj}/E_j$) and $\hat{W}_j \equiv W_j^{'}/W_j$ is the change in welfare. In the special case of a move from trade (λ_{jj}) to autarky ($\lambda_{jj}^{'} = 1$), the gains from trade (G_j) can be expressed as:

$$G_j = 1 - \lambda_{jj}^{1\big/\left[\theta\left(1+\frac{1}{\gamma}\right)\right]} = 1 - \lambda_{jj}^{1/\varepsilon_\tau}, \tag{17}$$

which is identical to Eq. (12) in Costinot and Rodriguez-Clare (2014). These results imply that, conditional on the trade elasticity, the impact of trade shocks on welfare are independent of the structure of marginal costs. At the same time, note that the definition of the trade elasticity itself is different in our model. In the presence of IMC, the larger trade elasticity implies (for a given λ_{jj}) a smaller welfare effect than in the constant marginal cost case, which we will term in this paper the "welfare diminution effect."

Second, to understand intuitively this welfare diminution effect, consider the benchmark Melitz model with CMC. The change in welfare $\left(\hat{W}_j\right)$ from a reduction in variable trade costs is directly proportionate to the change in average productivity $\left(\hat{\tilde{\varphi}}_{ij}\right)$ and

[19] In the CMC case, the assumption that $\frac{\theta}{\sigma - 1} > 1$ is necessary to solve the Melitz model. However, some empirical researchers have found evidence that estimates of θ are often below estimates of $\sigma - 1$, violating a necessary assumption of this model, cf., Feenstra (2016), page 168. Our results in equation (15) shed light on this finding. Our Melitz model under IMC requires only that $\theta > \frac{\gamma}{\sigma+\gamma}(\sigma - 1)$. Hence, θ can be less than $\sigma - 1$ as long as θ exceeds $\frac{\gamma}{\sigma+\gamma}(\sigma - 1)$, where $0 < \frac{\gamma}{\sigma+\gamma} < 1$.

[20] In Online Appendix C, we solved the model for the case of increasing marginal costs in *total firm output* (instead of destination-specific output). In the case of marginal costs depending on total firm output, we must assume symmetric countries and symmetric trade costs to obtain closed-form solutions. Since overall output is endogenous to the set of countries to which firms export, one cannot solve the model analytically with asymmetric country sizes and asymmetric trade costs. Yet, in the symmetric world, we can solve for analogous trade elasticities, cf., footnote 10.

Table 1
Elasticities and Welfare Measures by Model.

Model	Intensive margin elast.	Var. trade elast. (ε_τ)	Fixed trade elast. (ε_f)	Welfare measure
Armington differentiation (Anderson, 1979)	$\sigma - 1$	$\sigma - 1$	n.a.	$\hat{\lambda}_{jj}^{-\frac{1}{\sigma - 1}}$
Armington differentiation and CET (Bergstrand, 1985)	$\frac{1+\gamma}{\sigma+\gamma}(\sigma - 1)$	$\frac{1+\gamma}{\sigma+\gamma}(\sigma - 1)$	n.a.	$\hat{\lambda}_{jj}^{-\frac{1}{\frac{1+\gamma}{\sigma+\gamma}(\sigma - 1)}}$
Monopolistic Competition (Krugman, 1980)	$\sigma - 1$	$\sigma - 1$	n.a.	$\hat{\lambda}_{jj}^{-\frac{1}{\sigma - 1}}$
Heterogeneity without fixed trade costs (Eaton-Kortum, 2002)	n.a.	θ	n.a.	$\hat{\lambda}_{jj}^{-\frac{1}{\theta}}$
Heterogeneity with fixed trade costs and Pareto (Chaney, 2008)	$\sigma - 1$	θ	$\frac{\theta}{\sigma - 1} - 1$	$\hat{\lambda}_{jj}^{-\frac{1}{\theta}}$
Heterogeneity with fixed trade costs, Pareto, and IMC (BCG)	$\frac{1+\gamma}{\sigma+\gamma}(\sigma - 1)$	$\theta\left(\frac{1+\gamma}{\gamma}\right)$	$\frac{\theta\left(\frac{1+\gamma}{\gamma}\right)}{\frac{1+\gamma}{\sigma+\gamma}(\sigma - 1)} - 1$	$\hat{\lambda}_{jj}^{-\frac{1}{\theta\left(\frac{1+\gamma}{\gamma}\right)}}$

Notes: This table reports the (positively-defined) *ad valorem* variable trade-cost intensive margin elasticity, the *ad valorem* variable trade-cost trade elasticity, the fixed trade-cost trade elasticity, and the measure of welfare effects, under various theoretical assumptions as indicated in the first column's papers. The trade and intensive margin elasticities reported here for Bergstrand (1985) assume the case in that paper of $\sigma = \mu$ and $\gamma = \eta$. See Online Appendix B for an explanation; CET denotes constant elasticity of transformation. n.a. denotes not applicable.

the change in the number of varieties $\left(\hat{M}_{ij}\right)$, cf., Melitz (2003), Eq. (17). Feenstra (2010) shows also that the change in welfare can be simplified further to be proportionate to the change in output of the ZCP firm $\left(\widehat{q_{ij}\left(\varphi_{ij}^*\right)}\right)$ and to $\hat{\varphi}_{ij}^*$ (see his page 52). As seen in Eq. (8), under IMC the output of the cutoff productivity firm is proportional instead to $\left(\varphi_{ij}^*\right)^{\frac{\gamma}{1+\gamma}}$, due to diminishing marginal returns. In the limit, as γ approaches ∞, the relationship between $q_{ij}\left(\varphi_{ij}^*\right)$ and φ_{ij}^* becomes linear, as in the benchmark Melitz model. As a result, a given change in φ_{ij}^* has a smaller effect on output under IMC than CMC. This is the intuition underlying the "welfare diminution effect" from increasing marginal costs.[21]

Finally, it will be useful to summarize in a table the differences between the various "trade" elasticities and welfare-change effects of our model relative to those of the main models in the trade literature. Adapting Table 3.1 in Head and Mayer (2014), Table 1 contrasts the *ad valorem* variable trade-cost intensive margin elasticities, *ad valorem* variable trade-cost trade elasticities, fixed trade-cost trade elasticities, and welfare effects from the large class of models addressed in Arkolakis et al. (2012) with those from this paper.

4. Estimation methodology, specifications, and data

In order to conduct numerical analyses of the welfare gains from fixed- versus variable-trade-cost changes under increasing versus constant marginal costs in section 6, we need to estimate all three main structural parameters of the model: σ, γ, and θ.[22] To do so, in this section we introduce a two-pronged estimation method that consists of two reduced-form equations, both derived from our theoretical model. As is well known, properly specified and estimated gravity equations can generate estimates of the (variable trade-cost) trade elasticity, ε_τ. In our model, this elasticity is the product of θ and $\frac{1+\gamma}{\gamma}$. Consequently, to identify θ, we generate estimates of γ and σ by estimating an extension of the F/BW reduced form equation. Because of firm heterogeneity, our reduced form equation depends upon a large number of variables not included in the standard F/BW estimating equation.[23] Using the gravity equation implied by our theoretical model to estimate the trade elasticity and an estimate of γ from our F/BW reduced form, we can recover an estimate of θ.[24]

In section 4.1, we derive our extension of the F/BW reduced-form estimating equation accounting for firm heterogeneity. We begin by summarizing the key aspects of the F/BW methodology in section 4.1.1. In section 4.1.2, we derive the bilateral import

[21] We formalize this intuition using the constant-elasticity-of-transformation (CET) approach of Feenstra (2010) in sections 14 and 15 of Online Appendix A.

[22] In recent work, Fajgelbaum et al. (2020) use a setup similar to Feenstra (1994) to estimate both the bilateral import demand and the bilateral export supply elasticities using disaggregated trade data. Their approach is quite different from ours. First, they do not include firm heterogeneity in their theoretical framework; hence, estimating equations differ across the two studies. Second, they identify both elasticities using a single instrumental variable, tariff rates. Third, they estimate one demand parameter and one supply parameter common to all industries; by contrast, we estimate hundreds of industry-specific demand and supply parameters.

[23] Although the coefficients of our reduced-form extension of F/BW depend only on the three parameters of the model (σ, γ, and θ), the large number of non-linear restrictions precludes identification of θ from that regression. We return to this issue later.

[24] Due to our goal here of providing a novel methodological approach to estimate all three parameters (σ, γ, and θ) under our modified Melitz model framework, we omit allowing for heterogeneous bilateral export supply elasticities – across exporter-importer pairs – as addressed recently in Soderbery (2018) and Farrokhi and Soderbery (2020). Soderbery (2018), though still relying upon the same assumed bilateral export supply function as in the F/BW models, moves the literature in a different direction from our paper by exploring how heterogeneous (by exporter-importer pair) bilateral supply elasticities can help explain importers' market power and be adapted to evaluate optimal trade policy. Farrokhi and Soderbery (2020) extend Soderbery (2018) further. Their section 3 shows that the F/BW approach is a restricted version of a more general model allowing external economies of scale and labor mobility across industries. Specifically, they argue that the F/BW approach constrains the bilateral export supply elasticities to have positive slopes and assumes demand is not "convoluted by supply when using unit values."

demand (structural) equation from our model with firm heterogeneity. In section 4.1.3, we derive the (inverse) bilateral export-supply (structural) equation from our theoretical model. In section 4.1.4, we show that F/BW corresponds to a special case of our model and we discuss specifications and data requirements. In section 4.1.5, we discuss the moment and identification conditions in the context of our theoretically-based extended model. In section 4.2, we develop the other reduced-form estimating equation (based on the gravity equation implied by our model) to estimate the trade elasticity, which we will then use to generate estimates of θ. While we implement the model using data for hundreds of industries (as discussed in section 4.1.4), we omit the industry subscripts in what follows to simplify notation.

4.1. F/BW estimation methodology accounting for firm heterogeneity

In this section, we derive a (structural) nominal bilateral import-demand-share (X_{ij}^D/E_j) equation that motivates an estimable bilateral trade-flow-share equation, and a nominal bilateral export-supply-share (X_{ij}^S/E_j) equation that motivates an estimable bilateral import-unit-value equation, akin to F/BW. We will show that two error terms surface in these equations; one error term accounts for the role of deviations from the Pareto assumption for productivities (for small exporters) addressed in Arkolakis (2010) influencing the bilateral trade-share (demand) equation and the other error term accounts for the role of deviations from the Pareto assumption for productivities influencing the bilateral import unit-value (supply) equation. We then derive the reduced-form estimating equation that controls *explicitly* for firm heterogeneity, and we demonstrate that both the moment and identification conditions addressed in F/BW are satisfied.

4.1.1. The basic F/BW approach

To understand our contribution, we first provide a brief summary of the F/BW methodology. The F/BW approach entails a bilateral nominal import-demand-share equation:

$$\Delta^k \ln\left(X_{ijt}^D/E_{jt}\right) = (1-\sigma)\Delta^k \ln \overline{p}_{ijt}^c + \epsilon_{ijt} \tag{18}$$

where $\overline{p}_{ijt}^c$ is the observed bilateral import unit value, $t = 1, \ldots, T$ indexes time periods, $\Delta^k \ln$ refers to the *double difference* of a variable with respect to both time and a "reference" exporting country k, e.g., $\Delta^k \ln\overline{p}_{ijt}^c = (\ln\overline{p_{ijt}^c} - \ln\overline{p}_{ij,t-1}^c) - (\ln\overline{p}_{kjt}^c - \ln\overline{p}_{kj,t-1}^c)$, and ϵ_{ijt} is an error term that will be discussed shortly.

Rather than estimating the demand equation using instrumental variables to address simultaneity, F/BW introduce an ad hoc "supply" equation and rely on orthogonal supply shocks. Their method proceeds in three steps. First, F/BW assume monopolistically competitive firms face upward sloping bilateral export supply to each market, implying a (inverse supply) function in terms of a nominal bilateral export-supply share (X_{ijt}^S/E_{jt}):

$$\Delta^k \ln \overline{p}_{ijt}^c = \frac{1}{1+\gamma}\Delta^k \ln\left(X_{ijt}^S/E_{jt}\right) + \psi_{ijt} \tag{19}$$

where ψ_{ijt} is an error term that will be discussed shortly. Second, F/BW combine these demand and supply equations in a particular manner. They rewrite Eqs. (18) and (19) with ϵ_{ijt} and ψ_{ijt}, respectively, on the LHS, take the latter two terms' product, and rearrange terms to obtain:

$$\begin{aligned}\left(\Delta^k \ln \overline{p}_{ijt}^c\right)^2 &= \frac{1}{(\sigma-1)(1+\gamma)}\left(\Delta^k \ln s_{ijt}\right)^2 \\ &+ \frac{\sigma-\gamma-2}{(\sigma-1)(1+\gamma)}\left(\Delta^k \ln s_{ijt}\Delta^k \ln \overline{p}_{ijt}^c\right) + \epsilon_{ijt}\psi_{ijt},\end{aligned} \tag{20}$$

where s_{ijt} denotes the (partial equilibrium) trade share, but is measured using the actual bilateral trade share. Third, under the assumption that the demand and supply error terms are orthogonal, F/BW use the moment condition $\mathbb{E}\left(\epsilon_{ijt}\psi_{ijt}\right) = 0$ (where $\mathbb{E}$ denotes the expectation operator) to derive a reduced-form equation, averaging each of the variables over all T observations. Letting $\overline{Y}_{ij}$, $\overline{Z}_{1ij}$, $\overline{Z}_{2ij}$, and $\overline{\epsilon_{ij}\psi_{ij}}$ denote the time-averaged means of the respective variables in equation (20), consistent estimates of the coefficients are obtained by estimating:

$$\overline{Y}_{ij} = \beta_0 + \beta_1\overline{Z}_{1ij} + \beta_2\overline{Z}_{2ij} + \overline{\epsilon_{ij}\psi_{ij}} \tag{21}$$

separately for each industry. As evident from equation (20), the empirical model identifies the coefficients from the second moments of the data (i.e., variances and covariances). Identification therefore relies on the presence of heteroskedasticity such that $\overline{Z}_{1ij}$ and $\overline{Z}_{2ij}$ are not perfectly collinear, cf., Feenstra (1994), page 164.

