Global Gains from Reduction of Trade Costs[*]

Edwin L.-C. LAI[†] Haichao FAN[‡] Han (Steffan) QI[§]

This draft: 29 April 2019

Abstract

We develop a simple formula for computing the global welfare effect of reduction of bilateral trade costs, such as shipping costs or the costs of administrative barriers to trade. The formula is applicable to a broad class of perfect competition and monopolistic competition models and settings, including perfect competition with multi-stage production and Melitz's (2003) model with general firm productivity distribution. We prove that the underlying mechanism is the envelope theorem. We then extend our analysis to models with non-constant markups. Finally, we carry out some empirical applications to show the user-friendliness of the formula.

Keywords: global welfare, trade cost, gains from trade

JEL Classification codes: F10, F12, F13

[*]An earlier version of this paper was circulated under the title "Global Gains from Trade Liberalization" (CESifo working paper no. 3775, March 2012). We would like to thank Davin Chor, Gene Grossman, Stephen Yeaple, Tim Kehoe, Jaume Ventura, Jonathan Vogel, Hamid Sabourian, Ralph Ossa, Arnaud Costinot and seminar and conference participants in University of Melbourne, University of New South Wales, University of Hong Kong, City University of Hong Kong, HKUST, Shanghai University of Finance and Economics, National University of Singapore, Singapore Management University, Asia-Pacific Trade Seminars, AEA Annual Meeting in Philadelphia in 2014, and World Congress of Econometric Society in Montreal in 2015, for their helpful comments. Naturally, all errors remain ours. The work in this paper has been supported by the Research Grants Council of Hong Kong, China (General Research Funds Project no. 642210 and 691813), the Natural Science Foundation of China (No.71603155), and by the self-supporting project of Institute of World Economy at Fudan University..

[†]Lai: Corresponding author. CESifo Research Network Fellow. Department of Economics, Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong. Phone: +852-2358-7611; Fax: +852-2358-2084; Email: elai@ust.hk or edwin.l.lai@gmail.com

[‡]Fan: Institute of World Economy, School of Economics, Fudan University, Shanghai, China. Email: fan_haichao@fudan.edu.cn

[§]Qi: Department of Economics, Hong Kong Baptist University, Kowloon Tong, Kowloon, Hong Kong. Phone: +852-3411-5682; Email: steffan@hkbu.edu.hk

1 Introduction

One major theme of international trade economics is the gains from trade. There is a presumption in all trade models that the more integrated is the world, the larger are the gains from trade. Therefore, the lower the trade barriers, the greater global welfare should be. But, how much do trade barriers matter to the world quantitatively? For example, what is the global benefit in monetary terms of a one percent reduction in all bilateral trade costs worldwide? How sensitive is the answer to this question to the trade model being used? Answers to these questions would help us to quantify the global benefit of improvement of transportation technology or that of the reduction of administrative barriers to global trade. Understanding the global welfare impact of such changes is important as higher global GDP means that there is more room to make every country better off by proper side payments. In fact, international organizations recognize the importance of evaluating the global welfare gains from the reduction of administrative trade barriers and shipping time. To this end, OECD, World Bank and World Economic Forum found that trade facilitation could yield large economic gains to the world.[1] This paper derives a simple equation for computing the global welfare effect of simultaneous reduction of bilateral trade costs of multiple country pairs, such as shipping costs or the costs of administrative barriers to trade. We find that the equation is applicable to a broad class of models and settings. We then carry out some empirical applications. We find the estimates from our formula to be reasonable and consistent with others' in the literature.

We derive a measure of the percentage change in global welfare based on the concept of equivalent variation and Kaldor-Hicks' concept of welfare change for a group. We rigorously demonstrate that the expression $\sum_{i=1}^{n} \frac{E_i}{Y^w} \widehat{U}_i$ (the expenditure-share-weighted average percentage change of country welfare) is a reasonable measure of the percentage change in global welfare, where E_i is the aggregate expenditure of country i, Y^w is the global GDP of the n countries in the world, and $\widehat{U}_i$ is a small percentage change in welfare of country i. This expression for the percentage change in global welfare has been used in the literature, such as in Hsieh and Ossa (2016), Atkeson and Burstein (2010) (hereinafter abbreviated as AB) and Burstein and Cravino (2015) (hereinafter abbreviated as BC). However, as far as we are aware, we are among the first to present a justification for its use.

Then, we derive a simple equation for computing the total global welfare effect of simultaneous small reduction in bilateral trade costs of multiple country pairs. We make three simple assumptions: 1. the level of trade balance is fixed in each country; 2. price is constant markup over marginal cost; 3. there are no externalities. We find that as long as these assumptions are satisfied the percentage change in global welfare is given by

$$-\sum_{j=1}^{n} \sum_{i=1}^{n} \frac{X_{ij}}{Y^w} \widehat{\tau}_{ij} \ , \tag{1}$$

where X_{ij} is the total value of exports from country i to country j, and $\widehat{\tau}_{ij} \equiv \Delta\tau_{ij}/\tau_{ij}$ is a small percentage change in the iceberg cost, τ_{ij}, of exporting from i to j. This turns out to be just the total

[1] See for example, World Economic Forum (2013) and OECD (2003).

saving in trade costs divided by global GDP, keeping the values of all bilateral trade flows unchanged. In other words, only the direct effect matters, as it is first order. The indirect effects, such as changes in allocation of resources to different goods and the resulting changes in input costs, do not matter as they are second order. The key reason is that the allocation of resources to different goods was already optimally chosen before the changes in trade costs take place.

Expression (1) reflects the fact that, in the absence of externalities and price distortions, the market response to changes in bilateral trade costs is identical to the optimal response of a global planner who maximizes global income. Therefore, one can invoke the envelope theorem when evaluating the effect of changes in bilateral trade costs on global welfare. The envelope theorem turns out to be a very powerful tool for evaluating the global welfare impact of changes in bilateral trade costs under rather general condition.

The envelope theorem can be applied to many models and settings. Examples are 1. models of perfect competition (hereinafter PC): Dixit and Norman (1980, Chapters 3-5) with trade costs, the Heckscher-Ohlin model with trade costs, Dornbusch-Fisher-Samuelson (DFS, 1977 and 1980), the Armington (1969) model, Eaton-Kortum (2002) (hereinafter EK2002), Melitz and Redding (2014), Yi (2003), Yi (2010); 2. monopolistic competition (hereinafter MC) models with constant markup: Krugman (1980) (hereinafter K1980), Melitz (2003) with general firm productivity distribution (hereinafter M-g). Yet, the envelope theorem does not apply to MC models with variable markup, such as Melitz and Ottaviano (2008).

The reason that the envelope theorem applies to the above models is that the market is efficient as there are no externalities and no price distortions. This is clear for the PC models. For MC models, we extend the proof of Dhingra and Morrow (2014) that the Melitz model is efficient in the global economy.

Our paper distinguishes from the literature for its generality, as reflected in its applicability to M-g and PC with multi-stage production. Intuitively, the reason that equation (1) applies to M-g is that as cutoff productivities change due to changes in trade costs, the effect on the average productivity of firms serving each market (productivity effect) and the effect on the mass of firms serving each market (firm mass effect) completely offset each other from the point of view of global welfare, regardless of the distribution of firm productivity. The reason is that, for each exporting country, the labor constraint dictates that the change in average productivity and change in firm mass in each market go in opposite directions, and they offset each other when summing up over all markets.

The intuition for equation (1) to hold for PC with multi-stage production (i.e. fragmentation) is that whenever there is an amount of terms of trade gain by an exporter of a good at any stage, there is an equal amount of terms of trade loss by an importer of the same good. Thus, the global effect of terms of trade changes is nil. Therefore, only the direct effect, which is the total saving in trade costs at all stages, matters for global welfare change. Furthermore, any saving in trade cost at any stage is eventually passed on to the final stage. As a result, the saving in trade cost at each stage shows up as the gain in global income in the final stage. Given that fragmentation is pervasive in practice, the fact that our formula is applicable to such a setting is important, as it demonstrates that our formula is relevant to the real world.

Ossa (2015) found that assuming symmetric trade elasticity across sectors leads to a gross underestimation of the gains from trade. In the extensions section, we show that if the sectors with low elasticities of substitution (i.e. low trade elasticities) tend to be associated with positive weighted sum of the firm mass effect and productivity effect, then there would be additional global gains from the reduction of trade costs beyond the benchmark result given by (1). Apparently, the finding of Ossa (2012) shows that this is true empirically.

Our work is inspired by Arkolakis, Costinot & Rodriguez-Clare (2012) (hereinafter abbreviated as ACR). They show that for a class of trade models which include Armington (1969), Krugman (1980) (hereinafter K1980), Melitz (2003) with Pareto distribution of firm productivity, and EK2002, a country's gains from trade are always given by the same simple formula that contains two sufficient statistics: (i) the share of expenditure on domestic goods, and (ii) the trade elasticity. The focus of our analysis is, however, different from that of ACR. In contrast to ACR, our goal is to calculate the global welfare change caused by small changes in bilateral trade costs. Nonetheless our two papers are related in some way. The mapping between our equation (1) and the gains from trade equation in ACR is that if one adopts the assumption that the import demand system is CES (i.e. Assumption R3 in ACR), then our global gains formula (1) is precisely equal to an expenditure-share-weighted average percentage change of individual country's gains from trade as given by the ACR formula. Therefore, ACR's equation implies our equation if their assumptions are adopted. Indeed, there is a mapping between our main result and that of ACR, which is presented in Appendix D. Thus, our work complements that of ACR in that while ACR's equation can be used to calculate the change in welfare of an individual country based on a certain set of assumptions, our equation can be used to calculate the global welfare change based on a set of less restrictive assumptions. We require a less restrictive set of assumptions because the indirect effects (which are important for calculating welfare change of an individual country) cancel each other in aggregation and we do not have to account for them when we calculate the global welfare impact. Thus, our equation is applicable to a broader set of models and settings than ACR, e.g. our equation can be applied to M-g and PC with multi-stage production. Our result is valid as long as there is constant markup and there are no externalities arising from the actions of economic agents.

Our work is also inspired by Atkeson and Burstein (2010) (AB), who prove that the details of firms' responses are of secondary importance to the estimation of the welfare impact of trade costs reduction for an individual country. Assuming that countries are symmetric,[2] they find that though changes in trade costs can have a substantial impact on heterogeneous firms' exit, export, and process innovation

[2] That is, their expenditures in all periods and productivity distributions of operating firms in all periods are the same.

decisions, the impact of these changes on a country's welfare largely offset each other. In the end, only the "direct effect" of trade cost reduction matters. Our paper follows a similar line of thinking as AB, but our focus is very different. Whereas AB focus on proving that the individual and global welfare gains from changes in trade costs depend only on the direct effect based on a particular two-country model, we focus on establishing a simple general formula for the global welfare gains induced by changes in bilateral trade costs that can be applied to as general a condition as possible. Thus, our work complements that of AB.

Our work has some overlaps with the independent work of BC. They find that changes in world real GDP in response to changes in variable trade costs coincide with changes in theoretical consumption, up to a first order approximation.[3] Like us, they have an equation that associates changes in world real GDP with two sets of sufficient statistics, namely, changes in bilateral variable trade costs and export shares of continuing exporting producers. Thus, our work complements each other — we corroborate each other's finding, though we start from different model environments. In contrast with BC, we show that the formula applies to Melitz with general firm productivity distribution, not just Pareto distribution; moreover, we show that it applies to multi-stage production under perfect competition as well. Distinct from both AB and BC, we rigorously prove that the underlying mechanism for the result is that in the absence of externalities and price distortions, the envelope theorem can be applied to the maximization problem of a global planner to obtain the market outcome.

Further distinguishing our paper from the literature, later in the paper, we allow for non-constant markup under MC. We examine two cases. The first is a MC model with multiple sectors and different elasticities of substitution across sectors. We find that there is an extra term which depends on the combination of the firm mass effect and productivity effect. Unlike in the case of MC with constant markup, these two effects do not cancel each other. The intuition is that the sectors with lower elasticities of substitution would set higher markups, and if these sectors tend to have negative (positive) combined productivity effect and firm mass effect, then there would be negative (positive) overall impact on global welfare gains from reduction of trade costs. This makes sense as sectors with higher markups are associated with greater distortion. The effect of sectors with greater distortion would dominate the effect of sectors with smaller markups (and hence smaller distortion).

The second case is a one-sector MC model with variable markups. We use the simplest possible model to illustrate the effect of the existence of variable markups on the global gains formula (1). In this case, we find that there is an extra term which depends on the changes in the markups of firms. If firms with large market shares tend to reduce (raise) their markups following reduction of trade costs, the global gains would be larger (smaller) than the benchmark case. This makes sense as markups are distortions, and lower markups lead to high efficiency and thus higher global welfare gains. This result is consistent with the empirical finding of, for example, Edmond, Midrigan and Xu (2015), who report

[3] "Theoretical consumption" is a welfare measure based on consumption of goods.

that there are additional gains from reduction of trade costs due to the existence of variable markup as firms with larger market shares (i.e. domestic firms) tend to lower their markups.

Clearly, once we depart from constant markup, the global gains formula becomes a lot more complicated, and one needs to use more information and a more complicated computation method to calculate the extra effect due to the existence of non-constant markups. We describe the method and the additional data needed to carry out that task in each of the two cases.

Finally, we carry out two empirical applications. In the first empirical application, we calculate that a reduction of border-procedure related trade transaction costs by one percent of the value of world trade in 2003 would increase global income by USD 44.3 billion, roughly the same estimate by OECD. In the second empirical application, we calculate that the reduction of shipping time during 1960-2010 has cumulatively increased global income by somewhere between 2.7 to 9.8 percent, a magnitude consistent with the literature.

The structure of this paper is as follows. In section 2, we state the general setting and provide a justification of the definition of the percentage change in global welfare. Then, we state and explain the general result of the paper. We state three assumptions and two propositions and present the sketches of the proofs. The two propositions together indicate that the underlying mechanism for formula (1) is the envelope theorem. Section 3 presents the extensions to a multi-sector M-g model and a one-sector MC model with variable markups, together with some empirical applications of the formula. The last section concludes.

2 General Results

2.1 General Setting

Suppose in the world economy there are n countries (the set of countries is denoted by $\mathcal{N}$) that are capable of producing final goods $\omega \in \Omega^F$ where the superscript "F" is assigned to variables pertaining to "final good" whenever it is necessary to avoid confusion. The set of final goods that country i is capable of producing is Ω_i^F. Assume that all goods are tradable, and that there is complete or incomplete specialization (complete specialization means that a country cannot import the same good from more than one country) in all sectors for all countries, and variable extensive margins of trade. The extensive margin of country i's exports to country j is denoted by Ω_{ij}^F. Therefore, $\Omega_{ij}^F \subseteq \Omega_i^F \subseteq \Omega^F$.

Define E_i, Y_i, P_i and U_i as the expenditure, income, exact price index and welfare of country i, respectively. Define $\mathbf{q_j^F} \equiv \left\{ q_{ij}^F(\omega) \middle| \omega \in \Omega_{ij}^F,\, i \in \mathcal{N} \right\}$ as a vector of quantities of final goods consumed in country j (where $q_{ij}^F(\omega)$ is the quantity of final good ω consumed in j that is imported from country i), $\mathbf{p_j^F} \equiv \left\{ p_{ij}^F(\omega) \middle| \omega \in \Omega_{ij}^F,\, i \in \mathcal{N} \right\}$ as a vector of the corresponding prices (where $p_{ij}^F(\omega)$ is the price of final good ω consumed in country j that is imported from country i). We shall assume that ω is continuous unless otherwise stated.[4] The utility function (or welfare function) of country i, given by $U_i\left(\mathbf{q_i^F}\right)$, is assumed to be homogeneous of degree one in $\mathbf{q_i^F}$. Consequently, we can define an exact price index P_i, which stands for the cost a consumer has to pay to obtain one unit of utility. Therefore, the total utility of all consumers in country i (i.e. welfare of country i) is given by $U_i = E_i/P_i$ for all i.

We assume that labor is the only factor input, and marginal cost of production is assumed to be invariant with output. The variable L_j denotes country j's fixed labor supply while w_j denotes its labor wage. There is an iceberg trade cost such that τ_{ij} units are shipped from the source country i for one unit to arrive at the destination country j (assume that $\tau_{ii} = 1$).[5] Therefore,

$$p_{ij}^F(\omega) = p_{ii}^F(\omega)\,\tau_{ij} \text{ for } \omega \in \Omega_{ij}^F.$$

Define $\widehat{x} \equiv dx/x$, which we call the percentage change of x. Since $\widehat{p_{ii}^F}(\omega) = \widehat{w_i}$ for all ω, we have

$$\widehat{p_{ij}^F}(\omega) = \widehat{w_i} + \widehat{\tau}_{ij} \text{ for } \omega \in \Omega_{ij}^F.$$

The variable $y_{ij}^F(\omega)$ denotes the quantity of good ω produced in i that is exported to j. The iceberg trade cost links $q_{ij}^F(\omega)$ with $y_{ij}^F(\omega)$:

$$\tau_{ij} q_{ij}^F(\omega) \equiv y_{ij}^F(\omega) \quad \text{for } \omega \in \Omega_{ij}^F$$

which implies that

$$p_{ij}^F(\omega)\, q_{ij}^F(\omega) \equiv p_{ii}^F(\omega)\, y_{ij}^F(\omega) \equiv x_{ij}^F(\omega) \quad \text{for } \omega \in \Omega_{ij}^F$$

where $x_{ij}^F(\omega)$ denotes the exports of final good ω from i to j. Aggregate exports of final goods from i to j is denoted by $X_{ij}^F \equiv \int_{\omega\in\Omega_{ij}^F} x_{ij}^F(\omega)\, d\omega$.

Definition of percentage change in global welfare

Next, we present a justification for an expression that we use to measure the percentage change in global welfare. We want to define a measure of percentage change in global welfare resulting from small changes (infinitesimal ones in the formal analysis) in bilateral trade costs. We would like to have a concept of change of global welfare such that an increase in global welfare signifies an enlargement of global GDP so that *potentially* every country can be made better off by some proper income transfers between countries. Note that income is transferable but utility is not transferable. Therefore, the sum of utility of all countries is not a good measure of global welfare based on this concept.

[4] Calling $\mathbf{q_j^F}$ and $\mathbf{p_j^F}$ "vectors" is a slight abuse of language when ω is continuous. But since there is no ambiguity, we shall use it for simplicity of exposition.

[5] The amount $\tau_{ij} - 1$ can be called the "wastage due to shipping" per unit arriving at the destination, but it should also include the ad-valorem trade cost equivalent of any administrative delay or other non-tariff barriers.

We define the percentage increase in global welfare (following trade costs reduction) as the maximum potential equiproportional increase in welfare of all countries after some proper lump sum income transfers between countries. It measures the potential amount of Pareto improvement to the countries of the world as a whole. In principle, this amount can be negative. The above concept of the change in global welfare is consistent with that of Kaldor and Hicks (see, for example, Feldman 1998).[6] Consistent with Kaldor-Hicks' concept of efficiency, an outcome is more efficient if those that are made better off could in principle compensate those who are made worse off, so that a Pareto improving outcome can potentially result. This concept of Pareto improvement does not require compensation actually be paid, but merely that the possibility for compensation exists.

