University of Zurich[UZH]

URPP Equality of Opportunity

Within-Country Inequality and the Patterns of Trade

Reto Foellmi
Marco Schmid
Josef Zweimüller

Equality of Opportunity Research Series #17
April 2023

URPP Equality of Opportunity

URPP Equality of Opportunity Discussion Paper Series No.17, April 2023

Within-Country Inequality and the Patterns of Trade

Reto Foellmi
University of St. Gallen
reto.foellmi@unisg.ch

Marco Schmid
University of Zurich
marco.schmid@econ.uzh.ch

Josef Zweimüller
University of Zurich
josef.zweimueller@econ.uzh.ch

URPP Equality of Opportunity, University of Zurich, Schoenberggasse 1, 8001 Zurich, Switzerland
info@equality.uzh.ch, www.urpp-equality.uzh.ch

Within-Country Inequality and the Patterns of Trade

Reto Foellmi*, Marco Schmid†, Josef Zweimüller‡

June 2022

Abstract

We introduce a demand-side trade model to shed light on the interaction between countries' income distributions, the patterns of international trade and parallel trade policy restrictions in general equilibrium. We provide a comprehensive theoretical foundation for the role of income inequality for trade that allows distinguishing within- and between-country inequality. To bring differences in the willingness to pay of differentially rich consumers to the forefront of the analysis, we deviate from the canonical model by replacing the standard CES preferences with non-homothetic 0-1 preferences. We characterize and numerically solve the trade equilibrium for a discrete labor endowment distribution with several consumer-income groups. Our model predicts trade intensity to be increasing in the income distribution overlap, which is our preferred proxy for demand similarity. Our model, therefore, provides a theoretical foundation for the Linder hypothesis, that bilateral trade volume is increasing in the similarity of demand structure between two countries. Furthermore, our model predicts a *Manhattan effect,* capturing that poor consumers are badly off if they are a small minority in a predominantly rich country due to the high price level for basic products.

JEL classification: F10, F12, F19

Keywords: Within-country inequality, non-homothetic preferences, parallel imports, arbitrage, extensive margin, intensive margin, export zeros, North-South trade, income distribution similarity

*University of St. Gallen and CEPR, SIAW-HSG, Bodanstrasse 8, CH-9000 St. Gallen, e-mail: reto.foellmi@unisg.ch

†University of Zurich, Department of Economics, Schönberggasse 1, CH-8001 Zürich, e-mail: marco.schmid@econ.uzh.ch

‡University of Zurich and CEPR, Department of Economics, Schönberggasse 1, CH-8001 Zürich, e-mail: josef.zweimueller@econ.uzh.ch

We thank Kiminori Matsuyama, Ralph Ossa, David Hémous and participants of the Macro Doctoral Seminar at the University of Zurich for helpful comments and suggestions. We thank Claudia Bernasconi for making her code to generate income distributions and similarity measures available.

1 Introduction

Going back to Engel (1857), economists recognized that consumption patterns change with income expansion. Rich consumers do not behave like scaled versions of poor consumers. High income households consume a wider range of varieties and spend relatively less of their income on basic goods. Hence, consumer preferences are inherently non-homothetic. Nevertheless, even though the key implications of homothetic preferences are plainly rejected in the data, constant elasticity of substitution (CES) preferences, the most common form of homothetic preferences, are omnipresent in trade models.[1] CES preferences arguably own their popularity to tractability due to convenient aggregation properties rather than conviction of providing a good approximation to reality.[2] There is an entire literature in macroeconomics focusing on GDP per capita[3] and over recent years a prominent discussion of income inequality emerged.[4] However, demand-side heterogeneity is largely absent in theoretical trade models.[5]

We suggest a theoretical framework to improve on this by explicitly modeling countries' income distributions and studying their effect on trade patterns. We introduce a demand-side, general equilibrium, two-country model of international consumer goods trade featuring non-homothetic preferences and discus the role of parallel trade policy restrictions. Parallel trade occurs when a good protected by a patent, copyright, or trademark, having been legally purchased in one country, is exported to another without the authorization of the local owner of the intellectual property rights in the importing market (see Maskus, 2000). Hence, parallel imports are goods imported by unauthorized resellers. To bring differences in the willingness to pay of differentially rich consumers to the forefront of the analysis, we deviate from the canonical model by replacing the standard CES preferences with non-homothetic 0-1 preferences following Foellmi, Hepenstrick, and Zweimüller (2018; henceforth FHZ). We extend their representative agents model by introducing heterogeneous endowment levels *within* countries. With 0-1 preferences a households decides for every available variety whether to consume 1 unit of it or not to consume this variety at all. Such preferences combined with a non-degenerate income distribution generate an extensive consumption margin within countries and thus a role for income inequality. These simple 0-1 preferences, only having an extensive margin, are in stark contrast to CES preferences, which only have an intensive margin as CES consumers decide how much of each variety to buy but always consume all varieties.[6] We thus make progress over trade models relying on CES preferences by adding the extensive consumption margin.

The key innovation of our model is that it not only allows for between-country but also for within-county inequality. This enables us to study the effect of empirically relevant income distribution constellations on the patterns of bilateral trade. We are especially interested in the patterns of trade between poor and rich countries – in an average GDP per capita sense – and varying degrees of inequality within those countries as this can create export exclusion effects, where some firms choose to not supply all markets due to the threat of international arbitrage. While most of the literature on trade and inequality studies the effect of trade on inequality,[7] we address the opposite direction: the effect of inequality on trade. Matsuyama (2019)

[1] For an overview of empirical tests of CES see Diebot and Kyrtsou (2005).

[2] See Allen and Arkolakis (2016) p.17.

[3] For example GDP per capita features prominently in the strand of the growth literature building on the Solow-Swan (1956) model. One of the fundamental questions of economic growth is why some countries are so much richer than others. Hall and Jones (1999) address this from an output per worker perspective. Also see Jones (2016) for a review.

[4] See Piketty (2014), Piketty and Saez (2003).

[5] One of the few exceptions are Behrens and Murata (2012) and Kichko et al. (2014).

[6] The consumption of all varieties under CES is due to the property that the marginal utility of a variety goes to infinity when its quantity goes to zero. Hence, there is a infinite reservation price for every variety yielding love for variety .

[7] Fajgelbaum and Khandelwal (2016) or Galle-Rodriguez and Clare (2017) are recent examples for theoretical work and Autor et al. (2013) set off a series of empirical papers on the effect of the China trade shock on U.S. workers.

introduces a new class of non-homothetic preferences – referred to as isoelastically non-homothetic CES – creating a role for GDP per capita. He employs these preferences to study the intersectoral patterns of trade across countries but adheres to homogeneous consumers within countries.

Our model provides a theoretical foundation for the classical Linder (1961) hypothesis, that bilateral trade volume is increasing in the similarity of the demand structure between two countries. Most previous studies use the similarity of *average* income levels to approximate the similarity of demand patterns. See for example Hallak (2010), which uses GDP per capita to capture demand similarity. However, only relying on average income levels masks important heterogeneity in how incomes are distributed within countries. Bernasconi (2013) shows empirically that about one third of the variation in the income distribution overlap, which is her preferred measure for demand similarity, cannot be explained by the variation in the GDP per capita ratio. She documents substantial variation in similarity of income distributions for a given GDP per capita ratio. Moreover, she finds that while the similarity in average income levels approximates the similarity in income distributions not too well for North-North and North-South country pairs, it performs especially poorly for South-South country pairs.[8] Hence, to understand empirical trade patterns we need a theoretical framework to explicitly model within-country inequality rather than just summarizing all consumers of a country into a representative agent. We are the first to provide such a theoretical foundation for the role of income distribution similarity for trade that allows distinguishing within- and between-country inequality.[9] Our perspective of letting the demand-side heterogeneity shape the patterns of trade should be considered complementary to the heterogeneous firm approach initiated by Melitz (2003). In our model, all firms have identical Dixit-Stiglitz style production technology. Hence, the patterns of trade are entirely shaped by the heterogeneity of consumers.

We can characterize and numerically solve the trade equilibrium for a discrete labor endowment distribution with several consumer-income groups and a variety of bilateral income distribution constellations. Our model predicts trade intensity to be increasing in the income distribution overlap and is hence consistent with the Linder hypothesis. We show that for sufficiently high endowment differences between poor and rich consumers, world trade volume is strictly increasing in income distribution similarity. Furthermore, our model predicts a *Manhattan effect,* capturing that poor consumers are extremely badly off if they are a small minority in a rich location because of the high price level for basic products. This is in line with the empirical findings of Handbury (2021), who finds that when accounting for non-homothetic preferences poor households face higher food prices in rich relative to poor cities. In our model, rich consumers in both countries always fare equally well irrespective of their domicile, while poor consumers' welfare strongly depends on where they live. We show how trade patterns are driven by within- versus between-country income inequality, parallel import policy and the equilibrium response of firms to the demand structure.

We also provide further empirical evidence for the Linder hypothesis using more recent income distributions data from the World Income Inequality Database and the World Bank's Exporter Dynamics Database. Distinguishing between the intensive and extensive margin of trade, we show that both margins are increasing in income similarity, which confirms the findings of Bernasconi (2013). Our paper is related to Fernandes et al. (2018), who emphasize the empirical importance of the intensive trade margin using firm-level data from the World Bank's Exporter Dynamics Database. To improve upon the standard Pareto-Melitz model that only features an extensive margin, they suggest using a lognormal firm productivity distribution. They

[8] Other, empirical studies on how within-country income distribution shapes trade patterns are Dalgin et al. (2008) and Fieler (2011).

[9] Related theoretical frameworks are Mitra and Trindade (2005), which also features demand-side heterogeneity but uses asset ownership inequality rather than income inequality and relies on a two goods two factors of production framework. Also related is Matsuyama (2000), which develops a Ricardian trade model and thus relies on heterogeneous production technology.

use maximum likelihood methods to estimate this generalized Melitz model and show that this allows to match the intensive margin observed in the data. Our paper suggests an alternative mechanism to create the intensive margin of trade purely through demand-side forces.

Pricing-to-market due to heterogeneous consumers within and across countries can create international arbitrage opportunities.[10] FHZ elaborate that in a representative agent setting, featuring a poor and a rich country, the threat of parallel imports plays a crucial role in a firm's decision over which markets to serve.[11] Firms are confronted with a price versus market size trade-off. If a firm charges a high price, it can only sell its products in the rich North and therefore has a small market. This exclusion strategy gives rise to an export zero for this particular product. If the firm wants to offer its product to the large global market, it has to take potential parallel imports into account. If the firm were to charge a high price in the rich North but simultaneously a low price in the poor South, arbitrageurs would purchase the good cheaply in the South and ship it back to the North, where they can market it for a price marginally below the one the official dealer charges and make a profit. In order to avoid loosing the Northern markets to arbitrageurs, the firm has to set a reduced price in the North that is constrained by the level of trade costs. A classic example is a US located pharmaceutical firm choosing not to export expensive medication to China since parallel imports from China would undermine the firm's pricing in developed countries. Hence, firms applying pricing-to-market and the threat of parallel imports generates export zeros for varieties targeted at rich consumers when trade costs are low and income differences are high. This provides an alternative to Melitz's (2003) productivity cutoff explanation for trade zeros on the product level. In his heterogeneous firm model only firms with a sufficiently high productivity export their product. We extend this discourse by introducing within-country inequality, meaning there are poor and rich consumers living in both countries. The threat of parallel imports is much *less* likely to give rise to export zeros in this setting since the richest consumers in both countries will always be served unless the richest consumers in the two countries strongly differ in endowment or trade costs are prohibitively high. Nevertheless, we show that also in this setting arbitrage constraints play a crucial role in firms' marketing decisions.

Our model provides a framework to study the welfare effects of parallel trade policy restrictions under income inequality in general equilibrium and is therefore related to the literature on the welfare effects of the exhaustion of intellectual property rights. When parallel imports are allowed, international arbitrage opportunities can be exploited and discipline prices. When parallel imports are prohibited by law, arbitrage is ruled out and firms can set discriminatory prices. Malueg and Schwartz (1994) present a supply-side partial equilibrium model with a single monopolistic producer per country to study the welfare effects of parallel import policy. They conclude that allowing parallel imports and therefore enforcing uniform pricing is not necessarily beneficial for a country's welfare because of the variety losses due to regional exclusion. Prohibiting parallel imports is beneficial for poor consumers as it allows them to be served despite high prices for the rich. On the other hand, a uniform pricing requirement is beneficial for the rich as arbitrage is being exploited, allowing them to enjoy lower prices. They find that prohibiting parallel imports and therefore allowing for discriminatory prices improves world welfare if demand dispersion across markets is large as it allows to serve larger markets. Our model predictions are in line with their findings that global welfare is higher when parallel imports are prohibited compared to when they are allowed as products reach a larger number of consumers when prices can be set discriminatorily. Grossman and Lai (2008) discuss the effect of parallel imports on innovation in the context of government price controls applied to the pharmaceutical

[10] Pricing-to-market refers to the situation where for the same good, producers choose markups that vary across markets to exploit differences in willingness to pay.

[11] Malueg and Schwartz (1994) define parallel imports as "goods imported by unauthorized resellers".

market. They show that incentives for product innovation in the North can be higher when the North permits parallel imports but the South will suffer welfare losses. Their dynamic perspective on the effect of parallel import policy can be seen as complementary to our static treatment focusing on inequality.

Roy and Saggi (2012) discuss the effect of parallel import policies on international market structure and welfare in partial equilibrium. They model a North-South quality differentiated duopoly competing in prices. They do not allow for firm entry. They endogenize the government's choice of parallel import policies and study its effect on firms' price competition and the interdependence of national parallel import policies. Consumers are heterogeneous in their quality preference and make a 0-1 purchase decision. They show that for high between-country inequality the North has an incentive to forbid parallel imports since Northern consumers are harmed by the high price of the Northern duopolist abstaining from exporting to the South due to the threat of arbitrage. However, the social planner wants to ensure that the Northern duopolist sells to the South. For low between-country inequality the North permits parallel imports since there is no export exclusion. Hence, there is uniform pricing, which is the North's preferred market outcome. Southern consumers are harmed by the above monopoly pricing of the Southern duopolist under uniform pricing. Hence, the South's preferred market outcome is international price discrimination and hence forbidding parallel imports. In out model the North always prefers to allow parallel imports as it can benefit from an increased domestic variety when export exclusion occurs. Our model's different preferred parallel trade prediction compared to Roy and Saggi (2012) is rooted in the different market structure. It is the possibility of entry that under monopolistic competition makes exclusion attractive for the rich country and hence results in the rich country preferring to allow parallel trade. Allowing parallel trade induces some rich country firms to exclude the poor country market which retains more resources in the rich country and hence allows setting up more firms. Our paper is the first to study parallel trade in general equilibrium with monopolistic competition and pointing out this resource feedback effect.

This paper derives how trade patterns are shaped by within- and between-country inequality in the presence of international arbitrage constraints. Understanding how demand shapes trade, we can evaluate how trade translates the endowment distribution into the distribution of welfare and who benefits or looses from trade. The presented framework can also be used to evaluate how a trade liberalization may affect rich and poor consumers or welfare between countries.

The remainder of this paper is organized as follows. In Section 2, we explain how to obtain empirical income distributions and provide further empirical support for the Linder hypothesis via regressing the trade margins on the overlap in the income distributions. Section 3 introduces 0-1 preferences and provides the model setup. In Section 4, we first develop our model for two endowment groups to build intuition and then extend it to three endowment groups in Section 5. Ultimately, we are pursuing to formulate the model with a general endowment distribution, so that it can be easily matched to empirically observed income distributions. Section 6 gives an outlook on how continuous labor endowment distributions behave and Section G in the appendix suggest how the model can be taken to the data. Finally, Section 7 concludes and outlines the road ahead for further research.

2 Empirical Motivation: Income Distribution Similarity and Trade Margins

To motivate the demand for our model we first provide some empirical evidence for the relationship between countries' income distribution similarity and international trade margins. The empirical part of this paper is

strongly building on Bernasconi (2013).[12] It provides an empirical motivation why within-country inequality matters for trade patterns and presents a number of stylized facts that our theoretical model can address. Bernasconi (2013) documents robust empirical evidence for the Linder hypothesis that bilateral trade volume is increasing in the similarity of demand structure between two countries and that both the intensive and the extensive margin of trade contribute to this about equally. She suggests using the overlap in the income distribution as an intuitive measure to approximate the degree to which the demand structure of country o resembles the demand structure of country d.[13] However, a theoretical framework that models the relation between income similarity and trade patterns is lacking. We are the first to present a theoretical trade model that allows distinguishing within and between-country inequality and its effect on trade patterns.

In the following we document some stylized facts about the relationship between income distribution similarity and international trade patters. We use the World Income Inequality Database by UNU-WIDER (WIID) combined with GDP and population data from Penn World Table (PWT) to generate countries' income distributions. We use two alternative sources of trade flows: CEPII's BACI database and the World Banks's Exporter Dynamics Database (EDD).

2.1 Income distributions and overlap

The World Income Inequality Database (WIID)[14] contains measures of income inequality for almost every country in the world. Data has been collected through surveys and from reports. The available inequality measures include Gini coefficient, quintile and decile shares, survey means and income shares of the richest and poorest 5%. The database covers 189 countries and contains a total of 11'101 inequality measurements, of which 3'500 are unique country-year observations, for the timespan 1867-2017. There are several estimates reported per country and year due to several sources based on different surveys and concepts.

A detailed protocol on how to calculate a country's empirical income distribution from the decile and quintile income shares data available in WIID combined with GDP per capita and population data from PWT is provided in Appendix A.1.[15] After applying a resistant, nonlinear smoother to smooth out discontinuities, this yields continuous empirical income distributions as depicted for the US in Figure 1.

Using the timespan 2003-2013, there are income distributions available for 145 countries in WIID3.4. Hence, they combine to 10'440 possible country pairs (145*144/2). One possible measure to capture income distributions similarity between two countries is the size of the area where their income distributions overlap. We follow the approach suggested by Bernasconi (2013) on how to calculate the empirical income distribution overlap. Let $f_a(x)$ denote the density function of income x of country $a \in \{c, n\}$. The overlap O_{od} is defined as the minimum sum of the income distributions of two countries o and d on the 5k interval $\breve{x}$.

$$O_{od} = \sum_{\breve{x}} min\left\{f_o(\breve{x}), f_d(\breve{x})\right\}, \quad \in [0; 1]$$

The income distribution overlap for the two major world economies – the USA and China – is depicted as the gray shaded area in in Figure 2. For 2013, the overlap area is 0.35 and concentrates at low and middle incomes. Bernasconi (2013) uses the average income level of the overlap area as an additional similarity

[12] We thank Claudia Bernasconi for making her code to generate income distributions and similarity measures available. This allowed a close replication of her empirical analysis.

[13] Choi et al. (2009), Bohman and Nilsson (2007) or Martinez-Zarzoso and Vollmer (2016) are other examples from the literature that use the overlap as an empirical measure of similarity of income distributions.

[14] The current version WIID4 was released in December 2018 and is available under www.wider.unu.edu/database/world-income-inequality-database-wiid4

[15] We use Penn World Table version 9.0 available under www.rug.nl/ggdc/productivity/pwt/

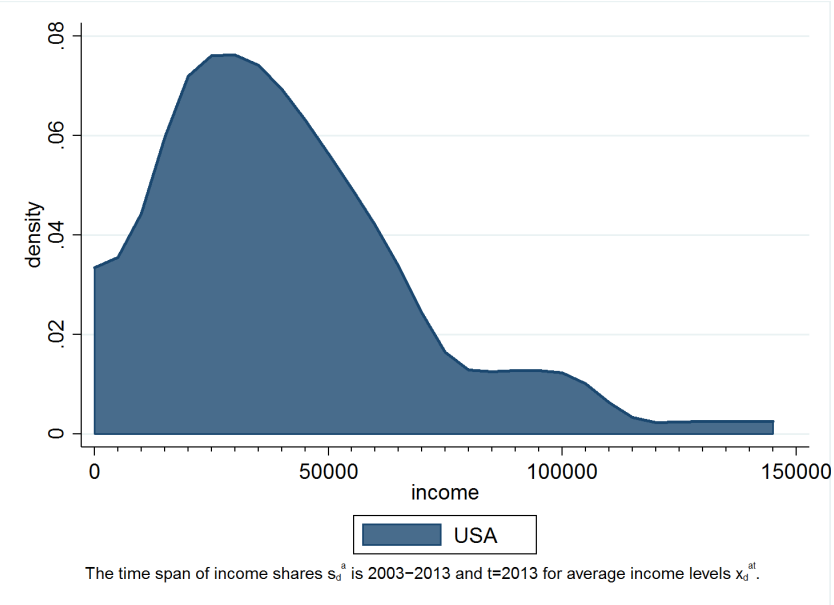


Figure 1: US Income distribution

measure intended to capture the location of the overlap.

2.2 Trade data and trade margins

As our primary trade data source, we use trade flow data from CEPII's BACI (see Gaulier and Zignago (2010)),[16] which is a consolidated version of UN COMTRADE. The unit of observation is the annual bilateral trade flow value on HS6-digit level. BACI data covers the timespan 1995-2017 and allows to match 93 origins and 93 destinations in 2002. An alternative trade data source is the World Bank's Exporter Dynamics Database (EDD) introduced in Fernandes et al. (2018),[17] which is based on firm-level customs data. The publicly available data used here are annual average firm-level statistics on HS2-digit level. EDD covers the timespan 1997-2014 and allows to match 37 origins and 104 destinations in 2012. The observed measures include the number of exporting firms, the average value of exports per firm, the average number of varieties exported per firm, the average export value per product per firm and the average unit price.

We decompose the value of aggregate bilateral trade flows into the intensive and extensive margin as suggested by Hummels and Klenow (2005). We follow Bernasconi (2013) to construct aggregate trade margins from 6-digt product data. Let v_{odi} denote the US dollar value of product category i exported from country o to country d during a given year. We define V_{od} to be the US dollar value of country d's overall imports from country o normalized by d's imports from the rest of the world

$$V_{od} = \frac{\sum_{i \in I} v_{odi}}{\sum_{i \in I} v_{rdi}}, \quad v_{rdi} = \sum_{\hat{o} \in C_{-o}} v_{\hat{o}di}$$

The extensive margin captures the number of HS 6-digit categories with positive trade flows between origin o and destination d. Hence, it reflects the diversity of the trade flow. The extensive margin EM_{od} is defined as the count of traded product categories relative to the number of varieties destination d imports

[16] Available under www.cepii.fr/cepii/en/bdd_modele/presentation.asp?id=1
[17] Available under datacatalog.worldbank.org/dataset/exporter-dynamics-database

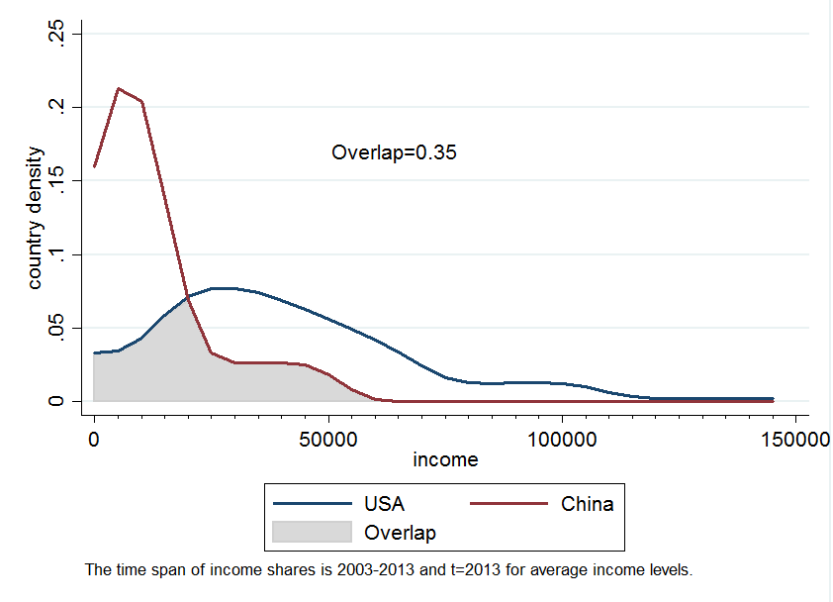


Figure 2: Income distributions and overlap for USA and China

from the rest of the world

$$EM_{od} = \frac{\sum_{i \in I} 1_{odi}}{\sum_{i \in I} 1_{rdi}}$$

On the other hand, the intensive margin captures the average value traded over all product categories and hence addresses trade intensity. The intensive margin IM_{od} is defined as the value of d's imports from o in the set of goods I_{od} relative to the value of destination d's imports from the rest of the world in the same set

$$IM_{od} = \frac{\sum_{i \in I_{od}} v_{odi}}{\sum_{i \in I_{od}} v_{rdi}}.$$

This standardization of bilateral trade values by the imports from rest of world only makes sense if there more than two countries. Hence, for our two-country model we define trade margins unstandardized as in Fernandes et al. (2018). See Appendix A.2 for the definitions.

2.3 Trade margin regressions

Reproducing Bernasconi (2013) with more recent income distribution data and alternative trade data, we test the Linder hypothesis that countries with more similar demand structures trade more with each other. We use the income distribution overlap to approximate similarity of demand structures among trading partners. We apply Bernasconi (2013) augmented gravity equation, exploiting cross-sectional variation, to estimate the association between income distribution overlap and the margins of international trade.