In the remaining subsections of section 4.1, we show first that our model delivers both the bilateral trade-flow-share and bilateral import-unit-value analogue equations to those in F/BW, but based upon micro-foundations from our general equilibrium model of trade. Second, we show that the moment condition requires the inclusion of additional controls suggested by our theory. Third, in the context of our general equilibrium framework, the model calls for a reinterpretation of the error terms used for identification of the coefficients.

4.1.2. Bilateral import demand

In this section, we motivate our analogue to Eq. (18).[25] Using product-level bilateral import demand Eq. (2), aggregate bilateral (quantity) import demand (for each industry), Q_{ij}^D, is:

$$Q_{ij}^D = M_{ij}\int_{\varphi_{ij}^*}^{\infty} c_{ij}(\varphi)\mu_{ij}(\varphi)d\varphi = M_{ij}b_i^{1-\sigma}E_jP_j^{\sigma-1}\int_{\varphi_{ij}^*}^{\infty} p_{ij}^c(\varphi)^{-\sigma}\mu_{ij}(\varphi)d\varphi, \tag{22}$$

and the *value* of aggregate bilateral import demand, X_{ij}^D, is:

$$X_{ij}^D = M_{ij}\int_{\varphi_{ij}^*}^{\infty} p_{ij}^c(\varphi)c_{ij}(\varphi)\mu_{ij}(\varphi)d\varphi = M_{ij}b_i^{1-\sigma}E_jP_j^{\sigma-1}\int_{\varphi_{ij}^*}^{\infty} p_{ij}^c(\varphi)^{1-\sigma}\mu_{ij}(\varphi)d\varphi. \tag{23}$$

Note that the bilateral import unit value, denoted $\overline{p}_{ij}^c$, can be expressed as a function solely of the two unobservable price integrals in Eqs. (23) and (22):

$$\overline{p}_{ij}^c \equiv \frac{X_{ij}^D}{Q_{ij}^D} = \frac{\int_{\varphi_{ij}^*}^{\infty} p_{ij}^c(\varphi)^{1-\sigma}\mu_{ij}(\varphi)d\varphi}{\int_{\varphi_{ij}^*}^{\infty} p_{ij}^c(\varphi)^{-\sigma}\mu_{ij}(\varphi)d\varphi} \tag{24}$$

Recall from section 2, we introduced a Pareto distribution for heterogeneous productivities in order to obtain *closed-form* solutions, as common to theoretical Melitz models. As shown in Arkolakis (2010) and Eaton et al. (2011), empirically the Pareto distribution does not approximate firms' sales distribution very well for small exporters, with heterogeneity in this effect across country-pairs. Hence, we introduce two multiplicative error terms e_{ij}^{P1} and e_{ij}^{P2} for $\int_{\varphi_{ij}^*}^{\infty} p_{ij}^c(\varphi)^{1-\sigma}\mu_{ij}(\varphi)d\varphi$ and $\int_{\varphi_{ij}^*}^{\infty} p_{ij}^c(\varphi)^{-\sigma}\mu_{ij}(\varphi)d\varphi$, respectively. As we explain below, the deviations from the Pareto distribution will play a central role in the identification of the structural parameters of the model.

We proceed in two steps. First, as shown in section 1 of Online Appendix D, we can solve for $p_{ij}^c(\varphi)$ as a function of its productivity threshold φ_{ij}^*. Recalling $c_{ij}(\varphi) = q_{ij}(\varphi)/\tau_{ij}$ and $p_{ij}^c(\varphi) = \tau_{ij}p_{ij}(\varphi)$, we can use optimal demand Eq. (2), optimal pricing rule (6), and the Pareto distribution *allowing deviations* to yield:

$$\int_{\varphi_{ij}^*}^{\infty} p_{ij}^c(\varphi)^{1-\sigma}\mu_{ij}(\varphi)d\varphi = \left[\frac{\theta(\sigma+\gamma)}{\theta(\sigma+\gamma)-\gamma(\sigma-1)}\right]\left[p_{ij}^c\left(\varphi_{ij}^*\right)\right]^{1-\sigma}e_{ij}^{P1}, \tag{25}$$

$$\int_{\varphi_{ij}^*}^{\infty} p_{ij}^c(\varphi)^{-\sigma}\mu_{ij}(\varphi)d\varphi = \left[\frac{\theta(\sigma+\gamma)}{\theta(\sigma+\gamma)-\gamma\sigma}\right]\left[p_{ij}^c\left(\varphi_{ij}^*\right)\right]^{-\sigma}e_{ij}^{P2}. \tag{26}$$

Using Eq. (24) and some algebra yields:

$$p_{ij}^c\left(\varphi_{ij}^*\right) = \left[\frac{\theta(\sigma+\gamma)-\gamma(\sigma-1)}{\theta(\sigma+\gamma)-\gamma\sigma}\right]\overline{p}_{ij}^c\left(\frac{e_{ij}^{P2}}{e_{ij}^{P1}}\right). \tag{27}$$

Second, we can combine the results in Eqs. (23), (25) and (27) – after first substituting Eq. (9) to replace the productivity threshold φ_{ij}^* and an extended version of Eq. (11) to allow for deviations (e_{ij}^{P3}) from Pareto for the endogenous mass of firms M_{ij} – to express the share of aggregate nominal bilateral trade flow in total expenditures (s_{ij}) as a function of bilateral import unit value. Following F/BW, we eliminate time-invariant factors by first differencing the resulting structural equation and then we eliminate importer-specific variables by taking a difference with respect to a reference exporting country k, obtaining:

$$\Delta^k \ln s_{ijt} = (1-\sigma)\Delta^k \ln \overline{p}_{ijt}^c + \delta_{ijt}. \tag{28}$$

[25] In this section and the next one, for brevity we omit the time subscript, t (as well as the industry subscript, as earlier). In section 4.1.4, we reintroduce the time subscript.

where we define a new term δ_{ijt} as:

$$\begin{aligned}\delta_{ijt} = {} & \left[1+\theta\left(\frac{1+\gamma}{\gamma}\right)\left(\frac{\sigma}{\sigma-1}\right)\right]\Delta^k \ln A_{it} + \Delta^k \ln L_{it} - \theta\left(\frac{1+\gamma}{\gamma}\right)\left(\frac{\sigma}{\sigma-1}\right)\Delta^k \ln w_{it} \\ & - \theta\left(\frac{1+\gamma}{\gamma}\right)\Delta^k \ln b_{it} - \theta\left(\frac{1+\gamma}{\gamma}\right)\Delta^k \ln \tau_{ijt} - \frac{\theta\left(\frac{1+\gamma}{\gamma}\right)}{\frac{1+\gamma}{\sigma+\gamma}(\sigma-1)}\Delta^k \ln f_{ijt} + \epsilon_{ijt}.\end{aligned} \tag{29}$$

where (error term) $\epsilon_{ijt} \equiv \Delta^k \ln e_{ijt}^{P1} + (1-\sigma)\Delta^k \ln e_{ijt}^{P2} + \Delta^k \ln e_{ijt}^{P3}$. The general form of Eq. (28) corresponds to the benchmark F/BW structural bilateral trade-flow-share Eq. (18) in section 4.1.1 with three notable differences: (i) δ_{ijt} includes a host of additional variables (beyond $\Delta^k \ln \overline{p}_{ijt}^c$ in Eq. (18)); (ii) error term ϵ_{ijt} in Eq. (18) now has a clear interpretation; and (iii) some of the additional variables are unobservable (e.g., A_{it} and b_{it}).

First, in the context of our general equilibrium Melitz model, numerous determinants of the mass of varieties exported from i and the cutoff productivity need also to be accounted for in the trade-flow-share equation ($A_{it}, L_{it}, w_{it}, b_{it}, \tau_{ijt},$ and f_{ijt}). In their absence, coefficient estimates in benchmark F/BW reduced forms may be substantially biased.

Second, the literature beginning with Feenstra (1994) has assumed that the error term in the basic F/BW structural trade-flow-share equation can be interpreted simply as a "taste shock." However, in the context of our general equilibrium framework, b_{it} is a determinant of the trade-flow share and so cannot represent the error term. By contrast, in our framework ϵ_{ijt} is driven by deviations from the Pareto distribution for productivities across country pairs that influence $[\int_{\varphi_{ij}^*}^{\infty} p_{ij}^c(\varphi)^{1-\sigma}\mu_{ij}(\varphi)d\varphi]$, $[\int_{\varphi_{ij}^*}^{\infty} p_{ij}^c(\varphi)^{-\sigma}\mu_{ij}(\varphi)d\varphi]$, and M_{ij}. Note that $\Delta^k \ln e_{ijt}^{P1}$, $\Delta^k \ln e_{ijt}^{P2}$, and $\Delta^k \ln e_{ijt}^{P3}$ all have expected values of zero.

Third, several of the additional variables – $A_{it}, L_{it}, w_{it},$ and b_{it} – are exporter-specific variables but unobservable (A_{it}, b_{it}) or difficult-to-measure across countries and over time at the industry level (L_{it}, w_{it}). At the same time, as discussed below, the coefficients on these variables will not be relevant to estimating $\sigma, \gamma,$ and θ in section 5 and conducting our counterfactual exercises in section 6. Consequently, we can hold constant the influences of these four variables by employing an exporter-year fixed effect, labeled α_{it}^1, allowing us to rewrite Eq. (29) more succinctly as:

$$\delta_{ijt} = \alpha_{it}^1 - \theta\left(\frac{1+\gamma}{\gamma}\right)\Delta^k \ln \tau_{ijt} - \frac{\theta\frac{1+\gamma}{\gamma}}{\frac{1+\gamma}{\sigma+\gamma}(\sigma-1)}\Delta^k \ln f_{ijt} + \epsilon_{ijt}. \tag{30}$$

4.1.3. Bilateral export supply

Turning our attention to the bilateral export supply equation, we can invert the optimal pricing function (6) to get an analytical expression for output as a function of the price. Using the result, we can define average bilateral export supply (in physical units) as:

$$\overline{q}_{ij}^S \equiv \int_{\varphi_{ij}^*}^{\infty} q_{ij}(\varphi)\mu_{ij}(\varphi)d\varphi = \left[\left(\frac{\gamma}{1+\gamma}\right)\left(\frac{\sigma-1}{\sigma}\right)\frac{A_i}{w_i}\right]^{\gamma}\int_{\varphi_{ij}^*}^{\infty}\left[\varphi p_{ij}(\varphi)\right]^{\gamma}\mu_{ij}(\varphi)d\varphi. \tag{31}$$

Defining industry bilateral export supply (in physical units) as $Q_{ij}^S \equiv M_{ij}\overline{q}_{ij}^S$, using Eq. (31) yields any industry's bilateral export supply:

$$Q_{ij}^S = M_{ij}\left[\left(\frac{\gamma}{1+\gamma}\right)\left(\frac{\sigma-1}{\sigma}\right)\frac{A_i}{w_i}\right]^{\gamma}\int_{\varphi_{ij}^*}^{\infty}\left[\varphi p_{ij}(\varphi)\right]^{\gamma}\mu_{ij}(\varphi)d\varphi. \tag{32}$$

Because the integral over firm-level prices and productivity is not observable, we need to solve for it as a function of the observed bilateral import unit value.