Following the reduction of trade costs, the vector of price-cum-welfare of the countries changes from $(P_1, ..., P_n; U_1, ..., U_n)$ to $(P_1 + dP_1, ..., P_n + dP_n; U_1 + dU_1, ..., U_n + dU_n)$. Let μ be the potential equiproportional increase in welfare of all countries following the reduction of trade costs. Hereinafter, $\sum_i \equiv \sum_{i=1}^n$ to simplify notation. Then, $\sum_i (P_i + dP_i)(U_i + dU_i) = \sum_i (P_i + dP_i)(\mu + 1) U_i$. The LHS is the total global expenditure before lump-sum transfers while the RHS is the total global expenditure after lump-sum transfers that lead to an equiproportional increase in welfare for all countries by a fraction μ. Note that this scheme of lump-sum transfers is equivalent to summing up the compensating variations of all countries and then distribute them equiproportionally across countries. However, because the changes in P_i and U_i are infinitesimal, this scheme is the same as summing up the equivalent variations of all countries and then distribute them equiproportionally across countries.[7] That is,

$$\sum_i P_i (U_i + dU_i) = \sum_i P_i (\mu + 1) U_i$$

In the rest of the paper, we shall use the concept of equivalent variation to evaluate the percentage change in global welfare. Re-arranging the above equation and simplifying, we have

$$\mu = \frac{1}{Y^w} \sum_i E_i \widehat{U}_i$$

where $\widehat{U}_i \equiv \frac{dU_i}{U_i}$, $E_i = P_i U_i$ (by definition) and $Y^w \equiv \sum_k P_k U_k$ is the GDP of the world. Note that the sum of equivalent variations of all countries is equal to $\sum_{i=1}^n E_i \widehat{U}_i$.[8]

[6] One shortcoming of the Kaldor-Hicks compensation criterion is that it is possible to construct an example such that distribution x is Pareto-superior to distribution y and at the same time distribution y is Pareto-superior to distribution x using the Kaldor-Hicks criterion. This problem arises when the compensating variation of a person (or group) is different from the equivalent variation. Because we are considering small changes, compensating variation is equal to equivalent variation. Thus this shortcoming does not arise. See, for example, Feldman (1998).

[7] This is because $(P_i + dP_i)(U_i + dU_i) - (P_i + dP_i) U_i$ is the compensating variation of country i. Note also that the equivalent variation of country i, $P_i (U_i + dU_i) - P_i U_i$, is the same as the compensating variation in the current context, because the changes are infinitesimal. For the concepts of compensating variation and equivalent variation, see, for example, Varian (1992, pp.160-163)

[8] Note also that though our formal analysis is based on infinitesimal changes, the equation should be a sufficiently good approximation as long as all percentage changes of P_i and U_i are small. The approximation error increases with the size of the change.

Thus, $\sum_i \frac{E_i}{Y^w}\widehat{U}_i$ (or the expenditure-share-weighted average percentage change of welfare of all countries) is the percentage change in global welfare. This makes sense as the importance of a country as indicated by its size should be reflected in the calculation of the change in global welfare. Note that we do not have to define what the global welfare function is. We only need to define what the percentage change in global welfare is. Note also that utility is cardinal, not ordinal, in this model. Intuitively, the welfare impact of a fractional change in global welfare, μ, is equivalent to that of having all consumers in the world increasing their consumption of each good by a fraction of μ, if the same sets of goods are produced, traded and consumed by each country before and after the shock.[9]

Note that

$$E_j = P_j U_j = \mathbf{p_j^F} \cdot \mathbf{q_j^F}.$$

Moreover, the exact price index P_j is a function of $\mathbf{p_j^F}$. Thus, given $\mathbf{p_j^F}$ and $\mathbf{q_j^F}$, we can calculate P_j and U_j. In other words, there is a one-to-one mapping from $\left(\mathbf{p_j^F};\, \mathbf{q_j^F}\right)$ onto $(P_j,\, U_j)$, and therefore a one-to-one mapping from $\left(\mathbf{p_1^F}, ..., \mathbf{p_n^F};\ \mathbf{q_1^F}, ..., \mathbf{q_n^F}\right)$ onto $(P_1, ..., P_n; U_1, ..., U_n)$. The equivalent variation of j, given by $P_j dU_j$, is equal to $\mathbf{p_j^F} \cdot d\mathbf{q_j^F}$. So, the sum of equivalent variations of all countries, given by $\sum_j P_j\,(dU_j)$, is equal to $\sum_j \mathbf{p_j^F} \cdot d\mathbf{q_j^F}$. Thus, the percentage change in global welfare can also be written as

$$\mu = \left(\sum_j \mathbf{p_j^F} \cdot d\mathbf{q_j^F}\right) \Big/ \left(\sum_j \mathbf{p_j^F} \cdot \mathbf{q_j^F}\right). \tag{2}$$

2.2 Specific Setting

We consider two settings below. The environment stated in section 2.1 is satisfied in both settings. In addition, some more structure is imposed in each model.

In the rest of the paper, where it is useful to simplify notation, we shall use $\sum_{a,b,c}$ to denote $\sum_a \sum_b \sum_c$ where each summation is over all the possible values that the dummy variable can take, e.g. $\sum_i \equiv \sum_{i=1}^n$ if i is the dummy for a country and there are n countries in the world.

We make three assumptions, which are to be applied to each of the two settings:

Assumptions:

1. The level of trade balance is fixed in each country.

2. Price is constant markup over marginal cost.

3. There are no externalities.

[9] This is because we assume that the utility function of each country is homogeneous of degree one in quantities of all goods consumed in that country. Therefore, the percentage change in global welfare is homogeneous of degree one in the percentage change of quantities of all goods consumed in the world as a whole.

1. PC with multi-stage production. The multi-stage production case subsumes single-stage production as a special case.

Preferences. The utility function, which is homogeneous of degree one in the final goods consumed, is given by $U_j = U_j\left(\left\{q_{ij}^F(\omega)\,\middle|\,\forall\, i,\, \omega \in \Omega_{ij}^F\right\}\right)$, where $q_{ij}^F(\omega)$ is country j's consumption of the final good (which is also stage-F good) ω imported from i.

Technology. Goods produced in each stage other than the final stage are used as intermediate inputs in the production of goods in the next stage, and all intermediate goods and final goods are tradeable. The final good is assumed to be produced in F sequential stages.[10] (For a single-stage production model, $F = 1$.) For $s = \{2, 3, ..., F\}$, the output of stage-s production (which shall be called "stage-s good") requires the inputs of labor and the previous stage's output. The production of stage-1 good requires only labor. The outputs at all stages are tradeable, and all countries possess the technologies of production for all stages. The production function for the stage-s good is assumed to be constant returns to scale, and is given by:

$$y_j^s(\omega) = \begin{cases} \varphi_j^s(\omega)\, f\left(\left\{q_{ij}^{s-1}(\omega')\middle|\, i \in \mathcal{N},\, \omega' \in \Omega_{ij}^{s-1}\right\},\ l_j^s(\omega)\right) & \text{for } \ s = 2, 3, ..., F \\ \varphi_j^s(\omega)\, l_j^s(\omega) & \text{for } \ s = 1 \end{cases} \qquad \text{for } \omega \in \Omega_j^s \tag{3}$$

where $y_j^s(\omega)$ is country $j's$ output of the stage-s good ω; $q_{ij}^{s-1}(\omega') = y_{ij}^{s-1}(\omega')/\tau_{ij}^{s-1}$ is country $j's$ use of imported input of the stage-(s-1) good ω' from country i for producing stage-s good ω; Ω_{ij}^{s-1} is the extensive margin of exports of the stage-(s-1) goods from i to j; $l_j^s(\omega)$ is country $j's$ labor input in the production of the stage-s good ω; $\varphi_j^s(\omega)$ is productivity.

Market Structure. The market structure for all goods is assumed to be perfect competition.

Examples of this kind of model include: neoclassical model based on endowment-driven comparative advantage (e.g. Dixit and Norman (1980) with trade costs or the Heckscher-Ohlin model with trade costs), Armington (1969), DFS1977, DFS1980 and EK2002, Melitz and Redding (2014) and Yi (2003, 2010).

2. MC with heterogeneous firm productivity. We assume that there is a large number of firms producing differentiated goods so that any single firm's choice of price would not affect the demand curve faced by other firms.

Preferences. The utility function is homogeneous of degree one in the final goods consumed.

Technology. There is only one stage of production and all goods are final goods. For every final good $\omega \in \Omega_i$, there is a blueprint that has been acquired by a firm through R&D. If a firm from country

[10] Here, we assume all final goods are produced in F sequential stages. Even if we assume that the number of production stages for different countries is different, our results continue to hold.

i produces $\mathbf{y}_i^{\mathbf{F}} \equiv \left\{ y_{ij}^F(\omega) \middle| j \in \mathcal{N} \right\}$ units of good ω, its cost function is given by

$$C_i\left(w_i, \mathbf{y}_i^{\mathbf{F}}, \omega\right) = \sum_j \left[\tau_{ij} w_i a_i(\omega) q_{ij}^F(\omega) + \xi_{ij} w_i \cdot \mathbf{1}\left(y_{ij}^F(\omega) > 0\right)\right]$$

where $\mathbf{1}\left(y_{ij}^F(\omega) > 0\right)$ is an indicator function, $a_i(\omega)$ denotes the unit labor requirement in producing good ω in country i, and ξ_{ij} denotes the labor requirement that underlies the fixed cost of exporting from i to j, where $\xi_{ij} \geq 0 \; \forall \; i, \, j$.

Market Structure. N_i is the measure of the number of entrants (successful or not) in country i, which is endogenously determined by the free entry condition so that the expected net profit for any firm is equal to zero. A firm from country i needs to hire f_e units of labor to develop a blueprint, which confers it with monopoly power. In equilibrium, the entry cost, $w_i f_e$, is equal to the expected profit of each firm. The number of firms in country i serving market j, N_{ij}, is determined by the zero cutoff profit conditions. Labor productivity is a random variable denoted by $\varphi \equiv 1/a_i(\omega)$. From now on, φ and ω are used interchangeably to index goods. The functions $G_i(\varphi)$ and $g_i(\varphi)$ are the cdf and pdf respectively of φ. Define φ_{ij}^* as the cutoff productivity for a firm in country i that can profitably export to country j.

Examples of this kind of model include K1980 and M-g.

2.3 Results

Below we state two propositions. Proposition 1 provides the basis for proving Proposition 2, which is the key proposition of this paper.

Proposition 1 *(Market Efficiency) Under the specific settings stated in section 2.2, the market is efficient.*

Proof of Proposition 1

For the setting of PC with multi-stage production, Proposition 1 follows from the First Fundamental Theorem of Welfare Economics.

For the setting of MC with heterogeneous firm productivity, refer to Online Appendix A. In that appendix, we prove that the global market is efficient, in the sense that the market allocation of resources $\{l_{ij}(\varphi)\}$ is identical to the allocation of a global planner who maximizes global income subject to the labor constraint of each country. Conceptually, the proof is an extension of Dhingra and Morrow's (2016) analysis to the global economy. Intuitively, with constant markup, there is no distortion in the relative prices. So, in the absence of externalities, the market is efficient.■

Next, we state the core proposition of this paper.

Proposition 2 *(Global Gains) Under the specific settings stated in section 2.2, the percentage change in global welfare induced by changes in bilateral trade costs is given by expression (1). In other words,* $\frac{1}{Y^w}\sum_i E_i \widehat{U}_i = -\sum_{j,i}\frac{X_{ij}}{Y^w}\widehat{\tau}_{ij}$*, where* X_{ij} *is the total value of exports from country* i *to country* j*,* $\widehat{\tau}_{ij}$ *is a small percentage change in the iceberg trade cost, and* Y^w *is global GDP.*

Note that with multi-stage production, $X_{ij} = \sum_s X_{ij}^s$ where X_{ij}^s is the value of exports of the stage-s goods from i to j; and we assume that $\tau_{ij}^s = \tau_{ij}$ $\forall s$ in the above proposition. With single-stage production, $X_{ij} = X_{ij}^F$ and τ_{ij} applies to trade in final goods.

General Proof of Proposition 2

Proposition 1 implies that the effect of reduction of trade costs on global welfare under the market is the same as that under the setting when a global planner maximizes global income (call it W), by choosing the allocation of labor to the production of all goods in all countries and the consumption of all goods in all countries, subject to the set of all bilateral trade costs, the shadow prices of all goods (which are also the market prices), labor supplies of all countries and the production functions of all goods. Therefore, we can prove Proposition 2 by invoking Proposition 1 and calculate the effect of changes in trade costs on global welfare under the setting with a global planner. Based on the argument presented in section 1, we shall use the pre-change market prices and the concept of equivalent variation to evaluate the global welfare change. Below we give a general proof of Proposition 2 based on this approach. Specific proofs of the two models are relegated to Appendixes A and B.

From Proposition 1, we know that the equilibrium values of the endogenous variables under the market is identical to the optimal values of the choice variables chosen by the global planner.[11] As she chooses labor allocation to the production of each good to maximize global income, she would automatically maximize the income of each country subject to its resource constraint and the set of shadow prices. Thus, the national GDP resulting from the global planner's labor allocation is also the market-determined national GDP. Simultaneously, on the demand side, the combination of the quantities of goods consumed in each country (which are determined by the allocation of labor to the production of those goods) maximizes the utility of the country subject to the set of shadow prices and the GDP of the country.

Let $l_{ij}^s(\omega)$ be the variable labor input used to produce the stage-s good ω that is exported from i to j. For a single-stage model, $l_{ij}^F(\omega)$ is also denoted by $l_{ij}(\omega)$ for simplicity of notation. At the shadow prices $\mathbf{p}^s \equiv \{\mathbf{p_1^s}, ..., \mathbf{p_n^s}\}$ for all s, where $\mathbf{p_j^s} \equiv \left\{ p_{ij}^s(\omega) \middle| i \in \mathcal{N}, \omega \in \Omega_{ij}^s \right\}$, the quantities demanded and quantities supplied of all goods are equalized.

[11] The global planner maximizes the value of the global GDP function, given the set of shadow prices of all goods, the labor supplies of all countries and all bilateral trade costs. This is analogous to a country's social planner maximizing the value of the national GDP function subject to its labor supply, the shadow prices of all goods, and the production functions of all goods. (See, for example, Feenstra 2004, pp. 6-8, for the idea of GDP function used here.)

Define $\widetilde{W}$ as the maximized value of W after the global planner has optimally chosen the values of her choice variables. We shall show that the percentage change in $\widetilde{W}$ is equal to the percentage change in global welfare. Then, to prove Proposition 2, we shall calculate $\left(1/\widetilde{W}\right)\sum_{s,i,j}\left[\left(d\widetilde{W}/d\tau_{ij}^{s}\right)d\tau_{ij}^{s}\right]$, and show that it is the same as (1). As τ_{ij}^{s} changes, it affects the equilibrium values of all the endogenous variables, including the prices of all goods in all countries, denoted by $\mathbf{p}$, and the values of all the choice variables of the central planner, namely, the labor allocated to the production of all goods in all countries, and the bilateral exports of all goods for all country-pairs. This set of choice variables of the central planner is denoted by $\mathbf{\Phi}$. The proof of Proposition 2 hinges on invoking the envelope theorem in that the total derivative of $\widetilde{W}$ with respect to τ_{ij}^{s}, $d\widetilde{W}/\partial\tau_{ij}^{s}$, is just equal to the partial derivative $\partial W/\partial\tau_{ij}^{s}\Big|_{\mathbf{\Phi}=\widetilde{\mathbf{\Phi}}}$ (i.e. when $\mathbf{\Phi}$ is optimally chosen). In other words, only the direct effect of τ_{ij}^{s} on $\widetilde{W}$ matters; the indirect effects do not matter. The proof is completed when we show that $\left(1/\widetilde{W}\right)\sum_{s,i,j}\left[\left(\partial\widetilde{W}/\partial\tau_{ij}^{s}\right)d\tau_{ij}^{s}\right]$ is given by (1). The proof is presented in three steps.

Step 1: Maximization of global income by the global planner

The global planner maximizes W by choosing a vector of choice variables $\mathbf{\Phi}$, namely, the labor allocated to the production of all goods in each country, and the bilateral exports of all goods for all country-pairs (with the formal definition given below), taking the trade costs τ_{ij} for all i,j, and the shadow prices $p_{ij}^{F}(\omega)$ for all i,j,ω, as given. Define $\mathbf{l}\equiv\left(l_{ij}^{s}(\omega)\right),\forall i,j,s,\omega$ as a vector of all labor allocations; $\mathbf{p}\equiv\left(p_{ij}^{s}(\omega)\right),\forall i,j,s,\omega$ as a vector of all shadow prices; and $\boldsymbol{\tau}\equiv\left(\tau_{ij}^{s}\right),\forall i,j,s$ as a vector of all trade costs. The definitions of $\mathbf{l}$, $\mathbf{p}$ and $\boldsymbol{\tau}$ differ for different models as explained in the specific proofs in the appendix. Thus, she solves

$$\max_{\{\mathbf{\Phi}\}} W=\sum_{i,j}\frac{\int_{\omega\in\Omega_{ij}^{F}}p_{ij}^{F}(\omega)\,y_{ij}^{F}(\omega)\,(\mathbf{\Lambda_{ij}})\,d\omega}{\tau_{ij}^{F}}$$

subject to the labor constraint of each country and the production functions of all goods, $y_{ij}^{F}(\omega)(\mathbf{\Lambda_{ij}})\ \ \forall i,j,\omega$, where the vector of inputs $\mathbf{\Lambda_{ij}}$ and the vector of choice variables $\mathbf{\Phi}$ are different for different models. For PC with single-stage production, $\mathbf{\Lambda_{ij}}=l_{ij}(\omega)$; for PC with multi-stage production, $\mathbf{\Lambda_{ij}}=\{\mathbf{l},\boldsymbol{\tau}\}$; for M-g, $\mathbf{\Lambda_{ij}}=l_{ij}(\omega)$, which is also denoted by $l_{ij}(\varphi)$ (as φ and ω are used interchangeably to index goods).[12] $\mathbf{\Phi}=\left\{\mathbf{l},\left\{\Omega_{ij}^{s}\right\}\right\}$ for PC with single-stage production and PC with multi-stage production;[13] $\mathbf{\Phi}=\left\{\{l_{ij}(\varphi),\forall i,j,\varphi\},\left\{\varphi_{ij}^{*}\right\},\{N_i\}\right\}$ for M-g. The maximized value of W before the changes in trade costs is denoted by Y^{w}. In other words, $Y^{w}\equiv\widetilde{W}$, and we shall use them interchangeably.

[12] For PC with multi-stage production, $y_{ij}^{F}(\omega)$ is affected not only by labor input $l_{ij}^{F}(\omega)$ but by intermediate inputs (which are functions of $\mathbf{l}$ and $\boldsymbol{\tau}$) and labor inputs $(\mathbf{l})$ in all stages. Thus $y_{ij}^{F}(\omega)$ is a function of $\mathbf{l}$ and $\boldsymbol{\tau}$. For PC with single-stage production, $y_{ij}^{F}(\omega)$ is a function of $l_{ij}^{F}(\omega)$ only, and is independent of $\boldsymbol{\tau}$.