We use Bernasconi (2013) main specification, which is estimated with OLS

$$ln\left(Y_{od}\right) = \alpha + \beta O_{od} + \tau'_{od}\gamma + A_o + A_d + \epsilon_{od}, \quad Y \in \{V, EM, IM\}, \tag{1}$$

where Y_{od} represents the respective aggregate bilateral trade margin between exporter o and importer d.

Each logarithmized trade margin is regressed on the overlap O_{od}. The coefficient of interest β estimates the marginal effect of income distribution similarity on bilateral trade. This estimated effect is conditional on trade costs τ_{od} as well as importer and exporter fixed effects, A_d and A_o. We include a vector of standard controls to approximates bilateral trade costs τ_{od}. The used controls are geographic distance, indicators for free trade agreement, currency union, common border, common legal system, common language and colonial ties as well as dummies grouping North-North, South-South and North-South trade relationships. Including origin and destination fixed effects implies that the marginal effect of demand similarity on trade is identified using variation across exporters for a given importer, and using variation across importers for a given exporter. Hence, the fixed effects specification isolates the effect of demand similarity from origin and destination specific characteristics such as production technology and gravity forces such as trade costs and economic sizes. As the rationale for our hypothesis is driven by the demand side, we only include HS 6-digit codes corresponding to consumer goods. There are 1'263 HS 6-digit codes including consumer goods, which corresponds to 27% of the value of world trade, in 2002. Examples for such HS 6-digit product categories are cars with large cylinder capacity, microwave ovens or cashew nuts (Bernasconi (2013)).

BACI trade margins and overlap

To reconstruct Bernasconi's (2013) preferred sample with WIID version 3.4, we exclude countries with a population smaller than one million in order to avoid very small countries dominating the sample.[18] Our reconstructed sample accounts for 94% of the value of worldwide trade in consumer goods. Table 1 reports the results of regressing bilateral trade margins on the income distribution overlap as specified in equation (1) via OLS using BACI trade data for 93 countries in 2002 and overlaps constructed from WIID3.4 with timespan 1992-2002.[19] The mean overlap in this sample is 0.53, with a standard deviation of 0.31. Estimating the model in logs – to linearize the multiplicative gravity equation – runs into issues due to many bilateral trade flows being zero, $V_{od} = 0$. An easy fix to deal with these zeros is to apply the linear transformation $Y_{od} + 1$. A more sophisticated Pseudo Maximum Likelihood approach is discussed below.[20]

The three columns show that on average a higher income distribution overlap increases the value of bilateral trade V_{od}, increases the number of products traded EM_{od} and increases the average value of trade per product IM_{od}. The regressions are calculated in logs on the standardized trade margins introduced in Section 2.2. However, to get a sense of the magnitude, it is useful to be aware of the unstandardized levels. The average yearly export value per aggregate bilateral trade flow $\widetilde{V_{od}}$ is 163 million USD (SD 1446 million USD). The average number of HS6 codes with positive flows from origin to destination country $\widetilde{EM_{od}}$ is 112 (SD 220). The average yearly average export value per HS6 code with positive flows from origin to destination country $\widetilde{IM_{od}}$ is 300'000 USD (SD 1.5 million USD).

The first column implies that a one standard deviation increase in overlap (+0.31) is associated with a 66% increase in the bilateral trade value.[21] Respectively, a 0.1 SD increase in $ln(V_{od})$.[22] The second column implies that a 1 SD increase in overlap (+0.31) is associated with 31% more varieties traded between the two countries. Respectively, a 0.09 SD increase in $ln(EM_{od})$. The third column implies that a 1 SD increase in overlap (+0.31) is associated with a 27% increase in the average value of bilateral trade per product. Respectively, a 0.1 SD increase in $ln(IM_{od})$. Hence, effects are not only statistically signification but also

[18] We do not discard observations with a value of less than US$2'000 but instead include trade zeros in our main specification.

[19] This yields $93 * 92 = 8'556$ potential trade flow observations of which 6'455 have a positive value.

[20] Alternatively, we can also drop the zeros and estimate the model only for positive trade flows with OLS. Table 5 in the appendix shows that this does not alter main findings.

[21] $100\,(exp(\beta * \sigma(O_{nc})) - 1))\% = 100\,(exp(1.629 * 0.31) - 1)\,\% = 66\%$

[22] $\beta * \sigma(O_{nc})/\sigma(ln(V_{od})) = 1.629 * 0.31/5.3 = 0.1$

economically relevant.

These regressions confirm that the income distribution overlap is highly predictive for the intensive and extensive margin of trade. Due to the relationship $V_{od} = EM_{od}IM_{od}$, the coefficients on the extensive and intensive margin add up to the overall effect when estimating the model in logs. In our main specification the extensive margin contributes 53% and the intensive margin 47% to higher trade flow values due to more similar income distributions.

Table 1: Main specification – Trade margins BACI and income overlap

	$ln(V+1)$	$ln(EM+1)$	$ln(IM+1)$
Overlap	1.629***	0.866***	0.763***
	(0.170)	(0.092)	(0.098)
# obs	8,556	8,556	8,556
# regs	193	193	193
Adj R^2	0.805	0.825	0.693
Mean	5.294	2.047	3.247
SD	5.277	3.037	2.434

Notes: ***, **, * denote statistical significance on the 1%, 5%, and 10% level, respectively. Robust standard errors are given in parentheses. Controls: geographic distance, dummies for free trade agreement, currency union, common border, common legal system, common language, colonial ties, dummy variable allowing for a different intercept for NN, SS, NS and SN trade flows, importer and exporter fixed effects. Sample: countries with population > 1 million, HS6 codes which include consumer goods. Year=2002.

It is a well known fact that a substantial fraction of potential trade flows on product level are zero.[23] In our BACI sample the share of zero flows is 25%. The state of the art method to deal with zeros in bilateral trade flow data is to apply Poisson Pseudo Maximum Likelihood (PPML) estimation as introduced by Santos Silva and Tanreyro (2006). This takes care of the potential selection problem from omitting the zeros. There might be selection bias since country-pairs with very high trade costs do not trade or only trade very small values that are rounded down to zero.[24] Hence, omitting zeros underestimates trade costs as the highest trade costs are removed from the sample. Santos Silva and Tenreyro (2006) suggest a Poisson Pseudo-Maximum Likelihood estimator which proved to perform well even when the data has a large proportion of zeros. From econometric theory we know that PML, using a density from the linear exponential family – which Poisson satisfies – is consistent as long as the mean is correctly specified. Therefore, we should trust the PML estimates more compared to OLS. Table 4 in the appendix reports the estimates on the same data as used used for Table 1 estimating PPML on levels instead of OLS on logs. Surprisingly, the coefficients are not significant using the PPML estimation. However, using the alternative trade data source EDD the PPML estimates look more reasonable as shown in Section 2.3. Hence, this odd behavior might be an issue only occurring with the BACI data that has to be investigated further.

EDD trade margins and overlap

We use the EDD as an alternative trade flow data source and show that the main findings documented in the BACI data also hold in this data. Analogous to Table 1, Table 2 reports the results of regressing bilateral trade margins on the income distribution overlap via OLS, using EDD trade data for 37 origins and 104 destinations in 2012 and overlaps constructed from WIID3.4 with timespan 2002-2012.[25] We again use the

[23] See Hummels an Klenow (2005) for a discussion.
[24] Trade flows might also be zero due to misrecorded missing observations.
[25] This yields $37 * 104 = 3'848$ potential trade flow observations of which 2'379 have a positive value.

$Y_{od}+1$ transformation for our main specification to deal with the export zeros but below also report the PPML estimate.[26] We use the year 2012 instead of 2002 for the EDD data since the number of countries covered is highest in this year while it would be very low in 2002. The mean overlap in this sample is 0.52 (SD 0.27). The average yearly export value per aggregate bilateral trade flow V_{od} is 230 million USD (SD 3539 million USD). The average number of HS6 codes with positive flows from origin to destination country EM_{od} is 414 (SD 2018). The average yearly average export value per HS6 code with positive flows from origin to destination country IM_{od} is 260'000 USD (SD 1.3 million USD).

The first column implies that a one standard deviation increase in overlap (+0.27) is associated with a 166% increase in bilateral trade value.[27] Respectively, a 0.06 SD increase in $ln(\widetilde{V_{od}})$.[28] The second column implies that 1 SD increase in overlap (+0.31) is associated with 33% more varieties traded between the two countries. Respectively, a 0.03 SD increase in $ln(\widetilde{EM_{od}})$. The third column implies that 1 SD increase in overlap (+0.31) is associated with a 101% increase in the average value of bilateral trade per product. Respectively, a 0.06 SD increase in $ln(\widetilde{IM_{od}})$. Hence, also using the EDD trade data, estimated effects are not only statistically signification but also economically relevant. Moreover, the fact that we also find a strongly positive relationship between income similarity and trade margins a decade later suggest that this relationship persists over time.

Table 2: Main specification – Trade margins EDD and income overlap

	$ln(V_{od}+1)$	$ln(EM_{od}+1)$	$ln(IM_{od}+1)$
Overlap	3.630***	1.061*	2.569***
	(0.972)	(0.637)	(0.921)
# obs	3,848	3,848	3,848
# regs	149	149	149
Adj R^2	0.630	0.545	0.324
Mean	1.783	-1.616	3.399
SD	17.880	10.157	12.153

Notes: ***, **, * denote statistical significance on the 1%, 5%, and 10% level, respectively. Robust standard errors are given in parentheses. Controls: geographic distance, dummies for free trade agreement, currency union, common border, common legal system, common language, colonial ties, dummy variable allowing for a different intercept for NN, SS, NS and SN trade flows, importer and exporter fixed effects. Sample: countries with pop > 1 million, HS2 codes which include consumer goods. Year=2012.

Table 3 reports the estimates on the same EDD dataset as used for Table 2 but using PPML instead of OLS. All coefficients still have the same sign and a similar order of magnitude but are less statistically significant when using PPML estimation.[29]

Running the same augmented gravity regressions on disaggregated HS 6-digit product level, confirms the findings from the aggregate level regressions.[30] Our results are in line with Bernasconi (2013) and provide further support for the Linder hypothesis. Similarity of demand patterns, approximated by the income distribution overlap, is an important determinant of trade flows in consumer goods. This is driven both by the intensive and the extensive margin. The overlap in the income distribution of two countries has a

[26] Instead of applying the $Y_{od}+1$ transformation we can also drop the zeros and estimate the model only for positive trade flows with OLS. Table 6 in the appendix shows that this does not alter main findings.

[27] $100\,(exp(\beta * \sigma(O_{nc}))-1))\% = 100\,(exp(3.630 * 0.27)-1)\,\% = 166$

[28] $\beta * \sigma(O_{nc})/\sigma(ln(\widetilde{V_{od}})) = 3.630 * 0.27/17.9 = 0.06$

[29] To compare the magnitude one has to take into account that OLS is estimated in logs but PPML in levels. Hence, for the extensive margin a one standard deviation increase in the overlap increases the number of traded varieties by $(exp(\beta_{PMML} * \sigma(O_{nc}))-1) * \mu(EM_{od}) = 0.33 * 414 = 137$ using OLS and by $\beta_{OLS} * \sigma(O_{nc}) = 0.491 * 0.27 = 132$ using PPML.

[30] HS 6-digit results are available on request.

Table 3: Trade margins EDD and income overlaps, PPML estimation

	V_{od}	EM_{od}	IM_{od}
Overlap	0.514*	0.491***	0.327
	(0.275)	(0.173)	(0.322)
# obs	3,848	3,848	3,848
# regs	149	149	149
Mean	2.3e+08	413.7	256339.8
SD	3.5e+09	2017.6	1300881.3

Notes: ***, **, * denote statistical significance on the 1%, 5%, and 10% level, respectively. Robust standard errors are given in parentheses. Controls: geographic distance, dummies for free trade agreement, currency union, common border, common legal system, common language, colonial ties, dummy variable allowing for a different intercept for NN, SS, NS and SN trade flows, importer and exporter fixed effects. Sample: countries with pop $>$ 1 million, HS2 codes which include consumer goods. Year=2012.

statistically and economically significant positive impact on the margins of bilateral trade.

3 Model setup

Let us consider two countries trading consumption goods with each other. We refer to them as North and South, where the Southern country's variables will be denoted by an asterisk. The setup on the consumer and firm side is taken from FHZ with the key difference that we introduce within-country income inequality.

3.1 Consumers and labor market

The two economies' population sizes are $\mathcal{P}$ and $\mathcal{P}^*$. We assume an inelastic labor supply and hence this coincides with the economies' labor forces. Labor is perfectly mobile within countries and completely immobile across countries. The labor market is competitive. Hence, workers are paid their marginal product. Choosing the Northern wage to be the numéraire, we can define relative wages as $\omega = \frac{W^*}{W} = W^*$.

The crucial feature of our model is that consumers within a country vary in their labor endowment. A type θ consumer is endowed with θ units of labor. The poorest household in the country has an endowment of $\underline{\theta}$ and the richest of $\overline{\theta}$, with $\underline{\theta} < \overline{\theta}$. The labor endowment distribution can be discrete or continuous. To build intuition, we here present the discrete case for a world with two and three distinct endowment levels. Households have identical preferences and spend their entire income $W\theta$ on a continuum of differentiated consumption goods. Goods are indivisible and only yield positive utility for the first unit. This is a sensible assumption for durable consumer goods such as cars, bicycles, domestic appliances, computers, cellphones, furniture, etc. Hence, the consumption decision for each good is a binary choice whether to purchase one unit of the good or not to purchase it at all. Such 0-1 preferences can be expressed as

$$U = \int_0^\infty c(j)dj, \quad \text{where } c(j) \in \{0, 1\}.$$

These preferences are additively separable and symmetric as all goods provide the same amount of utility. 0-1 preferences only have an extensive but no intensive margin. If the budget of a household increases, she prefers to buy additional varieties not consumed so far rather than buying a larger quantity of the already consumed varieties. A household's utility is therefore simply given by the number of consumed varieties.

This allows to formulate the autarky consumer problem

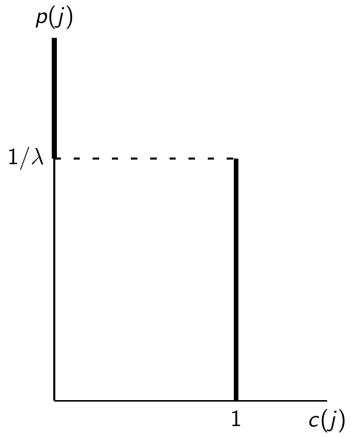


Figure 3: Individual demand function

$$max_{\{c(j)\}} \int_0^{\infty} c(j)dj \qquad s.t. \int_0^{N} p(j)c(j)dj \leq W\theta \qquad [\lambda(\theta)],$$

where $\lambda(\theta)$ is the Lagrange multiplier attached to the budget constraint. It can therefore be interpreted as household θ's marginal utility of income. Taking first order conditions, yields the simple binary purchasing decision rule

$$\begin{aligned} c(j) &= 1 \quad if \quad 1 \geq \lambda(\theta)p(j) \\ c(j) &= 0 \quad if \quad 1 < \lambda(\theta)p(j). \end{aligned}$$

The household will purchase one unit of good j if his willingness to pay $\frac{1}{\lambda(\theta)}$ is at least as high as the good's price $p(j)$. Figure 3 illustrates the individual demand function arising from this utility function. Individual demand is a one-step function of the price with the reservation price being the inverse of the consumer's marginal utility of income. Appendix B briefly discusses the extension of adding an intensive margin and a hierarchy in the order products are consumed. Non-homothetic preferences and demand side heterogeneity are needed to activate the extensive consumption margin. Under homothetic preferences, income inequality does not matter for aggregate outcomes since rich consumer behave like a scaled version of poor and the income distribution can simply be summarized by a representative agent.

3.2 Firms

The supply side is standard New Trade as in Krugman (1980). There is a continuum of firms each having a blueprint to produce a single, differentiated consumption good. Labor is the only factor of production. All firms have identical technology. It requires a fixed labor input F to set up a new firm and variable labor input $1/a$ to produce one unit of output. Firms can sell their product in the domestic and in the export market. To ship one unit of the consumption good to the other country iceberg trade costs τ accrue. The presence of increasing returns to scale in production technology combined with trade costs generate a Krugman-style home market effect. There is monopolistic competition among these homogeneous firms and free entry into the market. Hence, zero-profit conditions are endogenously determining the number of active

firms in equilibrium.

For simplicity let us first consider the autarky situation. Each firm chooses a marginal consumer θ by setting its price $p(\theta)$. All consumers with endowment at least as large as a particular firm's marginal consumer purchase one unit of the firm's product.[31] For a continuous labor endowment distribution, there always $\exists\theta$ *s.t.* $\frac{1}{\lambda(\theta)} = p(\theta)$. Take for example a Northern firm that sells its product to all households in the North. The profit of such a firm is $p(\underline{\theta})\mathcal{P} - W\left(F + \frac{\mathcal{P}}{a}\right)$. This firm sets its price such that it breaks-even in equilibrium and this price is equal to the marginal willingness to pay of the poorest Northern consumer, $p(\underline{\theta}) = \frac{aF+\mathcal{P}}{a\mathcal{P}} = \frac{1}{\lambda(\underline{\theta})}$. The marginal utility of income is therefore an equilibrium object that is determined by technology parameters. The marginal utility of income is high when products are cheap since then a consumer can buy many more varieties with one additional unit of income and thus far expand his consumption margin, which directly materializes in utility. Goods are cheap when the fixed cost of setting up a firm is low and the productivity of the firm, a, is high. Moreover, the price is decreasing in market size since fixed costs can then be spread over more consumers in larger markets. Thus, within-country inequality affects the demand structure via heterogeneous willingness to pay. Rich households have a higher willingness to pay for a product than poor households and therefore consume a larger variety of goods. This implies that the market size and hence the demand for cheap products is large. Vice versa, for expensive products the market size is small. We have shown that prices are determined by technology parameters and market size and that they relate to consumers marginal utility of income.

Note that our model features endogenous markups.[32] Markups vary in market size. The markup for products with small markets is larger (luxury products) than the markup for product with a large market (necessity products) since marginal costs are identical but the price decreases in market size. The markup of a firm serving the entire local market is $\mu = \frac{p(\underline{\theta})}{1/a} = \frac{aF+\mathcal{P}}{\mathcal{P}}$.

In the case of international trade, a firm has to choose a marginal home consumer θ and a marginal foreign consumer θ^* by setting its home price $p(\theta)$ and its export price $p^*(\theta^*)$. The ability to set different prices in the domestic and foreign market is often referred to as *pricing-to-market*. Monopolistically competitive firms use this price discrimination to take advantage of differences in willingness to pay across countries.[33]

4 Two endowment groups

The simplest case of a within-country income distribution is to have two distinct endowment groups: poor and rich. Poor households are endowed with $\theta_P = \underline{\theta}$ units of labor and rich households with $\theta_R = \bar{\theta}$. We keep the endowment levels of the rich and the poor identical across the two countries and fix the number of consumers in the two endowment groups on a global level. This implies world population and total world endowment to be constant.

To study within- versus between-country inequality, we only vary the share of the world population in the two groups that is allocated to the North and the South. We define α_P to be the Northern share of the world's poor consumers $\mathcal{P}_P$ and α_R to be the Northern share of the world's rich consumers $\mathcal{P}_R$. Hence, the corresponding Southern shares are $\alpha_P^* = 1 - \alpha_P$ and $\alpha_R^* = 1 - \alpha_R$. We assume that the Northern country is richer than the Southern country, i.e. the majority of poor consumers lives in the South $\alpha_P \leq 1/2$.

[31] The marginal consumer is the poorest consumer that purchases the product. Which marginal consumers can be targeted depends on the specific labor endowment distribution.

[32] There is an active literature on endogenous markups, the rise of market power and trade, see De Loecker and Eeckhout (2017), Melitz and Ottaviano (2008).

[33] Krugman (1986) was the first to study pricing-to-market. He documented that European firms do not fully pass through movements in the exchange rate to US importers but rather maintain their export prices tied to local US prices.

Additionally, we keep the *population* of the two countries constant and identical, $\mathcal{P} = \mathcal{P}^*$. This implies

$$\alpha_R = \frac{\mathcal{P} - \alpha_P \mathcal{P}_P}{\mathcal{P}_R}.$$

Hence, choosing α_P determines α_R.
Alternatively, we can also keep country *endowments* constant and identical $E = E^*$ such that we only manipulate population versus GDP per capita when changing the allocation shares. Country endowments are $E \equiv \alpha_P \mathcal{P}_P \underline{\theta} + \alpha_R \mathcal{P}_P \bar{\theta}$ and $E^* \equiv (1-\alpha_P)\mathcal{P}_P \underline{\theta} + (1-\alpha_R)\mathcal{P}_P \bar{\theta}$. This implies

$$\alpha_R = \frac{E - \alpha_P \mathcal{P}_P \underline{\theta}}{\mathcal{P}_P \bar{\theta}}, \quad \mathcal{P} = \alpha_P \mathcal{P}_P + \alpha_R \mathcal{P}_R \quad \text{and} \quad \mathcal{P}^* = \mathcal{P}_P + \mathcal{P}_R - \mathcal{P}.$$

Hence, choosing α_P determines $\mathcal{P}$, $\mathcal{P}^*$ and α_R. Keeping country endowments constant and identical allows for a direct comparison to the case of CES preferences, where the effects of population versus GDP per capita are indistinguishable due to homotheticity. Either countries' endowment or countries' population can be held constant but not both simultaneously.

In either setting, only one parameter, namely α_P, is sufficient to control within- versus between-country inequality. When $\alpha_P = 0$, all poor consumers live in the South and all rich consumers live in the North. Hence, there is only between- but no within-country inequality. This is the case of representative agents used in FHZ. When $\alpha_P = 1/2$, half of the poor and the rich consumers live in the South and the same mirrored in the North. The two countries are identical. Hence, there is only within- but no between-country inequality.

We use the Gini coefficient for summarizing the degree of within-country inequality. For a discrete probability distribution with probability mass function $f(y_i), i = 1, ..., n$, where $f(y_i)$ is the fraction of the country's population with income equal to $y_i > 0$, the Gini coefficient can be calculated as

$$G = 1 - \frac{\sum_{i=1}^{n} f(y_i)(S_{i-1} + S_i)}{S_n},$$

where $S_i = \sum_{j=1}^{i} f(y_j) y_j$ and $S_0 = 0$ (Silber (1999)).

There is a direct correspondence between within- versus between-country inequality and the income distribution overlap in our model. For identical county populations, setting parameters $\mathcal{P}_P = \mathcal{P}_R = \mathcal{P} = \mathcal{P}^* = 1$, or for identical country endowments, setting parameters $E = E^* = \mathcal{P}_R = \bar{\theta} = 1$, $\mathcal{P}_P = 2$, $\underline{\theta} = \frac{1}{2}$, it holds that $\alpha_R = 1 - \alpha_P$. Hence, the overlap between country o and country d can be expressed as

$$O_{od} = 2\alpha_P \quad \text{for } \alpha_P \leq 1/2.$$

The nominal GDP of the Northern country is $Y = E$. Similarly, nominal GDP of Southern country is $Y^* = \omega E^*$. We define trade intensity, ϕ, as the ratio between world trade value and world GDP

$$\phi = \frac{TVA}{Y + Y^*},$$

where TVA is the total value of goods traded worldwide. I denote the consumption of varieties by group i in country j by C_i^j. Furthermore, world trade volume TVO is defined as the unit count of goods traded worldwide.

4.1 Firm types

We distinguish different types of firms depending on which market segments they decide to serve. We denote firm types with two indices, ij. The first index i indicates the poorest consumer group served in the firm's local market and the second index j the poorest consumer group served in the firm's export market. For example, a PR firm sells to everyone in the local market but only to the rich in the export market.