Similar to the import demand equation, we proceed in several steps. First, we solve for the integral of firm-level prices as a function of the ZCP firm's productivity level and price. We can use optimal demand Eq. (2), the optimal pricing rule (6), and Eq. (8) for the output of the ZCP firm to derive an optimal pricing equation that is a function of $f_{ij}, \varphi_{ij}^*, w_i/A_i,$ and φ. Substituting this optimal price equation into the price integral in Eq. (32), assuming our Pareto distribution allowing for deviations e_{ij}^{P5}, and solving yields:

$$\int_{\varphi_{ij}^*}^{\infty}\left[\varphi p_{ij}(\varphi)\right]^{\gamma}\mu_{ij}(\varphi)d\varphi = \frac{\theta(\sigma+\gamma)}{\theta(\sigma+\gamma)-\gamma\sigma}\left[\varphi_{ij}^* p_{ij}\left(\varphi_{ij}^*\right)\right]^{\gamma}e_{ij}^{P5}. \tag{33}$$

Second, we can use Eq. (27) and the fact that $p_{ij}^c = \tau_{ij} p_{ij}$ to define $p_{ij}\left(\varphi_{ij}^*\right)$. Substituting with the result in Eq. (33) yields:

$$\int_{\varphi_{ij}^*}^{\infty} \left[\varphi p_{ij}(\varphi)\right]^{\gamma} \mu_{ij}(\varphi) d\varphi = \frac{\theta(\sigma+\gamma)}{\theta(\sigma+\gamma)-\gamma\sigma} \left[\frac{\theta(\sigma+\gamma)-\gamma(\sigma-1)}{\theta(\sigma+\gamma)-\gamma\sigma}\right]^{\gamma} \left(\overline{p}_{ij}\right)^{\gamma} \left(\varphi_{ij}^*\right)^{\gamma} e_{ij}^{P5}. \tag{34}$$

Third, substituting the RHS of Eq. (34) for the integral in Eq. (32) and solving for average price yields:

$$\overline{p}_{ij} = k_4^{-\frac{1}{\gamma}} \left(\frac{Q_{ij}^S}{M_{ij}}\right)^{\frac{1}{\gamma}} \frac{w_i}{A_i \varphi_{ij}^*} \left(e_{ij}^{P5}\right)^{-\frac{1}{\gamma}}. \tag{35}$$

where k_4 is a constant that depends only on the structural parameters σ, γ, and θ (defined in section 2 of Online Appendix D). Fourth, we make the industry bilateral export-supply Eq. (35) comparable to the industry bilateral trade-flow-share equation by eliminating Q_{ij}^S, M_{ij}, and φ_{ij}^*. The value of industry bilateral export supply (X_{ij}^S) equals the value of bilateral import demand (X_{ij}^D), such that $M_{ij}\overline{p}_{ij}\overline{q}_{ij}^S = \frac{X_{ij}^D}{E_j} E_j$ or $Q_{ij}^S = \tau_{ij} \frac{X_{ij}^D}{E_j} \frac{E_j}{\overline{p}_{ij}^c}$ (recalling that $\overline{p}_{ij}^c = \tau_{ij}\overline{p}_{ij}$). Substituting this expression for Q_{ij}^S in Eq. (35), substituting for φ_{ij}^* using Eq. (9) and for M_{ij} using an extended version of Eq. (11) allowing deviations from Pareto, substituting s_{ij} for X_{ij}^D/E_j as in the previous section, solving the resulting expression for $\overline{p}_{ij}^c$, and double-differencing the resulting equation yields:

$$\Delta^k \ln \overline{p}_{ijt}^c = \left(\frac{1}{1+\gamma}\right) \Delta^k \ln s_{ijt} + \eta_{ijt}, \tag{36}$$

where we define η_{ijt} as:

$$\eta_{ijt} = \alpha_{it}^2 + \frac{\theta-\gamma}{\gamma} \Delta^k \ln \tau_{ijt} + \frac{\frac{\theta-\gamma}{\gamma}}{\frac{1+\gamma}{\sigma+\gamma}(\sigma-1)} \Delta^k \ln f_{ijt} + \psi_{ijt} \tag{37}$$

where $\psi_{ijt} \equiv -\frac{1}{1+\gamma}\Delta^k \ \ln e_{ijt}^{P3} - \frac{1}{1+\gamma}\Delta^k \ \ln e_{ijt}^{P5}$ and α_{it}^2 is an exporter-year fixed effect.[26] The general form of Eq. (36) corresponds to the benchmark F/BW structural bilateral import unit value Eq. (19) in section 4.1.1 with the three notable differences analogous to Eq. (29): (i) η_{ijt} includes the same host of additional variables (beyond $\Delta^k \ln s_{ijt}$ in Eq. (19)); (ii) error term ψ_{ijt} now has a clear interpretation; and (iii) some of the additional variables are unobservable (e.g., A_{it}). First, if we ignore the additional covariates, the estimated coefficient will suffer from omitted variables bias. Second, the literature beginning with Feenstra (1994) has assumed that the error term in the structural bilateral import unit value equation can be interpreted simply as a "technology shock." However, our general equilibrium framework shows that A_{it} is a determinant of the bilateral import unit value and so cannot represent the error term. Third, several of these additional variables are exporter-specific and unobservable or difficult to measure, suggesting inclusion of an analogous exporter fixed effect term α_{ijt}^2.

It will be useful at this point to note that, in the trade literature, *ad valorem* variable trade costs τ_{ijt} typically reflect the product of gross tariff rates, labeled tar_{ijt}, and gross c.i.f.-f.o.b. transport-cost factors, labeled $trans_{ijt}$ (both of which are greater than 1). Consequently, in the following empirical specifications, we account for both components of variable trade costs separately.

4.1.4. Reduced-form specifications and data issues

Having motivated Eqs. (28), (30), (36), and (37), we are now in a position to derive our reduced-form specifications, following in the spirit of F/BW and section 4.1.1. First, we can substitute Eq. (30) for δ_{ijt} in Eq. (28), and then solve for ϵ_{ijt} on the LHS. Second, we can substitute Eq. (37) for η_{ijt} in Eq. (36), and then solve for ψ_{ijt} on the LHS. Third, we take the product of the two expressions and rearrange terms to yield:

$$Y_{ijt} = \sum_{k=1}^{20} \beta_k Z_{ijt,k} + \xi_{ijt}, \tag{38}$$

[26] For brevity, we omit here the analogue to equation (29); refer to Eq. (D.13) in Online Appendix D for guidance on this expression. Another error term e_{ijt}^{P4} is embedded inside e_{ijt}^{P3} due to the role of the deviation from Pareto influencing the mass for firms M_{ijt}; see section 1 of Online Appendix D.

where

$$
\begin{array}{llll}
\text{Group1}: & Y_{ijt} = \left(\Delta^k \ \ln \overline{p}^c_{ijt}\right)^2, & Z_{ijt,1} = \left(\Delta^k \ \ln s_{ijt}\right)^2, & Z_{ijt,2} = \Delta^k \ \ln s_{ijt} \Delta^k \ \ln \overline{p}^c_{ijt}, \\
\text{Group2}: & Z_{ijt,3} = \Delta^k \ \ln \overline{p}^c_{ijt} \Delta^k \ \ln tar_{ijt}, & Z_{ijt,4} = \Delta^k \ \ln s_{ijt} \Delta^k \ \ln tar_{ijt}, & Z_{ijt,5} = \left(\Delta^k \ \ln tar_{ijt}\right)^2, \\
 & Z_{ijt,6} = \Delta^k \ \ln \overline{p}^c_{ijt} \Delta^k \ \ln trans_{ijt}, & Z_{ijt,7} = \Delta^k \ \ln s_{ijt} \Delta^k \ \ln trans_{ijt}, & Z_{ijt,8} = \left(\Delta^k \ \ln trans_{ijt}\right)^2, \\
 & Z_{ijt,9} = \Delta^k \ \ln tar_{ijt} \Delta^k \ \ln trans_{ijt}, & & \\
\text{Group3}: & Z_{ijt,10} = \alpha_{it}, & Z_{ijt,11} = \alpha_{it} \Delta^k \ \ln \overline{p}^c_{ijt}, & Z_{ijt,12} = \alpha_{it} \Delta^k \ \ln s_{ijt}, \\
 & Z_{ijt,13} = \alpha_{it} \Delta^k \ \ln tar_{ijt}, & Z_{ijt,14} = \alpha_{it} \Delta^k \ \ln trans_{ijt}, & \\
\text{Group4}: & Z_{ijt,15} = \Delta^k \ \ln s_{ijt} \Delta^k \ \ln f_{ijt}, & Z_{ijt,16} = \Delta^k \ \ln \overline{p}^c_{ijt} \Delta^k \ \ln f_{ijt}, & Z_{ijt,17} = \Delta^k \ \ln tar_{ijt} \Delta^k \ \ln f_{ijt}, \\
 & Z_{ijt,18} = \Delta^k \ \ln trans_{ijt} \Delta^k \ \ln f_{ijt}, & Z_{ijt,19} = \left(\Delta^k \ \ln f_{ijt}\right)^2, & Z_{ijt,20} = \alpha_{it} \Delta^k \ \ln f_{ijt},
\end{array}
$$

where $\xi_{ijt} \equiv \epsilon_{ijt}\psi_{ijt}$ is a residual. We will explain shortly the relevance of the "Groups" for motivating the specifications.

The 20 β_k's are functions of only *three* structural parameters: $\sigma, \gamma,$ and θ. The first two coefficients, β_1 and β_2, are defined exactly as in F/BW:

$$\beta_1 = \frac{1}{(1+\gamma)(\sigma-1)}, \quad \text{and} \quad \beta_2 = \frac{\sigma-\gamma-2}{(1+\gamma)(\sigma-1)}. \tag{39}$$

Hence, we can use the same methodology as F/BW to back out structural parameters σ and γ from the reduced-form estimates of β_1 and β_2. Importantly, in the context of our Melitz (2003) model with firm heterogeneity and IMC, Eq. (38) makes it clear that estimation of the first two RHS variables will suffer from omitted variable bias (OVB) if variables in Groups 2–4 are not accounted for in the reduced-form specification.

Following F/BW, a consistent estimator of coefficients $\beta_1 - \beta_{20}$ can be obtained by averaging each of the variables in Eq. (38) over all $t = 1, \ldots, T$. Letting $\overline{Y}_{ij}, \overline{Z}_{1,ij}, \ldots, \overline{Z}_{2,ij}$, and $\overline{\xi}_{ij} = \overline{\epsilon_{ij}\psi_{ij}}$ denote the means, this yields the reduced form equation for estimation:

$$\overline{Y}_{ij} = \beta_0 + \sum_{k=1}^{20} \beta_k \overline{Z}_{k,ij} + \overline{\xi}_{ij}, \tag{40}$$

where the over-bar indicates that the variables are averages over time (e.g., $\overline{Z}_{ij} \equiv T^{-1}\sum_{t=1}^{T} Z_{ijt}$). In the remainder of this subsection, we describe the three specifications we estimate along with relevant data needs.

4.1.4.1. Specification 1: F/BW. As a benchmark, the first specification we estimate includes the three variables in Group 1 only. This is exactly the same specification as in F/BW. According to our model, the (reduced-form and structural) coefficient estimates will be biased because of omitted variables.

For estimation, we need data on trade flows in values and in quantities at the industry level. Data for trade flows come from the United Nations' Comtrade Database. This database collects bilateral f.o.b. export values that correspond to the transaction value of the goods, as well as bilateral c.i.f. import values which include the value of services performed to deliver goods to the border of the importing country. This database also contains information on the quantities exported and imported.[27] We combine the measures of trade flows and expenditures by industry to construct bilateral trade-flow shares and we combine measures of bilateral import values and quantities to construct bilateral import unit values. For our analysis, we define industries as four-digit Standard Industrial Trade Classification (SITC4) categories. Our sample covers the period from 1999 through 2010; after taking the time differences, we end up with years 2000–2010.

4.1.4.2. Specification 2: IMC-Partial. The second specification we estimate includes the variables in Groups 1 and 2. To generate a sense of the importance of the variables in Group 2 (all related to variable trade costs) for correcting for omitted variables bias, we provide a stand-alone specification including the nine RHS variables. For illustrative purposes, we provide in section 3 of Online Appendix D the derivations for the theoretical coefficients associated with this specification, labeled IMC-Partial. As explained earlier, the coefficients depend on only three structural parameters. However, the nine non-linear restrictions implied by the model prevent identification of all three parameters from this single reduced-form equation; specifically, the large number of restrictions preclude identification of θ. Nevertheless, the seven additional RHS variables are included to control for OVB, and estimates of σ and γ can *still* be determined from the estimates of β_1 and β_2.

Using the United Nations' Comtrade data discussed above, we construct *ad valorem* measures of gross transport costs factors

[27] When possible, we convert physical units of measurement to a common denominator (e.g., "Thousands of items" to "Items"). For industries with multiple units of measurement, we keep only the observations which report physical quantity in the unit of measurement that account for the largest value of import over the entire sample.

$trans_{ijt}$ from the ratios of the c.i.f. to the f.o.b. unit values. Feenstra and Romalis (2014) provide a database of *ad valorem* tariff rates based upon Most-Favored-Nation (MFN) status or any preferential status available, from which we construct tar_{ijt}.[28] The tariff rates are reported at the four-digit SITC level.

4.1.4.3. Specification 3: IMC-Full. The third specification will account fully for the variables *in all four groups*, and will be our preferred specification to address OVB. As with the IMC-Partial specification, the numerous non-linear restrictions preclude estimation of θ; however, we can still determine σ and γ from the estimates of β_1 and β_2. Later, with reduced-form gravity estimates of the trade elasticity, we will be able to determine θ. We discuss the motivation for this specification in two parts.