[13] $\left\{\Omega_{ij}^{s}\right\}$ is included in $\mathbf{\Phi}$ when extensive margins of trade are variable under PC; it is not included when extensive margins are fixed.

The expressions for (i) the exports of final goods from i to j, given by $\int_{\omega\in\Omega_{ij}^F} p_{ij}^F(\omega)\, y_{ij}^F(\omega)\,(\mathbf{\Lambda_{ij}})\, d\omega$, (ii) the labor constraint in each country, and (iii) the production function, given by $y_{ij}^F(\omega)(\mathbf{\Lambda_{ij}})$, differ for different models, as explained in the specific proofs in the appendix.

Note that we can also write global income in the following way:

$$W = \sum_{i,j} X_{ij}^F = \sum_j \mathbf{p_j^F} \cdot \mathbf{q_j^F}$$

We shall make use of these equalities later.

Let the optimal value of $\mathbf{\Phi}$ be denoted by $\widetilde{\mathbf{\Phi}}$. Noting that $\mathbf{p^F}$ is a function of $\boldsymbol{\tau}$, the above maximization problem of the global planner can be re-stated as

$$\begin{aligned}
&\max_{\{\mathbf{\Phi}\}} W\left(\boldsymbol{\tau}, \mathbf{p^F}(\boldsymbol{\tau}), \mathbf{\Phi}\right),\\
\Longrightarrow \qquad & \frac{\partial}{\partial \mathbf{\Phi}} W\left(\boldsymbol{\tau}, \mathbf{p^F}(\boldsymbol{\tau}), \mathbf{\Phi}\right) = 0 \text{ (first order condition)},\\
\Longrightarrow \qquad & \widetilde{\mathbf{\Phi}} = g\left[\boldsymbol{\tau}, \mathbf{p^F}(\boldsymbol{\tau})\right], \text{ where } g \text{ is some function,}
\end{aligned}$$

provided that the condition for the implicit function theorem is satisfied, which we assume to be the case. Thus, $\widetilde{\mathbf{\Phi}}$ is a function of $\boldsymbol{\tau}$ and $\mathbf{p^F}(\boldsymbol{\tau})$. Therefore, a change in $\boldsymbol{\tau}$ affects $\widetilde{\mathbf{\Phi}}$ directly as well as indirectly through $\mathbf{p^F}$. Consequently, the maximized value of W is given by

$$\widetilde{W} = W\left(\boldsymbol{\tau}, \mathbf{p^F}(\boldsymbol{\tau}), \mathbf{\Phi}\right)\big|_{\mathbf{\Phi}=\widetilde{\mathbf{\Phi}}} = W\left(\boldsymbol{\tau}, \mathbf{p^F}(\boldsymbol{\tau}), \widetilde{\mathbf{\Phi}}\left(\boldsymbol{\tau}, \mathbf{p^F}(\boldsymbol{\tau})\right)\right).$$

Hereinafter, for simplicity, we shall omit the arguments of $\widetilde{W}(.)$ and $W(.)$ unless there is a risk of confusion.

Step 2: Invoking the Envelope Theorem

Based on the last equation, in evaluating the total effect of τ_{ij}^s on $\widetilde{W}$, we have to evaluate the direct effect of τ_{ij}^s as well as the indirect effects of how $\mathbf{\Phi}$ and $\mathbf{p^F}$ are affected by the change of τ_{ij}^s. In other words, we have to take into account (1) the direct effect of $\tau_{ij}^s \longrightarrow \widetilde{W}$; (2) plus the indirect effect of $\tau_{ij}^s \longrightarrow \mathbf{\Phi} \longrightarrow \widetilde{W}$; (3) plus the indirect effect of $\tau_{ij}^s \longrightarrow \mathbf{p^F} \longrightarrow \mathbf{\Phi} \longrightarrow \widetilde{W}$; (4) plus the indirect effect of $\tau_{ij}^s \longrightarrow \mathbf{p^F} \longrightarrow \widetilde{W}$. However, as we explain below, only effects (1) through (3) are relevant for calculating the sum of equivalent variations of all countries. To see this, note that the total effect of τ_{ij}^s on $\widetilde{W}$ can be written as

$$\begin{aligned}
\frac{d\widetilde{W}}{d\tau_{ij}^s} &= \underbrace{\left.\frac{\partial W}{\partial \tau_{ij}^s}\right|_{\mathbf{\Phi}=\widetilde{\mathbf{\Phi}}}}_{\text{Effect (1)}} + \underbrace{\left.\frac{\partial W}{\partial \mathbf{\Phi}} \cdot \frac{\partial \mathbf{\Phi}}{\partial \tau_{ij}^s}\right|_{\mathbf{\Phi}=\widetilde{\mathbf{\Phi}}}}_{\text{Effect (2)}} + \underbrace{\left.\frac{\partial W}{\partial \mathbf{\Phi}} \cdot \frac{\partial \mathbf{\Phi}}{\partial \mathbf{p^F}} \cdot \frac{\partial \mathbf{p^F}}{\partial \tau_{ij}^s}\right|_{\mathbf{\Phi}=\widetilde{\mathbf{\Phi}}}}_{\text{Effect (3)}} + \underbrace{\left.\frac{\partial W}{\partial \mathbf{p^F}} \cdot \frac{\partial \mathbf{p^F}}{\partial \tau_{ij}^s}\right|_{\mathbf{\Phi}=\widetilde{\mathbf{\Phi}}}}_{\text{Effect (4)}}\\
&= \underbrace{\left.\frac{\partial W}{\partial \tau_{ij}^s}\right|_{\mathbf{\Phi}=\widetilde{\mathbf{\Phi}}}}_{\text{Effects (1)+(2)+(3)}} + \underbrace{\left.\frac{\partial W}{\partial \mathbf{p^F}} \cdot \frac{\partial \mathbf{p^F}}{\partial \tau_{ij}^s}\right|_{\mathbf{\Phi}=\widetilde{\mathbf{\Phi}}}}_{\text{Effect (4)}} \quad \text{since } \left.\frac{\partial W}{\partial \mathbf{\Phi}}\right|_{\mathbf{\Phi}=\widetilde{\mathbf{\Phi}}} = \mathbf{0} \text{ as } \mathbf{\Phi} \text{ has been optimally chosen.}
\end{aligned} \tag{4}$$

Accordingly, effects (2) and (3) are equal to zero, as they are second order, and the exogenous changes are infinitesimal. This is precisely the principle underlying the envelope theorem. On the other hand, since $\widetilde{W} = W|_{\mathbf{\Phi}=\widetilde{\mathbf{\Phi}}} = \sum_j \mathbf{p_j^F} \cdot \mathbf{q_j^F}\Big|_{\mathbf{\Phi}=\widetilde{\mathbf{\Phi}}}$ (where $\mathbf{q_j^F}\Big|_{\mathbf{\Phi}=\widetilde{\mathbf{\Phi}}}$ is the optimally chosen allocation of consumption goods), the total effect of τ_{ij}^s on $\widetilde{W}$ can also be written as

$$\frac{d\widetilde{W}}{d\tau_{ij}^s} = \underbrace{\sum_j \mathbf{p_j^F} \cdot \frac{d\mathbf{q_j^F}}{d\tau_{ij}^s}\Bigg|_{\mathbf{\Phi}=\widetilde{\mathbf{\Phi}}}}_{\text{Effects (1)+(2)+(3)}} + \underbrace{\sum_j \frac{d\mathbf{p_j^F}}{d\tau_{ij}^s} \cdot \mathbf{q_j^F}\Big|_{\mathbf{\Phi}=\widetilde{\mathbf{\Phi}}}}_{\text{Effect (4)}}.$$

Recall from subsection 2.1 and equation (2) that the first term on the RHS of the above equation corresponds to the sum of equivalent variations of all countries caused by a change in τ_{ij}^s, and it is the only term we care about in order to calculate the percentage change in global welfare. The rationale is that, as we are using pre-change market prices and concept of equivalent variation to evaluate the global welfare change, the direct effect of $\mathbf{p^F}$, i.e. Effect (4) above, can be ignored. Thus, comparing the last lines of the above two equations, we conclude that

$$\sum_j \mathbf{p_j^F} \cdot \frac{d\mathbf{q_j^F}}{d\tau_{ij}^s}\Bigg|_{\mathbf{\Phi}=\widetilde{\mathbf{\Phi}}} = \frac{d\widetilde{W}}{d\tau_{ij}^s} = \frac{\partial W}{\partial \tau_{ij}^s}\Bigg|_{\mathbf{\Phi}=\widetilde{\mathbf{\Phi}}}.$$

That is, the sum of equivalent variations of all countries induced by each unit of infinitesimal change of τ_{ij}^s is equal to the partial derivative $\partial W/\partial \tau_{ij}^s\Big|_{\mathbf{\Phi}=\widetilde{\mathbf{\Phi}}}$. Thus, we have the following key lemma of this paper.

Lemma 1 *(Irrelevance of Indirect Effects) The change in global welfare (measured by the sum of equivalent variations of all countries) induced by each unit of infinitesimal change in τ_{ij}^s is equal to $\partial W/\partial \tau_{ij}^s\Big|_{\mathbf{\Phi}=\widetilde{\mathbf{\Phi}}}$. In other words, the only effect is the direct effect of τ_{ij}^s on world income.*

Step 3: Invoking Lemma 1

Hence, according to (2), the percentage change in global welfare induced by changes in the set of trade costs $\left\{\tau_{ij}^s\right\}$ is given by

$$\begin{aligned}
\mu &= \sum_j \frac{\mathbf{p_j^F} \cdot d\,\mathbf{q_j^F}\Big|_{\mathbf{\Phi}=\widetilde{\mathbf{\Phi}}}}{\sum_j \mathbf{p_j^F} \cdot \mathbf{q_j^F}\Big|_{\mathbf{\Phi}=\widetilde{\mathbf{\Phi}}}} \\
&= \frac{1}{Y^w} \sum_{s,i,j} \frac{\partial W}{\partial \tau_{ij}^s}\Bigg|_{\mathbf{\Phi}=\widetilde{\mathbf{\Phi}}} d\tau_{ij}^s \\
&= -\sum_{s,i,j} \frac{X_{ij}^s \left(\tau_{ij}^s\right)^{-1}}{Y^w} d\tau_{ij}^s \\
&= -\sum_{i,j} \frac{X_{ij}\widehat{\tau_{ij}}}{Y^w} \qquad \text{if } \widehat{\tau_{ij}^s} = \widehat{\tau_{ij}} \text{ for all } s\text{, and } X_{ij} = \sum_s X_{ij}^s. \qquad (1)
\end{aligned}$$

where $W|_{\mathbf{\Phi}=\widetilde{\mathbf{\Phi}}}=\sum_{i,j}X_{ij}^F$ and $X_{ij}^s\equiv\left[\int_{\omega\in\Omega_{ij}^s}p_{ij}^s(\omega)\,y_{ij}^s(\omega)\,(\mathbf{\Lambda_{ij}})\,d\omega\right]/\tau_{ij}^s$, and $\int_{\omega\in\Omega_{ij}^s}p_{ij}^s(\omega)\,y_{ij}^s(\omega)\,(\mathbf{\Lambda_{ij}})\,d\omega$ is independent of τ_{ij}^s. Although the expression for $\int_{\omega\in\Omega_{ij}^s}p_{ij}^s(\omega)\,y_{ij}^s(\omega)\,(\mathbf{\Lambda_{ij}})\,d\omega$ is different for different models, expression (1) holds for all models. The second line of the above calculation stems from invoking Lemma 1, and then summing up the effects over all i,j,s. For single-stage production under PC and M-g, the third line of the above calculation is obvious. For multi-stage production under PC, the third line stems from the fact that the change in total global value of outputs at stage F is equal to the change in total global value of inputs at an earlier stage $s+1$ induced by a decrease in τ_{ij}^s (where $s<F$).[14] This completes our general proof of Proposition 2. ■

In Appendixes A and B, we provide specific proofs of Proposition 2 for PC with multi-stage production and MC with heterogeneous firms (M-g). Appendix A shows that the economic intuition for equation (1) to hold for PC with multi-stage production is that whenever there is an amount of terms of trade gain by an exporter of a good at any stage, there is an equal amount of terms of trade loss by an importer of the same good. Thus, the global effect of terms of trade changes is nil. Consequently, only the direct effect, which is the total saving in trade costs in all stages, matters for global welfare change. Furthermore, any saving in trade cost at any stage is eventually passed on to the final stage. As a result, the saving in trade cost in each stage shows up as the gain in global income in the final stage. The percentage change in global welfare is therefore given by $-\sum_{i,j,s}\frac{X_{ij}^s\widehat{\tau_{ij}^s}}{Y^w}=-\sum_{i,j}\frac{X_{ij}}{Y^w}\widehat{\tau_{ij}}$ if $\widehat{\tau_{ij}^s}=\widehat{\tau_{ij}}$ for all s, and $X_{ij}=\sum_s X_{ij}^s$. Note that a PC model with more stages of production and a PC model with fewer stages of production but with the same production function at each overlapping stage will in general give rise to different X_{ij} for the same set of bilateral trade costs $\{\tau_{ij}\}$ (with larger X_{ij} under the setting with more stages of production). Therefore, for the same percentage changes in bilateral trade costs, the global welfare gains from reduction of trade costs are higher when production is more fragmented internationally.

Appendix B shows that the intuition for equation (1) to hold for M-g is that as cutoff productivities change due to changes in trade costs, the effect on the average productivity of firms serving each market (productivity effect) and the effect on the mass of firms serving each market (firm mass effect) completely offset each other from the point of view of global welfare, regardless of the distribution of firm productivity. This is because, for each exporting country, the labor constraint dictates that the change in average productivity and change in firm mass in each market go in opposite directions, and they offset each other when summing up over all markets. As Melitz and Redding (2015) have pointed out, a K1980 model and a M-g model with the same deep parameters will in general give rise to different X_{ij} for the same set of bilateral trade costs (with M-g yielding larger X_{ij}). Therefore, for the same percentage changes in bilateral trade costs, the M-g model in general gives rise to larger global welfare gains than does the K1980 model with the same deep parameters.

[14] That is, $\frac{\partial W}{\partial\tau_{ij}^s}d\tau_{ij}^s=\frac{\partial}{\partial\tau_{ij}^s}\left(\sum_{i'}\sum_{j'}X_{i'j'}^F\right)d\tau_{ij}^s=\frac{\partial}{\partial\tau_{ij}^s}\left(\sum_{i'}\sum_{j'}X_{i'j'}^s\right)d\tau_{ij}^s=\frac{\partial X_{ij}^s}{\partial\tau_{ij}^s}d\tau_{ij}^s=-X_{ij}^s\left(\tau_{ij}^s\right)^{-1}d\tau_{ij}^s$, where $\sum_{i'}\sum_{j'}X_{i'j'}^s$ is the total global value of inputs at stage-$(s+1)$. See appendix A for more detail.

In Appendix C, we analyze how the market responds to exogenous changes in trade costs under the two models, viz. PC with multi-stage production and M-g. By analyzing the market's response instead of the global planner's response, we are able to identify effects that are canceled out when we consider the impact on global welfare instead of welfare of individual countries.

3 Extensions and Empirical Applications

In this section, we analyze two special cases for which Assumption 2 is violated so that prices are not constant markup over marginal cost. These are not general cases of variable markup, but they are interesting extensions to the general results reported in section 2. Following these two extensions, we carry out a couple of simple empirical applications of the baseline model.

3.1 M-g with multiple sectors

Suppose that the set of goods $\omega \in \Omega$ is separated into sectors denoted by Ω^z where sectors are indexed by $z = 1, ..., Z$. Consumers in country j have their preferences represented by the following utility function:

$$U_j = U\left(\{u_j(z) |\, z = 1, 2, .., Z\}\right) \text{ where } u_j\left(z\right) = \left[\sum_i \int_{\omega \in \Omega_{ij}^z} q_{ij}^z(\omega)^{\frac{\sigma_z - 1}{\sigma_z}} d\omega\right]^{\frac{\sigma_z}{\sigma_z - 1}}$$

where U_j is homogeneous of degree one in $u_j(z)$, and $q_{ij}^z(\omega)$ denotes the consumption in country j of variety ω in sector z originating from country i. In general, the elasticity of substitution σ_z can be different across sectors.

The fixed exporting cost from country i to j in sector z is equal to ξ_{ijz} in units of labor; the iceberg exporting cost from country i to country j in sector z is given by τ_{ij}^z. Each firm needs to pay a fixed entry cost equal to the cost of f_{ez} units of labor to acquire a blueprint to produce in sector z. The productivity of the firm, φ, is a random variable. The unit labor requirement of producing good ω is denoted by $a_i\left(\omega\right) \equiv 1/\varphi$. Thus, ω and φ can be used interchangeably to index a good. The functions $G_{iz}\left(\varphi\right)$ and $g_{iz}\left(\varphi\right)$ are the cdf and pdf respectively of φ for sector z in country i.

We prove the following proposition in Online Appendix D.

Proposition 3 *(Multiple Sectors) Suppose there are multiple sectors and the market structure of each sector is monopolistic competition as per Melitz (2003) with general firm productivity distribution, the percentage change in global welfare is given by*

$$\textstyle\sum_j \frac{E_j}{Y^w}\widehat{U}_j = -\sum_{j,i,z} \frac{X_{ij}^z}{Y^w}\widehat{\tau_{ij}^z} + \sum_{j,i,z} \frac{X_{ij}^z}{Y^w}\left(\widehat{\widetilde{\varphi}_{ijz}} + \frac{\widehat{N_{ij}^z}}{\sigma_z - 1}\right).$$

Moreover, $\sum_{j,i,z} X_{ij}^z \left[\left(\sigma_z - 1\right)\widehat{\widetilde{\varphi}_{ijz}} + \widehat{N_{ij}^z}\right] = 0.$

Besides the direct effect, $-\sum_{j,i,z}\frac{X_{ij}^z}{Y^w}\widehat{\tau}_{ij}^z$, there is one more term, $\sum_{j,i,z}\frac{X_{ij}^z}{Y^w}\left(\widehat{\widetilde{\varphi}_{ijz}}+\frac{\widehat{N_{ij}^z}}{\sigma_z-1}\right)$, which is the sum of the combination of firm mass effect $(\widehat{N_{ij}^z})$ and productivity effect $(\widehat{\widetilde{\varphi}_{ijz}})$, summing over all sectors and all country pairs. Moreover, it can be shown that the labor market clearing condition, together with the free entry condition, lead to $\sum_{j,i,z}X_{ij}^z\left[(\sigma_z-1)\widehat{\widetilde{\varphi}_{ijz}}+\widehat{N_{ij}^z}\right]=0$, i.e. the weighted sum of the combination of firm mass effect and productivity effect (with the weight being σ_z-1) is equal to zero when summing over all sectors and all country pairs.