For two endowment groups, there are four relevant types of firms potentially active in equilibrium.[34] (i) A *mass* producer PP sells to all consumers in both markets, sets prices p_P and p_P^*, respectively, and earns profits Π_{PP}. (ii) A *separating* producer PR sells to all consumers in the local market but only to the rich in the export market, sets prices p_P and $min\{\tau p_P p_R^*\}$, respectively, and earns profits Π_{PR}. (iii) An *exclusive* producer RR only sells to the rich in both markets, sets prices p_R and p_R^*, respectively, and earns profits Π_{RR}. (iv) Finally, a *domestic* rich firm R only sells to the rich in the local market and does not export, sets price p_R, and earns profits Π_R. The Northern firms' profits are

$$\begin{aligned}
\Pi_{PP} &= p_P\mathcal{P} + p_P^*\mathcal{P}^* - W\left(F + \frac{\mathcal{P} + \tau\mathcal{P}^*}{a}\right) \\
\Pi_{PR} &= p_P\mathcal{P} + min\left\{\tau p_P, p_R^*\right\}(1-\alpha_R)\mathcal{P}_R - W\left(F + \frac{\mathcal{P} + \tau(1-\alpha_R)\mathcal{P}_R}{a}\right) \\
\Pi_{RR} &= p_R\alpha_R\mathcal{P}_R + p_R^*(1-\alpha_R)\mathcal{P}_R - W\left(F + \frac{\alpha_R\mathcal{P}_R + \tau(1-\alpha_R)\mathcal{P}_R}{a}\right) \\
\Pi_R &= p_R\alpha_R\mathcal{P}_R - W\left(F + \frac{\alpha_R\mathcal{P}_R}{a}\right).
\end{aligned} \tag{2}$$

The conditions for Southern firms are :

$$\begin{aligned}
\Pi_{PP}^* &= p_P\mathcal{P} + p_P^*\mathcal{P}^* - \omega\left(F + \frac{\tau\mathcal{P} + \mathcal{P}^*}{a}\right) \\
\Pi_{PR}^* &= min\left\{\tau p_P^*, p_R\right\}\alpha_R\mathcal{P}_R + p_P^*\mathcal{P}^* - \omega\left(F + \frac{\tau\alpha_R\mathcal{P}_R + \mathcal{P}^*}{a}\right) \\
\Pi_{RR}^* &= p_R\alpha_R\mathcal{P}_R + p_R^*(1-\alpha_R)\mathcal{P}_R - \omega\left(F + \frac{\tau\alpha_R\mathcal{P}_R + (1-\alpha_R)\mathcal{P}_R}{a}\right) \\
\Pi_R^* &= p_R^*(1-\alpha_R)\mathcal{P}_R - \omega\left(F + \frac{(1-\alpha_R)\mathcal{P}_R}{a}\right).
\end{aligned} \tag{3}$$

Due to free firm entry, the firm strategies must break even in equilibrium. If mass production firms are present in both countries, the relative wage is given by

$$\omega = \frac{aF + \mathcal{P} + \tau\mathcal{P}^*}{aF + \tau\mathcal{P} + \mathcal{P}^*}. \tag{4}$$

The relative wage is determined by the population size only as the relative labor endowment does not play any role. Technology is identical in both countries. Hence, any difference in wage is only due to economies of scale depending on population size. If both countries are of equal size, $\mathcal{P} = \mathcal{P}^*$, factor price equalization holds and the relative wage equals one.

[34] In principle, a firm can target many possible market segments (various combinations of rich/poor and domestic/export). However, only the four presented strategies are relevant.

4.2 Parallel trade allowed

We next present the equilibria for the case when parallel imports are allowed and trade costs are relatively low such that firms have to take arbitrage considerations into account. This is the empirically relevant case as Simonovska (2015) found that international-arbitrage constraints are binding for more than 20 percent of bilateral trade flows, which account for over 45 percent of world trade volume. In Section 4.3, we briefly characterize the two-group equilibrium structure when parallel imports are prohibited and a hybrid case where there is imperfect enforcement of parallel import prohibition.

4.2.1 Symmetric countries (PI1)

Let us first look at the case of symmetric countries, $\alpha_P = \alpha_R = 1/2$, for constant and identical world income groups and country populations, $\mathcal{P}_P = \mathcal{P}_R = \mathcal{P} = \mathcal{P}^* = 1$. For endowment levels $\underline{\theta} = 1/2$ and $\bar{\theta} = 3/2$, this yields Gini coefficients of $G = G^* = 0.25$. Prices, wages and the number of varieties are identical between the two countries due to symmetry. We consider relatively low trade costs such that the arbitrage constraint is binding, $p_R \leq \tau p_P^*$.[35] The threat of parallel imports limits the price differentiation of separating firms PR to factor τ. This implies that $\Pi_{PR} < (\Pi_{PP} + \Pi_{RR})/2$, whenever τ is sufficiently low. Therefore, no separating firms are active in equilibrium. In equilibrium, all goods are traded and there are only mass PP and exclusive RR firms in both countries. There is factor price equalization, $\omega = 1$, due to the existence of mass firms in both countries and $\mathcal{P} = \mathcal{P}^*$.
Zero profit conditions are

$$2p_P = W\left(F + \frac{1+\tau}{a}\right) \quad \text{and} \quad p_R = W\left(F + \frac{1+\tau}{2a}\right).$$

Incremental budget constraints are

$$2p_P N_{PP} = \underline{\theta} W \quad \text{and} \quad 2p_R N_{RR} = (\bar{\theta} - \underline{\theta})W.$$

Trade balance is

$$p_P N_{PP}^* + p_R \frac{1}{2} N_{RR}^* = p_P^* N_{PP} + p_R^* \frac{1}{2} N_{RR}.$$

The resource constraint is

$$\frac{\underline{\theta} + \bar{\theta}}{2} = N_{PP}\left(F + \frac{1+\tau}{a}\right) + N_{RR}\left(F + \frac{1+\tau}{2a}\right).$$

This system of equations can be solved in closed form. The prices, obtained from the zero profit conditions, are determined by the respective average costs

$$\begin{aligned} p_P = p_P^* &= \frac{F}{2} + \frac{1+\tau}{2a} \\ p_R = p_R^* &= F + \frac{1+\tau}{2a}. \end{aligned}$$

[35] The sufficient condition on τ is given by equation (6) derived below.

The varieties, obtained from the budget constraints, are

$$\begin{aligned} N_{PP} = N^*_{PP} &= \frac{\underline{\theta}}{F + \frac{1+\tau}{a}} \\ N_{RR} = N^*_{RR} &= \frac{\bar{\theta} - \underline{\theta}}{2F + \frac{1+\tau}{a}}. \end{aligned}$$

Finally, welfare is given by

$$U_P = \frac{2\underline{\theta}}{F + \frac{1+\tau}{a}} \quad \text{and} \quad U_R = \frac{2\underline{\theta}}{F + \frac{1+\tau}{a}} + \frac{\bar{\theta} - \underline{\theta}}{F + \frac{1+\tau}{2a}}.$$

Location of production The location of production is indeterminate under perfect symmetry. While a symmetric division of PP and RR firms between the two identical countries is a natural outcome, also the continuation of the regimes prevailing under a infinitesimal deviation from $\alpha_P = 0.5$ is appealing. These regimes will be introduced in Sections C.2 and 4.2.2 below. Depending on whether $\bar{\theta}/\underline{\theta}$ is high or low the relevant asymmetric regime is PI4 or PI2, respectively (as seen in Figure 4).[36] Prices, trade and welfare are independent of the production location allocation.

Conditions for no deviation incentives We show that profits of all other strategies are weakly negative. The condition for the arbitrage constraint $p_R \leq \tau p^*_P$ to be binding is

$$\tau \leq \frac{1}{2}\left(\sqrt{a^2F^2 + 8aF + 4} - aF\right). \tag{5}$$

Binding arbitrage implies that separating firms PR cannot set domestic and export prices more than factor τ apart. However, for PR firms to be unprofitable, the arbitrage constraint has to be binding sufficiently strongly such that this pricing restriction is strong enough. The condition to avoid deviation to PR firms is

$$\tau \leq \frac{1}{2}\left(\sqrt{a^2F^2 + 10aF + 9} - aF - 1\right). \tag{6}$$

We also show that there are no domestic firms in equilibrium (P or R).

$$\begin{aligned} \Pi_P &= p_P\mathcal{P} - \left(F + \frac{\mathcal{P}}{a}\right) \leq 0 \qquad (7) \\ &if \quad \tau \leq aF + 1. \\ \Pi_R &= p_R\frac{\mathcal{P}}{2} - \left(F + \frac{\mathcal{P}}{2a}\right) \leq 0 \qquad (8) \\ &if \quad \tau \leq 2aF + 1. \end{aligned}$$

Conditions (7) and (8) are weaker conditions than (5) and hence always hold when arbitrage is binding. Thus, the condition for PI1 to be an equilibrium is given by (6).

Comparative statics for $\alpha_P = 0.5$ We provide comparative statics for symmetric countries with respect to the degree of within-country inequality, which in this case is controlled by $\underline{\theta}$ and trade costs, τ. Prices are constant in inequality since they are determined by endowment-invariant zero profit conditions.

[36] In fact, any convex combination between the symmetric and the asymmetric allocation of firms is an equilibrium for $\alpha_P = 0.5$.

Inequality: $\frac{\partial C_P}{\partial \underline{\theta}} = \frac{\partial C_P^*}{\partial \underline{\theta}} > 0$, $\frac{\partial C_R}{\partial \underline{\theta}} = \frac{\partial C_R^*}{\partial \underline{\theta}} < 0$, $\frac{\partial TVO}{\partial \underline{\theta}} > 0$, $\frac{\partial TVA}{\partial \underline{\theta}} = 0$, $\frac{\partial Y}{\partial \underline{\theta}} = \frac{\partial Y^*}{\partial \underline{\theta}} = 0$, $\frac{\partial \phi}{\partial \underline{\theta}} = 0$

Lower within-country inequality increases welfare for the poor but harms the rich. This is due to the poor gaining endowment while the rich loosing endowment. Lower within-country inequality increases trade volume but leaves trade value and GDP unaffected and hence also trade intensity.

Trade costs: $\frac{\partial C_P}{\partial \tau} = \frac{\partial C_P^*}{\partial \tau} < 0$, $\frac{\partial C_R}{\partial \tau} = \frac{\partial C_R^*}{\partial \tau} < 0$, $\frac{\partial TVO}{\partial \underline{\theta}} < 0$, $\frac{\partial TVA}{\partial \tau} = 0$, $\frac{\partial Y}{\partial \tau} = \frac{\partial Y^*}{\partial \tau} = 0$, $\frac{\partial \phi}{\partial \tau} = 0$

A trade liberalization increases welfare of both poor and rich. This is due to less of the gross output being lost during transportation which leaves more output for consumption. A trade liberalization increases trade volume while leaving trade value and GDP unaffected.

4.2.2 Mostly within-country inequality (PI2)

Let us next consider the case when most inequality is *within* countries. The Northern country has somewhat more rich consumers than the Southern country, $\alpha_P \in (\widetilde{\alpha}_P, \frac{1}{2})$, where $\widetilde{\alpha}_P$ denotes the cutoff below which Regime PI3 prevails.[37] For endowment levels $\underline{\theta} = 1/2$ and $\bar{\theta} = 3/2$ and $\alpha_P = 3/8$, this yields a Gini coefficient of $G = 0.21$ in the North and $G^* = 0.27$ in the South. Exclusive markets – those for products with high prices that only the rich can afford – are served only by Northern firms. This is due to more rich consumers living in the North (cf. Krugman home market effect). Mass markets – those for products with low prices that everyone can afford – are served by both Southern and Northern producers. All goods are traded and there are PP and RR firms in the North and only PP firms in the South.
Zero profit conditions (ZPs) are

$$\begin{aligned} p_P + p_P^* &= W\left(F + \tfrac{1+\tau}{a}\right) \\ p_P + p_P^* &= W^*\left(F + \tfrac{1+\tau}{a}\right) \\ p_R(1-\alpha_P) + p_R^*\alpha_P &= W\left(F + \tfrac{1-\alpha_P+\tau\alpha_P}{a}\right). \end{aligned}$$

Incremental budget constraints (BCs) are

$$\begin{aligned} p_P(N_{PP} + N_{PP}^*) &= \underline{\theta} \qquad \text{and} \qquad p_R N_{RR} = (\bar{\theta} - \underline{\theta})W, \\ p_P^*(N_{PP} + N_{PP}^*) &= \underline{\theta} W^* \qquad \text{and} \qquad p_R^* N_{RR} = (\bar{\theta} - \underline{\theta})W^*. \end{aligned}$$

Trade balance is

$$p_P N_{PP}^* = p_P^* N_{PP} + p_R^* \alpha_P N_{RR}.$$

The resource constraints (RCs) are

$$\begin{aligned} \alpha_P\underline{\theta} + (1-\alpha_P)\bar{\theta} &= N_{PP}\left(F + \tfrac{1+\tau}{a}\right) + N_{RR}\left(F + \tfrac{1-\alpha_P+\tau\alpha_P}{a}\right) \\ (1-\alpha_P)\underline{\theta} + \alpha_P\bar{\theta} &= N_{PP}^*\left(F + \tfrac{1+\tau}{a}\right). \end{aligned}$$

[37] We need to restrict the gap between rich and poor consumer endowment to not be too large. Otherwise, Regime PI4 prevails instead of Regime PI2.

The prices, obtained from the ZPs and BCs, are

$$\begin{aligned} \omega &= 1 \\ p_P &= p_P^* = \frac{F}{2} + \frac{1+\tau}{2a} \\ p_R &= p_R^* = F + \frac{1-\alpha_P+\tau\alpha_P}{a}. \end{aligned}$$

The varieties, obtained by combining RCs and BCs, are

$$\begin{aligned} N_{PP}^* &= \frac{(1-\alpha_P)\underline{\theta}+\alpha_P\bar{\theta}}{F+\frac{1+\tau}{a}} \\ N_{PP} &= \frac{(1+\alpha_P)\underline{\theta}-\alpha_P\bar{\theta}}{F+\frac{1+\tau}{a}} \\ N_{RR} &= \frac{\bar{\theta}-\underline{\theta}}{F+\frac{1-\alpha_P+\tau\alpha_P}{a}}. \end{aligned}$$

Similar to Regime PI1, there also is a restriction on τ for PI2 to be an equilibrium, which is $\tau \leq \frac{1}{6}\sqrt{9F^2a^2+72Fa+64} - \frac{Fa}{2} - 1/3$. See Appendix C.1 for the derivation of this condition.

4.2.3 Mostly between-country inequality (PI3)

Let us next consider the case when most inequality is *between* countries. The Northern country has mainly rich and the Southern country mainly poor consumers, $\alpha_P \in (0, \widetilde{\alpha}_P)$. For endowment levels $\underline{\theta} = 1/2$ and $\bar{\theta} = 3/2$ and $\alpha_P = 1/8$, this yields Gini coefficients $G = 0.08$ in the North and $G^* = 0.18$ in the South. Exclusive markets are still only served by Northern firms. This is due to more rich consumers living in the North. Mass markets are served by both Southern and Northern producers. Some Southern firms sell to all consumers at home but only to the rich abroad. The threat of parallel exports restrains their price setting. All goods are traded and there are PP and RR firms in the North and PP and PR firms in the South.
Zero profit conditions (ZPs) are

$$\begin{aligned} p_P + p_P^* &= W\left(F + \tfrac{1+\tau}{a}\right) \\ p_P + p_P^* &= W^*\left(F + \tfrac{1+\tau}{a}\right) \\ p_R(1-\alpha_P) + p_R^*\alpha_P &= W\left(F + \tfrac{1-\alpha_P+\tau\alpha_P}{a}\right) \\ \tau p_P^*(1-\alpha_P) + p_P^* &= W^*\left(F + \tfrac{1+\tau(1-\alpha_P)}{a}\right). \end{aligned}$$

Incremental budget constraints (BCs) are

$$\begin{aligned} p_P(N_{PP}+N_{PP}^*) &= \underline{\theta}W \quad \text{and} \quad p_R N_{RR} + \tau p_P^* N_{PR}^* = (\bar{\theta}-\underline{\theta})W, \\ p_P^*(N_{PP}+N_{PP}^*+N_{PR}^*) &= \underline{\theta}W^* \quad \text{and} \quad p_R^* N_{RR} = (\bar{\theta}-\underline{\theta})W^*. \end{aligned}$$

Trade balance is

$$p_P N_{PP}^* + \tau p_P^* N_{PR}^*(1-\alpha_P) = p_P^* N_{PP} + p_R^*\alpha_P N_{RR}.$$

The resource constraints (RCs) are

$$\alpha_P\underline{\theta} + (1-\alpha_P)\bar{\theta} = N_{PP}\left(F + \tfrac{1+\tau}{a}\right) + N_{RR}\left(F + \tfrac{1-\alpha_P+\tau\alpha_P}{a}\right)$$
$$(1-\alpha_P)\underline{\theta} + \alpha_P\bar{\theta} = N^*_{PP}\left(F + \tfrac{1+\tau}{a}\right) + N^*_{PR}\left(F + \tfrac{1+\tau(1-\alpha_P)}{a}\right).$$

This system of equations can be solved in closed form but in the interest of brevity we abstain from plugging in all terms. The prices, obtained from the ZPs and BCs, are

$$\begin{aligned}
\omega &= 1 \\
p_P &= F + \frac{\tau}{a} - \frac{F}{1+\tau(1-\alpha_P)} \\
p^*_P &= \frac{1}{a} + \frac{F}{1+\tau(1-\alpha_P)} \\
p_R &= \frac{\underline{\theta} - \bar{\theta} - \tau p^*_P N^*_{PR}}{N_R} \\
p^*_R &= \frac{\underline{\theta} - \bar{\theta}}{N_{RR}}.
\end{aligned}$$

The varieties, obtained by combining RCs and BCs, are

$$\begin{aligned}
N^*_{PR} &= \frac{\underline{\theta}}{p^*_P} - \frac{\bar{\theta}}{p_P} \\
N^*_{PP} &= \frac{(1-\alpha_P)\underline{\theta} + \alpha_P\bar{\theta} - N^*_{PR}\left(F + \frac{1+\tau(1-\alpha_P)}{a}\right)}{F + \frac{1+\tau}{a}} \\
N_{PP} &= \frac{\underline{\theta}}{p_P} - N^*_{PP} \\
N_{RR} &= \frac{\alpha_P\underline{\theta} + (1-\alpha_P)\bar{\theta} - N_{PP}\left(F + \frac{1+\tau}{a}\right)}{F + \frac{1-\alpha_P+\tau\alpha_P}{a}}.
\end{aligned}$$

The condition for PI3 to be an equilibrium is
$\tau \leq \frac{1}{2}\left(\sqrt{(\alpha_P a(F-1) - Fa + 2a - 1)^2 + 8a} + a(2-\alpha_P) + (\alpha_P - 1)aF - 1\right)$.

4.2.4 Representative agents (PI7)

Let us finally consider the case of only between-country inequality, $\alpha_P = 0$ and $\alpha_R = 1$, meaning there is a representative rich agent in the North and a representative poor agent in the South. This is the arbitrage equilibrium described in FHZ. This yields Gini coefficients $G = G^* = 0$. This regime emerges when Southern consumers are so poor that they can no longer afford to buy all products. Hence, some Northern firms decide to only operated domestically. Only a subset of all goods is traded and there are RP and R firms in the North and only PR firms in the South. The mass firms are constraint by the arbitrage constraint and set $p_P = \tau p^*_P$.
Zero profit conditions (ZPs) are

$$\begin{aligned}
\tau p^*_P + p^*_P &= W\left(F + \tfrac{1+\tau}{a}\right) \\
\tau p^*_P + p^*_P &= W^*\left(F + \tfrac{1+\tau}{a}\right) \\
p_R &= W\left(F + \tfrac{1}{a}\right).
\end{aligned}$$

Budget constraints (BCs) are

$$\tau p_P^*(N_{RP} + N_{PR}^*) + p_R N_R = \bar{\theta} W \quad \text{and} \quad p_P^*(N_{RP} + N_{PR}^*) = \underline{\theta} W^*.$$

Trade balance is

$$\tau p_P^* N_{PR}^* = p_P^* N_{RP}.$$

The resource constraints (RCs) are

$$\bar{\theta} = N_{RP}\left(F + \frac{1+\tau}{a}\right) + N_R\left(F + \frac{1}{a}\right) \quad \text{and} \quad \bar{\theta} = N_{PR}^*\left(F + \frac{1+\tau}{a}\right).$$

The prices, obtained from ZPs, are

$$\begin{aligned} \omega &= 1 \\ p_R &= F + \frac{1}{a} \\ p_P^* &= \frac{F}{1+\tau} + \frac{1}{a}. \end{aligned}$$

The varieties, obtained by combining RCs and BCs, are

$$\begin{aligned} N_{PR}^* &= \frac{\underline{\theta}}{F + \frac{1+\tau}{a}} \\ N_{RP} &= \frac{\tau\underline{\theta}}{F + \frac{1+\tau}{a}} \\ N_R &= \frac{\bar{\theta} - \tau\underline{\theta}}{F + \frac{1}{a}}. \end{aligned}$$

Finally, welfare is given by

$$U_P^* = \frac{\underline{\theta}}{\frac{F}{1+\tau} + \frac{1}{a}} \quad \text{and} \quad U_R = \frac{\underline{\theta}}{\frac{F}{1+\tau} + \frac{1}{a}} + \frac{\bar{\theta} - \tau\underline{\theta}}{F + \frac{1}{a}}.$$

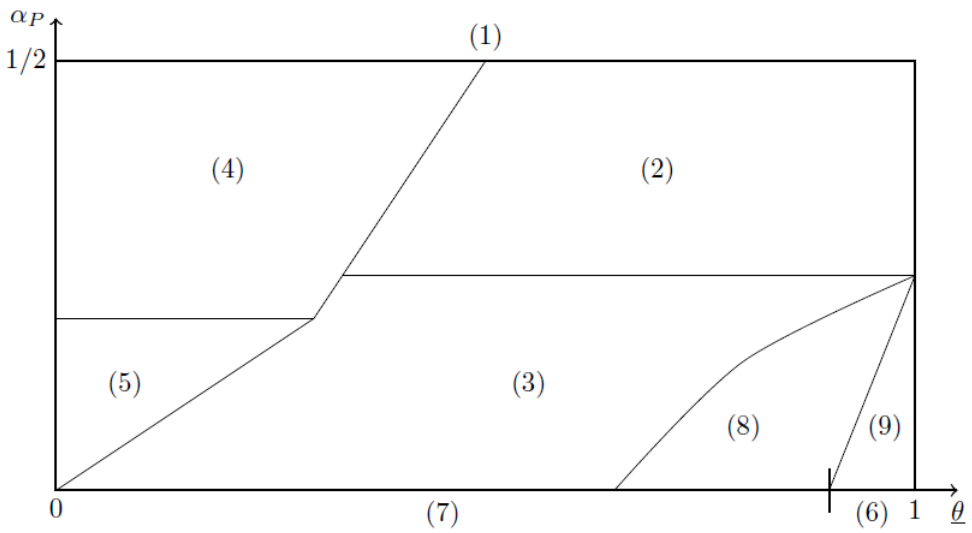
The proof that this is an equilibrium is provided in FHZ Lemma 2 (price setting) and FHZ Proposition 5 (trade condition).

4.2.5 Two-group parallel import equilibria

Which equilibrium regime prevails, also depends on the difference between poor group endowment $\underline{\theta}$ and rich group endowment $\bar{\theta}$. It is therefore useful to draw the different equilibrium regimes in a $(\underline{\theta}, \alpha_P)$ parameter space to get an understanding of how the structure of the equilibria behaves. Figure 4 depicts this when parallel imports are allowed, country population is constant and identical and $\tau = 1.1$, which satisfies all the conditions on τ for the above outlined regimes to be equilibria.[38] Sections C.2 and C.3 in the appendix provides the equilibrium derivation for the regimes PI4 and PI6, respectively.

However, for a closer analysis we focus on a fixed labor endowment level $\underline{\theta} = 1/2$, which covers the four equilibrium regimes PI1, PI2, PI3 and PI7 described in Subsections 4.2.1-4.2.4. Graphically this corresponds

[38] Figure 11 shows the same space with the equilibrium regimes that occur when parallel imports are prohibited.



(1) N: PP+RR S: PP+RR or (2)/(4)
(2) N: PP+RR S: PP
(3) N: PP+RR S: PP+PR
(4) N: RR S: PP+RR
(5) N: RR S: PP+PR+R
(6) N: RP S: PR
(7) N: RP+R S: PR
(8) N: PP+RR S: PP+PR+R
(9) N: PP S: PP+PR+R

Figure 4: Two-group parallel import regimes for constant country population

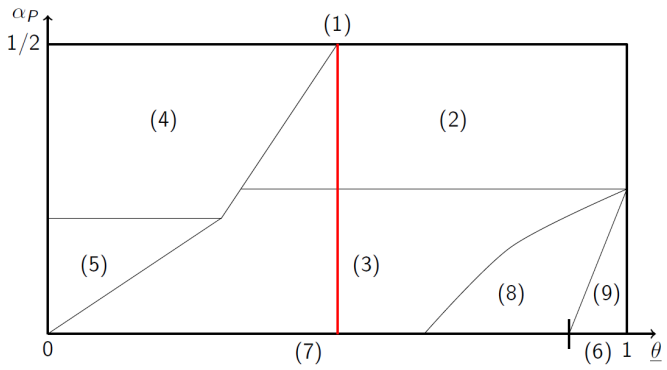


Figure 5: $\underline{\theta} = 1/2$ cross-section through two-group parallel import regimes for constant country population

to taking a cross-section through the $(\underline{\theta}, \alpha_P)$ regimes graph as illustrated by the red line in Figure 5.

Figure 6 shows the four equilibrium regimes occurring over the cross-section at $\underline{\theta} = 1/2$ for $\tau = 1.1$, as we move from only between-country inequality ,$\alpha_P = 0$ (PI7), to only within-country inequality, $\alpha_P = 0.5$ (PI1). This increase in within-country inequality is reflected in the Gini monotonously increasing from PI7 to PI1.

What is the intuition for switching between different regimes? Moving from PI1 to PI2 – from perfectly symmetric countries to the North having somewhat more rich consumers – RR firms disappear in the South and the corresponding number of firms N^*_{RR} relocates to the North, to become N_{RR}. This happens due to a home market effect: all rich-consumer varieties are produced in the location with the majority of rich households. At PI1, the production location of RR goods is indeterminate as firms are indifferent due to equally many rich consumers living in both counties. Moreover, $p_R = p^*_R$ decreases smoothly as α_P decreases, while $p_P = p^*_P$ is independent of α_P.