First, we address the Group 3 variables. A major benefit of specification IMC-Full is that the inclusion of the exporter fixed effects α_i (alongside the Groups 1 and 2 variables) in the reduced-form equations of the time-averaged variables *eliminates* having to include measures of A_i, L_i, w_i, and b_i; recall that A_i and b_i are unobservable. Nevertheless, for a robustness analysis, we will discuss later a specification including the variables in Groups 1 and 2 and crude proxies for L_i and w_i, but without exporter fixed effects and their interactions (which implies omitting controls for A_i and b_i).[29]

Second, we address the variables in Group 4, f_{ijt} and its interaction terms. While quality data exists on *ad valorem* tariff rates and transport costs, the international trade literature has so far struggled to construct and implement persuasive measures of bilateral fixed trade costs that affect only the decisions to export to a foreign market. To date, the most comprehensive effort to measure these fixed costs is the World Bank's *Doing Business* (DB) indicators. Covering a comprehensive swath of countries over multiple years, the DB indicators provide a widely respected quantification of the "ease of doing business" along numerous dimensions. However, unlike the theoretical variable, f_{ijt}, which is country-pair specific, the DB indicators are *country specific*.[30]

The World Bank also provides the *Deep Trade Agreements* database described in Hofmann et al. (2017) and Mattoo et al. (2020).[31] Hofmann et al. (2017) note that liberalizations of multilateral provisions are much more common, compared to liberalizations of bilateral provisions. From years 1980–84 to 2010–15, the average number of multilateral provisions has more than doubled from 4 to 9. By contrast, for the same 30 year period, the average number of bilateral provisions has increased by only 1, from 4 to 5. This all suggests that reductions in fixed trade costs due to deep trade agreements are largely captured by exporter-specific and importer-specific components of fixed trade costs.

Consequently, our third specification, IMC-Full, can capture the multilateral influences of f_{ij} owing to the inclusion of an exporter fixed effect and exporter fixed effects interacted with $\overline{\Delta^k \ln \overline{p}_{ij}^c}, \overline{\Delta^k \ln s_{ij}}, \overline{\Delta^k \ln tar_{ij}}$, and $\overline{\Delta^k \ln trans_{ij}}$. The rationale is the following. Since the discussion above suggests that most of the variation in $\ln f_{ijt}$ can be explained across country-pairs and over time by variation in $\ln f_{it}$ and $\ln f_{jt}$ – treating the remaining variation as a residual, $\ln f_{ijt}^R$ – the introduction of an exporter fixed effect, alongside the differencing with respect to reference exporting country k (which removes importer effects, such as the influences of E_{jt} and P_{jt}), can account for most of the variation in $\ln f_{ijt}$, as long as the exporter-fixed-effect interactions are present. To see this, consider the variable $\left(\Delta^k \ln s_{ijt}\right)\left(\Delta^k \ln f_{ijt}\right)$ from Eq. (38). The differencing with respect to exporting reference country k removes the variation and influence of $\ln f_{jt}$, leaving variation in $\ln f_{it}$ and $\ln f_{ijt}^R$. Assume $\Delta^k \ln f_{ijt}^R$ is randomly distributed with mean 0 and variance $\sigma^2_{\Delta \ln f_{ijt}^R}$; we address this later. Suppose $\Delta^k \ln f_{it}$ follows a random walk with a drift (Φ_i); hence, $\Delta^k \ln f_{it} = \Phi_i + u_{it}$. Substituting $\Phi_i + u_{it}$ for $\Delta \ln f_{it}$ yields:

$$\left(\Delta^k \ln s_{ijt}\right)\left(\Delta^k \ln f_{ijt}\right) = \left(\Delta^k \ln s_{ijt}\right)\Phi_i + \left(\Delta^k \ln s_{ijt}\right)u_{it} + \left(\Delta^k \ln s_{ijt}\right)\left(\Delta^k \ln f_{ijt}^R\right). \tag{41}$$

[28] This database combines information from the TRAINS data, the World Trade Organization's (WTO) Integrated Data Base, the International Customs Journal, and the texts of preferential trade agreements obtained from the WTO's website.

[29] In the robustness specification we will report later, we use per capita GDPs of countries as a proxy for w_{it}. Information on employment is not available at this level of detail. Instead, as a proxy for L_{it}, we obtain an estimate of employment. We follow Feenstra and Romalis (2014) and distribute employment across industries in proportion to export production. For each industry-country-year category, we measure employment as total employment multiplied by industry export value divided by GDP. Information on employment and GDP for each country-year is from the Penn World Tables (version 9.1).

[30] WorldBank (2020), Table 1.1 identifies 12 major country-specific categories of fixed costs that cover policy (artificial) and non-policy (natural) fixed costs associates with an importing country, ranging across base of "starting a new business, getting a location, accessing finance, dealing with day-to-day operations, and operating in a secure business environment." All such elements influence the decision of a potential exporter to enter a foreign market.

[31] This database is the first comprehensive source of information using dummy variables to indicate the presence or absence of each of 937 "deep" provisions within 219 preferential trade agreements (PTAs) between pairings of 189 countries annually from 1958 to 2017. Fortunately, Hofmann et al. (2017) identify so-called "core" provisions that dominate the DTAs. These core provisions are grouped in 16 "policy areas" (excluding industrial and agricultural tariff rates, common to 98% or of PTAs). The 16 policy areas are: competition policy, investment, movement of capital, intellectual property rights, customs (facilitation), technical barriers to trade, sanitary and photo-sanitary, state aid, GATS (services), TRIPS (intellectual property), state trading enterprises, TRIMS (investment measure state), export taxes, anti-dumping provisions, countervailing measures, and public procurement. As noted in Hofmann et al. (2017), liberalization in all 16 policy areas can be categorized into multilateral (or non-discriminatory) and bilateral (or preferential) liberalization. Of the 16 areas, 12 are considered multilateral in nature and only four are considered bilateral. Furthermore, the four "core-WTO-X" policy-areas (competition policy, intellectual property rights, investment, and movement of capital) that are considered "important features of DTAs" are multilateral in nature, with almost 90% of PTAs including at least one of them.

Summing both sides of Eq. (41) over $t = 1, \ldots, T$ and dividing both sides by T yields:

$$\overline{\left(\Delta^k \ln s_{ij}\right)\left(\Delta^k \ln f_{ij}\right)} = \Phi_i\left(\Delta^k \ln s_{ijt}\right) \tag{42}$$

because the terms $(1/T)\sum_{t=1}^{T}\left(\Delta^k \ln s_{ijt}\right)u_{it}$ and $(1/T)\sum_{t=1}^{T}\left(\Delta^k \ln s_{ijt}\right)\left(\Delta^k \ln f^R_{ijt}\right)$ are covariances and both covariances are zero.[32]

Finally, it is important to draw attention to the fact that ϵ_{ijt} is a linear function of $\Delta^k \ln e^{P1}_{ijt}, \Delta^k \ln e^{P2}_{ijt}$, and $\Delta^k \ln e^{P3}_{ijt}$, but ψ_{ijt} is a *different* linear function of $\Delta^k \ln e^{P3}_{ijt}$ and $\Delta^k \ln e^{P5}_{ijt}$. The necessary moment condition in this framework to employ the method-of-moments estimator is that $\mathbb{E}\left(\epsilon_{ijt}\psi_{ijt}\right) = 0$. However, in the presence of an intercept in the estimating reduced form, $\mathbb{E}\left(\epsilon_{ijt}\psi_{ijt}\right)$ need only be a constant; we will demonstrate this in the next subsection. Furthermore, identification requires that the relative variances (over time) of ϵ_{ijt} and ψ_{ijt} differ; we will demonstrate this in the next subsection as well.

4.1.5. Moment and identification conditions

Estimation of Eq. (40) produces consistent coefficient estimates under two conditions. The first is the "moment" condition, $\mathbb{E}\left(\xi_{ijt}\right) \equiv \mathbb{E}\left(\epsilon_{ijt}\psi_{ijt}\right) = 0$. Recalling $\epsilon_{ijt} \equiv \Delta^k \ln e^{P1}_{ijt} + (1-\sigma)\Delta^k \ln e^{P2}_{ijt} + \Delta^k \ln e^{P3}_{ijt}$ and $\psi_{ijt} \equiv -\frac{1}{1+\gamma}\Delta^k \ln e^{P3}_{ijt} - \frac{1}{1+\gamma}\Delta^k \ln e^{P5}_{ijt}$, we show in Online Appendix D that:

$$\mathbb{E}\left(\epsilon_{ijt}\psi_{ijt}\right) = -\left(\frac{1}{1+\gamma}\right) \operatorname{var}\left(\Delta^k \ \ln e^{P3}_{ijt}\right) = -4\left(\frac{1}{1+\gamma}\right) \operatorname{var}\left(\ \ln e^{P3}_{ijt}\right) \equiv -4\left(\frac{1}{1+\gamma}\right)\sigma^2_{\ln e^{P3}_{ij}} \tag{43}$$

is a constant, where $\sigma^2_{\ln e^{P3}_{ij}}$ denotes the variance over time of $\ln e^{P3}_{ijt}$.[33] Hence, the moment condition $\mathbb{E}\left(\epsilon_{ijt}\psi_{ijt}\right) = 0$ is met as long as Eq. (40) includes an intercept, β_0 (as in Feenstra (1994)).

The second is the "identification" condition, which requires that the relative variances of ϵ_{ij} across country-pairs differ from the relative variances of ψ_{ij} across country-pairs. Modifying Eq. (12) in Feenstra (1994), identification in our context requires:

$$\frac{\sigma^2_{\epsilon_{ij}} + \sigma^2_{\epsilon_{kl}}}{\sigma^2_{\epsilon_{mn}} + \sigma^2_{\epsilon_{kl}}} \neq \frac{\sigma^2_{\psi_{ij}} + \sigma^2_{\psi_{kl}}}{\sigma^2_{\psi_{mn}} + \sigma^2_{\psi_{kl}}} \tag{44}$$

where σ^2_z, as above, denotes the variance over time of variable z. In terms of our model, the equivalent expression for identification is:

$$\begin{aligned}&\frac{\left(\sigma^2_{\Delta^k \ \ln e^{P1}_{ij}}\right) + (1-\sigma)^2\left(\sigma^2_{\Delta^k \ \ln e^{P2}_{ij}}\right) + \left(\sigma^2_{\Delta^k \ \ln e^{P3}_{ij}}\right) + \left(\sigma^2_{\Delta^k \ \ln e^{P1}_{kl}}\right) + (1-\sigma)^2\left(\sigma^2_{\Delta^k \ \ln e^{P2}_{kl}}\right) + \left(\sigma^2_{\Delta^k \ \ln e^{P3}_{kl}}\right)}{\left(\sigma^2_{\Delta^k \ \ln e^{P1}_{mn}}\right) + (1-\sigma)^2\left(\sigma^2_{\Delta^k \ \ln e^{P2}_{mn}}\right) + \left(\sigma^2_{\Delta^k \ \ln e^{P3}_{mn}}\right) + \left(\sigma^2_{\Delta^k \ \ln e^{P1}_{kl}}\right) + (1-\sigma)^2\left(\sigma^2_{\Delta^k \ \ln e^{P2}_{kl}}\right) + \left(\sigma^2_{\Delta^k \ \ln e^{P3}_{kl}}\right)}\\ &\neq \frac{\sigma^2_{\Delta^k \ \ln e^{P3}_{ij}} + \sigma^2_{\Delta^k \ \ln e^{P5}_{ij}} + \sigma^2_{\Delta^k \ \ln e^{P3}_{kl}} + \sigma^2_{\Delta^k \ \ln e^{P5}_{kl}}}{\sigma^2_{\Delta^k \ \ln e^{P3}_{mn}} + \sigma^2_{\Delta^k \ \ln e^{P5}_{mn}} + \sigma^2_{\Delta^k \ \ln e^{P3}_{kl}} + \sigma^2_{\Delta^k \ \ln e^{P5}_{kl}}}\end{aligned} \tag{45}$$

where *ij*, *kl*, and *mn* denote different country-pairs.

The condition above requires that there must be *some differences* in the relative variances of the "demand" (ϵ_{ij}) and "supply" (ψ_{ij}) disturbances. Although many factors can explain such differences, the key consideration is that the LHS and RHS of equation (45) include variances of time-differenced (as well as reference-exporting-country differenced) Pareto deviations of integrals over different variables. For instance, on the LHS $\Delta^k \ln e^{P1}_{ij}$ refers to the double-differenced deviations associated with the (demand-side) integral $\int_{\varphi^*_{ij}}^{\infty}[\tau_{ij}p_{ij}(\varphi)]^{1-\sigma}\mu_{ij}(\varphi)d\varphi$. By contrast, on the RHS $\Delta^k \ln e^{P5}_{ij}$ refers to the double-differenced deviations associated with the (supply-side) integral $\int_{\varphi^*_{ij}}^{\infty}\left[\varphi p_{ij}(\varphi)\right]^{\gamma}\mu_{ij}(\varphi)d\varphi$. Hence, the inequality condition (45) is likely to hold. It can be shown numerically that if the relative variances are different, condition (45) does hold.

4.2. Gravity equation with firm heterogeneity

As discussed throughout section 4 so far, our extension of the F/BW framework precludes estimation of θ using the F/BW reduced-form equation, due to the non-linear restrictions issue raised above. The previous sections motivate estimation of σ and γ, but only under explicit controls for exporter masses and export productivity cutoffs to avoid omitted variables bias as

[32] Note that $\Delta \ln f^R_{ijt}$ needs to be accounted for also in the moment condition and in the identification condition to be discussed shortly below. Online Appendix D addresses how each of these conditions discussed in section 4.1.5 is altered in an inconsequential manner.