If σ_z is the same across sectors, $\sum_{j,i,z}\frac{X_{ij}^z}{Y^w}\left(\widehat{\widetilde{\varphi}_{ijz}}+\frac{\widehat{N_{ij}^z}}{\sigma_z-1}\right)$ is equal to zero as the firm mass effect and productivity effect completely offset each other (as shown in equation (14) in section C.2). In that case, the change in global welfare is given by $\sum_j\frac{E_j}{Y^w}\widehat{U}_j=-\sum_{j,i,z}\frac{X_{ij}^z}{Y^w}\widehat{\tau}_{ij}^z$, which reduces to expression (1) when $\tau_{ij}^z=\tau_{ij}$ for all z. If σ_z are different across sectors, price markups are different across sectors. As a result, relative prices are distorted and market resource allocation is not efficient. If $\sum_{j,i}X_{ij}^z\left[(\sigma_z-1)\widehat{\widetilde{\varphi}_{ijz}}+\widehat{N_{ij}^z}\right]$ tends to be positive (negative) in the sector with small σ_z (i.e. higher markup), then $\sum_{j,i,z}X_{ij}^z\left(\widehat{\widetilde{\varphi}_{ijz}}+\frac{\widehat{N_{ij}^z}}{\sigma_z-1}\right)$ would be positive (negative), and the combination of the firm mass effect and productivity effect on global welfare would be positive (negative). In other words, the effect of the sectors with higher markups dominate. This makes sense as sectors with higher markups are associated with greater distortion. The effect of sectors with greater distortion would dominate over the effect of sectors with smaller markups (and hence smaller distortion). Ossa (2015) finds that assuming symmetric trade elasticity across sectors leads to a gross underestimation of the gains from trade. Thus, we can infer that Ossa's (2012) empirical finding implies that $\sum_{j,i}X_{ij}^z\left[(\sigma_z-1)\widehat{\widetilde{\varphi}_{ijz}}+\widehat{N_{ij}^z}\right]$ tends to be positive (negative) when σ_z is small (large).

3.2 MC with single sector and variable markups

To use the simplest possible model to illustrate how the existence of variable markups affect our benchmark result, we consider the case with heterogeneous firms and assume that the number of firms serving each market is discrete in equilibrium. As there is a discrete number of firms, and the changes in trade costs are infinitesimal, the number of firms that serve each market by each country is unchanged after the reduction of trade costs. Thus, there is neither firm mass effect nor productivity effect. We continue to assume that each firm only produces one variety. As we shall see, the result would be sharper if we consider a case when a significant market share is concentrated in a small number of firms from a small number of countries.

We assume that the utility in country j is given by:

$$U_j=\left[\sum_i\sum_{\omega\in\Omega_{ij}^F}q_{ij}^F(\omega)^{\frac{\sigma-1}{\sigma}}\right]^{\frac{\sigma}{\sigma-1}} \tag{5}$$

where $\sigma>1$ is the elasticity of substitution. Then, consumer optimization yields the demand function for good ω:

$$q_{ij}^F(\omega) = \frac{p_{ij}^F(\omega)^{-\sigma}}{P_j^{1-\sigma}} E_j \tag{6}$$

where $P_j = \left\{ \sum_i \sum_{\omega \in \Omega_{ij}^F} \left[p_{ij}^F(\omega) \right]^{1-\sigma} \right\}^{\frac{1}{1-\sigma}}$ is the consumer price index in country j. Let φ be the labor productivity associated with variety ω.

Given (6), profit maximization yields

$$\mu_{ij}(\varphi) = 1 + \frac{1}{(\sigma - 1)\left[1 - s_{ij}(\varphi)\right]} \tag{7}$$

where $\mu_{ij}(\varphi) = \frac{p_{ij}^F(\varphi)}{c_{ij}(\varphi)}$ is the markup by a firm and $c_{ij}(\varphi) = \frac{\tau_{ij} w_i}{\varphi}$ is the marginal cost, and $s_{ij}(\varphi) \equiv x_{ij}(\varphi)/E_j$ is the market share of a good produced by a firm with productivity φ from country i in country j, and $x_{ij}(\varphi)$ is the value of exports of the firm with productivity φ from country i to country j.

Clearly, $\mu_{ij}(\varphi)$ increases with $s_{ij}(\varphi)$, which in turn increases with φ. Obviously, $\widehat{P}_j$ would be affected not just by $\widehat{w}_i + \widehat{\tau}_{ij}$ (which affects $c_{ij}(\varphi)$), but also by $\widehat{\mu}_{ij}(\varphi)$.

We prove the following proposition in online appendix E.

Proposition 4 *(Variable Markups) Suppose there is a single sector and single stage of production and the market structure is monopolistic competition with variable markups, the percentage change in global welfare is given by*

$$\sum_j \frac{E_j \widehat{U}_j}{Y^w} = -\sum_{i,j} \frac{X_{ij}}{Y^w} \widehat{\tau}_{ij} - \sum_{i,j} \frac{E_j}{Y^w} \left[\sum_{\omega \in \Omega_{ij}^F} s_{ij}(\varphi) \widehat{\mu}_{ij}(\varphi) \right]$$

Compared with the benchmark result, there is an extra term $-\sum_{i,j} \frac{E_j}{Y^w} \left[\sum_{\omega \in \Omega_{ij}^F} s_{ij}(\varphi) \widehat{\mu}_{ij}(\varphi) \right]$ which depends on the changes in markups of firms. If firms with large market shares tend to reduce (raise) their markups following reduction of trade costs, the global gains would be larger (smaller) than the benchmark case. This makes sense as markups are distortions, and lower markups lead to high efficiency and thus higher global welfare gains. This is consistent with the finding of, for example, Edmond, Midrigan and Xu (2015), who report that reduction of trade costs tend to raise the markups of foreign firms but lower the markups of domestic firms, but since the market shares of domestic firms are larger, there are additional gains from reduction of trade costs due to the existence of variable markups. Note that in the case where countries are symmetric (with fixed exporting costs), meaning that for any given φ, $s_{ij}(\varphi)$ is the same for all i and j such that $i \neq j$, the extra term is trivial given that there is a large number of countries n. However, if we consider a highly asymmetric case where a significant market share is concentrated in a small number of firms from a small number of countries, then the extra term can be non-trivial compared with the first term.

In principle, we should be able to calculate the values of the extra terms in Propositions 3 and 4 respectively, but then we need a lot more data and have to go through much more nuanced computation in order to do that. This is discussed in more detail in subsection 3.3 below.

3.3 Estimating the extra terms due to variable markups in Propositions 3 and 4

In this subsection, we discuss the method and the data needed to compute the extra terms due to variable markup as stated in Propositions 3 and 4.

For Proposition 3

For simplicity, we assume that $U_j = \prod_z u_j(z)^{\alpha(z)}$ with $\sum_z \alpha(z) = 1$. Let $p_j(z)$ denote the exact price index for subutility $u_j(z)$ as in Online Appendix D. Then equations (28) (exact price index of $u_j(z)$), (29) (labor market clearing condition), (34) (free entry condition) and (30) (productivity effect) from Online Appendix D, together with $N_{ij}^z = N_i^z \left[1 - G_{iz}\left(\varphi_{ijz}^*\right)\right]$ which is true by definition, and the zero cutoff profit condition, represent a system of $3n^2Z + 2nZ + n$ equations with the same number of unknowns $\widehat{w}_i$, $\widehat{p}_j(z)$, $\widehat{\varphi_{ijz}^*}$, $\widehat{N_{ij}^z}$, $\widehat{\widetilde{\varphi}_{ijz}}$ and $\widehat{N_i^z}$ for all i, j, z. Under Pareto distribution, the system can be simplified to one with $n^2Z + 2nZ + n$ equations with the same number of unknowns $\widehat{w}_i$, $\widehat{p}_j(z)$, $\widehat{\widetilde{\varphi}_{ijz}}$ and $\widehat{N_i^z}$ for all i, j, z. Given bilateral trade data (X_{ij}^z) and the values of the parameters θ and σ_z (both can be obtained from the literature), we can solve for these $n^2Z + 2nZ + n$ unknowns given the changes in trade costs $\{\widehat{\tau_{ij}}\}$. Thus, $\widehat{\varphi_{ijz}^*}$ and $\widehat{N_{ij}^z}$ can be calculated as well. Therefore, the extra term in Proposition 3 can be computed. See Online Appendix F for the detail. Since all firms in the same sector have the same markup, only sectoral data are needed. No firm level data are needed.

For Proposition 4

Let the number of firms in the set Ω_{ij}^F be N_{ij}. Thus, there are $\sum_i N_{ij}$ firms serving market j. Log-linearizing (36) and (37) for each firm in Online Appendix E and the labor market clearing condition $w_i L_i = \sum_j X_{ij} = \sum_j \sum_{\omega \in \Omega_{ij}^F} s_{ij}(\varphi) w_j L_j$ for all i, we get a linear system with $2\sum_j \sum_i N_{ij} + n$ equations and $2\sum_j \sum_i N_{ij} + n$ unknowns, namely $\{\widehat{s_{ij}(\varphi)}\}$ and $\{\widehat{p_{ij}^F(\varphi)}\}$ for all firms $\omega \in \Omega_{ij}^F, \forall i, j$ and $\{\widehat{w_i}\}, \forall i$. The system can be solved when the market share $s_{ij}(\varphi)$, and the markup $\mu_{ij}(\varphi)$ for each firm before the changes in trade costs are known. After we solve for the system given the changes in trade costs $\{\widehat{\tau_{ij}}\}$, we can compute $\widehat{\mu_{ij}(\varphi)}$ as a function of $s_{ij}(\varphi)$, $\mu_{ij}(\varphi)$ and $\widehat{s_{ij}(\varphi)}$ based on an equation obtained by totally differentiating (7). Then the extra term in Proposition 4 can be computed. For the detail, refer to Online Appendix G. In this case, firm-level data are needed, since different firm-destination pairs have different markups.

In practice, it is almost impossible to obtain the market shares of all the firms in all destination markets in the world. But we can focus on the largest firms instead as a reasonable approximation. In recent literature of granularity, people found that all types of international trade activities concentrate mainly in a small number of largest multinational firms.[15] Thus we can focus on the largest multinationals, whose financial and market share information are relatively easy to obtain. For example, we can focus on the largest 1000 or 5000 firms in the world, using the Fortune 1000 or Global 5000 database. At the same time, focusing on a relative small number of largest firms can greatly reduce the dimension of the linear system of equations mentioned above. As a result, the computational complexity can be greatly reduced as well.

3.4 Empirical Applications

To illustrate the user-friendliness of our formula (1), we offer two empirical applications.

Empirical Application One. First, we demonstrate that we can easily calculate the elasticity of global welfare with respect to a uniform percentage reduction of all bilateral iceberg trade costs. This reduction can be due to technological improvements or other exogenous changes such as implementation of trade facilitation measures. Based on our theory, we only need to know the share of total trade value in world GDP in order to calculate the impact on global welfare. For example, in 2003, the data indicate that total value of world merchandise trade was approximately equal to 20.0% of world GDP. It follows from (1) that the elasticity of global welfare with respect to a uniform percentage reduction in all bilateral iceberg trade costs is equal to 0.200. In other words, 1% reduction in all bilateral iceberg trade costs would increase global welfare by 0.200% of world GDP in 2003.

How is the global welfare change estimated from our formula (1) compared with other estimates in the profession? The OECD (2003, p.4) estimates that, in 2003, assuming that trade facilitation leads to a reduction in border procedure-related trade transaction costs (TTCs) by 1 per cent of the value of world trade, global welfare gains would be about USD 40 billion. The analysis by OECD was carried out by using the GTAP database and model, which could account for changes in production, consumption, trade and economic welfare of countries. Let us compare the two estimates.

By definition, the bilateral TTCs (denoted by T_{ij}) between i and j as a fraction of the value of exports from i to j can be written as $T_{ij} = \tau_{ij} - 1 - t_{ij}$, where t_{ij} denotes other sources of bilateral trade costs (such as transport costs, tariffs, etc) as a fraction of the value of bilateral exports. Therefore, a reduction of bilateral TTCs from i to j by 1% of the value of exports from i to j means that $dT_{ij} = -0.01$, which implies that $d\tau_{ij} = -0.01$ when t_{ij} stays unchanged. Anderson and van Wincoop (2004) estimate that the tax equivalent total international trade cost is about 74%, which implies that $\tau_{ij} = 1.74$. With $d\tau_{ij} = -0.01$, we have $d\tau_{ij}/\tau_{ij} = -0.00575$. The value of world trade in 2003, $\sum_{i,j} X_{ij}$, is approximately USD 7.7 trillion. Thus, in 2003, a uniform reduction of T_{ij} by 1% of the value of exports from i to j for

[15] For example, Bernard, Jensen, Redding, and Schott (2016) found that "The largest decile of firms accounts for over 95 percent of total trade, exports and imports, and over 99 percent of related-party trade in 2007" for the U.S..

all i, j leads to an increase in global income of $\sum_{i,j} X_{ij}\widehat{\tau_{ij}}$, which is equal to USD 44.3 billion, according to our theory. Therefore, our estimate is not that different from that of OECD, though they have used a much more nuanced approach than ours.

Empirical Application Two. A notable component of shipping costs is shipping time. We want to find the cumulative global welfare impact of the reduction of international shipping time in the fifty-year period 1960-2010. Hummels and Schaur (2013) estimate that each additional day in transit is equivalent to an increase in [0.6%, 2.1%] of ad valorem trade cost. We need to estimate how much bilateral shipping time has been reduced each year over this period, and it is not easy to find data. But we can do a rough estimate. Based on Hummels' (2001) estimate that "the introduction of containerization in the late 1960s and 1970s results in a doubling of the average ocean fleet speed", we assume that average ocean shipping time was 40 days for international trade in 1960, compared to the 20 days in 2010 cited by him. In 1960, almost all international shipments were through ocean freight. We also assume that there was gradual reduction of ocean shipping time and a gradual substitution towards air freight since 1960. The average international ocean shipping time in 2010 was 20 days and air-shipped trade rose from (approximately) 0% to 50% from 1960 to 2010.[16] Thus, the average shipping days dropped from 40 to 0.5*20 (by sea) + 0.5*1 (by air) = 10.5 days during the period 1960-2010 (assuming that air freight takes just one day). If we assume the cost reduction process to be gradual (i.e. the annual reduction of shipping days is constant throughout the fifty years), the number of shipping days dropped by (40-10.5)/50 = 0.59 per year during the period 1960-2010, which is equivalent to a reduction of ad-valorem trade cost by [0.354%, 1.239%] per year. Based on the data of the share of merchandise trade in GDP from World Development Indicator (WDI) published by the World Bank for each of the years 1960-2010, we calculate from equation (1) the cumulative global welfare gains in the five decades 1960-2010 from the saving in shipping time to be [2.71%, 9.81%], which is equivalent to an increase in global income in the range USD [1709, 6183] billion in 2010, with the mid-point being USD 3946 billion. These estimates can be considered reasonable in the sense that they are consistent with the finding of, say, Ossa (2014), who finds that the average welfare gains of moving from autarky to free trade for Brazil (10%), China (13.1%), European Union (12.6%), India (11.2%), Japan (15.4%) and US (14.2%) is equal to about 11.0%. Our estimate should be interpreted as the average percentage welfare gains of all countries in the world from 1960 to 2010 due to reduction of trade costs. Since trade barriers were less restrictive than autarky in 1960, and they were more restrictive than free trade in 2010, our estimate should be less than the average percentage welfare gains from autarky to free trade as estimated by Ossa (2014). Since the range [2.71%, 9.81%] is less than 11%, we can say that our estimate is within reasonable bounds.

[16] These numbers are based on U.S. trade statistics. Given that the shipping industry has been very competitive, we think it is reasonable to assume that these numbers also apply to all other countries of the world. A sharp speeding up of ocean transport followed from the introduction of containerization in the late 1960s and 1970s. To simplify the calculation, we assume that the annual rate of reduction of trade cost is constant during this period.

4 Conclusion and Caveats

Our paper is motivated by the following question: How much does trade facilitation matter to the world? Guided by this question, based on a set of simple assumptions, we derive a simple equation to evaluate the quantitative impact on global welfare of small reduction of bilateral trade costs, such as shipping costs or the costs of administrative barriers to trade. Although our equation cannot evaluate the distribution of gains for different countries, it informs us of the magnitude of increase in global GDP. If global gains resulting from bilateral trade costs reduction are found to be large, then it would provide stronger support to the advocates of global trade facilitation such as WTO, OECD, and the World Bank.

Surprisingly, the equation is very general and is applicable to a broad class of models and settings. We also carry out some extensions by relaxing the assumption of constant markup. We illustrate the user-friendliness of the formula by carrying out a couple of simple empirical applications. We find the estimates obtained from the empirical applications to be reasonable and consistent with other estimates in the literature.

Our paper distinguishes from other works in the literature in a few key aspects. First, not only have we proved that only the direct effect matters, but we have also proved that the underlying mechanism driving the result is the envelope theorem. Thus, the formula is applicable to a broad variety of models and settings as long as there are no externalities or price distortions. This intuition has not been explained clearly in the literature. Second, we rigorously justify the use of the expenditure-share-weighted average percentage change of country welfare as a measure of the change of global welfare, based on the concept of equivalent variation. Third, we investigate the implications of non-constant markups on our benchmark result by carrying out a couple of extensions. In each case, we describe the additional data needed and the method of computing the extra term due to non-constant markups. The extensions further deepen our understanding of how to evaluate quantitatively the global gains from reduction of trade costs.

Appendix

A Specific Proof of Proposition 2: PC with multi-stage production

The setting, preferences, technology and market structure are as described in sections 2.1 and 2.2. This specific proof is a specific case (PC with multi-stage production) of the general proof of Proposition 2 presented in section 2.3. The multi-stage production model subsumes the single-stage production model. So, we do not provide a separate proof for the single-stage production case. Recall that the pre-change vector of market prices of final goods sold in country j is $\mathbf{p_j^F} \equiv \left\{ p_{ij}^F(\omega) \middle| i \in \mathcal{N},\, \omega \in \Omega_{ij}^F \right\}$ and pre-change vector of market quantity of final goods consumed by country j is $\mathbf{q_j^F} \equiv \left\{ q_{ij}^F(\omega) \middle| i \in \mathcal{N},\, \omega \in \Omega_{ij}^F \right\}$. The sum of equivalent variations of all countries is equal to $\sum_j \mathbf{p_j^F} \cdot d\mathbf{q_j^F} = \sum_{i,j} \int_{\omega\Omega_{ij}^F} p_{ij}^F(\omega)\, dq_{ij}^F(\omega)\, d\omega$.