Moving from PI2 to PI3 – from the North having somewhat more rich consumers to the clear majority of rich consumers living in the North – PR firms appear in the South, whereas under PI2 all Southern firms were PP firms. As α_P decreases, the poor-consumer market size in the North is no longer sufficient for all Southern firms to pursue the PP strategy. Some Southern start to only export to the rich, PR firms. They prefer charging a higher price in the export market – τp^*_P rather than p_P – since there now is a large rich-consumer market in the North. Hence, the price versus market size trade-off tips in favor of the price as the size difference between serving the entire export market and serving only the rich-consumer segment is small now. The logic being: charging high prices from the many rich, while sacrificing the few poor.

Moving from PI3 to PI7 – from the majority of rich consumers living in the North to all rich consumers being located in the North – PP firms disappear in the South and the the corresponding mass of firms N^*_{PP} reallocates to N^*_{PR}. As all rich consumers are now located in the North, a Southern firm that used to serve everyone in the export market – a PP firm – is now equivalent to one only serving the rich in the export market, a PR firm. The same is true for a Northern firm only selling to the rich worldwide – a RR firm – and one only selling domestically, a R firm.

To understand the mechanisms at work, we generated numerical solutions along this cross-section. Figure 7 shows world trade volume TVO, which is defined as the unit count of goods traded worldwide. Trade volume is monotonously increasing in within-country inequality. This is due to varieties targeted to rich consumers having to be shipped to both countries under symmetric countries, whereas they are produced locally when all rich are located in one country. Figure 8 shows world net output, which is the unit count of

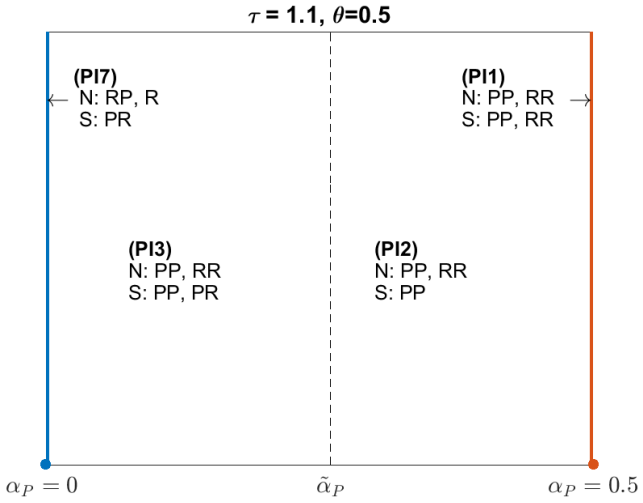


Figure 6: Four regimes of $\underline{\theta} = 1/2$ cross-section

goods produced worldwide that reaches consumers.[39] World net output instead is monotonously decreasing in within-country inequality. This is due to efficiency losses caused by the costly trade of rich-consumer varieties dissipating global resources. Figure 9 shows welfare for poor and rich consumers. Poor-consumer welfare in the North is increasing in income distribution similarity, while it is decreasing in the South. This is due to the poor in the North being excluded from Southern PR products in regime PI3. Rich-consumer welfare is identical in both countries and non-linearly related to income similarity. Rich-consumer welfare is identical to the number of varieties available worldwide since rich consumers purchase all products. Figure 10 shows prices for poor and rich consumers.

Two-group results

These results are derived from numerical simulations. For special, cases there are proofs provided in the appendix.

Result 1. *For sufficiently high endowment differences, a sufficient condition being* $\frac{\bar{\theta}}{\underline{\theta}} > 2\tau - 1$*, world trade volume is increasing in income distribution similarity. However, world net output is always decreasing in income distribution similarity and hence so is world welfare.*

Similar income distributions increase the volume of goods traded since both countries also want to consumer the varieties produced in the other country. The model predicts that trade volume (and also trade intensity) is monotonously increasing in the overlap of the income distribution (recall $O_{od} = 2\alpha_P$), which is in line with the empirical findings. The decrease in world net output is due to the resources lost because of intense costly trade of rich-consumer varieties if there are rich consumers in both countries. For only within-country inequality, many resources are wasted shipping N_{RR} and N^*_{RR} across boarders, while for only between-country inequality, N_R are produced and consumed locally. World welfare is identical to total net production since that is the number of goods consumed globally and each good yields one unit of utility. Total welfare is maximized under representative agents as it minimizes the iceberg losses. Trade is maximized under identical countries, when there is full overlap. The proof for the case of only within- versus only

[39] The gross production includes the part of production lost in iceberg trade.

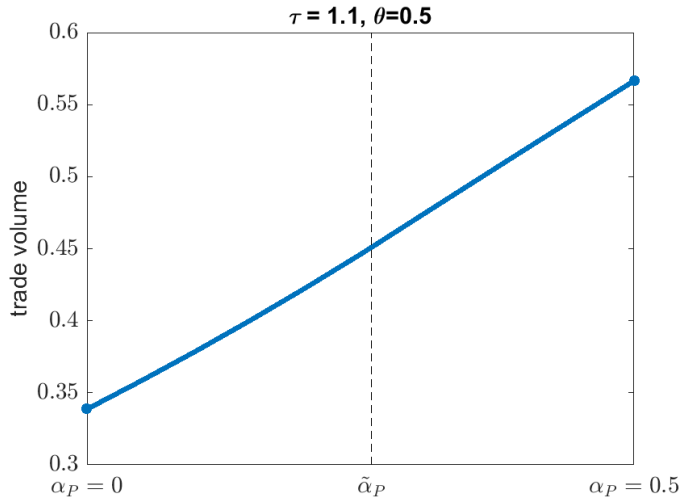


Figure 7: World trade volume for $\underline{\theta} = 1/2$ cross-section

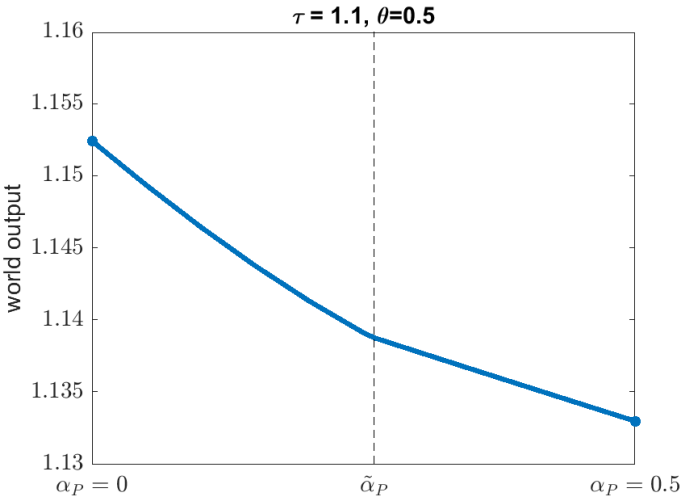


Figure 8: World net output for $\underline{\theta} = 1/2$ cross-section

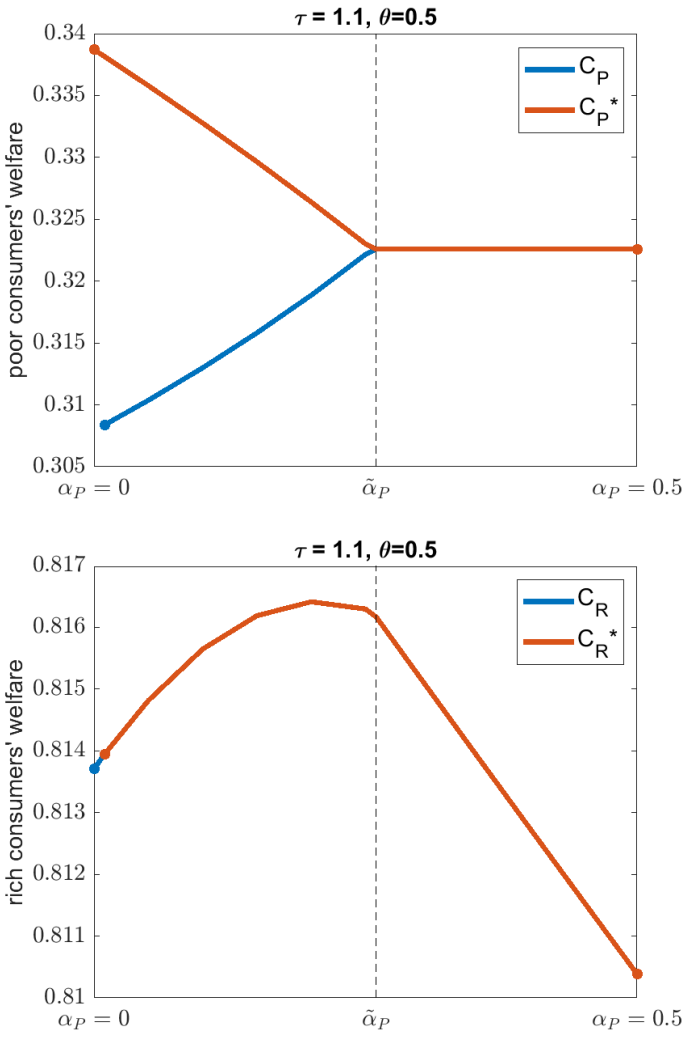


Figure 9: Poor and rich consumers' welfare for $\underline{\theta} = 1/2$ cross-section

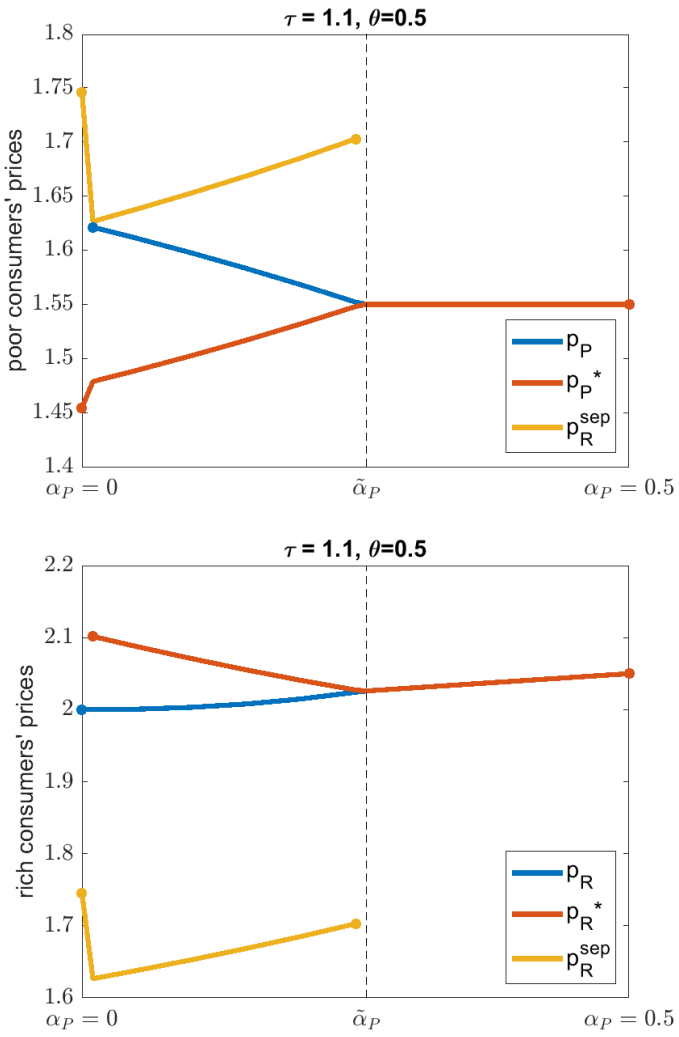


Figure 10: Poor and rich consumers' prices for $\underline{\theta} = 1/2$ cross-section

between-country inequality is provided in Appendix C.4.1. Section 4.4 discusses how trade intensity behaves in the entire inequality/endowment space of Figure 5.

Result 2. *Poor consumers are extremely badly off if they are a small minority in the rich country.*

This is due to poor Northern consumers being excluded from Southern PR products and having to pay higher prices for PP products in PI3, as seen in Figure 10. Poor consumers in the North do not have access to all products poor consumers in the South have, which results in lower welfare. As a real world example, we can think of China as the Southern and the USA as the Northern country. If a consumer with as little resources as the poor in China lives in the USA, he fares badly since he is confronted with high prices for basic products. On a local level, this can also be thought of as a small minority of poor households living in a location with almost exclusively rich households such as Manhattan. We therefore refer to this as the *Manhattan effect*.

Result 3. *Rich consumers in both countries always fare equally well, while for poor consumers it may matter where they live.*

Whenever PR firms exist (PI3), poor consumers are better off living in the country with many other poor consumers since this decreases prices for PP products, as elaborated in Result 2. Rich consumers always consume all varieties produced worldwide, so location of residence is irrelevant for them. If between-country inequality is low such that there are no PR firms in the South (PI2), changing α_P leaves welfare of the poor unchanged.

Result 4. *If $\frac{\bar{\theta}}{\underline{\theta}}$ is high, welfare of poor and rich consumers is lower under only within- compared to only between-country inequality.*

This is related to Result 1. The reason for lower welfare is the intense costly trade in rich-consumer varieties among identical countries. There are large iceberg losses from exporting the locally produced exclusive products RR in both directions. The proof is provided in Appendix C.4.2.

Result 5. *For low $\frac{\bar{\theta}}{\underline{\theta}}$, rich consumers are better off under only within- compared to only between-country inequality but poor consumers are always worse off under only within-country inequality.*

For low endowment differences, rich consumers benefit from a higher number of exclusive varieties produced under only within- compared to only between-country inequality. This is due to a higher price charged for poor varieties under only between-country inequality, τp_P, as producers can pass on trade costs to the rich. Hence, the rich have less resources left to buy exclusive products N_R. The small welfare improvement of the rich is never sufficient to compensate the welfare loss of the poor. Thus, global welfare is always decreasing in income distribution similarity due to inefficient trade.

Result 6. *Rich consumers are best off if there is some within-inequality but not too much.*

This can be seen from the inverse U shape in the bottom panel of Figure 9. In Regime PI3, global product variety is increasing in within-country inequality due to the variety of mass and exclusive products increasing more than the decrease in separating varieties. As world net output is decreasing in income similarity, the welfare gain of rich consumers must come at the cost of a welfare loss for the poor in the South. Poor consumers in the South get to consume less PR products and some of the saved resources are used for more RR firms. In Regime PI2, the number of mass firms remains constant but the number of RR decreases because of resources lost due to costly trade. Thus, rich welfare decreases for higher income similarity.

Result 7. *A trade liberalization always increases trade and benefits the rich. Whenever there are no separating firms PR, it also benefits the poor in both countries.*

When PR firms are active (PI3 and PI7), a trade liberalization can reduce welfare of the poor in the South and even decrease global welfare. This corresponds to the effect found in FHZ that the poor might loose from a trade liberalization due to the North focusing on exclusive markets RR and reducing the number of mass varieties PP produced due to the threat of arbitrage causing exclusion of poor consumers.

Result 8. *Poor consumers are best off if they all live in one country.*

When all poor consumers live in the South *($\alpha_P = 0$)* the large domestic market size allows for the cheapest prices p^* and the least amount of resources is wasted in trade. Hence, consumption of the poor is maximized.

4.3 Parallel trade prohibited

When parallel imports are prohibited by law, firms can freely set prices that differ between the domestic and the export market. Figure 11 depicts the regimes that occur when parallel imports are prohibited in the $(\underline{\theta}, \alpha_P)$ parameter space for constant country population, $\mathcal{P} = \mathcal{P}^*$, and for $\tau = 1.1$.[40]

4.3.1 Symmetric countries and low inequality (noPI1)

We first look at the case of low inequality. We indicate this regime with a L superscript. We guess that

$$\Pi_S = \Pi_M = 0 \tag{9}$$

is an equilibrium. Using (2) and $\Pi_S = \Pi_M = 0$, we solve for $p_P\mathcal{P}$. Insert this into $\Pi_S = 0$, which results in $\Pi_E < 0$. Hence, the guessed regime is indeed an equilibrium. Moreover, this is the only equilibrium with low inequality. In particular, it cannot hold that both mass production firms and exclusive firms are active. If $\Pi_M = \Pi_E = 0$ were an equilibrium, it is easy to show that $\Pi_S > (\Pi_M + \Pi_E)/2 = 0$, hence firms would have a deviation incentive. Due to the existence of M firms in both countries, we know that $\omega = 1$ if $\mathcal{P} = \mathcal{P}^*$. The equilibrium price of the poor equals average costs, which are the sum of fixed costs per capita $F/(2\mathcal{P})$ plus average marginal costs,

$$p_P^L = \frac{1}{2}\left(\frac{F}{\mathcal{P}} + \frac{1+\tau}{a}\right). \tag{10}$$

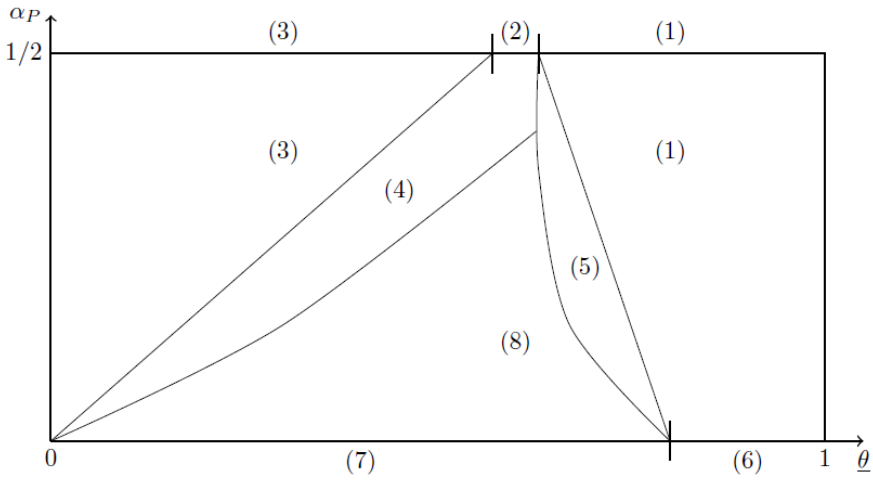
Using the zero profit constraint from Equation (9) and Equation (10), we get the price p_R^L that the separating firms charge from rich consumers in the exporting market

$$p_R^L = \frac{1}{2}\left(\frac{F}{\mathcal{P}(1-\alpha_P)} + \frac{1+\tau-(\tau-1)\alpha_P/(1-\alpha_P)}{a}\right) = \frac{1}{2(1-\alpha_P)}\left(\frac{F}{\mathcal{P}} + \frac{1+\tau-2\tau\alpha_P}{a}\right).$$

The varieties, obtained from the budget constraints, are

$$\begin{aligned} N_M = N_M^* &= \frac{\underline{\theta}}{\frac{F}{\mathcal{P}} + \frac{1+\tau}{a}} - \frac{\bar{\theta}-\underline{\theta}}{2p_R^L} \\ N_S = N_S^* &= \frac{\bar{\theta}-\underline{\theta}}{p_R^L}. \end{aligned}$$

[40] In area (8), no equilibrium could yet be identified.



(1) N: PP+PR S: PP+PR

(2) N: PR S: PR

(3) N: PR+RR S: PR+RR

(4) N: PR+RR S: PR

(5) N: PP+PR S: PR

(6) N: PP S: PP

(7) N: PP+RR S: PP

(8) UNSOLVED

Figure 11: Two-group no parallel import regimes for constant country population

4.3.2 Symmetric countries and medium inequality (noPI2)

With medium inequality all goods are traded and there are only separating firms in both countries. We indicate this regime with a M superscript. We have $\omega = 1$ due to perfect symmetry. The varieties, obtained from the resource constraints, are

$$N_S = N_S^* = \frac{a\mathcal{P}\left(\alpha_P \underline{\theta} + (1-\alpha_P)\bar{\theta}\right)}{aF + \mathcal{P} + \tau(1-\alpha_P^*)\mathcal{P}^*} = \frac{(\bar{\theta} - \underline{\theta})/2}{F/\mathcal{P} + 1 + \tau/2}.$$

This allows to express prices as

$$\begin{aligned} p_P^M &= \frac{\underline{\theta}}{N_S} \\ p_R^M &= \frac{\bar{\theta} - \underline{\theta}}{N_S^*}. \end{aligned}$$

4.3.3 Symmetric countries and high inequality (noPI3)

In the case of high inequality, only separating and exclusive producers are present. We indicate this regime with a H superscript. Hence, the equilibrium is given by $\Pi_S = \Pi_E = 0$, implying $\Pi_M < 0$. The price of the exclusive firms equals the average costs, recall that the fixed costs per capita equal $F/(2\mathcal{P}(1-\alpha_P))$ in that case,

$$p_R^H = \frac{1}{2}\left(\frac{F}{\mathcal{P}(1-\alpha_P)} + \frac{1+\tau}{a}\right) > p_R^L.$$

The price paid by the poor is lower than in the low-inequality case. The reason lies in the fact that the separating firms may charge a higher price from the rich consumers since exclusive firms are present, $p_R^H > p_R^L$. It equals

$$p_P^H = \frac{1}{2}\left(\frac{F}{\mathcal{P}} + \frac{1+\tau}{a} - (\tau - 1)\frac{\alpha_P}{a}\right) < p_P^L.$$

We have $\omega = 1$ due to perfect symmetry. The varieties, obtained from the resource constraints, are

$$\begin{aligned} N_S = N_S^* &= \frac{\underline{\theta}}{p_P^H} \\ N_E = N_E^* &= \frac{\bar{\theta} - \underline{\theta}}{2p_R^H} - \frac{\underline{\theta}}{2p_P^H}. \end{aligned}$$

4.3.4 Representative agent and full trade (noPI6)

Let us study the extreme case $\alpha_P = 0$. In the home country, there are only rich consumers, in the foreign country only poor ones. This is the FHZ case but we did not analyze the equilibrium without parallel imports in a case where the price differences would exceed τ. There are only mass firms active in both countries. This will prevail as long as $p_P > \tau\omega/a$ which holds if $\underline{\theta} > \frac{2\tau}{aF+1+\tau}$.

The expressions for prices and varieties, as well as the comparative statics are identical to the PI allowed case (see the PI6 regime described in Section C.3) as long as the arbitrage constraints are not binding yet ($p \leq \tau p^*$, which holds for $\underline{\theta} \geq \frac{2}{1+\tau}$). Hence, if parallel imports are forbidden, a weaker condition is sufficient to ensure full trade compared to the case where parallel imports are allowed.

The prices, obtained by combining budget constraints and zero profit conditions, are

$$\begin{aligned} p &= \frac{\left(\frac{F}{\mathcal{P}} + \frac{1+\tau}{a}\right)\bar{\theta}}{\bar{\theta}+\underline{\theta}} \\ p^* &= \frac{\left(\frac{F}{\mathcal{P}} + \frac{1+\tau}{a}\right)\underline{\theta}}{\bar{\theta}+\underline{\theta}}. \end{aligned}$$

The varieties, obtained from the resource constraints, are

$$\begin{aligned} N_M &= \frac{\bar{\theta}}{\frac{F}{\mathcal{P}} + \frac{1+\tau}{a}} \\ N_M^* &= \frac{\underline{\theta}}{\frac{F}{\mathcal{P}} + \frac{1+\tau}{a}}. \end{aligned}$$

4.3.5 Representative agent and Partial trade (noPI7)

What happens when the foreign country is so poor that the consumers can no longer afford to buy all products. In that case $p_P^* = \tau/a$ and the exporters in the rich country just break even in their exporting business. The relative wage equals

$$\omega = \frac{aF + \mathcal{P} + \tau\mathcal{P}^*}{aF + \tau\mathcal{P} + \mathcal{P}^*}$$

as above. From zero profits for mass firms

$$\Pi_M = p_P\mathcal{P} + p_P^*\mathcal{P}^* - \left(F + \frac{\mathcal{P} + \tau\mathcal{P}^*}{a}\right) = 0$$

we get

$$p_P = \frac{F}{\mathcal{P}} + \frac{1}{a}.$$

From zero profits for exclusive firms

$$\Pi_E = p_R\mathcal{P} - \left(F + \frac{\mathcal{P}}{a}\right) = 0$$

we get

$$p_R = \frac{F}{\mathcal{P}} + \frac{1}{a}.$$

The budget constraint of the consumers in foreign reads

$$p_P^*\left(N_M + N_M^*\right) = \omega\underline{\theta}$$

Welfare therefore is

$$U_P^* = \frac{\omega\underline{\theta}}{p_P^*} = \frac{a\underline{\theta}\omega}{\tau} = \frac{aF + \mathcal{P} + \tau\mathcal{P}^*}{aF + \tau\mathcal{P} + \mathcal{P}^*}\frac{a\underline{\theta}}{\tau}.$$

The resource constraint allows us to determine

$$N_M^* = \frac{a\underline{\theta}\mathcal{P}^*}{aF + \mathcal{P}^* + \tau\mathcal{P}}.$$

The amount of exports from the rich to the poor country, may be derived using the balanced trade

condition $p_P^* N_M = p_P N_M^*$,

$$N_M = \left(\frac{aF}{\mathcal{P}} + 1\right) \frac{1}{\tau} \frac{a\underline{\theta}\mathcal{P}^*}{aF + \mathcal{P}^* + \tau\mathcal{P}}.$$

When $\underline{\theta}$ goes to zero, N_M and N_M^* go to zero as well. Finally, plugging N_M into the resource constraint gives,

$$N_E = \frac{a\bar{\theta}\mathcal{P} - N_M}{aF + \mathcal{P}}.$$

Regime noPI7 differs from PI7 since here the trade condition, $p_P^* = \tau/a$, rather than the arbitrage condition, $p = \tau p^*$, becomes binding.[41]

Section D in the appendix provides some thoughts on the case when there is imperfect enforcement of parallel import prohibition.