[33] Note that σ still denotes the elasticity of substitution in consumption whereas σ^2_z denotes the variance of the variable in the subscript (e.g., z).

shown by Eqs. (38) and (40). However, our general equilibrium model suggests gravity Eq. (13), as developed in sections 2.4 and 2.5. Consistent with the gravity-equation literature, estimates of the "trade elasticity" – the elasticity of bilateral trade flows with respect to *ad valorem* variable trade costs – in the context of our model provide reduced-form estimates of $-\theta\left(\frac{1+\gamma}{\gamma}\right)$. Using the trade-elasticity estimates by sector along with estimates of γ discussed above (by sector), industry-specific estimates of θ are readily determined. With all three structural parameters, numerical counterfactuals then can be performed in section 6.

We follow the econometric literature for estimating trade elasticities in the presence of panel data, where here we use industry-level nominal bilateral trade flows. Much of the recent literature on estimation of trade-policy effects using panel data in gravity equations follows Baier and Bergstrand (2007) and Baier et al. (2008, 2014, 2018). Consistent with these papers, the trade elasticity can be identified using a log-linear regression equation of bilateral trade flows (X_{ijt}) on exporter-year fixed effects, importer-year fixed effects, and a measure of *ad valorem* bilateral trade costs τ_{ijt}. Using Eq. (13), such a specification is:

$$X_{ijt} = \phi_{it} + \Psi_{jt} - \theta\left(\frac{1+\gamma}{\gamma}\right) \ln \tau_{ijt} + \left[1 - \frac{\theta\left(\frac{1+\gamma}{\gamma}\right)}{\frac{1+\gamma}{\sigma+\gamma}(\sigma-1)}\right] \ln f_{ijt} + \vartheta_{ijt} \tag{46}$$

where ϕ_{it} are exporter-year fixed effects capturing the influences of $A_{it}, L_{it}, w_{it},$ and b_{it} in Eq. (13), Ψ_{jt} are importer-year fixed effects capturing the roles of $L_{jt}, w_{jt},$ and the multilateral price/resistance term of the importer in the denominator of the first RHS term (in brackets) in Eq. (13), and ϑ_{ijt} is an error term.

In estimating Eq. (46), three issues surface. The first is that τ_{ijt} is associated with both (gross) bilateral tariff rates, tar_{ijt}, and c.i.f.-f.o.b. transport-cost factors, $trans_{ijt}$. We introduce these variables separately. However, due to more confidence in the observed measures of tariff rates, we use estimates of the trade elasticity from the tariff-rate variable.

The second issue concerns the potential endogeneity of the tariff-rate variable. There is an extensive literature noting the potential endogeneity of tariff rates and/or dummy variables for economic integration agreements, cf., Trefler (1993) and Baier and Bergstrand (2007), respectively. Consequently, to anticipate this potential endogeneity of tariff rates, we estimate Eq. (46) using a two-stage instrumental variables approach. In the first stage, we regress the (gross) bilateral tariff rate (by industry), tar_{ijt}, on the mean over country *j*'s bilateral tariffs with *all other non-i countries*; this variable is likely to be insensitive to X_{ijt}. We then use the instrument constructed from the first stage, $t\hat{a}r_{ijt}$, in the second stage regression, the gravity equation in (46), alongside our measure of the c.i.f.-f.o.b. transport-cost factor.[34]

The third issue concerns accounting for variation in fixed trade costs, f_{ijt}. As we addressed above for the F/BW specifications, variation in f_{ijt} can be decomposed into three terms: an exporter component f_{it}, an importer component f_{jt}, and a residual bilateral term f_{ijt}^R. As summarized above, much of the observed policy-based and non-policy-based factors that influence fixed trade costs tend to be *multilateral* – or country-specific – in nature. Consequently, regarding Eq. (46) above, the exporter-year and importer-year fixed effects will capture the vast bulk of variation in $\ln f_{ijt}$ via $\ln f_{it}$ and $\ln f_{jt}$, respectively, leaving residual variation in $\ln f_{ijt}^R$ to be accounted for by the error term ϑ_{ijt}.

5. Estimation results

Section 5.1 presents the results from using our gravity equation to obtain estimates of the (positively defined) *ad valorem* variable trade-cost "trade elasticity," $\varepsilon_\tau = \theta\left(\frac{1+\gamma}{\gamma}\right)$. Section 5.2 provides the estimates of (structural) parameters σ and γ using our three specifications applying the F/BW methodology. Using these estimates, section 5.3 provides the implied estimates of θ and of the fixed trade-cost trade elasticity, $\varepsilon_f = \frac{\theta\frac{1+\gamma}{\gamma}}{\frac{1+\gamma}{\sigma+\gamma}(\sigma-1)} - 1$. These estimates will then be used in section 6 for two numerical counterfactual analyses.

5.1. Estimation of ε_τ

Because we obtain hundreds of estimates across industries, it would not be practical to report them all; instead, we present only the distributions of the estimated coefficients. Table 2 provides the distribution of estimates of the (*ad valorem* variable trade-cost) "trade elasticity" in columns 2 and 3 using our gravity Eq. (46). The only difference between the two specifications is that column 2 uses *observed* gross tariff rates (tar_{ijt}) whereas column 3 uses our (two-stage least squares) instrument for gross tariff rates. As evident, the distributions of the two sets of estimates are very similar. On econometric grounds, our preferred specification is that in column 3. Recalling that the data is at the 4-digit SITC industry level, the median estimated *ad valorem* trade elasticity is 10.08, which is in line with previous estimates; the 10th–90th percentile range is 3.82–17.35.[35]

[34] Note that endogeneity of tariff rates is not a concern in our F/BW specifications as those are reduced-form regressions using time-averaged variances and covariances of double-differences of the underlying variables.

[35] The 10th–90th percentile range of values found is consistent with other estimates using disaggregated bilateral cross-sectional/time-series trade data, cf., Hillberry and Hummels (2013). Aggregate trade data generates lower trade elasticity estimates in the range of 2–10, cf., Anderson and van Wincoop (2004).

Table 2
Estimated *Ad Valorem* Variable Trade-Cost Trade Elasticities.

Percentile	ε_τ	
	OLS	IV
1	−0.29	0.36
5	2.42	2.24
10	3.97	3.82
25	6.55	7.16
50	9.41	10.08
75	12.37	13.59
90	15.92	17.35
95	18.52	19.67
99	28.13	27.44

Notes: This table presents the distributions of the estimated structural parameters of the model obtained from estimating Eq. (46) separately for each of 568 industries in our sample.

Table 3
Estimated bilateral import demand and bilateral export supply elasticities.

Percentile	F/BW		IMC-Partial		IMC-Full	
	σ	γ	σ	γ	σ	γ
1	2.65	0.46	2.61	0.64	2.73	0.79
5	3.01	1.18	3.27	1.44	3.29	1.61
10	3.34	1.61	3.77	1.93	3.80	2.29
25	3.96	2.36	4.62	3.34	4.85	3.57
50	4.70	4.03	6.08	5.99	6.45	6.00
75	6.02	7.00	9.23	11.31	9.26	10.96
90	8.68	14.09	15.12	22.20	14.90	21.40
95	11.52	21.83	22.79	35.14	20.73	30.71
99	27.82	58.19	87.25	76.73	55.92	69.51

Notes: This table presents the distributions of the estimated structural parameters of the model obtained from estimating Eq. (40) separately for each of 568 industries in our sample using three different specifications (see main text for details). Parameter σ is the elasticity of substitution and parameter γ is the inverse marginal cost elasticity of output.

5.2. Estimation of σ and γ

Estimation results for σ and γ are reported in Table 3.[36] As addressed earlier, the first specification, labeled "F/BW," is the F/BW specification that includes only Group 1 variables and assumes that the coefficients on the remaining variables in Groups 2–4 are zero. At the median, the estimate for σ is 4.70, which is in the middle of the range of σ estimates from Feenstra (1994), Table 2 for the six manufactured goods of 2.96 to 8.38; the median estimate in that group in Feenstra (1994) is 5.0. Furthermore, our range of σ estimates for percentiles 1–99 of 2.65–27.82 is similar to the range of 2.96–42.9 for all eight goods in Feenstra (1994). Broda and Weinstein (2006) provide four-digit SITC estimates for two different (averaged) time periods, 1972–1988 and 1990–2001. Using their mean estimates for differentiated products, the σ estimates for 1972–1988 and 1990–2001 are 5.2 and 4.7, respectively. Hence, our benchmark F/BW σ estimate of 4.70 is in line with both of those sets of estimates.

Our median estimate of γ using our benchmark F/BW specification is 4.03, which is also in the middle of the range of (positive) γ estimates from Feenstra (1994), Table 2 for four manufactured goods of 1.94 to 6.58; the median estimate in that group in Feenstra (1994) is 2.43. Furthermore, our range of γ estimates for percentiles 1–99 of 0.46–58.19 is similar to the (positive) range of 1.94–27.8 for the six goods in Feenstra (1994). Unfortunately, Broda and Weinstein (2006) do not report their estimates of γ.

The second specification, labeled "IMC-Partial," employs Group 1 and Group 2 variables, and assumes the coefficients on the remaining variables in Groups 3–4 are zero. Of course, the *ad valorem* (variable) trade-cost variable τ_{ijt} reflects both *ad valorem* (gross) tariff rates ($tar_{ijt} > 1$) as well as (gross) transport-cost factors ($trans_{ijt} > 1$), as discussed above. Hence, in the context of Eq. (38), adding Group 2 variables adds seven more RHS variables.[37] Turning to this specification's median estimates, the estimate of σ of 6.08 is *29%* larger in value than that in specification 1, implying OVB in the F/BW specification. Similarly, the median estimate of γ in specification 2 is 5.99, which is *49%* larger than that in specification 1. These are notable differences.

[36] We keep only industries for which the parameters of the model conform to the theoretical restrictions (i.e., $\sigma > 1$ and $\gamma > 0$) for all three specifications, as common to the F/BW literature. About 25% of industries for which we have data are excluded from our sample; by comparison, Broda and Weinstein (2006) exclude about 35% of their industries.

[37] Recall that the explicit IMC-Partial specification is provided in section 3 of Online Appendix D.

The third specification, labeled "IMC-Full," includes the variables described earlier in Groups 1 and 2, but also includes an exporter fixed effect and exporter fixed effects *interacted* with $\overline{\Delta^k \ln \overline{p}^c_{ij}}, \overline{\Delta^k \ln s_{ij}}, \overline{\Delta^k \ln tar_{ij}}$, and $\overline{\Delta^k \ln trans_{ij}}$. Turning to this specification's median estimates, the estimate of σ of 6.45 is *37%* larger in value than that of σ in F/BW specification 1. Similarly, the median estimate of γ in specification 3 is 6.00, which is *49%* larger than that in specification 1. These estimates are similar to the respective estimates in specification 2 and are notably different from those in benchmark specification 1. In the context of our theoretical model, previous estimates reveal OVB.[38]

The results presented in Table 3 have four important implications. First, our IMC estimates are quite different from the benchmark estimates. Our richer specifications increase the estimated values of the elasticity of substitution and the bilateral export supply elasticity. Second, our IMC estimates are robust to changes in specifications. Third, the distributions of the estimates are quite similar across both IMC specifications. Fourth, the estimated parameters are distributed densely around the medians.

Although we are able to compare our specifications' estimates for the elasticities of substitution with those in Feenstra (1994), Broda and Weinstein (2006), and potentially other studies' previous empirical results using similar data, the literature provides few comparisons for our estimates of γ; only Feenstra (1994) provides estimates for comparison. As noted, Broda and Weinstein (2006) do not report estimates for γ. However, Hottman et al. (2016) report an (implied) median estimate of γ of 6.25 using U.S. barcode firm-level data. Interestingly, this median estimate lies near our median IMC estimates using industry-level international trade data, but controlling for firm heterogeneity. In the next subsection, we address the importance of precise and unbiased estimates of γ for estimating θ and estimating the fixed-trade-cost elasticity, ε_f. In section 6, we demonstrate the importance of γ estimates for relevant policy-oriented quantitative comparative statics.

5.3. Estimation of θ and ε_f

Armed with estimates of ε_τ and γ, we can use our model to compute values for θ and ε_f. From gravity Eq. (46), $\varepsilon_\tau = \theta(1+\gamma)/\gamma$. This implies that we can recover estimates for θ as follows:

$$\theta = \left(\frac{\gamma}{1+\gamma}\right)\varepsilon_\tau. \tag{47}$$

The fixed cost elasticity can then be recovered from our estimates of σ, γ and θ and Eq. (15):

$$\varepsilon_f = \left[\frac{\theta\left(\frac{1+\gamma}{\gamma}\right)}{\frac{1+\gamma}{\sigma+\gamma}(\sigma-1)} - 1\right] \tag{48}$$

The distribution of estimates are reported in Table 4. Using the estimated values of the trade elasticity using IV (reported in Table 2) and the estimated values of γ (reported in the last column of Table 3), the second column of Table 4 provides at various percentiles the estimated values of θ, as implied by Eq. (47). As reported in the table, the median estimate of θ is 8.50, which is close to Eaton and Kortum (2002)'s and Arkolakis (2010)'s preferred estimate of 8.28. The third column of Table 4 reports the estimated values of the fixed trade-cost trade elasticities, ε_f, at various percentiles using Eq. (48) and our estimated values of σ, γ, and θ. These elasticities, along with our estimates of the three structural parameters, will be useful for our numerical comparative statics in the next section. As just one clue to the importance of IMC in those analyses, note that the fixed trade-cost trade elasticity at the median is 2.39. However, under the case of CMC, the theoretical fixed trade-cost trade elasticity (defined positively) is $\frac{\theta}{\sigma-1} - 1$, which is (under CMC) the (*ad valorem* variable trade-cost) "trade elasticity"relative to $\sigma - 1$ (and then minus 1). Using the median trade elasticity of 10.08 (from our gravity estimation) and our IMC-Full estimate of σ of 6.00, the implied CMC fixed trade-cost trade elasticity is only 1.02. Hence, under IMC, fixed trade-cost reductions have a larger impact of 2.3 times that under CMC.