<u>**Maximization of global income**</u>

The global planner maximizes global income W by choosing $l_{jk}^s, \forall j,k,s$, and Ω_{jk}^s $\forall j,k,s$ taking trade costs τ_{ij}^s for all i,j,s and the shadow prices of final goods p_{ij}^F for all i,j as given:

$$\max_{\left\{l_{jk}^s, \forall j,k,s\right\}\left\{\Omega_{jk}^s, \forall j,k,s\right\}} W = \sum_{i,j} \int_{\omega\Omega_{ij}^F} \frac{p_{ij}^F(\omega)\, y_{ij}^F(\omega)}{\tau_{ij}^F} d\omega$$

s.t. (i) labor constraint $L_j = \sum_{k,s} \int_{\omega\in\Omega_{jk}^s} l_{jk}^s(\omega)\, d\omega$ for all j , (ii) the production function of stage-s good ω exported from j to k, given by

$$y_{jk}^s(\omega) = \begin{cases} \varphi_j^s(\omega)\, f\left(\left\{ \frac{y_{ij,k}^{s-1}(\omega')}{\tau_{ij}^{s-1}} \middle| i \in \mathcal{N},\, \omega' \in \Omega_{ij}^{s-1} \right\},\; l_{jk}^s(\omega)\right) & \text{for } s = 2,3,...,F \\ \varphi_j^s(\omega)\, l_{jk}^s(\omega) & \text{for } s = 1 \end{cases} \quad \text{for } \omega \in \Omega_{jk}^s \tag{8}$$

for all $j,k = 1,...,n$, where $l_{jk}^s(\omega)$ is the quantity of labor used in j in combination with the quantities of inputs imported from i to j, $y_{ij,k}^{s-1}(\omega')/\tau_{ij}^{s-1} \equiv q_{ij,k}^{s-1}(\omega')$, to produce stage-s output ω to be exported from j to k, $y_{jk}^s(\omega)$, and φ_j^s is the productivity, which is constant. Thus, $\sum_k \int_{\omega\in\Omega_{jk}^s} l_{jk}^s(\omega)\, d\omega = l_j^s$ for all j and s, and $\sum_s l_i^s = L_i$ for all i.

Suppose there is a change in τ_{ki}^s such that $\widehat{\tau_{ki}^s} < 0$. We want to evaluate the effect of this change on W. We start from a state where all the $l_{ij}^s(\omega)$ for all i,j,s,ω are optimally chosen given the exogenous variables. According to (4), we only need to evaluate $\partial W/\partial\tau_{ki}^s$, i.e. evaluate the effect of τ_{ki}^s on W keeping $\mathbf{l}$ and $\mathbf{p}$ unchanged, recalling that $\mathbf{l} \equiv \left(l_{ij}^s(\omega)\right), \forall i,j,s,\omega$ and $\mathbf{p} \equiv (p_{ij}^s(\omega)), \forall i,j,s,\omega$. Recall that $p_{ij}^s(\omega) = p_{ii}^s(\omega)\,\tau_{ij}^s$. Since the unit labor requirement for any $p_{ij}^1(\omega)$ at stage 1 is constant, w_i is unchanged as all $p_{ij}^1(\omega)$ are kept unchanged. We proof our result step by step below.

1. The total global value of inputs at stage $s+1$ is equal to

$$\sum_i \left(\sum_k X_{ki}^s + w_i l_i^{s+1} \right) = \sum_{i,k} X_{ki}^s + \sum_i w_i l_i^{s+1}$$

where

$$X_{ki}^s = \frac{\int_{\omega \in \Omega_{ki}^s} p_{ki}^s(\omega)\, y_{ki}^s(\omega)\, d\omega}{\tau_{ki}^s} \text{ for all } s = 1, 2, ..., F.$$

is the value of stage-s output exported from k to i which is used as stage-$(s+1)$ input.

As τ_{ki}^s varies while keeping $p_{ki}^s(\omega)$ and $l_{ki}^s(\omega)$ unchanged for all s, k, i ω, $y_{ki}^s(\omega)$ is also unchanged. Therefore, the effect of τ_{ki}^s on X_{ki}^s is given by

$$\underbrace{\frac{\partial X_{ki}^s}{\partial \tau_{ki}^s} d\tau_{ki}^s}_{\substack{\text{Change in the value of stage-s} \\ \text{input exported from k to i}}} = -X_{ki}^s \widehat{\tau_{ki}^s} \tag{9}$$

which is the direct saving in trade costs.

2. The value of stage-(s+1) output exported from i to j is given by

$$X_{ij}^{s+1} = \int_{\omega \in \Omega_{ij}^{s+1}} p_{ii}^{s+1}(\omega)\, y_{ij}^{s+1}(\omega)\, d\omega.$$

At stage $s+1$, the total global value of outputs is equal to the total global value of inputs:

$$\underbrace{\sum_{i,j} X_{ij}^{s+1}}_{\text{global value of outputs at stage } s+1} = \underbrace{\sum_{i,j} X_{ij}^s}_{\text{global value of inputs at stage } s+1} + \underbrace{\sum_i w_i l_i^{s+1}}_{\text{global value-added at stage } s+1}. \tag{10}$$

As τ_{ki}^s is reduced, it increases the global value of inputs at stage $s+1$, $\sum_{i,j} X_{ij}^s$, through the direct saving in trade costs, according to (8) for stage s+1 production. This in turn increases the global value of outputs at stage $s+1$, $\sum_{i,j} X_{ij}^{s+1}$, through the increases in $\left\{ y_{ij}^{s+1}(\omega) \right\}$ according to (8). At stage s+2, we have

$$\sum_{i,j} X_{ij}^{s+2} = \sum_{i,j} X_{ij}^{s+1} + \sum_i w_i l_i^{s+2}.$$

That is, increases in $\left\{ y_{ij}^{s+1}(\omega) \right\}$ in turn lead to increases in $\left\{ y_{ij}^{s+2}(\omega) \right\}$ according to (8) for stage s+2 production, which raises the global value of outputs at stage s+2, $\sum_{i,j} X_{ij}^{s+2}$. This process goes on until the final stage F, leading to an increase in $\sum_{i,j} X_{ij}^F$ through increases in $\left\{ y_{ij}^F(\omega) \right\}$. Thus, we have

$$\underbrace{\sum_{i,j} X_{ij}^F}_{\text{global income}} = \underbrace{\sum_{i,j} X_{ij}^s}_{\text{global value of inputs at stage s+1}} + \underbrace{\sum_i w_i \sum_{m=s+1}^{F} l_i^m}_{\text{cumulative global value-added from stages } s+1 \text{ to } F}.$$

Since w_i $\forall i$ and l_i^m $\forall i$, m are kept unchanged as τ_{ki}^s varies, $\sum_i w_i \sum_{m=s+1}^F l_i^m$ is unchanged as well. As $W = \sum_{i,j} X_{ij}^F$, we have

$$\frac{\partial W}{\partial \tau_{ki}^s} d\tau_{ki}^s = \frac{\partial}{\partial \tau_{ki}^s}\left(\sum_{i,j} X_{ij}^F\right) d\tau_{ki}^s = \frac{\partial}{\partial \tau_{ki}^s}\left(\sum_{i,j} X_{ij}^s\right) d\tau_{ki}^s = \frac{\partial X_{ki}^s}{\partial \tau_{ki}^s} d\tau_{ki}^s = -X_{ki}^s \widehat{\tau_{ki}^s}.$$

The economic intuition is as follows. The initial welfare impact of a small reduction of trade cost of $d\tau_{ij}^s$ at stage s is equal to a gain of $X_{ij}^s \widehat{\tau_{ij}^s}$ at stage $s+1$ received by country j due to the reduction in the cost of its intermediate good, where X_{ij}^s denote the value of exports of stage-s good from country i to country j. But since country j's stage-$s+1$ good is subsequently used by all other countries for their production of goods at stage-$s+2$, $s+3$ and so on, the cost saving is passed on fully to all countries in each later stage. This process will go on until the final stage. Eventually, the saving in trade cost shows up as the gains in global income received by all countries in the final stage F. Therefore, the global welfare effect of a change in trade cost τ_{ij}^s at stage s is equal to $-X_{ij}^s \widehat{\tau_{ij}^s}$. Thus, the percentage change in global welfare is equal to

$$\underbrace{-\sum_{i,j,s} \frac{X_{ij}^s \widehat{\tau_{ij}^s}}{Y^w} = -\sum_{i,j} \frac{X_{ij}}{Y^w} \widehat{\tau_{ij}}}_{\text{If } \widehat{\tau_{ij}^s} = \widehat{\tau_{ij}} \text{ for all } s, \text{ and } X_{ij} = \sum_s X_{ij}^s},$$

which is expression (1).■

B Specific Proof of Proposition 2: M-g

The setting, preferences, technology and market structure are as described in sections 2.1 and 2.2. This specific proof is a specific case of the general proof of Proposition 2 presented in section 2.3. Recall that the labor productivity $\varphi \equiv 1/a_i(\omega)$. So, each variety is indexed by φ or ω, with a unique mapping between the two. The two indexes are used interchangeably. Analogous to the general proof in section 2.3, we define $\mathbf{p_j^F} = \left\{p_{ij}^F(\varphi), \forall i, \varphi\right\}$ and $\mathbf{q_j^F} = \left\{q_{ij}^F(\varphi), \forall i, \varphi\right\}$. The sum of equivalent variations of all countries is equal to $\sum_j \mathbf{p_j^F} \cdot d\mathbf{q_j^F} = \left[\sum_{j,i} N_i \int_{\varphi_{ij}^*}^{\infty} p_{ij}^F(\varphi)\, dq_{ij}^F(\varphi)\, g_i(\varphi)\, d\varphi\right]$.

Maximization of global income

The global planner maximizes global income W by choosing $\{l_{ij}(\varphi), \forall i, j, \varphi\}, \left\{\varphi_{ij}^*\right\}$ and$\{N_i\}$, taking trade costs τ_{ij} for all i, j and the shadow (market) prices $p_{ij}^F(\varphi)$ for all i, j, φ as given. Thus, she solves

$$\max_{\{l_{ij}(\varphi), \forall i,j,\varphi\}, \left\{\varphi_{ij}^*\right\}, \{N_i\}} \quad W = \sum_{i,j} N_i \int_{\varphi_{ij}^*}^{\infty} \frac{p_{ij}^F(\varphi)\, f^{\varphi}(l_{ij}(\varphi))\, g_i(\varphi)}{\tau_{ij}} d\varphi$$

$$\text{s.t. } N_i \left\{ f_e + \sum_j \xi_{ij} \left[1 - G_i\left(\varphi_{ij}^*\right)\right] + \sum_j \int_{\varphi_{ij}^*}^{\infty} l_{ij}(\varphi)\, g_i(\varphi)\, d\varphi \right\} = L_i \quad \text{for all } i$$

where $y_{ij}^F(\varphi) = f^{\varphi}(l_{ij}(\varphi))$ is the production function for producing $y_{ij}^F(\varphi) = \tau_{ij} q_{ij}^F(\varphi)$ units of good by a firm with productivity φ in country i for sales to country j.[17]

Analogous to the general proof in section 2.3, we define $\mathbf{l} \equiv (l_{ij}(\varphi)), \forall i, j, \varphi$ as a vector of all labor allocation; $\mathbf{p} \equiv (p_{ij}^F(\varphi)), \forall i, j, \varphi$ as a vector of all shadow prices; and $\boldsymbol{\tau} \equiv (\tau_{ij}), \forall i, j$ as a vector of all trade costs.[18]

Then,

$$\frac{\partial W}{\partial \tau_{ij}} = -N_i \int_{\varphi_{ij}^*}^{\infty} \frac{p_{ij}^F(\varphi) f^{\varphi}(l_{ij}(\varphi)) g_i(\varphi)}{\tau_{ij}^2} d\varphi = -\frac{X_{ij}}{\tau_{ij}}$$

as $N_i \int_{\varphi_{ij}^*}^{\infty} \frac{p_{ij}^F(\varphi) f^{\varphi}(l_{ij}(\varphi)) g_i(\varphi)}{\tau_{ij}} d\varphi = X_{ij}$. As $\{l_{ij}(\varphi), \forall i, j, \varphi\}$, $\left\{\varphi_{ij}^*\right\}$ and$\{N_i\}$ have been optimally chosen, their effects on W are second order. Thus, we can analogously invoke (4) to compute the percentage change in global welfare as

$$\frac{1}{Y^w} \sum_{i,j} \frac{\partial W}{\partial \tau_{ij}} d\tau_{ij} = -\sum_{i,j} \frac{X_{ij} \widehat{\tau_{ij}}}{Y^w}$$

which is expression (1).■

C Analysis of market response to reduction of trade costs

From now on, in order to simplify notation, we shall omit the superscript "F" for most of the variables in the context of discussion of single-stage production (PC or MC) whenever such omission would not cause any confusion.

C.1 PC with multi-stage production

The utility function and the production at each stage are as given in section 2.2. Let $E_j^s \equiv \sum_i X_{ij}^s$ denote the total expenditure on stage-s good in country j. (For the final stage, $E_j^F \equiv E_j$ is country j's total expenditure on final goods). Define $\theta_s^j \equiv \sum_i X_{ij}^{s-1} / \sum_k X_{jk}^s$ the cost share of intermediate goods in the total cost of stage-s output produced by country j. Note that 1. $X_{ij}^s = \int_{\omega \in \Omega_{ij}^s} p_{ij}^s(\omega) q_{ij}^s(\omega) d\omega$, where $p_{ij}^s(\omega)$ is the import price in country j of stage-s good ω from country i; 2. the total value of production of stage-s good in country j is equal to $\sum_i X_{ji}^s = \int_{\omega \in \Omega_j^s} p_{jj}^s(\omega) y_j^s(\omega) d\omega$ for $s = 1, 2, ..., F$, where $p_{jj}^s(\omega)$ is the unit cost of stage-s good ω produced by country j; 3. $p_{ij}^s(\omega) = p_{ii}^s(\omega) \tau_{ij}^s$; 4. $q_{ij}^s(\omega) \tau_{ij}^s = y_{ij}^s(\omega)$ which is the quantity of stage-s good ω exported from i to j measured at the origin; 5. $y_j^s(\omega) \equiv \sum_i y_{ji}^s(\omega)$.

[17] Here we assume that the production function is the same in all countries. We could assume that the production function is different across countries, but the conclusion will remain the same.

[18] We abuse the notation a little here as the vector $\mathbf{l}$ and $\mathbf{p}$ includes infinite elements (the upper bound for φ is ∞).

In addition, the total value of production of stage-s good (for $s = 2, 3, ..., F$) in country j, given by $\sum_i X_{ji}^s$, can be decomposed into two parts: first, the value-added by country j labor in the production of stage-s good, given by $V_j^s = \left(1 - \theta_s^j\right) \sum_i X_{ji}^s$; second, the total expenditure on intermediate inputs for the production of stage-s good in country j, given by $E_j^{s-1} = \theta_s^j \sum_i X_{ji}^s$. Moreover, $V_j^1 = \sum_i X_{ji}^1$ as no intermediate input is required for the first stage. GDP of j, Y_j, is equal to the sum of value-added over all stages:

$$Y_j = \sum_s V_j^s \text{ for } j = 1, 2, ..., n$$

The percentage change in welfare of country j as τ_{ij} changes is given by

$$\widehat{U}_j = \widehat{E}_j - \widehat{P}_j.$$

Intuitively, this says that, when j exports a good, an increase in w_j tends to raise U_j through the increase in E_j resulting from the increases in the prices of its exports, i.e. the terms of trade (TOT) effect on j as an exporter. On the other hand, when j imports a good from a foreign country i, the prices of its imports are affected in two ways. First, an increase in foreign wage w_i tends to reduce U_j through the increase in P_j resulting from the TOT effect on j as an importer. Second, P_j is further affected by the change in trade cost τ_{ij}. Thus, the expenditure-weighted change in welfare of country j is given by

$$E_j \widehat{U}_j = \underbrace{E_j \widehat{E}_j}_{\substack{\text{TOT effect on j when} \\ \text{it is an exporter}}} - \underbrace{E_j \widehat{P}_j}_{\substack{\text{Welfare effect on j when} \\ \text{it is an importer}}}.$$

Fixed level of trade balance implies that $dE_j = dY_j = L_j dw_j = w_j L_j \widehat{w_j} = Y_j \widehat{w_j} = \sum_s V_j^s \widehat{w}_j$, as $Y_j = \sum_s V_j^s$. Thus, the TOT effect on j when it is an exporter is given by:

$$E_j \widehat{E}_j = \sum_s V_j^s \widehat{w}_j$$

Thus, the global welfare effect resulting from changes in prices of exports is given by:

$$\sum_j E_j \widehat{E}_j = \underbrace{\sum_j \sum_s V_j^s \widehat{w}_j}_{\text{TOT effect on exporters over all stages}} \tag{11}$$

whereas the global welfare effect resulting from changes in prices of imports is given by

$$\sum_j E_j \widehat{P}_j = \underbrace{\sum_{s,j,i} X_{ij}^s \widehat{\tau}_{ij}^s}_{\text{Direct effect on trade in goods over all stages}} + \underbrace{\sum_{j,s} V_j^s \widehat{w}_j}_{\text{TOT effect on importers over all stages}} \tag{12}$$

This is the TOT effect on all the importers, plus the direct effect of changes in trade costs borne by the importers. The derivation of (12) is given in Online Appendix B.[19]

The global welfare effect is the difference between the global effect resulting from changes in prices of exports and the global effect resulting from changes in prices of imports. From (11) and (12), we conclude that the global welfare effect is

$$\sum_j E_j \widehat{U}_j = \underbrace{-\sum_{s,j,i} X_{ij}^s \widehat{\tau}_{ij}^s}_{\text{Direct effect on trade in goods over all stages}}$$

At each stage, the TOT effect on exporters exactly offsets the TOT effect on importers. This is no surprise, as for each amount of TOT gain by an exporter, there is an equal amount of TOT loss by the importer. Thus, at each stage, the only welfare impact from the global point of view is the direct effect. If $\widehat{\tau_{ij}^s} = \widehat{\tau_{ij}}$ for all s, we have: $\sum_j \frac{E_j}{Y^w} \widehat{U}_j = -\sum_{s,i,j} \frac{X_{ij}^s}{Y^w} \widehat{\tau_{ij}^s} = -\sum_{i,j} \frac{X_{ij}}{Y^w} \widehat{\tau_{ij}}$, (where $X_{ij} = \sum_s X_{ij}^s$), which is expression (1). Here we see that we need to take into account the indirect effect, namely the net TOT gains of the country, when calculating individual country's gains, but not when calculating global gains.