4.4 Effect of within- and between-country inequality on trade

Next we study the effect of within- and between-country inequality on trade volume and trade intensity. We compare trade between different within- and between-inequality constellations holding world endowment constant and for the sake of illustration focus on the simple two income groups case.[42]

First, recall that α_P controls the similarity of the two countries' income distributions and $\underline{\theta}$ controls the group endowment distance. In Section 4.2.5, we showed that trade volume is monotonously increasing in income similarity when parallel imports are allowed. FHZ, on the other hand, show in their Propositions 1 and 3 that for RA countries trade volume is monotonously decreasing in endowment difference when parallel imports are allowed. Hence, both countries' distribution similarity and endowment level similarity contribute to increasing trade.

Figure 12 shows a heat map of trade intensity ϕ in the group endowment versus country similarity space $(\underline{\theta}, \alpha_P)$. In order to keep world endowment constant, we define $\bar{\theta} = 2 - \underline{\theta}$ and keep $\mathcal{P}_P = \mathcal{P}_R = \mathcal{P} = \mathcal{P}^* = 1$.[43] Country distribution similarity is increasing in α_P and hence increasing from bottom to top, while group endowment level similarity is increasing in $\underline{\theta}$ and hence increasing from left to right. Warmer colors represent a higher trade intensity (green: lowest trade intensity, red: highest trade intensity). Trade intensity is maximal for identical countries ($\alpha_P = 0.5$ or $\underline{\theta} = 1$) and minimal for maximally differing representative agents ($\alpha_P = 0$ and $\underline{\theta} \to 0$). Thus, the model predicts that trade intensity is monotonously increasing in the overlap of the income distribution (increasing α_P) at every endowment cross-section, which is in line with the empirical findings. Two countries' income similarity and therefore also trade intensity is maximized when group endowment differences are minimized and global within-country inequality is maximized as then poor and rich are equally distributed over countries. Trade intensity is also increasing in endowment similarity as seen for RA countries in the bottom row of Figure 12. Starting from only-between country inequality ($\alpha_P = 0$) and the poor having a third of the endowment of the rich ($\underline{\theta} = 0.5$), we see that either eliminating group endowment differences ($\underline{\theta} \to 1$) or eliminating countries' distribution differences ($\alpha_P \to 0.5$) both double trade. A special case arises when the two countries are identical, then trade intensity is independent of the endowment gap as always half of each countries production is exported.

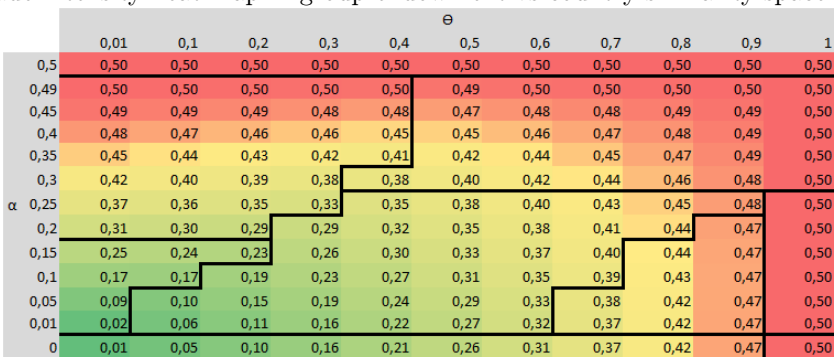
In order to isolate the difference in trade intensity predictions under 0-1 and CES preferences, it is instructive to hold constant country endowment instead of country population. Figure 13 plots trade intensity

[41] Recall that the trade condition in the RA case is $\tau \leq \sqrt{aF/\mathcal{P}^R + 1}$

[42] Note that like in the derivation of the equilibrium regimes in Sections 4.2 and 11, we hold country population and world endowment constant but country endowment varies.

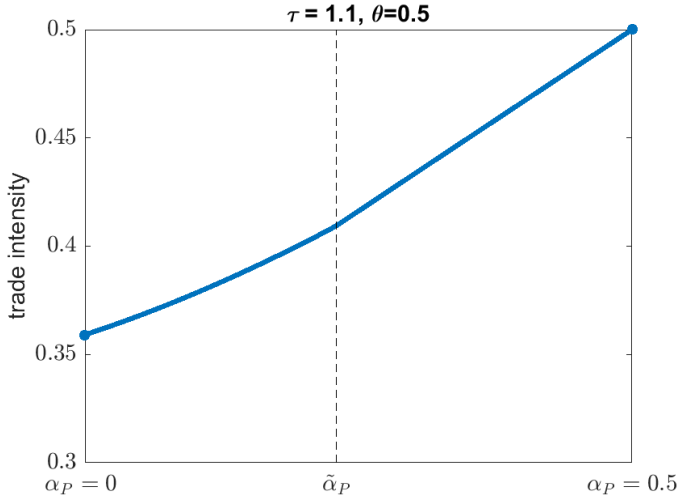
[43] Hence, world endowment is $E + E^* = \bar{\theta}\mathcal{P}_R + \underline{\theta}\mathcal{P}_P = 2$ for $\underline{\theta} \in (0, 1]$.

Figure 12: Trade intensity heat map in group endowment vs country similarity space for PI allowed



α \ Θ	0,01	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	1
0,5	0,50	0,50	0,50	0,50	0,50	0,50	0,50	0,50	0,50	0,50	0,50
0,49	0,50	0,50	0,50	0,50	0,50	0,49	0,50	0,50	0,50	0,50	0,50
0,45	0,49	0,49	0,49	0,48	0,48	0,47	0,48	0,48	0,49	0,49	0,50
0,4	0,48	0,47	0,46	0,46	0,45	0,45	0,46	0,47	0,48	0,49	0,50
0,35	0,45	0,44	0,43	0,42	0,41	0,42	0,44	0,45	0,47	0,49	0,50
0,3	0,42	0,40	0,39	0,38	0,38	0,40	0,42	0,44	0,46	0,48	0,50
0,25	0,37	0,36	0,35	0,33	0,35	0,38	0,40	0,43	0,45	0,48	0,50
0,2	0,31	0,30	0,29	0,29	0,32	0,35	0,38	0,41	0,44	0,47	0,50
0,15	0,25	0,24	0,23	0,26	0,30	0,33	0,37	0,40	0,44	0,47	0,50
0,1	0,17	0,17	0,19	0,23	0,27	0,31	0,35	0,39	0,43	0,47	0,50
0,05	0,09	0,10	0,15	0,19	0,24	0,29	0,33	0,38	0,42	0,47	0,50
0,01	0,02	0,06	0,11	0,16	0,22	0,27	0,32	0,37	0,42	0,47	0,50
0	0,01	0,05	0,10	0,16	0,21	0,26	0,31	0,37	0,42	0,47	0,50

Figure 13: Trade intensity vs country similarity for constant country endowment and PI allowed



for constant country endowment against the share of the world's poor living in the North. We see that under 0-1 preferences trade intensity is monotonously increasing in distribution similarity. Whereas under CES preferences, trade intensity would be completely independent of α_P as for constant country endowment a change in distribution similarity only changes the composition of GDP (population vs GDPpc) but leaves the size of countries' GDP and hence trade unaffected.[44]

4.5 Preferred parallel trade policy

We now discuss which parallel trade policy countries prefer assuming their objective is to maximize the utility of their inhabitants. A country's optimal parallel trade policy differs depending on the supply side market structure – e.g. duopoly versus monopolistic competition – and on the level of inequality. Let us first consider the case of only-between country inequality (representative agents). In partial equilibrium duopoly models such as Roy and Saggi (2012), the North prefers to *forbid* parallel trade when between-country inequality is high.[45] Whereas in the FHZ general equilibrium model with many firms, the North prefers to *allow* parallel imports when between-country inequality is high as shown in Proposition 1.

[44] Of course the CES trade intensity level will depend on the substitution elasticity sigma.
[45] Roy and Saggi (2012) Proposition 6 (ii).

Proposition 1. *When there is only between-country inequality ($\alpha_P = 0$) and group endowment differences are large ($\underline{\theta} \leq \frac{2\tau}{aF+1+\tau}$), rich consumers prefer to* ***allow*** *parallel trade but poor consumers prefer to* ***forbid*** *it. The social planner prefers to forbid parallel trade.*

Proof. For poor consumers

$$U_P^{*noPI7} = \frac{\underline{\theta}}{\frac{\tau}{a}} > \frac{\underline{\theta}}{\frac{F}{1+\tau} + \frac{1}{a}} = U_P^{*PI7}$$

if $\tau < \sqrt{aF+1}$, which is simply the trade condition. Hence, whenever there is trade the poor consumers prefer to forbid parallel imports. The same types of firms are active in PI7 and in noPI7 but $p_P^{*PI7} > p_P^{*noPI7}$ since when parallel imports are prohibited, mass firms can just set the break-even price in the poor country $p_P^{*noPI7} = \frac{\tau}{a}$ (lowest possible price where trade still worthwhile), while when parallel imports are allowed, mass firms' price setting is constraint by binding arbitrage $p_P^{*PI7} \geq \frac{p_P}{\tau}$, resulting in lower welfare for the poor.

For rich consumers

$$U_R^{noPI7} = \frac{\bar{\theta}}{F + \frac{1}{a}} < \frac{\underline{\theta}}{\frac{F}{1+\tau} + \frac{1}{a}} + \frac{\bar{\theta} - \tau\underline{\theta}}{F + \frac{1}{a}} = U_R^{PI7}$$

also if $\tau < \sqrt{aF+1}$. Hence, whenever there is trade the rich consumers prefer to allow parallel imports. When parallel imports are allowed, rich consumers can benefit from the lower price paid on mass products due to binding arbitrage $p_P^{PI7} = \tau p_P^{*PI7} = \frac{\tau F}{1+\tau} + \frac{\tau}{a} < p_P^{noPI7} = F + \frac{1}{a}$, while the price for the exclusive goods is the same in both parallel trade regimes $p_R^{PI7} = p_R^{noPI7} = F + \frac{1}{a}$.

The social planner, maximizing global welfare, prefers to forbid parallel trade since

$$U^{noPI7} = \frac{\underline{\theta}}{\frac{\tau}{a}} + \frac{\bar{\theta}}{F + \frac{1}{a}} > \frac{2\underline{\theta}}{\frac{F}{1+\tau} + \frac{1}{a}} + \frac{\bar{\theta} - \tau\underline{\theta}}{F + \frac{1}{a}} = U^{PI7},$$

which again holds if $\tau < \sqrt{aF+1}$. □

Our model's different preferred parallel trade prediction compared to Roy and Saggi (2012) is rooted in the different market structure. The key difference is a resource feedback effect operating through the number of firms, present in our model with free entry but absent in their duopoly model. In our model, the North has an incentive to allow parallel imports since Northern consumers benefit from Northern firms' export exclusion of the South. Exclusion implies less Northern resources being used for producing goods for the South. Hence, these saved resources can be used to finance the set up costs for more varieties in the North through more firm entry. In an oligopoly model on the other hand, Northern consumers do not benefit from export exclusion as this only increases prices but there are no new varieties as firm entry is absent. Hence, the Northern consumers are harmed by the high prices set by the duopolist, who chooses to abstain from exporting when parallel imports are allowed. Hence, in a duopoly model, the North prefers to forbid parallel imports in order to ensure that the Northern duopolist sells to the South by applying international price discrimination.

For low between-country inequality, the North prefers to allow parallel imports also in the duopoly model since there is no more export exclusion of the South when inequality is low.[46] In our model, the North is indifferent about parallel trade policy as shown in Proposition2.

[46] Roy and Saggi (2012) Proposition 6 (i).

Proposition 2. *When there is only between-country inequality ($\alpha_P = 0$) and group endowment differences are small ($\underline{\theta} > \frac{2}{1+\tau}$), both poor and rich consumers are indifferent about parallel trade policy and so it the social planner.*

Proof. As long as the arbitrage constraint is not binding $p < \tau p^*$, which holds for $\underline{\theta} > \frac{2}{1+\tau}$, the equilibrium under parallel import allowed and prohibited are identical. Hence, all consumers are indifferent about parallel trade rules and so is the social planner.

Note that also for $\frac{2\tau}{aF+1+\tau} < \underline{\theta} < \frac{2}{1+\tau}$ rich consumers prefer to allow parallel trade but poor consumers prefer to forbid it. □

The South always (weakly) prefers to prohibit parallel imports in both market structures. In the oligopoly model, the South is harmed by above monopoly pricing of the Southern duopolist when parallel imports are allowed due to inability to price discriminate. In our model, the South suffers from being excluded from Northern varieties when parallel imports are allowed. The Souths preferred market outcome is international price discrimination, which obtains when parallel imports are forbidden.

From a global perspective, it is (weakly) welfare maximizing to prohibit parallel trade as shown in Proposition 1 and 2. The reason is that whenever fixed costs for a product are incurred, the utility gotten out of it is maximized, when distributing these fixed costs over as many individuals as possible. Hence, export probability should be maximized. In the full trade case, the most efficient allocation is achieved.

4.5.1 Parallel trade with within-country inequality

Let us now extend this analysis to a world featuring within-country inequality to study the preferred parallel trade policy of rich versus poor consumers within a country.

Proposition 3. *When there is only within-country inequality ($\alpha_P = 0.5$) and group endowment differences are large ($\underline{\theta} \leq \frac{2\tau}{aF+1+\tau}$), both rich and poor consumers prefer to* ***forbid*** *parallel trade and so does the social planner.*

Proof. For the poor consumers

$$U_P^{noPI3} = \frac{2\underline{\theta}}{F + \frac{3+\tau}{2a}} > \frac{2\underline{\theta}}{F + \frac{1+\tau}{a}} = U_P^{PI1}$$

since $\tau > 1$. In PI1 there are mass firms, while in noPI3 there are separating firms. $p_P^{*noPI3} < p_P^{*PI1}$ since separating firms can efficiently price discriminate and therefore offer the product for a lower price at home.

For the rich consumers

$$U_R^{noPI3} = U_P^{noPI3} + \frac{\bar{\theta} - \underline{\theta}}{F + \frac{1+\tau}{2a}} > U_P^{PI1} + \frac{\bar{\theta} - \underline{\theta}}{F + \frac{1+\tau}{2a}} = U_R^{PI1}$$

as the extra varieties consumed by the rich are the same in both regimes. Hence, also $U^{noPI3} > U^{PI1}$. As prices are weakly lower when there are no parallel imports, it is easy to see that - for given τ and given level of labor endowment - forbidding parallel imports leads to a Pareto improvement. A social planner would always choose to have a positive mass of separating firms present. However, with PI binding, such firms cannot exist, because price discrimination is limited. □

Proposition 4. *When there is only within-country inequality ($\alpha_P = 0.5$) and group endowment differences are small ($\underline{\theta} > \frac{2}{1+\tau}$), rich consumers prefer to* ***forbid*** *parallel imports while poor consumers are indifferent. The social planner prefers to forbid parallel imports*

Proof. For poor consumers

$$U_P^{noPI1} = \frac{\underline{\theta}}{\frac{F}{2} + \frac{1+\tau}{2a}} = \frac{\underline{\theta}}{\frac{F}{2} + \frac{1+\tau}{2a}} = U_P^{PI1}.$$

For rich consumers

$$U_R^{noPI1} = \frac{\underline{\theta}}{\frac{F}{2} + \frac{1+\tau}{2a}} + \frac{\bar{\theta} - \underline{\theta}}{F + \frac{1}{a}} > \frac{\underline{\theta}}{\frac{F}{2} + \frac{1+\tau}{2a}} + \frac{\bar{\theta} - \underline{\theta}}{F + \frac{1+\tau}{2a}} = U_R^{PI1}$$

since $\tau > 1$ and therefore also $U^{noPI1} > U^{PI1}$. In PI1 there are exclusive firms while in noPI1 there are separating firms. $p_R^{*noPI1} < p_R^{*PI1}$ since separating firms set lower prices than exclusive due to serving a larger market. □

We see that in contrast to when there is only between-country inequality, where the rich prefer to allow parallel imports, for only within-country inequality also the rich prefer to forbid parallel imports. See Appendix C.5 for cases with both within-and between-country inequality.

5 Three endowment groups

As a step towards a continuous labor endowment distribution, we next extend our two-group model to three groups. There are three endowment levels in the world: poor $\theta_P = \underline{\theta}$, middle $\theta_M = 1$, and rich $\theta_R = \bar{\theta}$. We now additionally define α_M to be the Northern share of the world's middle income consumers $\mathcal{P}_M$. We still keep the number of consumers per endowment group constant on a global level to have constant world population and world endowment. We now vary the share of the world population in the three groups that is allocated to the North and the South to study different income distribution constellations between the two countries.[47]

Again, keeping the population of the two countries constant and identical, $\mathcal{P} = \mathcal{P}^*$, implies

$$\alpha_M = \frac{\mathcal{P} - \alpha_P \mathcal{P}_P - \alpha_R \mathcal{P}_R}{\mathcal{P}_M}.$$

Hence, choosing the Northern share of the world's poor consumers α_P and the Northern share of the world's rich consumers α_R determines the Northern share of the world's middle income consumers α_M.
Alternatively, we can also keep country endowments constant and identical $E = E^*$. Country endowments are $E \equiv \alpha_P \mathcal{P}_P \underline{\theta} + \alpha_M \mathcal{P}_M + \alpha_R \mathcal{P}_R \bar{\theta}$ and $E^* \equiv (1-\alpha_P)\mathcal{P}_P \underline{\theta} + (1-\alpha_M)\mathcal{P}_M + (1-\alpha_R)\mathcal{P}_R \bar{\theta}$. This implies

$$\alpha_R = \frac{E - \alpha_P \mathcal{P}_P \underline{\theta} - \alpha_R \mathcal{P}_R \bar{\theta}}{\mathcal{P}_M}, \quad \mathcal{P} = \alpha_P \mathcal{P}_P + \alpha_M \mathcal{P}_M + \alpha_R \mathcal{P}_R \quad \text{and} \quad \mathcal{P}^* = \mathcal{P}_P + \mathcal{P}_M + \mathcal{P}_R - \mathcal{P}.$$

Hence, choosing α_P and α_R determines $\mathcal{P}$, $\mathcal{P}^*$ and α_M.

[47] It is convenient to express the endowment distribution in world shares rather than country shares because this directly determines how many consumers of each endowment group are located in the other country. If we instead were to manipulate country shares given identical country population, this would not only change the within-country allocation but also global group size. Hence, so would world endowment unless we somehow compensate for the change in other country but then it would be easier to just use global shares directly.

In either setting, only two parameters, namely α_P and α_R, are sufficient to control the income distributions in both countries. For three endowment groups the overlap can be expressed as

$$O_{od} = \sum_{i \in \{P,M,R\}} min\{\alpha_i, (1-\alpha_i)\}.$$

We still assume that the Northern country is richer than the Southern country, i.e. for a constant country population, we require the number of rich consumers in the North to be larger than the number of poor consumers, $\alpha_R \geq \alpha_P$, and for constant country endowment we require the majority of rich consumers to live in the North, $\alpha_R \geq 1/2$.

For three endowment groups, there are six relevant types of firms potentially active in equilibrium. The Northern firms' profits when allowing for parallel imports, using $\mathcal{P}_P = \mathcal{P}_M = \mathcal{P}_R = 1$, are

$$\begin{aligned}
\Pi_{PP} &= p_P(\alpha_P + \alpha_M + \alpha_R) + p_P^*((1-\alpha_P) + (1-\alpha_M) + (1-\alpha_R)) - C_{PP} \\
\Pi_{MM} &= p_M(\alpha_M + \alpha_R) + p_M^*((1-\alpha_M) + (1-\alpha_R)) - C_{MM} \\
\Pi_{RR} &= p_R\alpha_R + p_R^*(1-\alpha_R) - C_{RR} \\
\Pi_{PM} &= p_P(\alpha_P + \alpha_M + \alpha_R) + min\{\tau p_P, p_M^*\}((1-\alpha_M) + (1-\alpha_R)) - C_{PM} \\
\Pi_{MR} &= p_M(\alpha_M + \alpha_R) + min\{\tau p_M, p_R^*\}(1-\alpha_R) - C_{MR} \\
\Pi_R &= p_R\alpha_R - W\left(F + \frac{\alpha_R}{a}\right),
\end{aligned}$$

where costs are $C_{ij} = W\left(F + \frac{\sum_{n=i}^{R}\alpha_n + \tau\sum_{n=j}^{R}(1-\alpha_n)}{a}\right)$. Profits of Southern firms are calculated analogously.

The relative wage if there are PP firms in both countries is

$$\omega = \frac{W^*}{W} = \frac{aF + (1-\tau)(\alpha_P\mathcal{P}_P + \alpha_M\mathcal{P}_M + \alpha_R\mathcal{P}_R) + \tau(\mathcal{P}_P + \mathcal{P}_M + \mathcal{P}_R)}{aF + (\tau-1)(\alpha_P\mathcal{P}_P + \alpha_M\mathcal{P}_M + \alpha_R\mathcal{P}_R) + \mathcal{P}_P + \mathcal{P}_M + \mathcal{P}_R} \in (\frac{1}{\tau}, \tau).$$

The wage is higher in the country with the larger population because a larger home market implies higher efficiency due to less iceberg losses.

5.1 Symmetric countries (3PI1)

Let us first look at the case of symmetric countries, $\alpha_P = \alpha_M = \alpha_R = 1/2$, for constant and identical world groups,$\mathcal{P}_\mathcal{P} = \mathcal{P}_M = \mathcal{P}_R = 1$, and country population $\mathcal{P} = \mathcal{P}^*$. The corresponding labor endowment distribution is depicted in Figure 14. For endowment levels $\underline{\theta} = 1/5$ and $\bar{\theta} = 9/5$, this yields Gini coefficients $G = G^* = 0.36$. Prices, wages and the number of varieties are identical between the two countries due to symmetry. We consider relatively low trade costs $\tau = 1.1$ such that both arbitrage constraints are binding, $p_M \leq \tau p_P^*$ and $p_R \leq \tau p_M^*$.[48] The threat of parallel imports limits the price differentiation of the two separating firm types PM and MR to factor τ. For symmetric countries prices are independent of endowments, hence whether arbitrage is binding or not only depends on τ. The condition $p_R \leq \tau p_M^*$ is already binding at higher levels of τ than $p_M \leq \tau p_P^*$ such that $p_M \leq \tau p_P^*$ binding is sufficient for both firm types to be constraint in their pricing. This implies that there are no separating firms active in equilibrium whenever τ is sufficiently low. In equilibrium, all goods are traded and there are PP, MM and RR firms

[48] Appendix E.2 briefly discusses the equilibria when arbitrage constraints are non-binding due to high trade costs.

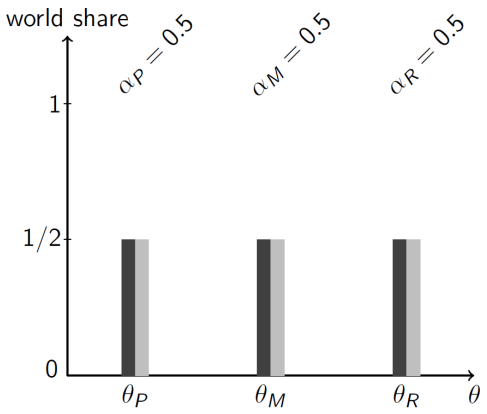


Figure 14: Labor endowment distribution for symmetric countries (3PI1)

in both countries. Factor prices equalize, $\omega = 1$, due to the existence of mass firms in both countries and $\mathcal{P} = \mathcal{P}^*$.

The prices, obtained from the zero profit conditions, are

$$\begin{aligned} \omega &= 1 \\ p_P &= p_P^* = \frac{F}{3} + \frac{1+\tau}{2a} \\ p_M &= p_M^* = \frac{F}{2} + \frac{1+\tau}{2a} \\ p_R &= p_R^* = F + \frac{1+\tau}{2a}. \end{aligned}$$

The varieties, obtained from the budget constraints, are

$$\begin{aligned} N_{PP} &= N_{PP}^* = \frac{\underline{\theta}}{\frac{2F}{3} + \frac{1+\tau}{a}} \\ N_{MM} &= N_{MM}^* = \frac{1-\underline{\theta}}{F + \frac{1+\tau}{a}} \\ N_{RR} &= N_{RR}^* = \frac{\bar{\theta} - 1}{2F + \frac{1+\tau}{a}}. \end{aligned}$$

As in the two-group case, there is indeterminacy of production location due to perfect symmetry. Another possible equilibrium structure is to only have MM, RR firms in the North and PP, MM firms in the South.

5.2 Max differentiated countries (3PI2)

Let us next consider maximally differentiated countries with overlap only for the middle income group, $\alpha_P = 0$, $\alpha_M = 1/2$ and $\alpha_R = 1$. The corresponding labor endowment distribution is depicted in Figure 15. We consider a wide endowment level range ($\underline{\theta} = 1/5$, $\bar{\theta} = 9/5$) such that both arbitrage constraints are binding, $p_M \leq \tau p_P^*$ and $p_R \leq \tau p_M^*$. There are MM and R firms active in the North and PP and MM firms in the South. This yields Gini coefficients $G = 0.12$ in the North and $G^* = 0.38$ in the South.[49]

[49]The Gini coefficient is not monotonous in income similarity for three endowment groups. Hence, it is not insightful for comparing within-income inequality in the three-group setting.