6. Numerical analyses

Having established in the previous section strong empirical evidence of increasing marginal costs using international data, we provide in this section two numerical analyses to illustrate the importance of allowing for IMC. First, for a given set of parameters, we quantify the impact of allowing for increasing marginal costs in welfare calculations. Second, we use our estimates to show that the necessary changes to fixed trade costs, to obtain the welfare-equivalent of (small) changes to variable trade costs, are *much smaller* in the case of empirically-justified increasing marginal costs than in the case of constant marginal costs, helping to explain the increasing prominence of deep trade agreements in the world economy.

[38] In a robustness check, using data and proxies discussed earlier for w_{it} and L_{it} alongside variables in Groups 1 and 2, but ignoring unobservable variables A_i and b_i and omitting exporter fixed effects, our median σ estimate is 6.50 and our median γ estimate is 6.34, both similar to the respective estimates for IMC-Partial and IMC-Full.

Table 4
Estimated Pareto parameters and fixed trade costs elasticities.

Percentile	θ	ε_f
1	0.27	−0.91
5	1.54	−0.40
10	3.03	0.09
25	5.73	1.08
50	8.50	2.39
75	11.34	4.04
90	15.13	5.88
95	17.32	7.89
99	24.39	11.95

Notes: This table presents the distributions of the Pareto parameters and the elasticities of trade estimated separately for each of the 568 industries in our dataset.

6.1. Counterfactual 1: Welfare gains from trade

We provide in this section a numerical analysis in the spirit of Feenstra (2010) and Costinot and Rodriguez-Clare (2014) to illustrate the importance of allowing for IMC in welfare calculations. We show using representative values of the (inverse) index of the heterogeneity of firms' productivities (θ) and of the inverse marginal cost elasticity of output (γ) that the welfare gains from trade are reduced (at the median) by about one percentage point (or by approximately 15%) in the case of IMC relative to the case of CMC.

From Eq. (17), the percentage change in real income associated with moving from the initial equilibrium (with trade) to autarky for country j is given by (100 times):

$$G_j = 1 - \lambda_{jj}^{1/\varepsilon_\tau}, \tag{49}$$

where λ_{jj} is the domestic absorption share of GDP and $\varepsilon_\tau = \theta\left(\frac{1+\gamma}{\gamma}\right)$.[39] Consequently, the only additional data needed for this numerical exercise is trade shares. As in Feenstra (2010), we use information on nominal exports and nominal GDPs from the Penn World Tables to calculate export shares.[40] A key consideration here is comparing the gains from trade with CMC versus the gains from trade with IMC. Consequently, we also calculate the gains from trade assuming a value of $\gamma = \infty$ to obtain a benchmark value.

As explained earlier, welfare gains from trade depend on two sufficient statistics: the trade share and the trade elasticity. We explore the impact of variation in each separately, beginning with changes in the trade elasticity. In our sample, the mean trade share is 39.1%, so we set $\lambda_{jj} = 60.9$. Conditional on that trade share, Table 5 presents the distribution across industries of the gains from trade (relative to autarky) under the assumption of CMC ($\gamma = \infty$) and IMC as indicated at the top of each column. Our median estimate under IMC is 4.78%, which is a reduction of 15.4% from the welfare gain of 5.65% in the benchmark case of constant marginal costs ($\gamma = \infty$). These values are consistent with the “welfare-diminution” effect discussed in section 3.

Table 6 presents the results for the impact of changes in the trade share, holding the elasticity constant at the median values. It reports calculations of the gains from trade for 20 countries of various levels of per capita real GDP, similar to Table 3.1 in Feenstra (2010). As expected, countries with larger export shares have larger gains from opening up from autarky. For instance, the United States has a small export share; consequently, the gains from trade are smaller. However, the presence of IMC still has a substantive effect for the United States; the reduction of welfare of 0.24 from 1.53 to 1.29 owing to increasing marginal costs is 15.6%. Overall, the results presented in this section suggest that increasing marginal costs have substantive effects on welfare calculations.

6.2. Counterfactual 2: Welfare-equivalent changes and deep trade agreements

As discussed in the introduction, the “new millennium” has also introduced “new types of trade agreements.” The stark contrast between shallow versus deep trade agreements is essentially the difference between reducing *ad valorem* tariff rates on international trade versus reducing “regulatory heterogeneity”:

> *Former WTO Director General Pascal Lamy put it this way: “TTIP isn't about trade trade-offs, but a process of regulatory convergence, which is a totally different ball game.”* Norberg (2015), p.1.

[39] In Feenstra (2010), p. 53, G_j is defined as $\left[\left(1-ExportShare_j\right)^{-1/\theta}-1\right]/\left[\left(1-ExportShare_j\right)^{-1/\theta}\right]$. However, using ACR notation and some algebra, this simplifies to $G_j = 1 - \lambda_{jj}^{1/\theta}$, which is identical to the measure of G_j in Costinot and Rodriguez-Clare (2014), p. 204.

[40] We could just as easily used the World Input-Output Database (WIOD) used in Costinot and Rodriguez-Clare (2014), but chose the set of countries in Feenstra (2010) largely due to the broader sample and wider variation in the levels of countries' per capita real GDPs.

Table 5
Welfare gains from trade, 2010.

Percentile	CMC	IMC
1	2.01	1.79
5	2.81	2.48
10	3.21	2.81
25	4.26	3.57
50	5.65	4.78
75	8.24	6.66
90	15.02	12.09
95	27.47	19.08
99	84.28	74.83

Notes: This table presents the absolute value of the percentage change in real income associated with moving from the initial equilibrium to autarky given by $1-\lambda_{jj}^{1/\varepsilon_T}$, where λ_{jj} is domestic absorption. In our sample, the mean trade share is 39.1, so we set $\lambda_{jj}=60.9$. We compute gains from trade separately for each of the 568 industries in our sample. In this section, we use s_{jj} to measure λ_{jj}.

Table 6
Welfare gains from trade for selected countries, 2010.

Name	GDPPC	Export Share	CMC	IMC
Guinea	1677	30.34	4.16	3.52
Mali	1736	22.84	3.00	2.54
Nepal	1807	9.58	1.18	0.99
Kyrgyzstan	2863	51.55	8.17	6.93
Republic of Moldova	3737	39.23	5.69	4.81
Congo	4709	65.81	11.86	10.09
Guatemala	6293	25.81	3.45	2.91
China	9423	26.27	3.52	2.97
Thailand	13,109	66.49	12.07	10.26
Gabon	13,151	57.66	9.62	8.16
Brazil	13,623	10.74	1.33	1.12
Malaysia	20,192	86.93	21.29	18.26
Israel	30,538	35.02	4.94	4.18
Bahamas	31,413	34.95	4.93	4.17
Italy	35,936	25.19	3.36	2.83
Germany	40,481	42.25	6.26	5.29
Saudi Arabia	41,482	49.57	7.74	6.56
United States	49,907	12.32	1.53	1.29
Norway	57,900	39.73	5.78	4.89
Bermuda	62,290	49.69	7.76	6.58

Notes: This table presents the absolute value of the percentage change in real income associated with moving from the initial equilibrium to autarky given by $1-\lambda_{jj}^{1/\varepsilon_T}$, where λ_{jj} is domestic absorption, computed for selected countries for year 2010. To the extent possible, we choose the same countries as in Table 3.1 of Feenstra (2010) to facilitate comparison.

As illustrated recently in the United States-Mexico-Canada Agreement, the successor to NAFTA, deep trade agreements embody a large increase in the number of chapters and the scope of the agreement. In reality, these developments essentially span three (partially overlapping) areas:

1. Modern trade agreements have been deepened to cover services trade flows, capital flows, migration flows, and idea flows;
2. Modern trade agreements aim to reduce barriers at the border and behind the border in terms of regulatory convergence, such as trade facilitation (customs administration), technical barriers to trade, sanitary and phytosanitary measures, and competition policy;
3. Modern trade agreements extend to addressing environmental policy and labor rights.

For our purposes, we are addressing the second category, where regulatory divergences create costs of trade unrelated to the level of output, i.e., fixed trade costs.

One of the earliest studies to document and categorize the degree to which European Union and United States' preferential trade agreements (PTAs) incorporated liberalizations beyond tariff-rate reductions that would reduce fixed trade costs is Horn et al. (2010), documenting such liberalizations beyond that established by the World Trade Organization (WTO). Evidence from the World Bank's DTA database suggests that several (inverse) indexes of fixed trade costs – legally enforceable provisions provided in DTAs such as trade facilitation, technical barriers to trade, sanitary and phytosanitary measures, and competition policy – have increased over time. For instance, Hofmann et al. (2017) note that the simple count of legally enforceable provisions

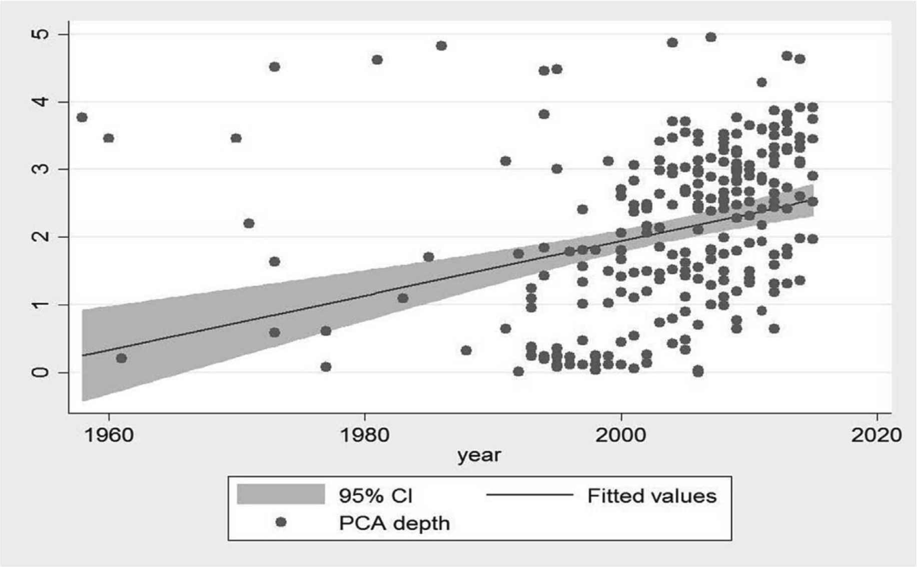

Fig. 2. Hofmann et al. (2017) PCA Measure of DTA "Depth" Over Time.

included in PTAs increased from 8 to 17 from the 1990s to 2015. More rigorously, Hofmann et al. (2017) created a measure of depth of PTAs using Principal Components Analysis (PCA); this PCA measure indicated that PTAs' depth has increased, on average, *150%* from the 1980 to 2015, as seen in Fig. 2.[41]

While empirical studies are now starting to flourish given the World Bank's new data base on the DTAs, the theoretical and quantitative welfare effects of deep trade agreements have been scarcely examined, especially in the context of the new trade theory with heterogeneous firms. Specifically, to the authors' knowledge only four papers address systematically quantifying the trade and welfare effects of bilateral (*ad valorem*) variable trade-cost liberalizations *relative to* fixed trade-cost changes. As mentioned in the introduction, Zhai (2008) is among the earliest of these rare studies that have introduced a Melitz model into a CGE model to calculate the trade and welfare effects of three types of policy simulations: a 50% tariff-rate cut, a 5% reduction in variable trade costs, and a 50% reduction in fixed trade costs.[42] Using a multi-country framework, a value of σ of 5, and a value of θ of 6.2, Zhai (2008) found for the United States, for example, that a 5% reduction in variable trade costs increased welfare by 32.8 billion (US) dollars. In the context of his model, a 50% reduction in fixed trade costs increased welfare by 44.8 billion (US) dollars. Hence, the welfare-equivalent reduction in fixed trade costs would be 36 (29) percent, to match a 5 (4) percent reduction in variable trade costs (or a ratio of approximately 7.25:1). This accords quantitatively to the notion that, for the same percent reduction in the cutoff productivity φ_{ij}^*, the fall in f_{ij} would need to be about 7 times, since φ_{ij}^* adjusts in proportion to $f_{ij}^{1/(\sigma-1)}$ in the case of CMC. In CGE analyses of the TTIP, a reduction of 36% in non-tariff measures was considered "very ambitious," and such a differential suggests *against* the proliferation of deep trade agreements.