C.2 MC with heterogeneous firm productivity (M-g)

According to Melitz (2003), the average productivity of a firm in country i serving market j is given by $\widetilde{\varphi}_{ij} \equiv \left[\frac{1}{1-G_i\left(\varphi_{ij}^*\right)} \int_{\varphi_{ij}^*}^{\infty} (\varphi)^{\sigma-1} g_i(\varphi)\, d\varphi\right]^{\frac{1}{\sigma-1}}$. Thus $\widetilde{\varphi}_{ij}$ is a function of φ_{ij}^*. The number of firms in country i serving market j, $N_{ij} = N_i \left[1 - G_i\left(\varphi_{ij}^*\right)\right]$, is a function of φ_{ij}^* and N_i. The values of $\widetilde{\varphi}_{ij}$ and N_{ij}, in turn, directly affect P_j, as shown below. Constant markup implies that the average price of a good sold in j imported from i is given by $\left(\frac{\sigma}{\sigma-1}\right)\frac{w_i \tau_{ij}}{\widetilde{\varphi}_{ij}}$. Therefore, the expected price index in country j is given by

$$P_j = \left\{\sum_i N_{ij}\left[\left(\frac{\sigma}{\sigma-1}\right)\frac{w_i \tau_{ij}}{\widetilde{\varphi}_{ij}}\right]^{1-\sigma}\right\}^{\frac{1}{1-\sigma}}$$

Totally differentiating the logarithm of the above equation and re-arranging yield:

$$E_j \widehat{P}_j = \sum_i X_{ij}\left(\widehat{w}_i + \widehat{\tau}_{ij} - \frac{1}{\sigma-1}\widehat{N}_{ij} - \widehat{\widetilde{\varphi}_{ij}}\right) \tag{13}$$

We call $\widehat{\widetilde{\varphi}_{ij}}$ the "productivity effect" — the effect of the change in cutoff productivity φ_{ij}^* on the average productivity $\widetilde{\varphi}_{ij}$. On the other hand, we call $\widehat{N}_{ij}$ the "firm mass effect" — the effect of the

[19] When there is single-stage production, (11) and (12) become $\sum_j E_j \widehat{E}_j = \sum_j \sum_i X_{ji} \widehat{w}_j$ and $\sum_j E_j \widehat{P}_j = \underbrace{\sum_j \sum_i X_{ij} \widehat{\tau}_{ij}}_{\text{Direct effect of trade costs}} + \underbrace{\sum_j \sum_i X_{ij} \widehat{w}_i}_{\text{TOT effect on importers}}$ respectively.

change in cutoff productivity φ_{ij}^* on firm mass N_{ij}. It turns out that, given any exporting country i, the two effects are related in the following way when summing over all the importing countries:

$$\sum_j \left(-\frac{X_{ij}\widehat{N}_{ij}}{\sigma - 1} \right) = \sum_j \left(X_{ij}\widehat{\widetilde{\varphi}_{ij}} \right) \qquad \text{i.e. firm mass effect plus productivity effect} = 0. \tag{14}$$

Please refer to Online Appendix C for detailed derivation of (14).The intuition of the equation is: A change in τ_{ij} for any i, j will have effect on φ_{ij}^* through its effect on wages. For any given i, from country j's welfare point of view, an increase in φ_{ij}^* leads to an increase in $\widetilde{\varphi}_{ij}$ (average productivity of each firm from i serving j is higher — RHS of the above equation without the summation over j) but a decrease in N_{ij} (fewer firms from i serving the market in j — LHS of the above equation without the summation over j), leading to counteracting (but not completely offsetting) effects on $E_j\widehat{P}_j$, according to (13). When summing over all j, the two effects offset each other completely from a global welfare point of view.[20] Thus, the productivity effect and firm mass effect completely offset each other from the global welfare point of view, when summing over all i and j.

Summing up (13) for all possible destination j, global welfare change is given by

$$\begin{aligned}
\sum_j E_j\widehat{U}_j &= \sum_j E_j \left(\widehat{E}_j - \widehat{P}_j \right) \\
&= \underbrace{\overbrace{\sum_{j,i} X_{ji}\widehat{w}_j}^{E_j\widehat{E}_j} - \sum_{j,i} X_{ij}\widehat{w}_i}_{\text{TOT effect on all importers and exporters} = 0} \quad \underbrace{-\sum_{i,j} X_{ij}\widehat{\tau}_{ij}}_{\text{Direct effect}} + \underbrace{\sum_{j,i} \left(\frac{X_{ij}\widehat{N}_{ij}}{\sigma - 1} + X_{ij}\widehat{\widetilde{\varphi}_{ij}} \right)}_{\text{Global firm mass effect plus productivity effect} = 0} \\
\implies \quad & \sum_j \frac{E_j}{Y^w}\widehat{U}_j = -\sum_{i,j} \frac{X_{ij}}{Y^w}\widehat{\tau}_{ij}
\end{aligned}$$

which is expression (1). Note that the same equation holds under K1980 simply because both $\widehat{N}_{ij}$ and $\widehat{\widetilde{\varphi}_{ij}}$ are equal to zero for any i and j. Again, here we see that we need to take into account the indirect effect, namely the sum of the TOT effect, productivity effect and the firm mass effect on the individual country when calculating its gains, but not when calculating global gains.

[20] For example, in the symmetric two-country case (with countries 1 and 2) in Melitz (2003), a reduction of $\tau = \tau_{12} = \tau_{21}$ raises the wage in each country. Suppose we focus on $i = 1$. This raises the domestic productivity cutoff φ_{11}^* (thus raising $\widetilde{\varphi}_{11}$ and lowering N_{11}) but lowers exporting productivity cutoff φ_{12}^* (thus lowering $\widetilde{\varphi}_{12}$ and raising N_{12}) . Moreover, $-\frac{X_{11}\widehat{N}_{11}}{\sigma-1} - \frac{X_{12}\widehat{N}_{12}}{\sigma-1} = X_{11}\widehat{\widetilde{\varphi}_{11}} + X_{12}\widehat{\widetilde{\varphi}_{12}}$.

D Mapping between our expression (1) and ACR's equation (1)

Assumption R3 in ACR is equivalent to $\widehat{\lambda}_{ij} - \widehat{\lambda}_{jj} = d\ln\lambda_{ij} - d\ln\lambda_{jj} = d\ln X_{ij} - d\ln X_{jj} = \varepsilon d\ln\tau_{ij} = \varepsilon\widehat{\tau}_{ij}$. Thus

$$\begin{aligned}
-\sum_{j,i}\frac{X_{ij}}{Y^w}\widehat{\tau}_{ij} &= -\sum_{j,i}\frac{X_{ij}}{Y^w}\frac{1}{\varepsilon}\left(\widehat{\lambda}_{ij} - \widehat{\lambda}_{jj}\right) \\
&= -\sum_{j}\frac{E_j}{Y^w}\sum_{i}\frac{X_{ij}}{E_j}\frac{1}{\varepsilon}\left(\widehat{\lambda}_{ij} - \widehat{\lambda}_{jj}\right) \\
&= -\sum_{j}\frac{E_j}{Y^w}\sum_{i}\frac{\lambda_{ij}}{\varepsilon}\left(\widehat{\lambda}_{ij} - \widehat{\lambda}_{jj}\right) \\
&= \sum_{j}\frac{E_j}{Y^w}\frac{1}{\varepsilon}\widehat{\lambda}_{jj}
\end{aligned}$$

as $\sum_i \lambda_{ij} = 1 \Rightarrow \sum_i d\lambda_{ij} = 0$. In short, if one adopts Assumption R3 in ACR, then the global gains formula is precisely equal to an expenditure-share-weighted average percentage change of individual country's gains from trade (as given by the ACR formula) where the weights are the respective country's shares in world expenditure.

References

[1] Anderson, James E. and Eric Van Wincoop. 2004. "Trade Costs." *Journal of Economic Literature* 42(3), 691-751.

[2] Arkolakis, Costas, Arnaud Costinot and Andres Rodriguez-Clare. 2012. "New Trade Models, Same Old Gains?" *American Economic Review* 102(1), 94-130.

[3] Armington, Paul. 1969. "A Theory of Demand for Products Distinguished by Place of Production." *International Monetary Fund Staff Papers*, XVI (1969), 159-178.

[4] Atkeson, Andrew and Ariel Burstein. 2010. "Innovation, Firm Dynamics, and International Trade." *Journal of Political Economy* 118(3), 433-484.

[5] Burstein, Ariel and Javier Cravino. 2015. "Measured Aggregate Gains from International Trade," *American Economic Journal: Macroeconomics*, 7(2): 181-218.

[6] Dhingra, Swati and John Morrow. 2016. "Monopolistic Competition and Optimum Product Diversity Under Firm Heterogeneity." manuscript, London School of Economics, forthcoming, *Journal of Political Economy.*

[7] Dixit, Avinash and Victor Norman. 1980. *Theory of International Trade*. Cambridge University Press.

[8] Dornbusch, Rudiger, Stanley Fisher and Paul A. Samuelson. 1977. "Comparative Advantage, Trade, and Payments in Ricardian model with a Continuum of Goods." *American Economic Review* 67(5), 823–839.

[9] Dornbusch, Rudiger, Stanley Fisher and Paul A. Samuelson. 1980. "Heckscher-Ohlin Trade Theory with a Continuum of Goods." *The Quarterly Journal of Economics* 95(2), 203-224.

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

[11] Edmond, Chris, Virgiliu Midrigan and Daniel Yi Xu. 2015. "Competition, Markups, and the Gains from International Trade." *American Economic Review* 105(10), 3183-3221.

[12] Fan, Haichao, Edwin L.-C. Lai and Han Steffan Qi. 2012. "Global Gains from Trade Liberalization." *CESifo Working Paper* No. 3775 (March 2012).

[13] Feenstra, Robert C. 2004. *Advanced International Trade*. Princeton University Press: Princeton, New Jersey.

[14] Feldman, Allan M. 1998. "Kaldor-Hicks Compensation." in *The New Palgrave Dictionary of Economics and the Law*, Peter Newman (ed.), Palgrave Macmillan.

[15] Hsieh, Chang-Tai and Ralph Ossa. 2016. "A Global View of Productivity Growth in China." *Journal of International Economics* 102, 209-224.

[16] Hummels, David L. 2001. "Time as a Trade Barrier." Unpublished working paper, Purdue University.

[17] Hummels, David L. and Schaur Georg. 2013. "Time as a Trade Barrier." *American Economic Review* 103(7), 2935-2959.

[18] Krugman, Paul. 1980. "Scale Economies, Product Differentiation, and the Pattern of Trade." *American Economic Review* 70(5), 950-959.

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

[20] Melitz, Marc J. and Gianmarco Ottaviano. 2008. "Market Size, Trade, and Productivity." *Review of Economic Studies*, 75, 295-316.

[21] Melitz, Marc J. and Stephen J. Redding. 2014. "Missing Gains from Trade?", *American Economic Review Papers and Proceedings* 104(5), 317-321.

[22] Melitz, Marc J. and Stephen J. Redding. 2015. "New Trade Models, New Welfare Implications." *American Economic Review*, 105(3), 1105-1146.

[23] OECD, Working Party of the Trade Committee, "Quantitative Assessment of the Benefits of Trade Facilitation," TD/TC/WP(2003)31 (19 September 2003).

[24] Ossa, Ralph. 2012. "Profits in the 'New Trade' Approach to Trade Negotiations." *American Economic Review, Papers and Proceedings*, 102(3), May, 466-469.

[25] Ossa, Ralph. 2015. "Why Trade Matters After All?" *Journal of International Economics* 97(2), 266–277.

[26] Varian, Hal. 1992. *Microeconomic Analysis*. Third Edition, W.W. Norton: New York.

[27] World Economic Forum, "*Enabling Trade, Valuing Opportunities,*" 2013. http://www3.weforum.org/docs/WEF_SCT_EnablingTrade_Report_2013.pdf

[28] Yi, Kei-Mu. 2003. "Can Vertical Specialization Explain the Growth of World Trade?" *Journal of Political Economy* 111(1), 52-102.

[29] Yi, Kei-Mu. 2010. "Can Multistage Production Explain the Home Bias in Trade?" *American Economic Review* 100(1), 364-393.

Online Appendix (not to be published)

A Proof that the global market outcome is efficient under M-g

The Market Outcome

In order to satisfy Assumptions 1-3, we need price to be constant markup over marginal cost. Therefore, more structure has to be imposed on preferences. Specifically, we have to assume that the utility in country j is given by

$$U_j = \left[\sum_i \int_{\omega \in \Omega_{ij}^F} q_{ij}^F(\omega)^{\frac{\sigma-1}{\sigma}} d\omega \right]^{\frac{\sigma}{\sigma-1}} \tag{15}$$

where $\sigma > 1$ is the elasticity of substitution. Profit maximization by firms and utility maximization by consumers subject to the budget constraint implies that the demand for each variety is given by

$$q_{ij}^F(\varphi) = \left[\left(\frac{\sigma w_i}{\sigma - 1} \right) \frac{\tau_{ij}}{\varphi P_j} \right]^{-\sigma} \frac{E_j}{P_j} \tag{16}$$

where E_j denotes total expenditure in country j, and P_j denotes the exact price index in country j. Correspondingly, the labor used to produced $q_{ij}^F(\varphi)$ is given by $l_{ij}(\varphi) = \frac{q_{ij}^F(\varphi)\tau_{ij}}{\varphi}$.

The Global Planner's Problem

The global planner solves two problems simultaneously. First, she chooses $q_{ij}^F(\varphi)$ through the choice of labor allocation $l_{ij}(\varphi)$ to maximize the global GDP Y^w. Second, she chooses $q_{ij}^F(\varphi)$ to maximize the utility of each individual country j, U_j. In other words, she makes the production decision as well as the consumption decision. The shadow price $p_{ij}^F(\varphi)$ equalizes the quantity supplied and quantity demanded of good $q_{ij}^F(\varphi)$, for $\forall i, j, \varphi$.

Maximization of utility of each country

She chooses $q_{ij}^F(\varphi)$ to maximize U_j given N_i, φ_{ij}^*, E_j and $p_{ij}^F(\varphi)$. Thus, she solves

$$\underset{\{q_{ij}^F(\varphi)\}}{\text{Max}} \; U_j = \left[\sum_i N_i \int_{\varphi_{ij}^*}^{\infty} q_{ij}^F(\varphi)^{\frac{\sigma-1}{\sigma}} g_i(\varphi)\, d\varphi \right]^{\frac{\sigma}{\sigma-1}} \tag{17}$$

$$s.t. \; \sum_i N_i \int_{\varphi_{ij}^*}^{\infty} p_{ij}^F(\varphi) q_{ij}^F(\varphi) g_i(\varphi)\, d\varphi = E_j \tag{18}$$

Let λ_j be the Lagrange multiplier. Then, the Lagrangian function is given by

$$\mathcal{L}_j = U_j + \lambda_j \left[E_j - \sum_i N_i \int_{\varphi_{ij}^*}^{\infty} p_{ij}^F(\varphi) q_{ij}^F(\varphi) g_i(\varphi)\, d\varphi \right]$$

Substituting the F.O.C. w.r.t. $q_{ij}^F(\varphi)$ into budget constraint (18) yields

$$\lambda_j = \frac{1}{P_j}$$

where $P_j = \left\{\sum_i N_i \int_{\varphi_{ij}^*}^{\infty} \left[p_{ij}^F(\varphi)\right]^{1-\sigma} g_i(\varphi)\, d\varphi\right\}^{\frac{1}{1-\sigma}}$ is the exact price index in country j. Note that $U_j = E_j/P_j$. Substituting the above relationship back into the F.O.C. w.r.t. $q_{ij}^F(\varphi)$ we have

$$q_{ij}^F(\varphi) = \left[\frac{p_{ij}^F(\varphi)}{P_j}\right]^{-\sigma} U_j = \left[\frac{p_{ij}^F(\varphi)}{P_j}\right]^{-\sigma} \frac{E_j}{P_j} \tag{19}$$

This is a demand function, and it will be used in the global income maximization problem.

Maximization of global income

She chooses N_i, φ_{ij}^* and labor allocation $l_{ij}(\varphi)$ to maximize the global GDP Y^w, given τ_{ij}, L_i and $p_{ij}^F(\varphi)$, for $\forall i, j, \varphi$. The variables $q_{ij}^F(\varphi)$ and E_j are implicitly determined from the choices of N_i, φ_{ij}^* and $l_{ij}(\varphi)$. Since $q_{ij}^F(\varphi) = \frac{\varphi l_{ij}(\varphi)}{\tau_{ij}}$, the choice of $l_{ij}(\varphi)$ is the same as the choice of $q_{ij}^F(\varphi)$ when τ_{ij} and φ are given. Thus, her problem is stated as:

$$\max_{\left\{q_{ij}^F(\varphi),\, N_i,\, \varphi_{ij}^*\right\}} Y^w = \sum_{i,j} N_i \int_{\varphi_{ij}^*}^{\infty} p_{ij}^F(\varphi) q_{ij}^F(\varphi) g_i(\varphi)\, d\varphi \tag{20}$$

$$\text{s.t. } N_i \left\{ f_e + \sum_j \xi_{ij}\left[1 - G_i\left(\varphi_{ij}^*\right)\right] + \sum_j \int_{\varphi_{ij}^*}^{\infty} \frac{\tau_{ij}}{\varphi} q_{ij}^F(\varphi) g_i(\varphi)\, d\varphi \right\} = L_i \text{ for all } i \tag{21}$$

$$\text{and } q_{ij}^F(\varphi) = \left(\frac{p_{ij}^F(\varphi)}{P_j}\right)^{-\sigma} \frac{E_j}{P_j} \quad \text{from utility maximization}$$

$$\text{and} \quad E_j = \sum_i N_i \int_{\varphi_{ij}^*}^{\infty} p_{ij}^F(\varphi) q_{ij}^F(\varphi) g_i(\varphi)\, d\varphi,$$

where $q_{ij}^F(\varphi) = \frac{y_{ij}^F(\varphi)}{\tau_{ij}} = \frac{\varphi l_{ij}(\varphi)}{\tau_{ij}}$. Let η_i be the Lagrange multiplier. Substituting for $p_{ij}^F(\varphi)$ from (19) into (20), we get the Lagrangian function

$$\begin{aligned}
\mathcal{L} = &\sum_{i,j} N_i \int_{\varphi_{ij}^*}^{\infty} (E_j)^{\frac{1}{\sigma}} \left[P_j q_{ij}^F(\varphi)\right]^{\frac{\sigma-1}{\sigma}} g_i(\varphi)\, d\varphi \\
&+ \sum_i \eta_i \left\langle L_i - N_i \left\{ f_e + \sum_j \xi_{ij}\left[1 - G_i\left(\varphi_{ij}^*\right)\right] + \sum_j \int_{\varphi_{ij}^*}^{\infty} \frac{\tau_{ij}}{\varphi} q_{ij}^F(\varphi) g_i(\varphi)\, d\varphi \right\} \right\rangle
\end{aligned}$$

From the F.O.C. w.r.t $q_{ij}^F(\varphi)$, we get

$$q_{ij}^F(\varphi) = E_j \left[\left(\frac{\sigma \eta_i}{\sigma - 1}\right) \frac{\tau_{ij}}{\varphi}\right]^{-\sigma} P_j^{\sigma-1}$$

Substituting this equation into the labor constraint (21) yields

$$\eta_i^{-\sigma} = \frac{\frac{L_i}{N_i} - f_e - \sum_j \xi_{ij}\left[1 - G_i\left(\varphi_{ij}^*\right)\right]}{\left(\frac{\sigma}{\sigma-1}\right)^{-\sigma} \sum_j \int_{\varphi_{ij}^*}^{\infty} \left(\frac{\tau_{ij}}{\varphi P_j}\right)^{1-\sigma} E_j g_i(\varphi)\, d\varphi}$$

Combining the last two equations, we get an expression for $q_{ij}^F(\varphi)$ in terms of E_j, P_j and other exogenous variables. Substituting for $q_{ij}^F(\varphi)$ from the resulting expression into equation (20) and substituting for $p_{ij}^F(\varphi)$ from (19) into (20), we have:

$$Y^w = \sum_i N_i \left\{\frac{L_i}{N_i} - f_e - \sum_j \xi_{ij}\left[1 - G_i\left(\varphi_{ij}^*\right)\right]\right\}^{\frac{\sigma-1}{\sigma}} \left[\sum_j \int_{\varphi_{ij}^*}^{\infty} \left(\frac{\tau_{ij}}{\varphi P_j}\right)^{1-\sigma} E_j g_i(\varphi)\, d\varphi\right]^{\frac{1}{\sigma}} \tag{22}$$

Now we can maximize Y^w by choosing the optimal combination of N_i and φ_{ij}^* based on the above equation. The F.O.C. w.r.t N_i of above expression leads to

$$N_i = \frac{L_i}{\sigma\left\langle \sum_j \left\{\left[1 - G_i\left(\varphi_{ij}^*\right)\right]\xi_{ij}\right\} + f_e\right\rangle}\ , \tag{23}$$

which can be easily obtained from the free entry condition under the decentralized market. Substituting equation (23) into (21) yields $\sum_j \int_{\varphi_{ij}^*}^{\infty} \frac{\tau_{ij}}{\varphi} q_{ij}^F(\varphi) g_i(\varphi)\, d\varphi = \left(\frac{\sigma-1}{\sigma}\right)\frac{L_i}{N_i}$, which means that $\frac{1}{\sigma}$ of the labor will be allocated to cover the fixed market entry cost, while the rest will be allocated to production. Substituting this expression into the definition of each country's GDP yields

$$\begin{aligned} w_i L_i = E_i &= \sum_j N_i \int_{\varphi_{ij}^*}^{\infty} E_j^{\frac{1}{\sigma}} \left[P_j q_{ij}^F(\varphi)\right]^{\frac{\sigma-1}{\sigma}} g_i(\varphi)\, d\varphi \\ &= \eta_i L_i \end{aligned}$$

Thus $\eta_i = w_i$, which implies that $p_{ij}^F(\varphi) = \left(\frac{\sigma}{\sigma-1}\right)\frac{w_i \tau_{ij}}{\varphi}$, which is the same expression as the market outcome. Thus $q_{ij}^F(\varphi)$ and the corresponding

$$l_{ij}(\varphi) = \frac{q_{ij}^F(\varphi)\tau_{ij}}{\varphi} = E_j \left(\frac{\sigma w_i}{\sigma - 1}\right)^{-\sigma} \left(\frac{\tau_{ij}}{\varphi P_j}\right)^{1-\sigma}$$

both have the same expressions as under market equilibrium (refer to (16)).