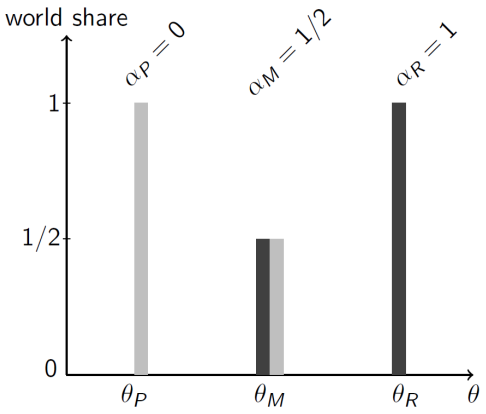


Figure 15: Labor endowment distribution for maximally differentiated countries (3PI2)

The system of equilibrium equations can be solved in closed form but we abstain from plugging in all terms for brevity. The prices, obtained from the zero profit conditions and the budget constraints, are

$$\begin{aligned}
\omega &= \frac{2aF+3+\tau}{2aF+1+3\tau} < 1 \\
p_P^* &= \frac{2\omega F}{3(1+\tau)} + \frac{\omega}{a} \\
p_P &= \tau p_P^* = \frac{2\omega\tau F}{3(1+\tau)} + \frac{\omega\tau}{a} \\
p_M &= \frac{1-\underline{\theta}\tau}{N_{MM}+N_{MM}^*} \\
p_M^* &= \frac{\omega(1-\underline{\theta})}{N_{MM}+N_{MM}^*} \\
p_R &= F + \frac{1}{a}.
\end{aligned}$$

The varieties, obtained by combining resource constraints and budget constraints, are

$$\begin{aligned}
N_{PP}^* &= \frac{\underline{\theta}}{\frac{2\omega F}{3(1+\tau)} + \frac{\omega}{a}} \\
N_{MM} &= \frac{3}{2F + \frac{3+\tau}{a}} \\
N_{MM}^* &= \frac{1 + 2\underline{\theta} - \frac{\underline{\theta}3(1+\tau)a}{2\omega Fa+3\omega(1+\tau)}\left(2F + \frac{3(1+\tau)}{a}\right)}{2F + \frac{1+3\tau}{a}} \\
N_R &= \frac{\bar{\theta}-1}{F+\frac{1}{a}}.
\end{aligned}$$

5.3 Three-group parallel import equilibria

Figure 16 depicts the three-group equilibrium regimes in the (α_P, α_R) parameter space when parallel imports are allowed, holding countries' populations constant and identical $\mathcal{P} = \mathcal{P}^*$, and for low trade costs $\tau = 1.1$. Figure 25 in the appendix shows the equilibrium regimes for holding countries' endowments identical

instead.[50] The regimes are qualitatively similar to the constant-country-population case but the layout differs.

As in the two group case, for a closer analysis, we again look at a cross-section through the (α_P, α_R) regimes graph. We focus on a fixed share of middle income consumers $\alpha_M = 1/2$. This corresponds to the red line in Figure 17. Figure 18 shows the three equilibrium regimes occurring over the cross-section at $\alpha_M = 1/2$ for $\tau = 1.1$ and $\underline{\theta} = 1/5$, as we move from maximally polarized (3PI2) to identical countries (3PI1). What is the intuition for switching between the different regimes? In Regime 3PI6, there are PM firms active in the South since there is a large Southern poor-consumer market but the majority of Northern consumers has at least middle income. As $\alpha_P \to 1/2$, the two countries become identical. Hence, there is no longer an incentive to pursue a separating strategy as Northern and Southern markets are equally large. Furthermore, note that Regime 3PI2 is a special case of Regime 3PI6 since if $\alpha_P = 0$ there are no poor consumers located in the North and no rich consumers located in the South. Thus, Southern PM and PP firms coincide and so do Northern RR and R firms.

It is a well known fact that there are many trade zeros on product level in the data and that many firms fear parallel imports. The threat of parallel imports can cause a firm to abstain exporting its product to poor markets. 3PI2 provides an example of a regime under which such export exclusion occurs. The varieties targeted at the rich consumers in the North R are not exported to the South since the low price necessary to sell the product in the South would undermine the firm's pricing in the North. In this case, the price effect outweighs the market size effect and the firm finds it optimal to only offer its product in the local market.

Figure 19 shows world trade volume. Trade volume is monotonously increasing as the two countries become more similar. This is again due to varieties targeted at rich consumers having to be shipped to both countries, whereas they are produced and consumed locally when all the rich are located in one country. Figure 20 shows world net output. Also for the three-group setting world net output is monotonously decreasing in income distribution similarity due to efficiency losses causes by intense trade in rich-consumer varieties. This confirms two-group Result 1. Figure 23 in the appendix shows welfare for poor, middle and rich consumers. Rich-consumer welfare is identical to the number of varieties available worldwide since rich consumers purchase all products. Figure 24 in the appendix shows prices for poor, middle and rich consumers.

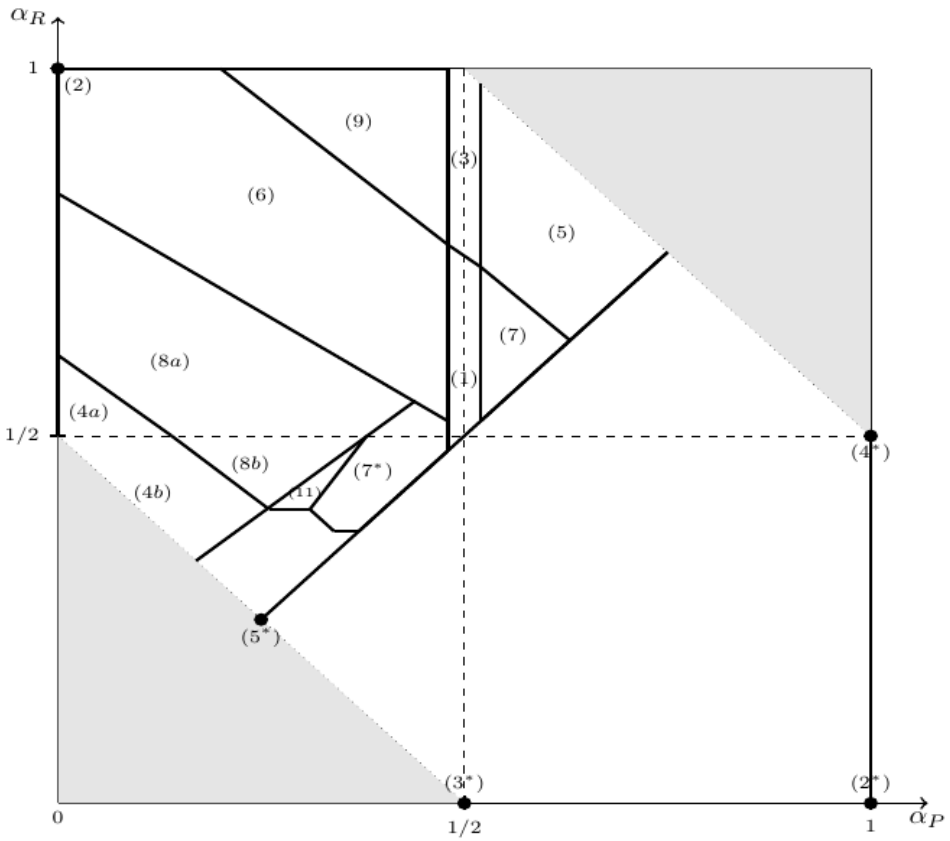
Poor-consumer welfare in the North is increasing in income distribution similarity, while it is decreasing in the South. This is due to the poor in the North being excluded from Southern PM products in Regime 3PI6. Poor consumers are better off if they are more numerous. Figure 24 shows that prices for basic goods PP are decreasing in the number of poor living in that country. This confirms two-group Result 2 – the Manhattan effect – also being active in the three-group case.

Middle- and rich-consumer welfare is identical across countries and decreasing in income similarity due to the decreasing total number of varieties produced. Thus, two-group Result 3, that rich consumers in both countries always fare equally well, while for poor consumers it matters where they live, also still holds in the three-group case. Furthermore, this confirms two-group Result 4, that for large endowment differences welfare for all consumers is highest for minimal income similarity.

Hence, two-group Results 1, 2, 3 and 4 continue to hold in the three-group case and are therefore likely to also hold for any higher number of endowment groups and eventually for a continuous distribution.

For practical applications, three endowment groups probably provide an insufficient approximation of empirical income distributions, as illustrated in Section G in the appendix. Hence, Section F in the appendix derives some general results for n endowment groups and the next Section presents the basic equilibrium for a continuous endowment distribution. However, there variations of the model are not fully solved yet except

[50] Recall that only either countries' populations or endowments can be held constant.



(1a)	N: MM + RR	S: PP + MM
(1b)	N: PP + MM + RR	S: MM
(2)	N: MM + R	S: PP + MM
(3a)	N: MM + R(R)	S: PP + MM + MR
(3b)	N: PP + MM + RR	S: MM + MR
(4a)	N: MM + RR + MR	S: PP + RR + PM
(4b)	N: PP + MM + MR	S: PP + RR + PM
(5)	N: PP + MM + RR + PM	S: MM + MR
(6)	N: MM + R(R)	S: PP + MM + PM
(7)	N: PP + MM + RR + PM	S: MM
(7*)	N: MM	S: PP + MM + RR + PM
(8a)	N: MM + RR	S: PP + RR + PM
(8b)	N: PP + MM + RR	S: PP + RR + PM
(9)	N: MM + R(R)	S: PP + MM + PM + MR
(11)	N: MM	S: PP + RR + PM

Figure 16: Three-group parallel import regimes for wide endowment range and constant country population

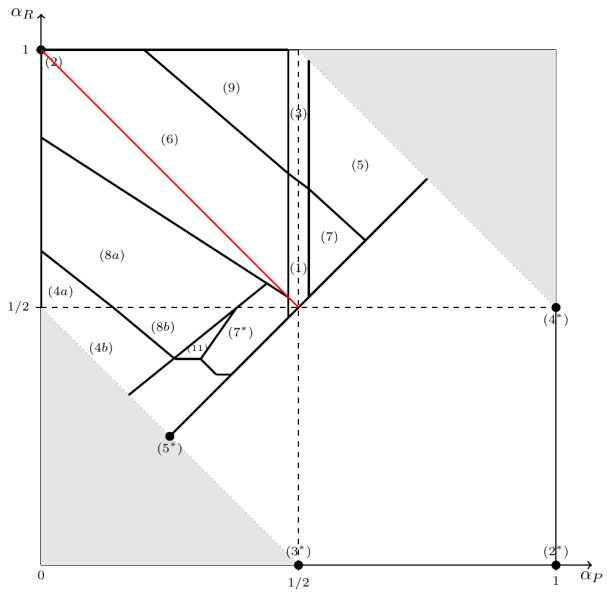


Figure 17: $\alpha_M = 1/2$ cross-section through three-group parallel import regimes for wide endowment range and constant country population

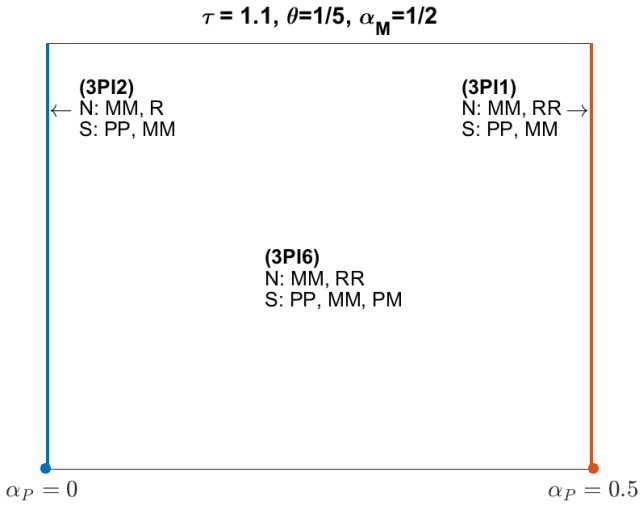


Figure 18: Three regimes of $\alpha_M = 1/2$ cross-section

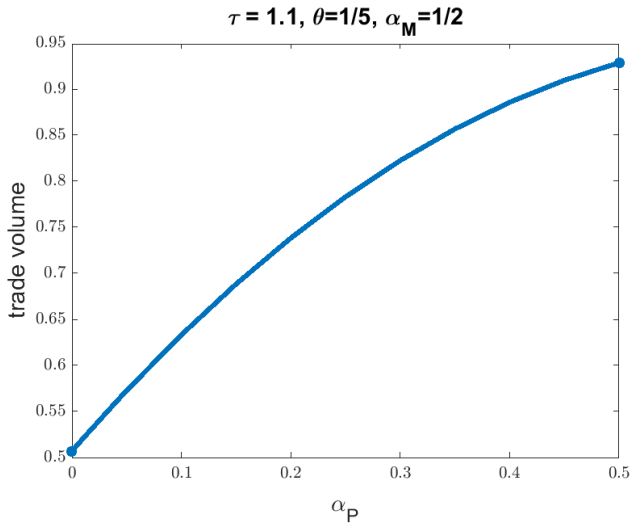


Figure 19: World trade volume for $\alpha_M = 1/2$ cross-section

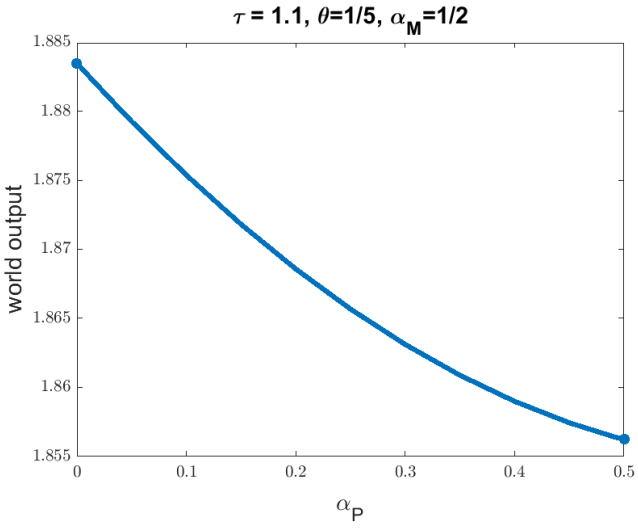


Figure 20: World net output for $\alpha_M = 1/2$ cross-section

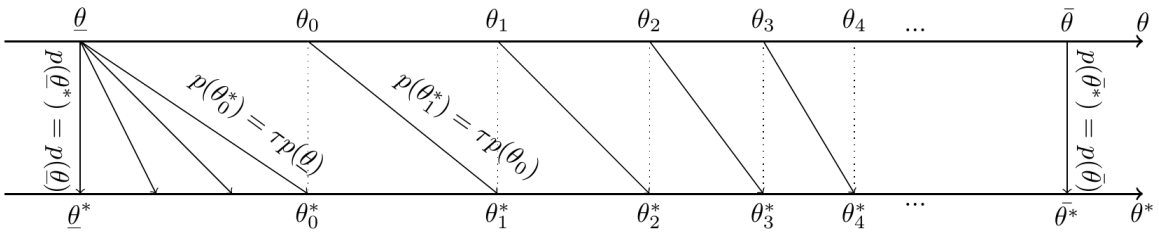


Figure 21: Choice of marginal home and foreign consumer when parallel imports allowed

for some special cases.

6 Continuous endowment distribution

We have presented the discrete case for two and three distinct labor endowment levels to build intuition. Solving the model for a continuous distribution would allow to fit the data without strongly stylizing features of any relevant income distribution . We are able to characterize the equilibrium structure for two *perfectly symmetric* countries with a continuous labor endowment distribution. For a continuous endowment distribution with cumulative density function $G(\theta)$, each firm chooses a marginal home consumer θ and a marginal foreign consumer θ^* by setting its home price $p(\theta)$ and its export price $p^*(\theta^*)$. The marginal consumer is the poorest consumer that purchases the product. Hence, the market size for a firm with marginal consumer θ is $(1 - G(\theta))$.

When parallel imports are allowed, the equilibrium structure is characterized by a family of arbitrage cutoffs and there is perfect sorting. When arbitrage is binding, firms choose to maximally differentiate prices between their home and export market, $p(\theta^*) = \tau p(\theta)$.[51] This shift between marginal home and foreign consumers when parallel imports are allowed is illustrated in Figure 21. Firms serve a higher marginal consumer in the export compared to the home market and the size of this shift is determined by the level of trade costs. This shifting arises due to a firm's comparative advantage in serving small export markets and large home markets due to higher production costs for exports.

The key insight from a continuous distribution is that due to costly trade, firms always choose a weakly smaller export market than the domestic market they serve. This confirms our findings from the discrete distribution that the targeted consumer group in the export market is always weakly richer than the targeted domestic consumer group.

When parallel imports are prohibited, the equilibrium structure is characterized by an export indifference cutoff $\hat{\theta}$. Firms are indifferent which marginal home consumer below the export cutoff and which which marginal foreign consumer above the cutoff to serve, as illustrated in Figure 22.

Finally, Section G in the appendix provides some thoughts of how our theoretical inequality-trade model can be taken to the data.

[51] There is one exception: home mass firms – firms with marginal home consumer $\underline{\theta}$ – are indifferent which marginal foreign consumer below the first arbitrage cutoff θ_0^* to serve.

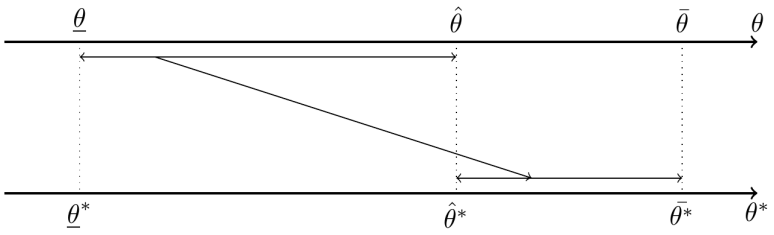


Figure 22: Choice of marginal home and foreign consumer for continuous endowment distribution when parallel imports prohibited

7 Conclusion

We present a demand-side, general equilibrium, two-country model of international consumer good trade featuring non-degenerate income distributions. This enables to distinguish between- versus within-country inequality when studying the effect of inequality on trade patterns in the presence of international arbitrage constraints. Using non-homothetic 0-1 preferences combined with a non-degenerate income distribution generates an extensive consumption margin within countries. We provide further empirical evidence for the Linder hypothesis that bilateral trade volume is increasing in the similarity of the demand structure between two countries. We are the first to provide a comprehensive theoretical foundation for the role of income distribution similarity for trade. Our model predicts trade intensity to be increasing in the income distribution overlap, which is our preferred proxy for demand similarity. This demand side heterogeneity shaping the patterns of trade should be considered as complementary to the heterogeneous firm approach common in the literature.

We show that for sufficiently high endowment differences between poor and rich consumers, world trade volume is strictly increasing in income distribution similarity. For high endowment differences, welfare of poor and rich consumers is lower under only within- compared to only between-county inequality. This effect is fully driven by costly trade dissipating resources. Furthermore, our model predicts a *Manhattan effect,* capturing that poor consumers are extremely badly off if they are a small minority in a rich country because of the high price level for basic products. Rich consumers always fare equally well irrespective of their domicile, while for poor consumers it can matter strongly where they live.

Having characterized the equilibrium for two and three labor endowment groups, the natural next step will be to generalize the model to n-groups or to a continuous endowment distribution. This would allow to easily map the model to empirically observed income distributions and take the model to the data to explain trade patterns such as export zeros and trade intensity. Our model can be used for welfare analysis such as studying the effect of trade liberalization or endowment redistribution and how trade costs and inequality interact. Furthermore, the model can provide policy recommendations on whether to allow or prohibit parallel imports. We find that generally the poor prefer to forbid parallel imports and so does the social planner. When there is only between-country inequality the rich prefer to allow parallel imports due to a resource feedback effect. In contrast, when there is only within-country inequality also the rich prefer to forbid parallel imports.

References

[1] Allen, T., and Arkolakis, C. Elements of Advanced International Trade. Lecture Notes, Northwestern University, 2016.

[2] Autor, D. H., Dorn, D., and Hanson, G. H. The China Syndrome: Local Labor Market Effects of Import Competition in the United States. *American Economic Review 103*, 6 (Oct. 2013), 2121–2168.

[3] Behrens, K., and Murata, Y. Globalization and individual gains from trade. *Journal of Monetary Economics 59*, 8 (Dec. 2012), 703–720.

[4] Bernasconi, C. Similarity of Income Distributions and the Extensive and Intensive Margin of Bilateral Trade Flows. *SSRN Electronic Journal* (2013).

[5] Bohman, H., and Nilsson, D. Market Overlap and the Direction of Exports - A New Approach of Assessing the Linder Hypothesis. CESIS Working Paper 86, 2007.

[6] Choi, Y. C., Hummels, D., and Xiang, C. Explaining import quality: The role of the income distribution. *Journal of International Economics 78*, 2 (July 2009), 293–303.

[7] Dalgin, M., Mitra, D., and Trindade, V. Inequality, Nonhomothetic Preferences, and Trade: A Gravity Approach. *Southern Economic Journal 74*, 3 (Jan. 2008), 747–774.

[8] De Loecker, J., and Eeckhout, J. The Rise of Market Power and the Macroeconomic Implications. NBER Working Paper 23687, National Bureau of Economic Research, Aug. 2017.

[9] Diebolt, C., and Kyrtsou, C., Eds. *New Trends in Macroeconomics*. Springer-Verlag, Berlin Heidelberg, 2005.

[10] Engel, Ernst. Die Productions- und Consumtionsverhältnisse des Königreichs Sachsen. *Zeitschrift des statistischen Bureaus des Königlich Sächsischen Ministerium des Inneren No. 8 and 9* (1857).

[11] Fajgelbaum, P. D., and Khandelwal, A. K. Measuring the Unequal Gains from Trade. *The Quarterly Journal of Economics 131*, 3 (Aug. 2016), 1113–1180.

[12] Fernandes, A. M., Klenow, P. J., Meleshchuk, S., Pierola, D., and Rodríguez-Clare, A. The Intensive Margin in Trade. NBER Working Paper 25195, National Bureau of Economic Research, Oct. 2018.

[13] Fieler, A. C. Nonhomotheticity and Bilateral Trade: Evidence and a Quantitative Explanation. *Econometrica 79*, 4 (2011), 1069–1101.

[14] Foellmi, R., Hepenstrick, C., and Zweimüller, J. International Arbitrage and the Extensive Margin of Trade between Rich and Poor Countries. *The Review of Economic Studies 85*, 1 (2018), 475–510.

[15] Galle, S., Rodríguez-Clare, A., and Yi, M. Slicing the Pie: Quantifying the Aggregate and Distributional Effects of Trade. NBER Working Paper 23737, National Bureau of Economic Research, Aug. 2017.

[16] Gaulier, G., and Zignago, S. BACI: International Trade Database at the Product-Level (the 1994-2007 Version). *SSRN Electronic Journal* (2010).

[17] Grossman, G. M., and Lai, E. L.-C. Parallel imports and price controls. *The RAND Journal of Economics 39*, 2 (June 2008), 378–402.

[18] Hall, R. E., and Jones, C. I. Why Do Some Countries Produce So Much More Output Per Worker Than Others? *The Quarterly Journal of Economics 114*, 1 (1999), 83–116.

[19] Hallak, J. A Product-Quality View of the Linder Hypothesis. *The Review of Economics and Statistics 92*, 3 (2010), 453–466.

[20] Handbury, J. Are Poor Cities Cheap for Everyone? Non-Homotheticity and the Cost of Living Across U.S. Cities. *Econometrica 89*, 6 (2021), 2679–2715. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.3982/ECTA11738.

[21] Hummels, D., and Klenow, P. J. The Variety and Quality of a Nation's Exports. *American Economic Review 95*, 3 (May 2005), 704–723.

[22] J. M. C. Santos Silva, and Tenreyro, S. The Log of Gravity. *The Review of Economics and Statistics 88*, 4 (Nov. 2006), 641–658.

[23] Jones, C. The Facts of Economic Growth. In *Handbook of Macroeconomics*, vol. 2. Elsevier, 2016, pp. 3–69.

[24] Kichko, S., Kokovin, S., and Zhelobodko, E. Trade patterns and export pricing under non-CES preferences. *Journal of International Economics 94*, 1 (Sept. 2014), 129–142.

[25] Krugman, P. Pricing to Market when the Exchange Rate Changes. NBER Working Paper 1926, National Bureau of Economic Research, May 1986.

[26] Linder, S. B. *An essay on trade and transformation.* Almqvist & Wiksell, Uppsala, 1961.

[27] Malueg, D. A., and Schwartz, M. Parallel imports, demand dispersion, and international price discrimination. *Journal of International Economics 37*, 3-4 (Nov. 1994), 167–195.

[28] Martínez-Zarzoso, I., and Vollmer, S. Bilateral Trade Flows and Income Distribution Similarity. *PLOS ONE 11*, 5 (May 2016).

[29] Maskus, K. E. Parallel Imports. *The World Economy 23*, 9 (2000), 1269–1284.