To the authors' knowledge, only three other papers have considered CGE analyses using a Melitz framework, Balisteri et al. (2011), Dixon et al. (2016), and Arkolakis et al. (2021). The structure of Balisteri et al. (2011) is similar in many respects to Zhai (2008), but differs in several other respects.[43] Another CGE model with a Melitz framework is Dixon et al. (2016); however, this study only examined relative impacts of reductions in (*ad valorem*) variable trade costs across Melitz and Krugman versions of their model. Finally, as noted in the introduction, Arkolakis et al. (2021) extends the Melitz model of trade to show that – for

[41] Moreover, recent empirical studies using gravity equations have demonstrated evidence of substantive effects of reductions in policy-related fixed trade costs on bilateral trade flows. First, Baier et al. (2014) documented using state-of-the-art panel techniques that deeper economic integration agreements (embodying fixed-trade-cost reductions) had larger partial effects on bilateral trade flows. Baier and Regmi (2022) was the first study using machine-learning techniques to show that deeper PTAs have larger trade-creation effects, noting substantive effects from provisions on anti-dumping, competition policy, customs harmonization, e-commere, export and import restrictions, sanitary and phytosanitary measures, and technical barriers to trade. Breinlich et al. (2022) similarly used machine-learning techniques to examine the effects of deep trade agreement provisions on trade flows, finding that provisions associated with technical barriers to trade, anti-dumping, and competition policy had significant effects. Another recent study finding significant effects of DTAs on trade flows is Fontagne et al. (2022).

[42] For purposes of this paper, we discuss the implications of the latter two simulations; the reason is that Zhai (2008) allows tariffs to generate income, whereas variable trade costs are "iceberg" trade costs, as in this paper. A 50% tariff-rate reduction in Zhai (2008) reduces disposable income, which has an offsetting effect on expenditures and trade; the model in this paper ignores this aspect, which is left for future research.

[43] Balisteri et al. (2011) actually estimate values for σ and even θ, and use exporter and importer fixed effects to estimate exporter- and importer-specific fixed trade costs (assuming CMC). The residuals in their approach are bilateral fixed trade costs, which adjust to match the simulated bilateral trade flows to actual trade flows. This method yields some difficult-to-rationalize bilateral fixed trade costs. For instance, the bilateral fixed trade cost of exports from the United States to Japan is twice as high as that from Canada to Japan; moreover, the fixed trade costs of intra-national Japanese trade are the same as fixed trade costs from Canada to Japan. Nevertheless, Balisteri et al. (2011) only compare a 50% reduction in tariff rates against a 50% reduction in fixed trade costs, which provides a non-comparable comparison to Zhai (2008) and our model, since tariff cuts in Balisteri et al. (2011) involve reductions in disposable income and cannot be compared to a 50% reduction in iceberg variable trade costs, as we know from Zhai (2008).

multiproduct firms facing constant marginal costs in producing their core product (though increasing marginal market-penetration costs) – additional products that are farther from the firm's core competency face increasing marginal production costs (despite economies of scope in market-access costs). Of particular relevance to this paper, the last substantive section of Arkolakis et al. (2021) conducts counterfactual experiments of reductions in market-access costs and, for comparison, tariff rates. In their baseline simulation, the elimination of the recently observed average 4% tariff rates in the world generates a welfare gain of 1.8%. In contrast, using their Table 6, Counterfactual 1 experiment of reducing total market-access costs, a 15% reduction in such fixed trade costs improves welfare by 2.0%. Hence, for comparison of the results in this model relative to Zhai (2008) (and later to our counterfactual), it would take a 13% reduction in fixed trade costs to generate the same welfare as a reduction in tariff rates of 4%, a ratio of 3.25:1.

In our second counterfactual, we are interested in measuring fixed trade-cost changes, $\hat{f}_{ij}$, that are equivalent in welfare to changing a given (*ad valorem*) variable trade cost, $\hat{\tau}_{ij}$. In our model, as seen in Eq. (13), we can write:

$$\phi_{ij} = \tau_{ij}^{-\varepsilon_\tau} f_{ij}^{-\varepsilon_f}, \tag{50}$$

such that for a given value of (the gross tariff rate) $\hat{\tau}_{ij}$, we define the welfare-equivalent fixed trade-cost change as $\hat{f}_{ij} = \hat{\tau}_{ij}^{\frac{\varepsilon_\tau}{\varepsilon_f}}$. This gives the increase in fixed trade costs that is equivalent to an increase in variable trade costs in terms of its impact on trade flows and welfare.[44]

Using results from section 3, the ratio of elasticities plays a critical role in defining welfare-equivalent trade-cost changes. From the theoretical model with IMC, we know that:

$$\frac{\varepsilon_\tau}{\varepsilon_f} \equiv \frac{\theta\left(\frac{1+\gamma}{\gamma}\right)}{\frac{\theta\left(\frac{1+\gamma}{\gamma}\right)}{\frac{1+\gamma}{\sigma+\gamma}(\sigma-1)} - 1}. \tag{51}$$

For any value of $\gamma < \infty$, this ratio is smaller than in the benchmark CMC case. In the limit, as $\gamma \rightarrow \infty$, the ratio converges to the benchmark. This implies that under IMC the welfare-equivalent change $\hat{f}_{ij}$ for a given $\hat{\tau}_{ij}$ is smaller than under CMC.

Consider the median values of our estimated parameters using the IMC-Full specification from section 5, $\sigma = 6.45$, $\gamma = 6.00$, and $\theta = 8.50$. Substituting in these values yields:

$$CMC: \frac{\varepsilon_\tau}{\varepsilon_f} = \frac{\theta}{\frac{\theta}{\sigma-1} - 1} = 15.18 \tag{52}$$

$$IMC: \frac{\varepsilon_\tau}{\varepsilon_f} = \frac{\theta\left(\frac{1+\gamma}{\gamma}\right)}{\frac{\theta^{\frac{1+\gamma}{\gamma}}}{\frac{1+\gamma}{\sigma+\gamma}(\sigma-1)} - 1} = 4.44 \tag{53}$$

Armed only with *observable* estimates of variable trade costs (using average MFN tariff rates), we can obtain a fixed-trade-cost change that is *equivalent in welfare* to introducing a country's (or an average of countries') MFN tariff rates. In our sample, the average tariff rates applied is about 4%. This implies that the welfare-equivalent fixed costs changes are:

$$CMC: \hat{f} = (1.04)^{15.18} = 1.81 \tag{54}$$

$$IMC: \hat{f} = (1.04)^{4.44} = 1.19. \tag{55}$$

These results make clear that the equivalent change is much larger under CMC.

The welfare-equivalent fixed trade-cost change depends on two sufficient statistics: the level of trade barriers and the ratio of elasticities. As we did for welfare, we explore the impact of each in turn. Table 7 reports the distribution of the ratio of welfare-equivalent fixed trade-cost changes across industries. To discipline the quantitative exercise, we use the mean tariff of 4% in each industry and consider the introduction of the tariff rate as our shock, $\hat{\tau}$.[45] We compute the welfare-equivalent change for two separate cases, the benchmark CMC case of $\gamma \rightarrow \infty$ and the IMC case, as indicated at the top of each column. The table shows that, for the median industry, the equivalent fixed trade-cost change under CMC is 38%, whereas under IMC it is only 15%. Both distributions of welfare-equivalent fixed trade-cost changes start at 0%, but the CMC distribution has a much thicker right tail. At the 90th

[44] We express the term as shown for expositional convenience. Mathematically, for values $\phi_{ij} = \tau_{ij}^{-\varepsilon_\tau} f_{ij}^{-\varepsilon_f}$ and $\tilde{\phi}_{ij} = \tilde{\tau}_{ij}^{-\varepsilon_\tau} \tilde{f}_{ij}^{-\varepsilon_f}$, if $\phi_{ij} = \tilde{\phi}_{ij}$ then $\hat{f}_{ij} = \hat{\tau}_{ij}^{-\frac{\varepsilon_\tau}{\varepsilon_f}}$ where $\hat{f}_{ij} \equiv \tilde{f}_{ij}/f_{ij}$ and $\hat{\tau}_{ij} \equiv \tilde{\tau}_{ij}/\tau_{ij}$.

[45] We are ignoring any change of tariff revenue, leaving this for future research.

Table 7
Equivalent fixed-trade-cost change.

Percentile	CMC		IMC	
	$\varepsilon_\tau/\varepsilon_f$	$\hat{f}$	$\varepsilon_\tau/\varepsilon_f$	$\hat{f}$
1	2.08	1.08	1.23	1.05
5	2.68	1.11	1.88	1.07
10	3.57	1.15	2.14	1.09
25	5.08	1.21	2.79	1.11
50	8.35	1.38	3.69	1.15
75	16.82	1.91	4.87	1.21
90	40.27	4.68	7.23	1.32
95	219.75	21.30	9.20	1.42
99	546.47	4550.28	16.90	1.91

Notes: This table presents the distribution of the average industry-level welfare-equivalent fixed-trade-cost changes ($\hat{f}$). We set $\tau = 1.04$ (the sample import-weighted mean) and let the elasticity of substitution (σ), the inverse elasticity of marginal costs (γ), and the Pareto parameter (θ) vary across industries. The equivalent fixed-trade-cost changes are obtained from $\hat{f}_{ij} = \hat{\tau}_{ij}^{\varepsilon_\tau/\varepsilon_f}$. We keep the 406 industries in the sample for which the fixed-trade-cost elasticities are positive.

Table 8
Average equivalent fixed-trade-cost changes for selected countries, 2010.

Name	GDPPC	Mean tariff	CMC		IMC	
			$\varepsilon_\tau/\varepsilon_f$	$\hat{f}$	$\varepsilon_\tau/\varepsilon_f$	$\hat{f}$
Guinea	1677	1.08	13.75	3.04	4.41	1.43
Mali	1736	1.09	15.80	4.18	5.09	1.59
Nepal	1807	1.12	16.03	5.86	5.11	1.76
Kyrgyzstan	2863	1.01	20.91	1.25	5.94	1.07
Moldova	3737	1.03	21.58	1.75	5.08	1.14
Congo	4709	1.15	14.60	8.16	4.13	1.81
Guatemala	6293	1.05	15.94	2.05	4.25	1.21
China	9423	1.08	21.46	5.50	4.82	1.47
Thailand	13,109	1.08	18.11	3.94	4.46	1.40
Gabon	13,151	1.15	15.21	8.86	3.89	1.75
Brazil	13,623	1.11	28.60	20.73	4.57	1.62
Malaysia	20,192	1.08	20.19	4.92	4.35	1.41
Israel	30,538	1.06	20.53	3.08	4.44	1.28
Bahamas	31,413	1.29	16.70	66.00	4.48	3.08
Italy	35,936	1.01	25.42	1.33	5.10	1.06
Germany	40,481	1.01	27.05	1.43	4.49	1.06
Saudi Arabia	41,482	1.09	19.79	5.58	5.38	1.60
United States	49,907	1.03	26.01	2.31	4.52	1.16
Norway	57,900	1.01	20.70	1.14	4.35	1.03
Bermuda	62,290	1.19	20.25	31.79	3.93	1.96

Notes: This table presents the distribution of the average country-level welfare-equivalent fixed-trade-cost changes ($\hat{f}$) for selected countries for year 2010. We set all parameters equal to the country's import-weighted averages. The equivalent fixed-trade-cost changes are obtained from $\hat{f}_{ij} = \hat{\tau}_{ij}^{\varepsilon_\tau/\varepsilon_f}$. To the extent possible, we choose the same countries as in Table 3.1 of Feenstra (2010) to facilitate comparison.

percentile of the distribution, the equivalent fixed trade-cost change is an increase of *368%*. But under IMC, it is a much more reasonable 32%.

Most importantly, note that – for the median industry – under CMC it would take a 28 [=100(1–1/1.38)] percent reduction in fixed trade costs to be welfare equivalent to a 4% reduction in variable trade costs. This result is similar to the finding mentioned earlier for Zhai (2008). By contrast, in our model with IMC, it would take only a 13% reduction in fixed trade costs to be welfare equivalent to a 4% reduction in variable trade costs, which accords more with the study of Brazilian exporters in Arkolakis et al. (2021), which allowed increasing marginal market-penetration costs.[46]

Table 8 reports the distribution of the ratio of welfare-equivalent fixed trade-cost changes for selected countries. This exercise aims to illustrate the impact of differences in trade barriers, so we set the ratio of elasticities at their median values. For each country, we compute the import-weighted average tariff. Again, we set the shock to introducing the country's average tariff rate. As in Table 7, we compute the welfare-equivalent fixed trade-cost changes for the CMC and IMC cases. The main point is that, even if parameters are the same across countries, changes in the compositions of trade flows have an impact.

[46] It is worth noting that our comparative static result is based on a firm-level model using 4-digit SITC industry data, whereas the economically similar result in Arkolakis et al. (2021) utilizes firm-level data.