According to equation (22), maximizing Y^w w.r.t. φ_{ij}^* is equivalent to maximizing

$$\frac{\sigma-1}{\sigma} \ln\left\{\frac{L_i}{N_i} - f_e - \sum_j \xi_{ij}\left[1 - G_i\left(\varphi_{ij}^*\right)\right]\right\} + \frac{1}{\sigma}\ln\left[\sum_j \int_{\varphi_{ij}^*}^{\infty} \left(\frac{\tau_{ij}}{\varphi P_j}\right)^{1-\sigma} E_j g_i(\varphi)\, d\varphi\right].$$

At the optimal φ_{ij}^*, we have

$$\left(\frac{\sigma-1}{\sigma}\right)\frac{\xi_{ij} g_i\left(\varphi_{ij}^*\right)}{\frac{L_i}{N_i} - f_e - \sum_j \xi_{ij}\left[1 - G_i\left(\varphi_{ij}^*\right)\right]} = \left(\frac{1}{\sigma}\right)\frac{\left(\frac{\tau_{ij}}{\varphi_{ij}^* P_j}\right)^{1-\sigma} E_j g_i\left(\varphi_{ij}^*\right)}{\sum_j \int_{\varphi_{ij}^*}^{\infty} \left(\frac{\tau_{ij}}{\varphi P_j}\right)^{1-\sigma} E_j g_i(\varphi)\, d\varphi}$$

which is equivalent to $\sigma\xi_{ij}w_i = E_j\left[\frac{p_{ij}^F(\varphi_{ij}^*)}{P_j}\right]^{1-\sigma} = p_{ij}^F\left(\varphi_{ij}^*\right)q_{ij}^F\left(\varphi_{ij}^*\right)$ by invoking (23). The marginal firm corresponding to the cutoff productivity will spend $\frac{1}{\sigma}$ of its revenue to pay for the market entry cost, which is the same as the market equilibrium outcome.

As the global planner's choices of N_i, φ_{ij}^* and $l_{ij}(\varphi)$ are exactly the same as the market equilibrium, we can conclude that the global planner's solution shown above is exactly the same as the equilibrium outcome.

B Derivation of Equation (12)

Note that $\widehat{p}_{jj}^s(\omega) = \widehat{p}_{jj}^s\ \forall j, s, \omega$. From the production function (3), we get

$$\widehat{p}_{jj}^s = \theta_s^j\left[\sum_i \frac{X_{ij}^{s-1}}{\sum_i X_{ij}^{s-1}}\left(\widehat{\tau}_{ij}^{s-1} + \widehat{p}_{ii}^{s-1}\right)\right] + \left(1-\theta_s^j\right)\widehat{w}_j, \text{ for } \ s = 2, 3, ..., F$$

$$\widehat{p}_{jj}^s = \widehat{w}_j, \text{ for } s = 1$$

This equation implies that

$$\begin{aligned}
\sum_i X_{ij}^{s-1}\widehat{p}_{jj}^s =& \theta_s^j\sum_i X_{ij}^{s-1}\left(\widehat{\tau}_{ij}^{s-1} + \widehat{p}_{ii}^{s-1}\right) + \left(1-\theta_s^j\right)\sum_i X_{ij}^{s-1}\widehat{w}_j \\
\Rightarrow\ \theta_s^j\sum_i X_{ji}^s\widehat{p}_{jj}^s =& \theta_s^j\sum_i X_{ij}^{s-1}\widehat{p}_{ii}^{s-1} + \theta_s^j\sum_i X_{ij}^{s-1}\widehat{\tau}_{ij}^{s-1} + \left(1-\theta_s^j\right)\theta_s^j\sum_i X_{ji}^s\widehat{w}_j \quad \text{as } \sum_i X_{ij}^{s-1} = \theta_s^j\sum_i X_{ji}^s
\end{aligned}$$

Dividing both sides by θ_s^j, summing over all j, and re-arranging terms, we have

$$\underbrace{\sum_{j,i} X_{ij}^s\widehat{p}_{ii}^s - \sum_{j,i} X_{ij}^{s-1}\widehat{p}_{ii}^{s-1}}_{\substack{\text{Increment of the cost effect for} \\ \text{importers from one stage to the next}}} = \underbrace{\sum_{j,i} X_{ij}^{s-1}\widehat{\tau}_{ij}^{s-1}}_{\text{increases in trade costs}} + \underbrace{\sum_{j,i}\left(1-\theta_s^j\right)X_{ji}^s\widehat{w}_j}_{\text{TOT effect on importers}}$$

In other words, the increment of the cost effect for importers from one stage to the next is the sum of the increases in trade costs and increases in the factor costs of the importers (i.e. TOT effect on importers).

Summing up the effect over all stages from stage 2 to stage F and noting that $\widehat{p}_{ii}^{1}=\widehat{w_i}$, we have

$$\begin{aligned}
\underbrace{\sum_{j,i} X_{ij}^{F}\widehat{p_{ii}^{F}}}_{\substack{\text{Cost effect on importers}\\ \text{of final goods}}} &= \sum_{s=1}^{F-1}\sum_{j,i} X_{ij}^{s}\widehat{\tau}_{ij}^{s}+\sum_{j,i} X_{ij}^{1}\widehat{w_i}+\sum_{s=2}^{F}\sum_{j,i}\left(1-\theta_s^{j}\right)X_{ji}^{s}\widehat{w}_j \\
&= \underbrace{\sum_{s=1}^{F-1}\sum_{j,i} X_{ij}^{s}\widehat{\tau}_{ij}^{s}}_{\substack{\text{Direct effect over all stages}\\ \text{other than the final one}}} + \underbrace{\sum_{s}\sum_{j} V_j^{s}\widehat{w}_j}_{\substack{\text{TOT effect on importers}\\ \text{over all stages}}} \qquad (24)
\end{aligned}$$

The global welfare effect resulting from changes in prices of imports is given by[21]

$$\sum_j E_j\widehat{P}_j=\sum_{j,i} X_{ij}^{F}\left(\widehat{p_{ii}^{F}}+\widehat{\tau}_{ij}^{F}\right).$$

Substituting from (24) into the above equation, we obtain (12).

C Derivation of Equation (14)

Define $R_{ij}(\varphi)$ as the revenue of a firm exporting from i to j, with productivity φ. Note that CES preferences implies that

$$\frac{R_{ij}\left(\widetilde{\varphi}_{ij}\right)}{R_{ij}\left(\varphi_{ij}^{*}\right)}=\frac{\left(\widetilde{\varphi}_{ij}\right)^{\sigma-1}}{\left(\varphi_{ij}^{*}\right)^{\sigma-1}} \qquad (25)$$

Total revenue is equal to total income:

$$w_iL_i=\sum_j X_{ij}=\sum_j N_{ij}R_{ij}\left(\widetilde{\varphi}_{ij}\right)$$

which, together with equation (25) and $R_{ij}\left(\varphi_{ij}^{*}\right)=\sigma\xi_{ij}w_i$ (zero cutoff profit condition), imply:

$$w_iL_i=\sum_j N_{ij}\frac{\left(\widetilde{\varphi}_{ij}\right)^{\sigma-1}}{\left(\varphi_{ij}^{*}\right)^{\sigma-1}}\sigma\xi_{ij}w_i$$

Differentiating the natural logarithm of this equation leads to

$$\widehat{w_i}=\sum_j X_{ij}\widehat{N}_{ij}+(\sigma-1)\sum_j X_{ij}\left(\widehat{\widetilde{\varphi}_{ij}}-\widehat{\varphi_{ij}^{*}}\right)+\widehat{w_i}$$

[21] Compare with $\sum_j E_j\widehat{P}_j=\sum_j\sum_i X_{ij}\left(\widehat{w}_i+\widehat{\tau}_{ij}\right)$ in the one-stage case.

which implies equation (14), since $\sum_j X_{ij}\left(\sigma-1\right)\widehat{\varphi_{ij}^*}=0$, which is a consequence of the free entry condition (FEC). To see this, first note that the FEC implies that the total expected fixed costs is equal to the expected net revenue:

$$f_e+\sum_j\left[1-G\left(\varphi_{ij}^*\right)\right]\xi_{ij}=\sum_j\frac{\left[1-G_i\left(\varphi_{ij}^*\right)\right]R_{ij}\left(\widetilde{\varphi}_{ij}\right)}{\sigma w_i}=\sum_j\left[1-G_i\left(\varphi_{ij}^*\right)\right]\xi_{ij}\frac{\left(\widetilde{\varphi}_{ij}\right)^{\sigma-1}}{\left(\varphi_{ij}^*\right)^{\sigma-1}}$$

Totally differentiating the LHS and RHS of the above equation leads to

$$\begin{aligned}\sum_j\xi_{ij}g\left(\varphi_{ij}^*\right)\varphi_{ij}^*\widehat{\varphi_{ij}^*}&=\sum_j\xi_{ij}g\left(\varphi_{ij}^*\right)\frac{\left(\widetilde{\varphi}_{ij}\right)^{\sigma-1}}{\left(\varphi_{ij}^*\right)^{\sigma-1}}\varphi_{ij}^*\widehat{\varphi_{ij}^*}\\&\quad-\sum_j\left[1-G\left(\varphi_{ij}^*\right)\right]\xi_{ij}\frac{\left(\widetilde{\varphi}_{ij}\right)^{\sigma-1}}{\left(\varphi_{ij}^*\right)^{\sigma-1}}\left[\left(\sigma-1\right)\widehat{\widetilde{\varphi}_{ij}}-\left(\sigma-1\right)\widehat{\varphi_{ij}^*}\right]\\ \Leftrightarrow\qquad&\sum_j\left[1-G\left(\varphi_{ij}^*\right)\right]\xi_{ij}\frac{\left(\widetilde{\varphi}_{ij}\right)^{\sigma-1}}{\left(\varphi_{ij}^*\right)^{\sigma-1}}\left(\sigma-1\right)\widehat{\varphi_{ij}^*}=0\\ \Leftrightarrow\qquad&\frac{1}{N_i}\sum_j N_{ij}\xi_{ij}\frac{\left(\widetilde{\varphi}_{ij}\right)^{\sigma-1}}{\left(\varphi_{ij}^*\right)^{\sigma-1}}\left(\sigma-1\right)\widehat{\varphi_{ij}^*}=0\qquad\text{as } N_{ij}=N_i\left[1-G\left(\varphi_{ij}^*\right)\right]\\ \Leftrightarrow\qquad&\sum_j X_{ij}\left(\sigma-1\right)\widehat{\varphi_{ij}^*}=0\end{aligned}$$

where the second equality stems from the effect of the change in cutoff productivity φ_{ij}^* on the average productivity $\widetilde{\varphi}_{ij}$, which is given by:

$$\widehat{\widetilde{\varphi}_{ij}}=\frac{\left[1-\left(\varphi_{ij}^*\right)^{\sigma-1}/\left(\widetilde{\varphi}_{ij}\right)^{\sigma-1}\right]g_i\left(\varphi_{ij}^*\right)\varphi_{ij}^*}{1-G_i\left(\varphi_{ij}^*\right)}\cdot\frac{\widehat{\varphi_{ij}^*}}{\sigma-1}\quad\text{“Productivity effect”}.\tag{26}$$

D Proof of Proposition 3

Step 1: *Derivation of the change in the exact price index.* The exact price index P_j of U_j can be written as:

$$P_j=\min\sum_z p_j\left(z\right)u_j\left(z\right)\ \text{ s.t. }\ U_j\left(\left\{u_j(z)\right|z=1,2,..,Z\}\right)\leq 1$$

where $p_j\left(z\right)$ is the exact price index for subutility $u_j\left(z\right)$. Since the utility function is homogeneous of degree one, totally differentiating P_j implies that $\widehat{P}_j=\sum_z\lambda_j\left(z\right)\widehat{p}_j\left(z\right)$ where $\lambda_j\left(z\right)=\frac{\sum_i X_{ij}^z}{\sum_{z,i}X_{ij}^z}$ denotes the expenditure share on sector-z goods in country j, where $X_{ij}^z\equiv\int_{\omega\in\Omega_{ij}^z}p_{ij}^z(\omega)q_{ij}^z(\omega)d\omega$ is the exports of sector-z goods from country i to country j, and

$$p_j\left(z\right)=\left\{\sum_i N_{ij}^z\left[\left(\frac{\sigma_z}{\sigma_z-1}\right)\frac{w_i\tau_{ij}^z}{\widetilde{\varphi}_{ijz}}\right]^{1-\sigma_z}\right\}^{\frac{1}{1-\sigma_z}}\tag{27}$$

is the aggregate price index for sector z, and $\widetilde{\varphi}_{ijz} \equiv \left[\frac{1}{1-G_{iz}\left(\varphi^*_{ijz}\right)} \int_{\varphi^*_{ijz}}^{\infty} (\varphi)^{\sigma_z - 1} g_{iz}(\varphi)\, d\varphi\right]^{\frac{1}{\sigma_z-1}}$ is the average productivity of a firm in country i serving market j in sector z, φ^*_{ijz} is the cutoff productivity for a firm in country i to profitably export to country j in sector z. Totally differentiating the consumer price index (27), we get

$$\widehat{p}_j(z) = \sum_i \frac{X^z_{ij}}{\sum_{\boldsymbol{i}} X^z_{ij}} \left[\widehat{w}_i + \widehat{\tau}^z_{ij} - \frac{1}{\sigma_z - 1}\widehat{N^z_{ij}} - \widehat{\widetilde{\varphi}_{ijz}}\right]. \tag{28}$$

Thus,

$$\begin{aligned} \widehat{P}_j &= \sum_z \frac{\sum_i X^z_{ij}}{\sum_{\boldsymbol{z},\boldsymbol{i}} X^z_{ij}} \widehat{p}_j(z) \\ \Rightarrow \quad E_j \widehat{P}_j &= \sum_{\boldsymbol{z},\boldsymbol{i}} X^z_{ij} \left[\widehat{w}_i + \widehat{\tau}^z_{ij} - \frac{1}{\sigma_z - 1}\widehat{N^z_{ij}} - \widehat{\widetilde{\varphi}_{ijz}}\right] \end{aligned}$$

Step 2: The change in global welfare is given by

$$\begin{aligned} &\sum_j E_j \widehat{U}_j = \sum_j E_j \left(\widehat{E}_j - \widehat{P}_j\right) \\ &= \sum_{j,i,z} X^z_{ji}\widehat{w}_j - \sum_{j,i,z} X^z_{ij}\left(\widehat{w}_i + \widehat{\tau}^z_{ij} - \frac{1}{\sigma_z - 1}\widehat{N^z_{ij}} - \widehat{\widetilde{\varphi}_{ijz}}\right) \\ &= \underbrace{\sum_{j,i,z} X^z_{ji}\widehat{w}_j - \sum_{j,i,z} X^z_{ij}\widehat{w}_i}_{\text{TOT effect on all importers and exporters} = 0} \underbrace{-\sum_{j,i,z} X^z_{ij}\widehat{\tau}^z_{ij}}_{\text{direct effect}} + \underbrace{\sum_{j,i,z}\left(X^z_{ij}\widehat{\widetilde{\varphi}_{ijz}} + \frac{X^z_{ij}\widehat{N^z_{ij}}}{\sigma_z - 1}\right)}_{\text{global firm mass effect plus productivity effect} \neq 0} \\ &= -\sum_{j,i,z} X^z_{ij}\widehat{\tau}^z_{ij} + \sum_{j,i,z}\left(X^z_{ij}\widehat{\widetilde{\varphi}_{ijz}} + \frac{X^z_{ij}\widehat{N^z_{ij}}}{\sigma_z - 1}\right) \end{aligned}$$

which leads to the main equation in Proposition 3.