[30] Matsuyama, K. Engel's Law in the Global Economy: Demand-Induced Patterns of Structural Change, Innovation, and Trade. *Econometrica 87*, 2 (2019), 497–528.

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

[32] Melitz, M. J., and Ottaviano, G. I. P. Market Size, Trade, and Productivity. *The Review of Economic Studies 75* (2008), 295–316.

[33] Mitra, D., and Trindade, V. Inequality and Trade. *The Canadian Journal of Economics 38*, 4 (Nov. 2005), 1253–1271.

[34] Piketty, T. *Capital in the Twenty-First Century.* Harvard University Press, Cambridge, MA, 2014.

[35] Piketty, T., and Saez, E. Income Inequality in the United States, 1913-1998. *The Quarterly Journal of Economics 118*, 1 (Feb. 2003), 1–41.

[36] Roy, S., and Saggi, K. Equilibrium parallel import policies and international market structure. *Journal of International Economics 87*, 2 (July 2012), 262–276.

[37] Silber, J., Ed. *Handbook of Income Inequality Measurement*. Springer, New York, 1999.

[38] Simonovska, I. Income Differences and Prices of Tradables: Insights from an Online Retailer. *The Review of Economic Studies 82*, 4 (Oct. 2015), 1612–1656.

[39] Solow, R. M. A Contribution to the Theory of Economic Growth. *The Quarterly Journal of Economics 70*, 1 (1956), 65–94.

Appendix

A Income distributions and overlap

A.1 Calculate countries' income distributions from WIID

Based on Bernasconi (2013), we calculate countries' income distributions using WIID data. We here provide a step by step protocol how to calculate a country's empirical income distribution from the decile and quintile income shares data available in WIID combined with GDP and population data from PWT.

The goal is to calculate the income distribution $f_{at}(x)$ of country a at time t from decile s_a^d information, where s_a^d denotes the average share of total income in country a earned by decile $d \in \{1, ..., 10\}$ during a 10 year timespan. As inequality changes slowly over time, the 'best' observation within a time span is used for all years. To improve data quality as consistent surveys as possible are selected, where the criteria to choose the best observation include decile rather than quintile data, net income rather than gross income, adjustment to person inequality, individual level rather than household level, high quality, etc. (see Bernasconi (2013) for details). If no decile data is available, quintile shares s_a^q are used instead, where $q \in \{1, ..., 5\}$.

1. Calculate the average per capita income level for each decile:

$$x_{at}^d = \frac{s_a^d \cdot GDP_{at}}{POP_{at}/10},$$

   where GDP and population are taken from PWT for year t.

2. Redistribute each 0.1 point mass of x_{at}^d uniformly over the interval $[\underline{\underline{x}}_{at}^d, \overline{\overline{x}}_{at}^d]$.

   - Calculate lower and upper bounds of decile's redistribution interval:

$$\underline{\underline{x}}_{at}^d = \begin{cases} x_{at}^d - (x_{at}^d - x_{at}^{d-1})/2 & if\ d > 1 \\ 1 & if\ d = 1 \end{cases},$$

$$\overline{\overline{x}}_{at}^d = \begin{cases} x_{at}^d + (x_{at}^{d+1} - x_{at}^d)/2 & if\ d < 10 \\ x_{at}^{10} + 2.7(x_{at}^{10} - x_{at}^9) & if\ d = 10 \end{cases},$$

     where scaling by 2.7 for the upper bound of the 10th interval is estimated from Luxembourg Income Study data.

   - Calculate density over decile's redistribution interval:

$$f_{at}^d = 0.1/(\overline{\overline{x}}_{at}^d - \underline{\underline{x}}_{at}^d)$$

3. Partition a country's income distribution into 100 income intervals $\breve{x}$ with length $h = 5'000$, $\breve{x} \in \{[1 - 5'000], ..., [495'001 - 500'000]\}$ to then calculate density $f_{at}(\breve{x})$ of these intervals.

   - Calculate the weights of deciles $w_{at}^d(\breve{x})$ within a certain interval (4 cases):
     - Entire decile lies within interval:

$$w_{at}^d(\breve{x}) = (\overline{\overline{x}}_{at}^d - \underline{\underline{x}}_{at}^d)/h \qquad if\ [\underline{\underline{x}}_{at}^d, \overline{\overline{x}}_{at}^d] \subset \breve{x}$$

     - Decile starts below interval and ends inside:

$$w_{at}^d(\breve{x}) = (\overline{\overline{x}}_{at}^d - \underline{\breve{x}})/h \qquad if\ \underline{\underline{x}}_{at}^d < \breve{x}\ and\ \overline{\overline{x}}_{at}^d \in \breve{x},$$

where $\underline{\underline{\breve{x}}}$ denotes the lower bound of the income interval $\breve{x}$.

- Decile starts in interval and ends above:

$$w_{at}^{d}(\breve{x}) = (\overline{\overline{\breve{x}}} - \underline{\underline{x}}_{at}^{d})/h \qquad if\, \underline{\underline{x}}_{at}^{d} \in \breve{x}\, and\, \overline{\overline{x}}_{at}^{d} > \breve{x},$$

where $\overline{\overline{\breve{x}}}$ denotes the upper bound of the income interval $\breve{x}$.

- Decile starts below interval and ends above:

$$w_{at}^{d}(\breve{x}) = 1 \qquad if\, [\underline{\underline{x}}_{at}^{d}, \overline{\overline{x}}_{at}^{d}] \supset \breve{x}$$

4. Calculate density of the income interval using weights of deciles:

$$f_{at}(\breve{x}) = h \cdot \sum_{d=1}^{10} w_{at}^{d}(\breve{x}) \cdot f_{at}^{d}$$

- By definition densities sum to 1:

$$\sum_{\breve{x}} f_{at}(\breve{x}) = 1$$

This yields a discrete empirical income distribution for each country and year.

A.2 Unnormalized trade margins

For our two-country model we define trade margins unstandardized as in Fernandes et al. (2018) and already use that in our model each variety will be produced by a different firm.

The value of total exports from o to d is

$$\widetilde{V_{od}} = \sum_{i \in I} v_{odi}$$

The extensive margin is the total number of firms in o exporting to d

$$\widetilde{EM_{od}} = \sum_{i \in I} 1_{odi}$$

The intensive margin is the average export value per firm in o exporting to d

$$\widetilde{IM_{nc}} = \frac{\widetilde{V_{nc}}}{\widetilde{EM_{nc}}} = \frac{\sum_{i \in I} v_{nci}}{\sum_{i \in I} 1_{nci}}$$

A.3 Additional trade margin regressions

BACI and EDD trade margins regressions excluding zeros

The shares of export zeros are 25% and 38% in the BACI and EDD trade data set, respectively. Tables 5 and 6 show trade margin regressions for these two alternative data sources when dropping zero trade flows. Comparison to Tables 1 and 2 confirms that the main findings are independent of how trade zeros are treated.

Table 4: Trade margins BACI and income overlaps, PPML estimation

	V_{od}	EM_{od}	IM_{od}
Overlap	-0.343	-0.005	0.605
	(0.239)	(0.072)	(0.399)
# obs	8,556	8,556	8,556
# regs	193	193	193
Mean	162469.7	111.9	297.4
SD	1445761.3	219.6	1490.5

Notes: ***, **, * denote statistical significance on the 1%, 5%, and 10% level, respectively. Robust standard errors are given in parentheses. Controls: geographic distance, dummies for free trade agreement, currency union, common border, common legal system, common language, colonial ties, dummy variable allowing for a different intercept for NN, SS, NS and SN trade flows, importer and exporter fixed effects. Sample: countries with population > 1 million, HS6 codes which include consumer goods. Year=2002.

Table 5: Trade margins BACI (excluding zeros) and income overlap

	$ln(V_{od})$	$ln(EM_{od})$	$ln(IM_{od})$
Overlap	1.250***	0.583***	0.667***
	(0.137)	(0.074)	(0.106)
# obs	6,455	6,455	6,455
# regs	193	193	193
Adj R^2	0.804	0.833	0.581
Mean	7.764	3.447	4.318
SD	3.471	2.037	1.810

Notes: ***, **, * denote statistical significance on the 1%, 5%, and 10% level, respectively. Robust standard errors are given in parentheses. Controls: geographic distance, dummies for free trade agreement, currency union, common border, common legal system, common language, colonial ties, dummy variable allowing for a different intercept for NN, SS, NS and SN trade flows, importer and exporter fixed effects. Sample: countries with population > 1 million, HS6 codes which include consumer goods. Year=2002.

Table 6: Trade margins EDD (excluding zeros) and income overlap

	$ln(V_{od})$	$ln(EM_{od})$	$ln(IM_{od})$
Overlap	1.256***	0.756***	0.500***
	(0.259)	(0.132)	(0.194)
# obs	2,379	2,379	2,379
# regs	149	149	149
Adj R^2	0.680	0.806	0.378
Mean	15.680	4.041	11.639
SD	3.327	2.243	1.796

Notes: ***, **, * denote statistical significance on the 1%, 5%, and 10% level, respectively. Robust standard errors are given in parentheses. Controls: geographic distance, dummies for free trade agreement, currency union, common border, common legal system, common language, colonial ties, dummy variable allowing for a different intercept for NN, SS, NS and SN trade flows, importer and exporter fixed effects. Sample: countries with pop > 1 million, HS2 codes which include consumer goods. Year=2012.

B Preferences

This section briefly discusses the possibility of adding an intensive margin and a hierarchy in the order products are consumed.

0-1 preferences can can be combined with CES preferences in order to generate an extensive *and* intensive margin. FHZ have shown that results remain qualitatively similar when adding an intensive margin. As long

as there is some extensive margin response, GDP per capita and the entire distribution of income will matter for the patterns of trade. Alternatively, one could consider hierarchical preferences that provide a ranking in which order the varieties are consumed as income increases. This would solve the ex-ante indeterminacy of who consumes which products. It is then intrinsic to the firm given its blueprint whether they produce a necessity or a luxury product.[52]

C Two endowment groups

This part provides the detailed deviations for the regimes when parallel imports are allowed and further results for the preferred parallel trade regimes when there this both within- and between-country inequality.

C.1 Conditions for no deviation incentives for PI2

We show that profits of all other strategies are weakly negative. The condition for the arbitrage constraint $p_R \leq \tau p_P^*$ to be binding is

$$\tau \leq \frac{1}{2}\left(\sqrt{a^2F^2 - 4\alpha_P(aF+3) + 10aF + 4\alpha_P^2 + 9} - aF + 2\alpha_P - 1\right). \tag{11}$$

Binding arbitrage implies that separating firms cannot set domestic and export prices freely. However, for PR firms to be unprofitable, the arbitrage constraint has to be binding sufficiently strongly such that the pricing restriction is strong enough. The condition to avoid deviation to PR firms is

$$\tau \leq \frac{1}{4}\sqrt{4F^2a^2 + 52Fa + 49)} - \frac{Fa}{2} - 3/4, \tag{12}$$

to avoid deviation to PR^* firms is

$$\tau \leq \frac{1}{6}\sqrt{9F^2a^2 + 72Fa + 64} - \frac{Fa}{2} - 1/3, \tag{13}$$

and to RR^* firms is

$$\Pi_{RR^*} = p_R - \left(F + \frac{\alpha_P + \tau(1-\alpha_P)}{a}\right) \leq 0,$$

which always holds for $\alpha_P \leq 1/2$.

[52] However, this indeterminacy is unproblematic as long as we only look at equilibrium outcomes. A hierarchy might be at work in background but does not need to be modeled explicitly.

We also need to show that there are no domestic firms (P, R, R^*).

$$\begin{aligned}
\Pi_P &= p_P\mathcal{P} - \left(F + \frac{\mathcal{P}}{a}\right) \leq 0 \\
&if \quad \tau \leq aF + 1 \\
\Pi_R &= p_R\frac{\mathcal{P}}{2} - \left(F + \frac{\mathcal{P}}{2a}\right) \leq 0 \\
&if \quad \tau \leq (\alpha_P - Fa - 1)/(\alpha_P - 1) \\
\Pi_R^* &= p_R\frac{\mathcal{P}}{2} - \left(F + \frac{\mathcal{P}}{2a}\right) \leq 0 \\
&if \quad \tau \leq (\alpha_P^2 - \alpha_P Fa + Fa)/\alpha_P^2
\end{aligned}$$

The most restrictive condition is the one for $\Pi_{PR}^* \leq 0$. Thus, the condition for PI2 to exist is given by (13).

C.2 Mostly within-county inequality with large endowment group differences (PI4)

All goods are traded and there are only E firms in home and M+E firms in foreign.
Zero profit conditions are

$$\begin{aligned}
p_P + p_P^* &= W^*\left(F + \tfrac{1+\tau}{a}\right) \\
p_R\left(1-\alpha_P\right) + p_R^*\alpha_P &= W\left(F + \tfrac{1-\alpha_P+\tau\alpha_P}{a}\right) \\
p_R\left(1-\alpha_P\right) + p_R^*\alpha_P &= W^*\left(F + \tfrac{\tau(1-\alpha_P)+\alpha_P}{a}\right)
\end{aligned}$$

Incremental budget constraints are

$$\begin{aligned}
p_P N_{PP}^* = \underline{\theta}W \qquad &\text{and} \qquad p_R\left(N_{RR} + N_{RR}^*\right) = (\bar{\theta} - \underline{\theta})W, \\
p_P^* N_{PP}^* = \underline{\theta}W^* \qquad &\text{and} \qquad p_R^*\left(N_{RR} + N_{RR}^*\right) = (\bar{\theta} - \underline{\theta})W^*.
\end{aligned}$$

Trade balance is

$$p_P N_{PP}^* + p_R N_{RR}^*(1-\alpha_P) = p_R^* N_{RR}\alpha_P.$$

The resource constraints are

$$\begin{aligned}
\alpha_P\underline{\theta} + (1-\alpha_P)\bar{\theta} &= N_{RR}\left(F + \tfrac{1-\alpha_P+\tau\alpha_P}{a}\right) \\
(1-\alpha_P)\underline{\theta} + \alpha_P\bar{\theta} &= N_{PP}^*\left(F + \tfrac{1+\tau}{a}\right) + N_{RR}^*\left(F + \tfrac{\alpha_P+\tau(1-\alpha_P)}{a}\right).
\end{aligned}$$

The prices, obtained from the ZPs and BCs, are

$$\begin{aligned}
\omega &= \frac{aF+1-\alpha_P+\tau\alpha_P}{aF+\tau-\alpha_P\tau+\alpha_P} \\
p_P &= \frac{F+\frac{1+\tau}{a}}{1+\frac{1}{\omega}} = \frac{F+\frac{1+\tau}{a}}{1+\frac{aF+\tau-\alpha_P\tau+\alpha_P}{aF+1-\alpha_P+\tau\alpha_P}} \\
p_P^* &= \omega p_P = \frac{aF+1-\alpha_P+\tau\alpha_P}{aF+\tau-\alpha_P\tau+\alpha_P}\frac{F+\frac{1+\tau}{a}}{1+\frac{aF+\tau-\alpha_P\tau+\alpha_P}{aF+1-\alpha_P+\tau\alpha_P}} \\
p_R &= \frac{F+\frac{1-\alpha_P+\tau\alpha_P}{a}}{1-\alpha_P+\omega\alpha_P} \\
p_R^* &= \omega p_R.
\end{aligned}$$

The varieties, obtained by combining RCs and BCs, are

$$\begin{aligned}
N_{PP}^* &= \frac{\underline{\theta}}{p_P} \\
N_{RR} &= \frac{\alpha_P\underline{\theta}+(1-\alpha_P)\bar{\theta}}{F+\frac{1-\alpha_P+\tau\alpha_P}{a}} \\
N_{RR}^* &= \frac{\bar{\theta}-\underline{\theta}}{p_R} - N_{RR}.
\end{aligned}$$

C.3 Only between-county inequality with small endowment group differences (PI6)

In the home country, there are only rich consumers, in the foreign country only poor ones and all goods are sold to all consumers. There are only M firms active in both countries. This will prevail as long as $p \leq \tau p^*$ which holds for $\underline{\theta} \geq \frac{2}{1+\tau}$. $\omega = 1$ if $\mathcal{P} = \mathcal{P}^*$ due to the existence of M firms in both countries. The prices, obtained by combining BCs and ZPs, are

$$\begin{aligned}
p &= \frac{\left(\frac{F}{\mathcal{P}}+\frac{1+\tau}{a}\right)\bar{\theta}}{\bar{\theta}+\underline{\theta}} \\
p^* &= \frac{\left(\frac{F}{\mathcal{P}}+\frac{1+\tau}{a}\right)\underline{\theta}}{\bar{\theta}+\underline{\theta}}.
\end{aligned}$$

The varieties, obtained from the RCs, are

$$\begin{aligned}
N_M &= \frac{\bar{\theta}}{\frac{F}{\mathcal{P}}+\frac{1+\tau}{a}} \\
N_M^* &= \frac{\underline{\theta}}{\frac{F}{\mathcal{P}}+\frac{1+\tau}{a}}.
\end{aligned}$$

C.4 Two group results

C.4.1 Proof of Result 1

Result 1 (polar cases). *For sufficiently high endowment differences, a sufficient condition being* $\frac{\bar{\theta}}{\underline{\theta}} > 2\tau-1$*, world trade volume is higher under only within compared to only between-country inequality. However, world net output is lower and hence so is world welfare.*

Proof. Under only within-country inequality, world trade volume is $TVO^{PI1} = 2N_{PP}^{PI1} + N_{RR}^{PI1}$. Under only between-country inequality, world trade volume is $TVO^{PI7} = N_{PR}^{*PI7} + N_{RP}^{PI7}$. Thus, there is higher trade under only within-country if

$$TVO^{PI1} = \frac{2\underline{\theta}}{F + \frac{1+\tau}{a}} + \frac{\bar{\theta} - \underline{\theta}}{2F + \frac{1+\tau}{a}} > \frac{(1+\tau)\,\underline{\theta}}{F + \frac{1+\tau}{a}} = TVO^{PI7}$$

$$\frac{\frac{\bar{\theta}}{\underline{\theta}} - 1}{2F + \frac{1+\tau}{a}} > \frac{\tau - 1}{F + \frac{1+\tau}{a}}. \tag{14}$$

Note that for $\frac{\bar{\theta}}{\underline{\theta}} = \tau$, which is the lowest value of $\frac{\bar{\theta}}{\underline{\theta}}$ where the arbitrage equilibrium (PI7) exists, condition (14) is violated. This means that for low endowment differences, trade with perfect overlap (PI1) is lower than with no overlap (PI7). However, trade is higher under only with-country inequality, when endowment differences are sufficiently high

$$\frac{\bar{\theta}}{\underline{\theta}} > \frac{2F + \frac{1+\tau}{a}}{F + \frac{1+\tau}{a}}(\tau - 1) + 1.$$

A sufficient condition for this is

$$\frac{\bar{\theta}}{\underline{\theta}} > 2(\tau - 1) + 1 = 2\tau - 1.$$

Under only within-country inequality, world net output is $4N_{PP}^{PI1} + 2N_{RR}^{PI1}$. Under only between-country inequality, world net output is $2N_{PR}^{*PI7} + 2N_{RP}^{PI7} + N_{R}^{PI7}$. World net output is higher under only within-country inequality if

$$\frac{4\underline{\theta}}{F + \frac{1+\tau}{a}} + \frac{2\left(\bar{\theta} - \underline{\theta}\right)}{2F + \frac{1+\tau}{a}} > \frac{2\left(1+\tau\right)\underline{\theta}}{F + \frac{1+\tau}{a}} + \frac{\bar{\theta} - \tau\underline{\theta}}{F + \frac{1}{a}}$$

$$\frac{\underline{\theta}\left(\tau - \frac{\bar{\theta}}{\underline{\theta}}\right)}{F + \frac{1}{a}} + \frac{2\underline{\theta}\left(\frac{\bar{\theta}}{\underline{\theta}} - 1\right)}{2F + \frac{1+\tau}{a}} > \frac{2\underline{\theta}\,(\tau - 1)}{F + \frac{1+\tau}{a}}. \tag{15}$$

Note that for $\frac{\bar{\theta}}{\underline{\theta}} = \tau$, we have

$$\frac{1}{2F + \frac{1+\tau}{a}} < \frac{1}{F + \frac{1+\tau}{a}},$$

which means that world net output is *larger* under only between-country compared to only under within-country. When endowment differences are larger than τ, inequality (15) is even more violated because the left hand side is decreasing in $\frac{\bar{\theta}}{\underline{\theta}}$

$$\frac{\partial}{\partial \frac{\bar{\theta}}{\underline{\theta}}}\left[\frac{\underline{\theta}\left(\tau - \frac{\bar{\theta}}{\underline{\theta}}\right)}{F + \frac{1}{a}} + \frac{2\underline{\theta}\left(\frac{\bar{\theta}}{\underline{\theta}} - 1\right)}{2F + \frac{1+\tau}{a}}\right] =$$

$$-\frac{\underline{\theta}}{F + \frac{1}{a}} + \frac{2\underline{\theta}}{2F + \frac{1+\tau}{a}} = -\frac{2\underline{\theta}}{2F + \frac{2}{a}} + \frac{2\underline{\theta}}{2F + \frac{1+\tau}{a}} < 0$$

since $\tau > 1$. Note that world net output is identical to world consumption. This implies that world net output and world welfare are always *lower* under only within-country inequality. ■ □

C.4.2 Proof of Result 4

Result 4. *If $\frac{\bar{\theta}}{\underline{\theta}}$ is high, welfare of poor and rich consumers is lower under only within- compared to only between-country inequality.*

Proof. Under only within-country inequality, welfare of the poor is $U_P^{PI1} = \frac{2\underline{\theta}}{F+\frac{1+\tau}{a}}$, while welfare of the rich is $U_R^{PI1} = \frac{2\underline{\theta}}{F+\frac{1+\tau}{a}} + \frac{\bar{\theta}-\underline{\theta}}{F+\frac{1+\tau}{2a}}$. Under only between-country inequality, welfare of the poor is $U_P^{*PI7} = \frac{\underline{\theta}(1+\tau)}{F+\frac{1+\tau}{a}}$, while welfare of the rich is $U_R^{PI7} = \frac{\underline{\theta}(1+\tau)}{F+\frac{1+\tau}{a}} + \frac{\bar{\theta}-\tau\underline{\theta}}{F+\frac{1}{a}}$. Welfare of the poor is lower under within-country inequality since $\tau > 1$. Welfare of the rich under only within-country inequality is *higher,* if

$$U_R^{PI1} = \frac{2\underline{\theta}}{F+\frac{1+\tau}{a}} + \frac{2\bar{\theta}-\underline{\theta}}{F+\frac{1+\tau}{2a}} > \frac{(1+\tau)\underline{\theta}}{F+\frac{1+\tau}{a}} + \frac{\bar{\theta}-\tau\underline{\theta}}{F+\frac{1}{a}} = U_R^{PI7}$$

$$\frac{\underline{\theta}\left(\tau-\frac{\bar{\theta}}{\underline{\theta}}\right)}{F+\frac{1}{a}} + \frac{\underline{\theta}\left(\frac{\bar{\theta}}{\underline{\theta}}-1\right)}{F+\frac{1+\tau}{2a}} > \frac{\underline{\theta}(\tau-1)}{F+\frac{1+\tau}{a}} \tag{16}$$

Consider the knife-edge case $\frac{\bar{\theta}}{\underline{\theta}} = \tau$, when the difference between rich and poor is at the edge between full and partial trade for only between-country inequality. Then we have

$$\frac{\underline{\theta}(\tau-1)}{F+\frac{1+\tau}{2a}} > \frac{\underline{\theta}(\tau-1)}{F+\frac{1+\tau}{a}},$$

which always holds. Thus, welfare of the rich under only within- is higher than under only-between country inequality. Now consider larger values of $\frac{\bar{\theta}}{\underline{\theta}}$ such that there is partial trade (PI7). The left hand side of inequality (16) is decreasing in $\frac{\bar{\theta}}{\underline{\theta}}$

$$\frac{\partial}{\partial\frac{\bar{\theta}}{\underline{\theta}}}\left[\frac{\underline{\theta}\left(\tau-\frac{\bar{\theta}}{\underline{\theta}}\right)}{F+\frac{1}{a}} + \frac{\underline{\theta}\left(\frac{\bar{\theta}}{\underline{\theta}}-1\right)}{F+\frac{1+\tau}{2a}}\right]$$

$$= -\frac{\underline{\theta}}{F+\frac{1}{a}} + \frac{\underline{\theta}}{F+\frac{1+\tau}{2a}} < 0$$

since $\tau > 1$. This means that $U_R^{PI1} < U_R^{PI7}$ as $\frac{\bar{\theta}}{\underline{\theta}}$ grows large. Eventually, rich welfare is *lower* under only within-country inequality. ■ □

C.5 Preferred parallel trade with within- and between-country inequality

Proposition 5. *When there is mostly within- but also between-country inequality ($\alpha_P \in (0.26, 0.5)$) and group endowment differences are large, both poor and rich consumers prefer to* ***forbid*** *parallel trade. The social planner prefers to forbid parallel trade.*

Proof. For poor consumers

$$U_P^{noPI3} = \frac{\underline{\theta}}{\frac{F}{2}+\frac{3+\tau}{4a}} > \frac{\underline{\theta}}{\frac{(aF+1-\alpha_P+\tau\alpha_P)(aF+1+\tau)}{a(2aF+1+\tau)}} = U_P^{PI4}$$

since $4\alpha_P(aF\tau - aF + \tau^2 - 1) > \tau^2 - 1$ for $\alpha > 0.26$.