We conclude by addressing a result for each of the United States and Germany. For the United States (Germany), the MFN tariff rate is only about 3 (1) percent, which conforms to most observers knowledge of it. While the initial value of bilateral fixed trade costs is unknown, the lack of that knowledge is immaterial for our calculations. All that is needed here is values of average tariff rates (or variable trade costs), the well-known (*ad valorem* variable-trade-cost) "trade elasticity," and a value for the fixed trade-cost trade elasticity. With little empirical knowledge of the *levels* of fixed trade costs, our estimates of σ, γ, and θ allow us to construct an estimate of $\frac{\theta^{\frac{1+\gamma}{\gamma}}}{\frac{1+\gamma}{\sigma+\gamma}(\sigma-1)} - 1$. Assuming IMC, we find that eliminating remaining U.S. tariffs of 3% are welfare-equivalent to a reduction in fixed trade costs of *only 14%* $[= 100(1 - 1/1.16)]$. For Germany, we find that eliminating their remaining tariffs of 1% are welfare-equivalent to a reduction in fixed trade costs of 6%. These results make deep trade agreements much more attractive to pursue, with a 14 (6) percent reduction well below the reductions of 25% used in earlier analyses of the Transatlantic Trade and Investment Partnership (TTIP) in Berden et al. (2010).

7. Conclusions

This paper has offered three contributions to the international trade literature. First, extending theoretically a standard (one-sector) Melitz model of international trade to *allow for* increasing marginal costs, we generated a gravity equation where the trade elasticity (θ) was magnified by one plus the marginal cost elasticity of output, implying that welfare gains from trade are reduced as diminishing marginal returns interact with the Pareto shape parameter to lower the average productivity gains from trade liberalizations.

Second, introducing a novel econometric extension of the Feenstra/Broda-Weinstein method to control explicitly for firm heterogeneity, we find evidence that increasing marginal costs exist. The median bilateral export supply elasticity estimate of 6.00 in our preferred specification is *far below* the value of ∞ assumed in the benchmark (CMC) trade models.

Third, we provided two numerical analyses. In the first, we examined the relative quantitative importance of increasing marginal costs for estimating the welfare gains from trade. Our second – and more novel – counterfactual provided insight into the increasing prominence of deep trade agreements in the world economy. Under constant marginal costs for the median industry, the reduction in fixed trade costs needed to be equivalent in welfare improvement to a 4% reduction in *ad valorem* variable trade costs was 28%, the latter considered "ambitious" in most CGE analyses of deep trade liberalizations. By contrast, under increasing marginal costs, the welfare-equivalent reduction in fixed trade costs is only *13%* for the median industry.

We offer two suggestions for future research. First, to reduce theoretical complexity, we have omitted disposable income changes associated with tariff revenues; future work could incorporate tariff revenue for computing the welfare effects of reducing tariff rates. Second, our framework could be extended in the future to incorporate the role of tastes for regulatory divergences addressed in Grossman et al. (2021) to better understand and potentially quantify the welfare-equivalent effects of fixed versus variable trade-cost reductions.

Data availability

EstimationGravity.dta EstimationStructural.dta AdValorem.dta (Original data) (Mendeley Data)

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We are grateful to editor Costas Arkolakis, two referees, Scott Baier, Matthew Clance, Anson Soderbery, Saad Ahmad, William Powers, David Riker, Wen Jin (Jean) Yuan, and conference and seminar participants at the U.S. International Trade Commission, the Midwest International Economics Group meetings at the University of Notre Dame, the University of Cape Town, and the University of Pretoria for very helpful comments on earlier drafts. All remaining errors are our own.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jinteco.2023.103774.

References

Anderson, J., 1979. A theoretical foundation for the gravity equation. Am. Econ. Rev. 69 (1), 106–116.
Anderson, J., 2011. The gravity model. Annual Review of Economics 3, 133–160.
Anderson, J., van Wincoop, E., 2003. Gravity with gravitas: a solution to the border puzzle. Am. Econ. Rev. 93 (1), 170–192.
Anderson, J., van Wincoop, E., 2004. Trade costs. J. Econ. Lit. 42 (3), 691–751.
Arkolakis, C., 2010. Market penetration costs and the new consumers margin in international trade. J. Polit. Econ. 118 (6), 1151–1199.
Arkolakis, C., Costinot, A., Rodriguez-Clare, A., 2012. New trade models, same old gains? Am. Econ. Rev. 102 (1), 94–130.

Arkolakis, C., Ganapati, S., Muendler, M.-A., 2021. The extensive margin of exporting products: a firm-level analysis. Am. Econ. J. Macroecon. 13 (4), 182–245.

Bagwell, K., 2007. The economic analysis of advertising. In: Armstrong, M., Porter, R. (Eds.), Handbook of Industrial Organization. North Holland, Amsterdam, Netherlands, pp. 1701–1844.

Baier, S., Bergstrand, J., 2007. Do free trade agreements actually increase members' international trade? J. Int. Econ. 71 (1), 72–95.

Baier, S.L., Regmi, N.R., 2022. Using machine learning to capture heterogeneity in trade agreements. Open Econ. Rev. 1–32.

Baier, S., Bergstrand, J., Egger, P., McLaughlin, P., 2008. Do economic integration agreements actually work? Issues in understanding the causes and consequences of the growth of regionalism. World Econ. 31 (4), 461–497.

Baier, S., Bergstrand, J., Feng, M., 2014. Economic integration agreements and the margins of international trade. J. Int. Econ. 93 (2), 339–350.

Baier, S., Bergstrand, J., Clance, M., 2018. Heterogeneous effects of economic integration agreements. J. Dev. Econ. 135, 587–608.

Balisteri, E., Hillberry, R., Rutherford, T., 2011. Structural estimation and solution of international trade models with heterogeneous firms. J. Int. Econ. 83, 95–108.

Berden, K., Francois, J., Tamminen, S., Thelle, M., Wymenga, P., 2010. Non-Tariff Measures in EU-US Trade and Investment. ECORYS, The Netherlands.

Bergstrand, J., 1985. The gravity equation in international trade: some microeconomic foundations and empirical evidence. Rev. Econ. Stat. 67 (3), 474–481.

Bernard, A., Redding, S., Schott, P., 2011. Multiproduct firms and trade liberalization. Q. J. Econ. 126, 1271–1318.

Breinlich, H., Corradi, V., Rocha, N., Ruta, M., Santos Silva, J., Zylkin, T., 2022. Machine learning in international trade research-evaluating the impact of trade agreements. CEPR Discussion Paper No. DP17325.

Broda, C., Weinstein, D., 2006. Globalization and the gains from variety. Q. J. Econ. 121 (2), 541–585.

Broda, C., Limao, N., Weinstein, D., 2008. Optinmal tariffs and market power: the evidence. Am. Econ. Rev. 98 (5), 2032–2065.

Chaney, T., 2008. Distorted gravity: the intensive and extensive margins of international trade. Am. Econ. Rev. 98 (4), 1707–1721.

Costinot, A., Rodriguez-Clare, A., 2014. Trade theory with numbers. In: Gopinath, G., Helpman, E., Rogoff, K. (Eds.), Handbook of Interantional Economics. vol. 4. Elsevier, Amsterdam.

Dai, M., Muitra, M., Yu, M., 2016. Unexceptional exporter performance in China? The role of processing trade. J. Dev. Econ. 121, 177–189.

Dixon, P., Jerie, M., Rimmer, M., 2016. Modern trade theory for CGE modelling: the Armington, Krugman and Melitz models. Journal of Global Economic Analysis 1 (1), 1–110.

Eaton, J., Kortum, S., 2002. Technology, geography, and trade. Econometrica 70 (5), 1741–1779.

Eaton, J., Kortum, S., Kramarz, F., 2011. An anatomy of international trade: evidence from french firms. Econometrica 79 (5), 1453–1498.

Fajgelbaum, P.D., Goldberg, O.K., Kennedy, P.J., Khandelwal, A.K., 2020. The return of protectionism. Q. J. Econ. 135 (1), 1–55.

Farrokhi, F., Soderbery, A., 2020. Trade Elasticities in General Equilibrium. Working Paper.

Feenstra, R., 1994. New product varieties and the measurement of international prices. Am. Econ. Rev. 84 (1), 157–177.

Feenstra, R., 2010. Product Variety and the Gains from International Trade. MIT Press, Cambridge, MA.

Feenstra, R., 2016. Advanced International Trade: Theory and Evidence. Second edition. Princeton University Press, Princeton, New Jersey.

Feenstra, R., Romalis, J., 2014. International prices and endogenous quality. Q. J. Econ. 129 (2), 477–527.

Feenstra, R.C., Weinstein, D.E., 2017. Globalization, markups, and U.S. welfare. J. Polit. Econ. 125 (4), 1040–1074.

Feenstra, R.C., Luck, P., Obstfeld, M., Russ, K.N., 2018. In search of the Armington elasticity. Rev. Econ. Stat. 100 (1), 135–150.

Flach, L., Unger, F., 2022. Quality and gravity in international trade. J. Int. Econ. 137, 103578.

Fontagne, L., Rocha, N., Ruta, M., Santoni, G., 2022. The economic impact of deep trade agreements. CESifo Working Paper (9529).

Gervais, A., 2015. Product quality and firm heterogeneity in international trade. Canadian Journal of Economics/Revue canadienne d'économique 48 (3), 1152–1174.

Goldberg, P.K., Pavcnik, N., 2016. The effects of trade policy. In: Bagwell, K., Staiger, R. (Eds.), The Handbook of Commercial Policy. vol. 1. Elsevier, Amsterdam, pp. 161–206.

Grossman, G., McCalman, P., Staiger, R., 2021. The new economics of trade agreements: from trade liberalization to regulatory convergence? Econometrica 89 (1).

Head, K., Mayer, T., 2014. Gravity equations: Workhorse, toolkit, and cookbook. In: Gopinath, G., Helpman, E., Rogoff, K. (Eds.), Handbook of Interantional Economics. vol. 4. Elsevier, Amsterdam.

Hillberry, R., Hummels, D., 2013. Trade elasticity parameters for a computable general equilibrium model. In: Dixon, P.B., Jorgenson, D.W. (Eds.), Handbook of Computable General Equiilibrium Modeling. Elsevier, Amsterdam, pp. 1213–1269.

Hofmann, C., Osnago, A., Ruta, M., 2017. Horizontal depth: a new database on the content of preferential trade agreements. World Bank Working Paper 7981.

Horn, H., Mavroidis, P., Sapir, A., 2010. Beyond the WTO? An anatomy of eu and us preferential trade agreements. World Econ. 33 (11), 1565–1588.

Hottman, C., Redding, S., Weinstein, D., 2016. Quantifying the sources of firm heterogeneity. Q. J. Econ. 131 (3), 1291–1364.

Jones, J.P., 1995. New Proof Tat Advertising Triggers Sales. Lexington, New York, NY.

Kohl, T., Brakman, S., Garretsen, H., 2016. Do trade agreements stimulate international trade differently? Evidence from 296 trade agreements. World Econ. 39 (1), 97–131.

Krugman, P., 1980. Scale economies, product differentiation, and the pattern of trade. Am. Econ. Rev. 70 (5), 950–959.

Lu, D., 2010. Exceptional Exporter Performance? Evidence from Chinese Manufacturing Firms. manuscript,University of Chicago.

Mattoo, A., Rocha, N., Ruta, M., 2020. The evolution of deep trade agreements. In: Mattoo, A., Rocha, N. (Eds.), Handbook of Deep Trade Agreements. World Bank.

Melitz, M., 2003. The impact of trade on intra-industry reallocations and aggregate industry productivity. Econometrica 71 (6), 1695–1725.

Norberg, H., 2015. TTIP is not your father's free trade agreement: it has potential to benefit all. CATO Online Forum, October 27, 2015.

Ossa, R., 2016. Quantitative models of commerical policy. In: Bagwell, K., Staiger, R. (Eds.), Handbook of Commercial Policy. vol. 1A. Elsevier, Amsterdam, The Netherlands, pp. 207–259.

Redding, S.J., 2011. Theories of heterogeneous firms and trade. Annu. Rev. Econ. 3 (1), 77–105.

Saunders, J., 1987. The specification of aggregate market models. Eur. J. Mark. 21, 5–47.

Simonovska, J., Waugh, J., 1980. The shape of the advertising response function. J. Advert. Res. 20, 767–784.

Soderbery, A., 2015. Estimating import supply and demand elasticities: analysis and implications. J. Int. Econ. 96, 1–17.

Soderbery, A., 2018. Trade elasticities, heterogeneity, and optimal tariffs. J. Int. Econ. 114, 44–62.

Sutton, J., 1991. Sunk Costs and Market Structure. MIT Press, Cambridge, MA.

Trefler, D., 1993. Trade liberalization and the theory of endogeneous protection: an econometric study of u.s. import policy. J. Polit. Econ. 101 (1), 138–160.

United States International Trade Commission, 2019. U.S.-Mexico-Canada Trade Agreement: Likely impact on the U.S. economy and on specific industry sectors. Publication number 4889. United States International Trade Commission, Washington, DC.

Vannoorenberghe, G., 2012. Firm-level volatility and exports. J. Int. Econ. 86 (1), 57–67.

WorldBank, 2020. Doing Business. World Bank, Washington, DC.

Zhai, F., 2008. Armington meets Melitz: introducing firm heterogeneity in a global CGE model of trade. J. Econ. Integr. 23 (3), 575–604.

Further reading

Baier, S., Kerr, A., Yotov, Y., 2018. Gravity, distance, and international trade. In: Wesley, W., Blonigen, B. (Eds.), Handbook of International Trade and Transportation. Edward Elgar Publishing.