Step 3: Proof of the following relationship stated in Proposition 3:

$$-\sum_{z,j} X^z_{ij}\widehat{N^z_{ij}} = \sum_{z,j} (\sigma_z - 1) X^z_{ij}\widehat{\widetilde{\varphi}_{ijz}} \tag{29}$$

Proof: The effect of the change in cutoff productivity φ^*_{ijz} on the average productivity $\widetilde{\varphi}_{ijz}$ is given by:

$$\widehat{\widetilde{\varphi}_{ijz}} = \frac{\left[1 - \left(\varphi^*_{ijz}\right)^{\sigma_z - 1} / \left(\widetilde{\varphi}_{ijz}\right)^{\sigma_z - 1}\right] g_{iz}\left(\varphi^*_{ijz}\right)\varphi^*_{ijz}}{1 - G_{iz}\left(\varphi^*_{ijz}\right)} \cdot \frac{\widehat{\varphi^*_{ijz}}}{\sigma_z - 1} \quad \text{“The productivity effect”}. \tag{30}$$

Define $R^z_{ij}(\varphi)$ as the revenue of a firm with productivity φ exporting from i to j in sector z. CES preferences implies that:

$$\frac{R^z_{ij}\left(\widetilde{\varphi}_{ijz}\right)}{R^z_{ij}\left(\varphi^*_{ijz}\right)} = \frac{\left(\widetilde{\varphi}_{ijz}\right)^{\sigma_z - 1}}{\left(\varphi^*_{ijz}\right)^{\sigma_z - 1}} \tag{31}$$

Total income is equal to total revenue:

$$w_i L_i = \sum_{z,j} X_{ij}^z = \sum_{z,j} N_{ij}^z R_{ij}^z \left(\widetilde{\varphi}_{ijz}\right)$$

which, together with equation (31) and $R_{ij}^z \left(\varphi_{ijz}^*\right) = \sigma_z \xi_{ijz} w_i$ (zero cutoff profit condition), imply:

$$w_i L_i = \sum_{z,j} N_{ij}^z \frac{\left(\widetilde{\varphi}_{ijz}\right)^{\sigma_z - 1}}{\left(\varphi_{ijz}^*\right)^{\sigma_z - 1}} \sigma_z \xi_{ijz} w_i \tag{32}$$

Differentiating the natural logarithm of this equation leads to

$$\widehat{w}_i = \sum_{z,j} X_{ij}^z \widehat{N_{ij}^z} + \sum_{z,j} \left(\sigma_z - 1\right) X_{ij}^z \left(\widehat{\widetilde{\varphi}_{ijz}} - \widehat{\varphi_{ijz}^*}\right) + \widehat{w}_i \tag{33}$$

which implies equation (29), since $\sum_{z,j} \left(\sigma_z - 1\right) X_{ij}^z \widehat{\varphi_{ijz}^*} = 0$, which is a consequence of the free entry condition (FEC), as shown below.

Proof that $\sum_{z,j} \left(\sigma_z - 1\right) X_{ij}^z \widehat{\varphi_{ijz}^*} = 0$. First, note that the FEC implies that the total expected fixed costs is equal to the expected net revenue:

$$f_{ez} + \sum_j \left[1 - G_{iz}\left(\varphi_{ijz}^*\right)\right] \xi_{ijz} = \sum_j \left[1 - G_{iz}\left(\varphi_{ijz}^*\right)\right] \frac{R_{ij}^z \left(\widetilde{\varphi}_{ijz}\right)}{\sigma_z w_i} = \sum_j \left[1 - G_{iz}\left(\varphi_{ijz}^*\right)\right] \frac{\left(\widetilde{\varphi}_{ijz}\right)^{\sigma_z - 1}}{\left(\varphi_{ijz}^*\right)^{\sigma_z - 1}} \xi_{ijz}.$$

Second, totally differentiating the LHS and RHS of the above equation leads to

$$\begin{aligned}\sum_j \xi_{ijz} g_{iz}\left(\varphi_{ijz}^*\right) \varphi_{ijz}^* \widehat{\varphi_{ijz}^*} &= \sum_j \xi_{ijz} g_{iz}\left(\varphi_{ijz}^*\right) \frac{\left(\widetilde{\varphi}_{ijz}\right)^{\sigma_z - 1}}{\left(\varphi_{ijz}^*\right)^{\sigma_z - 1}} \varphi_{ijz}^* \widehat{\varphi_{ijz}^*} \\ &\quad - \sum_j \left[1 - G_{iz}\left(\varphi_{ijz}^*\right)\right] \xi_{ijz} \frac{\left(\widetilde{\varphi}_{ijz}\right)^{\sigma_z - 1}}{\left(\varphi_{ijz}^*\right)^{\sigma_z - 1}} \left[\left(\sigma_z - 1\right) \widehat{\widetilde{\varphi}_{ijz}} - \left(\sigma_z - 1\right) \widehat{\varphi_{ijz}^*}\right]\end{aligned}$$

which, together with equation (30), imply:

$$\begin{aligned} & \sum_j \left[1 - G_{iz}\left(\varphi_{ijz}^*\right)\right] \xi_{ijz} \frac{\left(\widetilde{\varphi}_{ijz}\right)^{\sigma_z - 1}}{\left(\varphi_{ijz}^*\right)^{\sigma_z - 1}} \widehat{\varphi_{ijz}^*} = 0 \\ \Leftrightarrow \quad & \frac{1}{N_i^z} \sum_j N_{ij}^z \xi_{ijz} \frac{\left(\widetilde{\varphi}_{ijz}\right)^{\sigma_z - 1}}{\left(\varphi_{ijz}^*\right)^{\sigma_z - 1}} \widehat{\varphi_{ijz}^*} = 0 \qquad \text{as } N_{ij}^z = N_i^z \left[1 - G_{iz}\left(\varphi_{ijz}^*\right)\right] \\ & \Leftrightarrow \sum_j X_{ij}^z \widehat{\varphi_{ijz}^*} = 0 \qquad (34) \\ & \Leftrightarrow \sum_{z,j} \left(\sigma_z - 1\right) X_{ij}^z \widehat{\varphi_{ijz}^*} = 0. \end{aligned}$$

E Proof of Proposition 4

The expressions for U_j, $q_{ij}^F(\omega)$ and $\widehat{P_j}$ have been given in subsection 3.2. Given (6), the profit maximization problem of a product ω is given by:

$$\max_{p_{ij}^F(\omega)} \quad \left(p_{ij}^F(\omega) - \frac{\tau_{ij} w_i}{\varphi}\right) \frac{\left[p_{ij}^F(\omega)\right]^{-\sigma}}{P_j^{1-\sigma}} E_j.$$

Since the number of varieties is finite, a change in $p_{ij}^F(\omega)$ will affect the price index P_j. The first-order condition is therefore given by:

$$1 - \sigma \left(\frac{p_{ij}^F - \frac{\tau_{ij} w_i}{\varphi}}{p_{ij}^F}\right) + (\sigma - 1)\left(\frac{p_{ij}^F - \frac{\tau_{ij} w_i}{\varphi}}{p_{ij}^F}\right) \frac{\left(p_{ij}^F\right)^{1-\sigma}}{\sum_i \sum_{\omega \in \Omega_{ij}^F} \left[p_{ij}^F(\omega)\right]^{1-\sigma}} = 0 \tag{35}$$

where $p_{ij}^F \equiv p_{ij}^F(\omega)$. Reorganizing (35), we have:

$$p_{ij}^F(\varphi) = c_{ij}(\varphi)\left\{1 + \frac{1}{(\sigma-1)\left[1 - s_{ij}(\varphi)\right]}\right\}. \tag{36}$$

Note also that

$$s_{ij}(\varphi) \equiv \frac{p_{ij}^F(\varphi)\, q_{ij}^F(\varphi)}{\sum_i \sum_{\omega \in \Omega_{ij}^F} p_{ij}^F(\omega)\, q_{ij}^F(\omega)} = \frac{\left[p_{ij}^F(\varphi)\right]^{1-\sigma}}{\sum_i \sum_{\omega \in \Omega_{ij}^F} \left[p_{ij}^F(\omega)\right]^{1-\sigma}}. \tag{37}$$

Hence,

$$\begin{aligned} \widehat{P}_j &= \sum_i \left\{ \sum_{\omega \in \Omega_{ij}^F} s_{ij}(\varphi)\left[\widehat{\mu}_{ij}(\varphi) + \widehat{w}_i + \widehat{\tau}_{ij}\right] \right\} \\ &= \sum_i \left[\sum_{\omega \in \Omega_{ij}^F} s_{ij}(\varphi)\, \widehat{\mu}_{ij}(\varphi) \right] + \sum_i \frac{X_{ij}}{\sum_i X_{ij}} \left(\widehat{w}_i + \widehat{\tau}_{ij}\right) \end{aligned}$$

where $\sum_{\omega \in \Omega_{ij}^F} s_{ij}(\varphi) = \frac{X_{ij}}{\sum_i X_{ij}}$ is the import share from country i in country j.

Therefore, global welfare change is given by:

$$\begin{aligned} &\sum_j E_j \widehat{U}_j = \sum_j E_j \left(\widehat{E}_j - \widehat{P}_j\right) \\ &= \sum_{j,i} X_{ji} \widehat{w}_j - \sum_{i,j} \left[\sum_{\omega \in \Omega_{ij}^F} E_j s_{ij}(\varphi)\, \widehat{\mu}_{ij}(\varphi) \right] - \sum_{i,j} X_{ij} \left(\widehat{w}_i + \widehat{\tau}_{ij}\right) \\ &= -\sum_{i,j} X_{ij} \widehat{\tau}_{ij} - \sum_{i,j} \left[\sum_{\omega \in \Omega_{ij}^F} x_{ij}(\varphi)\, \widehat{\mu}_{ij}(\varphi) \right] \end{aligned}$$

where $x_{ij}(\varphi) = E_j s_{ij}(\varphi)$. Thus, we have the equation stated in Proposition 4.■

F Estimating the extra term in Proposition 3

For simplicity, we assume that $U_j = \prod_z u_j(z)^{\alpha(z)}$ with $\sum_z \alpha(z) = 1$. Let $p_j(z)$ denote the exact price index for subutility $u_j(z)$ as in Online Appendix D. Then the following equations represent a system of $3n^2Z + 2nZ + n$ equations with the same number of unknowns $\widehat{w}_i$, $\widehat{p}_j(z)$, $\widehat{\varphi^*_{ijz}}$, $\widehat{N^z_{ij}}$, $\widehat{\widetilde{\varphi}_{ijz}}$ and $\widehat{N^z_i}$ for all i, j, z:

$\widehat{p}_j(z) = \sum_i \frac{X^z_{ij}}{\sum_i X^z_{ij}} \left[\widehat{w}_i + \widehat{\tau}^z_{ij} - \frac{1}{\sigma_z - 1}\widehat{N^z_{ij}} - \widehat{\widetilde{\varphi}_{ijz}}\right]$	Price index	for all j and z
$0 = \sum_{z,j} X^z_{ij}\widehat{N^z_{ij}} + \sum_{z,j}(\sigma_z - 1) X^z_{ij}\widehat{\widetilde{\varphi}_{ijz}}$	Labor market clearing condition	for all i
$0 = \sum_j X^z_{ij}\widehat{\varphi^*_{ijz}}$	Free entry condition	for all i and z
$\widehat{\widetilde{\varphi}_{ijz}} = \frac{\left[1-\left(\varphi^*_{ijz}\right)^{\sigma_z-1}/\left(\widetilde{\varphi}_{ijz}\right)^{\sigma_z-1}\right] g_{iz}\left(\varphi^*_{ijz}\right)\varphi^*_{ijz}}{1-G_{iz}\left(\varphi^*_{ijz}\right)} \cdot \frac{\widehat{\varphi^*_{ijz}}}{\sigma_z - 1}$	Productivity effect	for all i, j and z
$\widehat{N^z_{ij}} = \widehat{N^z_i} - \frac{g_{iz}\left(\varphi^*_{ijz}\right)\varphi^*_{ijz}}{1-G_{iz}\left(\varphi^*_{ijz}\right)}\widehat{\varphi^*_{ijz}}$		for all i, j and z
$\widehat{w}_i = (1-\sigma_z)\left(\widehat{w}_i + \widehat{\tau}^z_{ij} - \widehat{\varphi^*_{ijz}} - \widehat{p}_j(z)\right) + \widehat{w}_j$		for all i, j and z

where the first four sets of equations are, respectively, (28), (29), (34) and (30) from Online Appendix D. The fifth set of equations stems from $N^z_{ij} = N^z_i\left[1 - G_{iz}\left(\varphi^*_{ijz}\right)\right]$ which is true by definition, and the sixth one stems from $\sigma_z \xi_{ijz} w_i = \frac{p\left(\varphi^*_{ijz}\right)^{1-\sigma_z}}{p_j(z)^{1-\sigma_z}}\alpha(z) w_j L_j = \left(\frac{\sigma_z}{\sigma_z - 1}\frac{w_i \tau^z_{ij}}{\varphi^*_{ijz}}\right)^{1-\sigma_z}\frac{\alpha(z) w_j L_j}{p_j(z)^{1-\sigma_z}}$ (zero cutoff profit condition).

When the distribution is Pareto such that $G_{iz}(\varphi) = 1 - \left(\frac{\varphi}{\varphi_{\min}}\right)^{-\theta}$ for all i and z, we have

$$\frac{g_{iz}\left(\varphi^*_{ijz}\right)\varphi^*_{ijz}}{1 - G_{iz}\left(\varphi^*_{ijz}\right)} = \theta$$

$$\text{and} \quad \frac{\left(\varphi^*_{ijz}\right)^{\sigma_z - 1}}{\left(\widetilde{\varphi}_{ijz}\right)^{\sigma_z-1}} = \frac{\left(\varphi^*_{ijz}\right)^{\sigma_z - 1}}{\frac{\theta}{\theta - (\sigma_z - 1)}\left(\varphi^*_{ijz}\right)^{\sigma_z - 1}} = \frac{\theta - (\sigma_z - 1)}{\theta}$$

Thus, under Pareto distribution, the system can be simplified to the following system of $n^2Z + 2nZ + n$ equations and the same number of unknowns $\widehat{w}_i$, $\widehat{p}_j(z)$, $\widehat{\widetilde{\varphi}_{ijz}}$ and $\widehat{N^z_i}$ for all i, j, z:

$\widehat{p}_j(z) = \sum_i \frac{X^z_{ij}}{\sum_i X^z_{ij}}\left[\widehat{w}_i + \widehat{\tau}^z_{ij} - \frac{1}{\sigma_z - 1}\widehat{N^z_i} + \frac{\theta - (\sigma_z - 1)}{\sigma_z - 1}\widehat{\widetilde{\varphi}_{ijz}}\right]$	for all j and z
$0 = \sum_{z,j} X^z_{ij}\left(\widehat{N^z_i} - \theta\widehat{\widetilde{\varphi}_{ijz}}\right)$	for all i
$0 = \sum_j X^z_{ij}\widehat{\widetilde{\varphi}_{ijz}}$	for all i and z
$\widehat{\widetilde{\varphi}_{ijz}} = \widehat{w}_i + \widehat{\tau}^z_{ij} - \widehat{p}_j(z) + \frac{\widehat{w}_i - \widehat{w}_j}{\sigma_z - 1}$	for all i, j and z

Given bilateral trade data (X^z_{ij}) and the values of the parameters θ and σ_z, we can solve for $\widehat{w}_i$, $\widehat{p}_j(z)$, $\widehat{\widetilde{\varphi}_{ijz}}$ and $\widehat{N^z_i}$ from the above $n^2Z + 2nZ + n$ equations. Thus, $\widehat{\varphi^*_{ijz}}$ and $\widehat{N^z_{ij}}$ for all i, j, z can be calculated from the first six sets of equations above. Therefore, the extra term in Proposition 3 can be calculated.

G Estimating the extra term in Proposition 4

The market share in country j of a firm producing variety $\omega \in \Omega_{ij}^F$ with productivity draw φ is

$$s_{ij}(\varphi) \equiv \frac{p_{ij}^F(\varphi)\, q_{ij}^F(\varphi)}{E_j} = \frac{p_{ij}^F(\varphi)^{1-\sigma}}{\sum_i \sum_{\omega \in \Omega_{ij}^F} \left[p_{ij}^F(\varphi)\right]^{1-\sigma}} \tag{38}$$

which is the same as (37). At the same time, according to (7), the markup $\mu_{ij}(\varphi)$, and therefore the price $p_{ij}^F(\varphi)$, is a function of market share $s_{ij}(\varphi)$:

$$p_{ij}^F(\varphi) = \mu_{ij}(\varphi)\, c_{ij}(\varphi) = \left\{1 + \frac{1}{(\sigma-1)\left[1 - s_{ij}(\varphi)\right]}\right\} \frac{\tau_{ij} w_i}{\varphi} \tag{39}$$

which is (36) with $c_{ij}(\varphi) = \frac{\tau_{ij} w_i}{\varphi}$.

The labor market clearing condition in each country is

$$w_i L_i = \sum_j X_{ij} = \sum_j \sum_{\omega \in \Omega_{ij}^F} x_{ij}(\varphi) = \sum_j \sum_{\omega \in \Omega_{ij}^F} s_{ij}(\varphi)\, w_j L_j \tag{40}$$

Letting the number of firms in the set Ω_{ij}^F be N_{ij}, then there are $\sum_i N_{ij}$ firms serving market j. Equation (38) and (39) for each firm and equation (40) for each country i together form a system of $2\sum_j \sum_i N_{ij} + n$ equations with the same number of unknowns, namely $\{s_{ij}(\varphi)\}$ and $\{p_{ij}^F(\varphi)\}$ for all firms $\omega \in \Omega_{ij}^F, \forall i, j$ and the wages $\{w_i\}, \forall i$. When there are infinitesimal changes in trade costs $\{\tau_{ij}\}$, the above system yields a new equilibrium. Log-linearizing the system leads to

$$\widehat{s_{ij}(\varphi)} = (1-\sigma)\left[\widehat{p_{ij}^F(\varphi)} - \sum_i \sum_{\omega \in \Omega_{ij}^F} s_{ij}(\varphi)\, \widehat{p_{ij}^F(\varphi)}\right] \quad \text{for all } i,\ j, \quad \text{from (37)} \tag{41}$$

$$\begin{aligned}
\widehat{p_{ij}^F(\varphi)} &= \widehat{\mu_{ij}(\varphi)} + \widehat{c_{ij}(\varphi)} \\
&= \frac{1}{\mu_{ij}(\varphi)} - \frac{\mu_{ij}(\varphi)-1}{\mu_{ij}(\varphi)}\left(\frac{1}{1-s_{ij}(\varphi)} - \frac{s_{ij}(\varphi)}{1-s_{ij}(\varphi)}\widehat{s_{ij}(\varphi)}\right) + \left(\widehat{\tau_{ij}} + \widehat{w_i}\right) \quad \text{for all } i,\ j, \\
&\quad \text{from (36) and (7)}
\end{aligned} \tag{42}$$

$$\widehat{w_i} = \sum_j \sum_{\omega \in \Omega_{ij}^F} \frac{x_{ij}(\varphi)}{w_i L_i}\left(\widehat{s_{ij}(\varphi)} + \widehat{w_j}\right) \quad \text{for all } i,\ j, \quad \text{from (40)} \tag{43}$$

The above linear system has $2\sum_j \sum_i N_{ij} + n$ equations and $2\sum_j \sum_i N_{ij} + n$ unknowns, namely $\{\widehat{s_{ij}(\varphi)}\}$ and $\{\widehat{p_{ij}^F(\varphi)}\}$ for all firms $\omega \in \Omega_{ij}^F, \forall i, j$ and $\{\widehat{w_i}\}, \forall i$. The system can be solved when the market share $s_{ij}(\varphi)$, and the markup $\mu_{ij}(\varphi)$ for each firm before the changes in trade costs are known. After we solve for the system given the changes in trade costs $\{\widehat{\tau_{ij}}\}$, we can obtain

$$\widehat{\mu_{ij}(\varphi)} = \frac{1}{\mu_{ij}(\varphi)} - \frac{\mu_{ij}(\varphi)-1}{\mu_{ij}(\varphi)}\left(\frac{1}{1-s_{ij}(\varphi)} - \frac{s_{ij}(\varphi)}{1-s_{ij}(\varphi)}\widehat{s_{ij}(\varphi)}\right) \quad \text{for all } i,\ j, \quad \text{from (7)}$$

This would enable us to compute the extra term in Proposition 4.