For rich consumers

$$U_R^{noPI3} = U_P^{noPI3} + \frac{\bar{\theta} - \underline{\theta}}{F + \frac{1+\tau}{2a}} > U_P^{PI4} + \frac{\bar{\theta} - \underline{\theta}}{\frac{F + \frac{1-\alpha_P+\tau\alpha_P}{a}}{1-\alpha_P+\alpha_P \frac{aF+1-\alpha_P+\tau\alpha_P}{aF+\tau-\alpha_P\tau+\alpha_P}}} = U_R^{PI4}$$

since $U_P^{noPI3} >> U_P^{PI4}$ while $\frac{\bar{\theta}-\underline{\theta}}{F+\frac{1+\tau}{2a}} < \frac{F+\frac{1-\alpha_P+\tau\alpha_P}{a}}{1-\alpha_P+\alpha_P\frac{aF+1-\alpha_P+\tau\alpha_P}{aF+\tau-\alpha_P\tau+\alpha_P}}$. □

Proposition 6. *When there is mostly within- but also between-country inequality ($\alpha_P \in (0.26, 0.5)$) and group endowment differences are small, rich consumers prefer to* ***allow*** *parallel trade while poor consumers are indifferent. The social planner prefers to allow parallel trade.*

Proof. For poor consumers

$$U_P^{noPI1} = \frac{2\underline{\theta}}{F + \frac{1+\tau}{a}} = \frac{2\underline{\theta}}{F + \frac{1+\tau}{a}} = U_P^{PI2}.$$

For rich consumers

$$U_R^{noPI1} = U_P^{noPI1} + \frac{\bar{\theta} - \underline{\theta}}{\frac{1}{2(1-\alpha_P)}\left(F + \frac{1+\tau-2\tau\alpha_P}{a}\right)} < U_P^{PI2} + \frac{\bar{\theta} - \underline{\theta}}{F + \frac{1-\alpha_P+\tau\alpha_P}{a}} = U_R^{PI2}$$

since $\tau > 1$ and therefore also $U^{noPI1} > U^{PI2}$. □

Proposition 7. *When there is mostly between- but also within-country inequality ($\alpha_P \in (0, 0.26)$) and group endowment differences are large, rich consumers prefer to* ***forbid*** *parallel trade while poor consumers in both countries prefer to allow it. The social planner prefers to forbid parallel trade.*

Proof. Verified by simulation. □

Proposition 8. *When there is mostly between- but also within-country inequality ($\alpha_P \in (0, 0.26)$) and group endowment differences are small, rich consumers prefer to* ***forbid*** *parallel trade while poor consumers in home also prefer to* ***forbid*** *but poor consumer in foreign prefer to* ***allow****. The social planner prefers to forbid parallel trade.*

Proof. Verified by simulation. □

D Imperfect enforcement of parallel import prohibition

An interesting exercise is also to consider imperfect enforcement of parallel import prohibition. Assume that prior to entry, firms only know the probability with which the prohibition of parallel imports will be enforced but not whether it will be enforced for their product specifically. A firm learns whether parallel import prohibition is enforced for the product it produces only after entry. The policy enforcement parameter ρ captures the fraction of products for which parallel import prohibition is enforced. In other words, there is imperfect enforcement of the parallel import prohibition law. To start, take the simplest case of identical countries ($\alpha_P = 0.5$), low within-country inequality and parallel import prohibition enforcement for half of the products ($\rho = 0.5$). In the pure cases, where parallel import prohibition is either fully enforced ($\rho = 1$) or parallel import are possible for all products ($\rho = 0$), the regimes noPI1 (M+S) and PI1 (M+E) would prevail, respectively. The enforcement of parallel import prohibition is only relevant for separating firms,

since only their price setting is affected by the arbitrage threat. In the hybrid case, profits for separating firms only need to be zero in expectation. The expected profits for separating firms and for exclusive firms in the home country are

$$\begin{aligned} E[\Pi_S] &= p_P\mathcal{P} + (\rho p_R^* + (1-\rho)\tau p_P)\,\alpha\mathcal{P}^* - \left(F + \frac{\mathcal{P} + \tau\alpha\mathcal{P}^*}{a}\right) \\ \Pi_E &= p_R\mathcal{P}\,(1-\alpha) + p_R^*\mathcal{P}^*\,(1-\alpha^*) - \left(F + \frac{\mathcal{P}\,(1-\alpha) + \tau\mathcal{P}^*\,(1-\alpha^*)}{a}\right). \end{aligned}$$

It turns out that $E[\Pi_S] = 0$ and $\Pi_E = 0$ both pin down a solution for p_R and these two solutions are typically inconsistent. Hence, separating and exclusive firms can never both be present in an equilibrium. Note that $p_P = \frac{F}{2} + \frac{1+\tau}{2a}$ is determined by $\Pi_M = 0$ since mass firms always exist irrespective of parallel import prohibition enforcement. Given $\rho = \alpha = 0.5$, there is a critical level of trade costs τ^* which determines whether only separating or only exclusive firms exist. If $\tau < \tau^* = 1/2(\sqrt{16 + 12aF + a^2F^2} - 2 - aF)$ then there are only exclusive (and mass) firms. Whereas, if $\tau > \tau^*$ then there are only separating (and mass) firms. This is intuitive since lower trade cost imply less scope for price discrimination. Hence, separating firm's profits decrease in trade costs. The critical trade cost level τ^* is decreasing in ρ as a higher probability of parallel import prohibition enforcement makes separating firms more attractive due to more often being able to set the unconstrained price.

E Three endowment groups

E.1 Welfare and prices

E.2 Equilibrium for three groups arbitrage non-binding

When τ is sufficiently large, arbitrage constraints become non-binding. For identical countries ($\alpha_P = \alpha_M = \alpha_R = 1/2$), prices are independent of endowment levels. Hence, whether arbitrage is binding or not only depends on τ. When τ is large, poor/middle arbitrage becomes non-binding, $p_M < \tau p_P$. In equilibrium, all goods are traded and there are MM, RR and PM firms in both countries. When τ is even larger, both poor/middle and middle/rich arbitrage become non-binding, $p_M < \tau p_P$ and $p_R < \tau p_M$. In equilibrium, all goods are traded and there are MM, PM and MR firms in both countries.[53]

F n endowment groups

Next, we generalize the 3-group model to n endowment groups. There are n endowment levels in the world: $\theta_1, \theta_2, \ldots, \theta_n$. As before, we denote Northern share of the world's group i consumers $\mathcal{P}_i$ by α_i. [54] Firm types are denoted with two indices ij with $i, j = 1, \ldots, n$. We keep the global number of consumers per income group constant: $\mathcal{P}_1 = \ldots = \mathcal{P}_n = (\mathcal{P} + \mathcal{P}^*)/n$. We impose a wide endowment level range such that arbitrage constraints are binding, $p_{i+1} > \tau p_i$. Countries' population is constant and identical, $\mathcal{P} = \mathcal{P}^*$. $(n-1)$ α parameters need to be set to define a global income distribution. One caveat to bear in mind is that a model

[53] Even though prices are independent of endowment levels, for the "poor/middle and middle/rich arbitrage non-binding" equilibrium to exist endowment levels need to be sufficiently *apart*. Otherwise, $N_{MM} < 0$.

[54] Expressing the endowment distribution in world shares rather than country shares is convenient because it directly determines how many consumers of each endowment group are located in the other country.

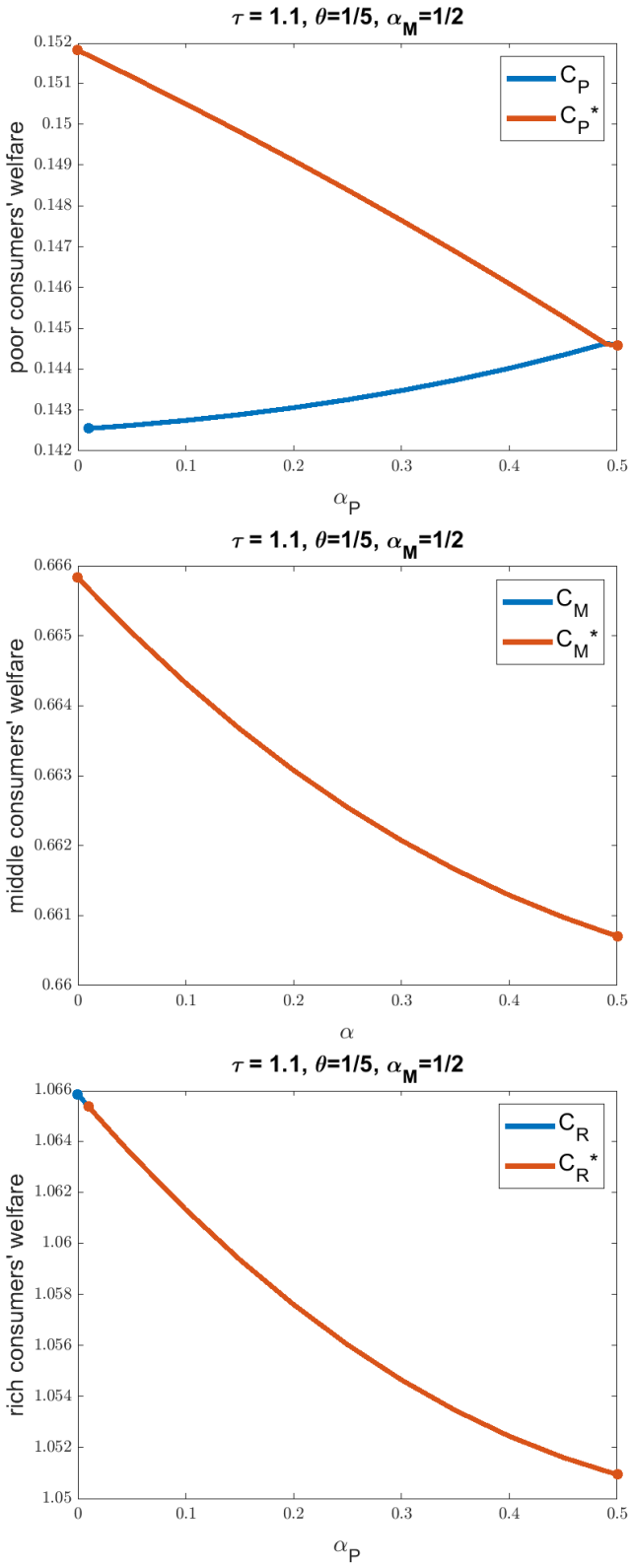


Figure 23: Poor, middle and rich consumers' welfare for $\alpha_M = 1/2$ cross-section

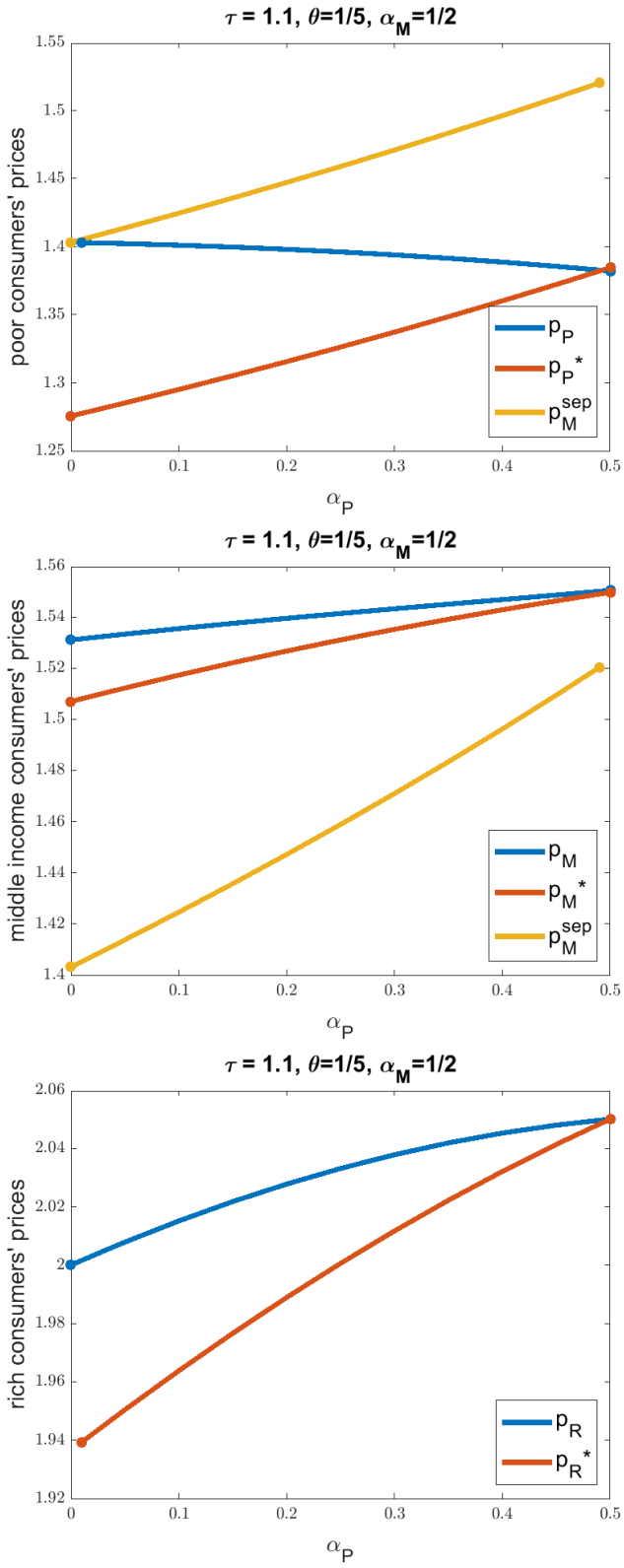
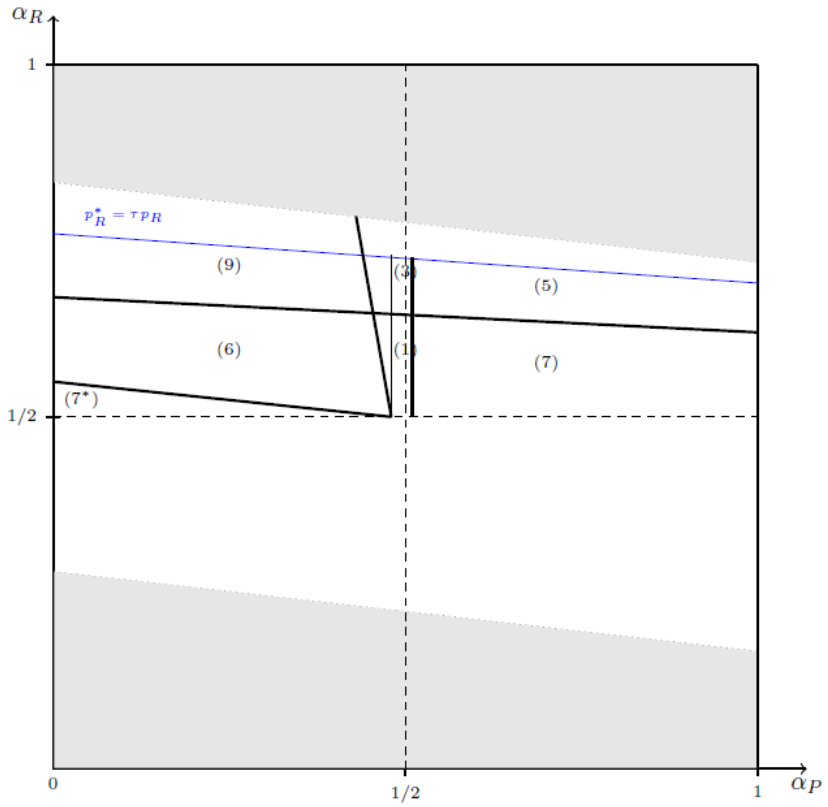


Figure 24: Poor, middle and rich consumers' prices for $\alpha_M = 1/2$ cross-section



(1a)	N: MM + RR	S: PP + MM
(1b)	N: PP + MM + RR	S: MM
(3a)	N: MM + RR	S: PP + MM + MR
(3b)	N: PP + MM + RR	S: MM + MR
(5)	N: PP + MM + RR + PM	S: MM + MR
(6)	N: MM + RR	S: PP + MM + PM
(6I)	N: PP + MM + RR	S: PP + MM + PM
(7)	N: PP + MM + RR + PM	S: MM
(7*)	N: MM	S: PP + MM + RR + PM
(9)	N: MM + RR	S: PP + MM + PM + MR

Figure 25: Three-group parallel import regimes for wide endowment range and constant country endowment

with more income groups will tend to predict fewer export zeros since there are few consumers in the richest group in the North. Hence, there is a smaller share of varieties produced exclusively for the richest groups.[55]

F.1 n identical countries

If the n counties are identical the system of equations is

$$ZP_{ii}: \quad 2p_i \sum_{k=i}^{n} \alpha_k = W\left(F + \frac{(1+\tau)\sum_{k=i}^{n}\alpha_k}{a}\right)$$

$$BC_i: \quad (\theta_i - \theta_{i-1})W = p_i(N_i + N_i^*), \quad \text{where } \theta_0 = 0$$

$$RC: \quad \sum_{i=1}^{n} \theta_i \alpha_i = \sum_{i=1}^{n} N_i \left(F + \frac{(1+\tau)\sum_{k=i}^{n}\alpha_k}{a}\right)$$

We therefore have n prices, n firm densities and the relative wage. Hence, $2n+1$ unknowns. There are n ZPs, n BCs, 1 RC equations. Hence, $2n+1$ conditions. Simulating this for n groups delivers some general results:

- For identical countries ($\alpha_i = 1/2$ $i = 1, \ldots n$), prices are *independent* of endowment. Hence, whether arbitrage binding or not only depends on τ.
- As the number of groups increases, poor prices get closer together such that arbitrage is non-binding anymore and incentive to serve smaller export than home market (separating firms) increase.
- Trade costs determine the max number of groups under which there are only non-separating firms. For 2 groups $\tau \leq 1.23$, for 3 groups $\tau \leq 1.1$, for 4 groups $\tau \leq 1.05$, for 9 groups $\tau \leq 1.01$. For a higher number of groups arbitrage becomes non-binding and separating firms emerge.
  - E.g. for 4 groups: arbitrage between group 1 and 2 becomes non-binding: $\tau p_1 > p_2$ and 12 separating firms emerge. All goods are traded. There are 22, 33, 44 and 12 firms in both countries

G Taking the model to the data

This section outlines the basis for taking our theoretical inequality-trade model to the data. It serves as an outlook what ideally can be achieved by further refining the model. Ultimately we are pursuing to formulate the model with a general endowment distribution allowing for asymmetry between countries, which can then directly be matched to the empirically observed income distributions as described in Section 2.

One way to move towards a general distribution is to generalize our three-group model to a n-group model as sketched in appendix Section F or to use a continuous distribution as suggested in Section 6.

An alternative to keeping country population constant and identical is to allow for migration between countries. When poor consumers move from one country to the other, this changes countries' populations and endowments. Furthermore, in light of the global reduction in poverty over recent decades, one might also want to study changes in the global income distribution. A natural application of our two-country setting would be to study the development of the interaction between the two major world economies – China and the Untied States. Over the past decades, China has emerged from a poor country to a middle-income country,

[55] Unless the Southern country has zero mass in many of the top income groups present in the North, then export zeros may occur similarly often as with few groups.

which resulted in a larger world middle class and less poor consumers in China and worldwide. To a first order, global population remained constant and hence China nowadays accommodates a higher share of the world's middle class than in 1980.

G.1 Empirically relevant income distributions

In this section we illustrate how our three group model can be applied to empirically relevant income distributions. We to use the World Bank's GNI per capita based income groups classification to split the world income distribution into three groups.[56] The poor group consists of WB's low-income and lower middle-income consumers. For 2013, this corresponds to consumers with a yearly income below 4'125$. The middle income group consists of WB's upper middle-income consumers, which corresponds to consumers with an income below 12'745$. Finally the rich income group consists of WB's high-income consumers.

The empirical income distributions of the two major world economies – USA and China – depicted in Figure 2 can be represented in this three group categorization. Figure 26 shows the world shares $\alpha_P, \alpha_P^*, \alpha_M, \alpha_M^*, \alpha_R, \alpha_R^*$ representation, where α_P is be the Northern share of the world's poor consumers $\mathcal{P}_P$. This illustrates that due to China's immense size (1'363 million inhabitants versus only 317 million in the US in 2013) its share of the world population is higher for all three groups. However, a three-group representation of the income distribution masks considerable heterogeneity within the rich income group. The average rich in the US are much richer than the average rich in China. The US population mass within the rich group is shifted right compared to China, which reconciles that the two countries have comparable GDPs (US: 16.3 trillion $, CN: 15.9 trillion $). This is a challenge for the three-group representation as with fixed levels of $\underline{\theta}, \theta_M, \bar{\theta}$, China has a substantially higher GDP than the US since the fixed income level $\bar{\theta}$ cannot address the within-group heterogeneity. In Figure 26 this is transformed into a 3-group income distribution. The three-group characterization also suggests a considerable higher income overlap (0.46) between the USA and China than the 5k bin income distribution in Figure 2 (0.35). As the number of income groups increases such approximation bias decreases.

[56] The classification is available under datahelpdesk.worldbank.org/knowledgebase/articles/906519-world-bank-country-and-lending-groups.

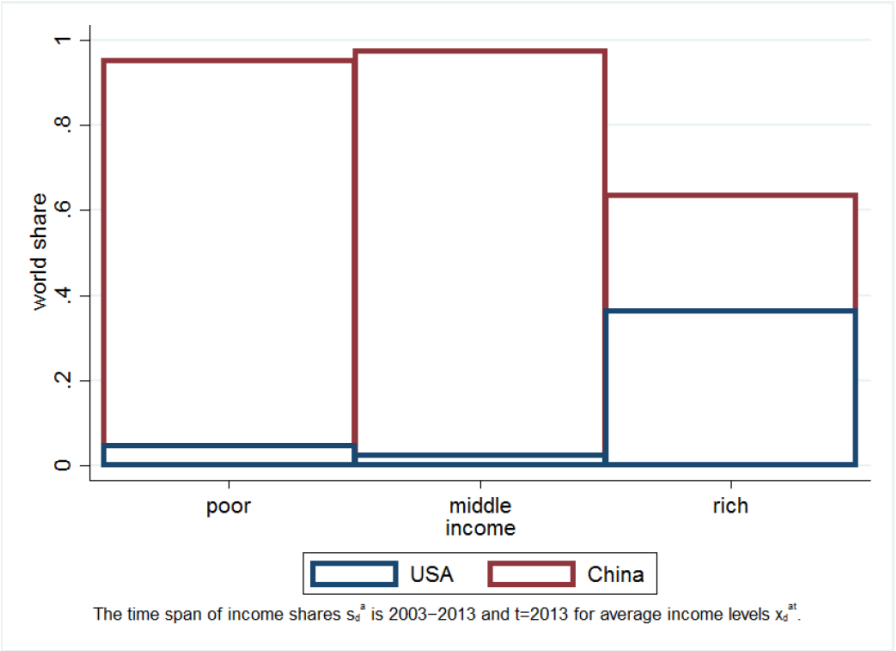


Figure 26: Three income groups for USA and China

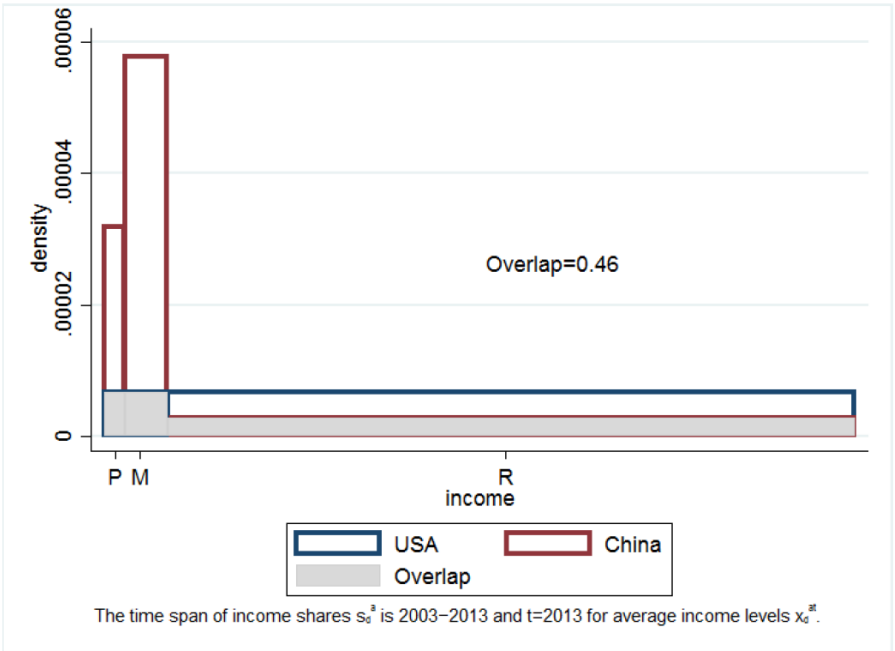


Figure 27: Three-group income distribution for USA and China