General-equilibrium studies on binary demand, emission regulation policies, and fertility rate

Kefu Lin

March 6, 2023

Acknowledgment

I want to show great gratitude and respect to my academic supervisor, Prof. Dao-Zhi Zeng, for providing me with invaluable guidance in all aspects during my master's and doctoral course. His patient and inspiring teaching has taken me from a freshman to the completion of my dissertation. His recognition and suggestions are always one of my motivations to move forward. He will always be a role model to me in my academic career.

I also want to thank Prof. Tatsuhito Kono, Assoc. Prof. Ryo Itoh, Assoc. Prof Naoya Fujiwara, and Dr. Tomokatsu Onaga for their constructive comments and criticism of this thesis. Moreover, I want to thank them for their guidance all the time during seminars and daily communication.

I especially appreciate Assoc. Prof. Jian Wang recommended me to Prof. Zeng's Lab. He is the initiator of my academic career, and I really received a lot of help from him. He always has some interesting ideas, and discussion with him inspires me a lot. I am looking forward to having more chances to cooperate with him in the future.

I also want to show my special thanks to two anonymous referees and the editor of Economic Modelling for their detailed comments. Taiji Furusawa, Prof. Yasuhiro Sato, and Prof. Hajime Takatsuka for their insightful suggestions at the conferences.

My appreciation goes to Mrs. Namie Miura for the various help. Without her, my lab life would become much tougher. I also want to express my sincere thanks to all the members who have supported me in the lab. Particularly, I want to thank Ms. Xinmeng Li, Mr. Rui Pan, Mr. Kangzhe Ding, Mr. Yunhan Du, and Mr. Kazufumi Tsuboi for their encouragement and accompany.

I am grateful for the financial support and various precious opportunities from the SyDE Program. The lectures and activities I took there were really meaningful and unforgettable for me. I appreciate all the professors and staff I met in the SyDE Program. In particular, I want to thank my mentor, Prof. Akira Hibiki, for his patient discussions and inspiring comments to me. He is also one of the main reasons I chose the environmental topic in Chapter 3.

Finally, I want to show my great appreciation to my parents for their support and love all the time. I haven't been back home for more than three years due to the covid-19. I really hope to reunite with them again as soon as possible.

Contents

List of Figures

List of Tables

List of variables

- $U_i,\ (i = c, 1, 2)$ utility in a closed country or country i
- $x(j)$ consumption of variety j
- $p(j)$ price of variety j
- $L_i,\ (i = c, 1, 2)$ population endowment in a closed country or country i
- l relative population endowment between two countries
- θ relative technology between two countries
- $\theta_i,\ (i = c, 1, 2)$ labor supply of each individual in a closed country or country i
- $X_c(j)$ aggregate demand of variety j in a closed country
- f_e entry costs of the entrants
- $G(\psi)$ marginal cost distribution function of entrants
- $g(\psi)$ marginal cost density function of entrants
- $\mu_i(\psi),\ (i = c, 1, 2)$ marginal cost density function of active firms in a closed country or country i
- $\bar{\psi}$ upper limitation of the marginal cost level
- $\pi_c(\psi)$ profit of firms with marginal cost level ψ in a closed country
- $\pi_{iN},\ (i = 1, 2)$ profit of non-traded firms in country i
- $\pi_{iT},\ (i = 1, 2)$ profit of traded firms in country i
- π_{2O} profit of export-only firms in country 2
- ψ_c^* marginal cost cut-off of production in a closed country
- $n_i,\ (i = c, 1, 2)$ mass of available varieties in a closed country or country i
- $N_i,\ (i = c, 1, 2)$ mass of active firms in a closed country or country i
- $N_{ei},\ (i = c, 1, 2)$ mass of entrants in a closed country or country i
- τ iceberg trade costs

- $\tilde{\tau}$ trade costs cut-off between no-arbitrage and arbitrage equilibrium
- $p_i,\ (i = c, 1, 2)$ market price in a closed country or country i
- $w_i,\ (i = c, 1, 2)$ wage rate in a closed country or country i
- p relative price between two countries
- w relative wage rate between two countries
- $\psi_i^*,\ (i = 1, 2)$ marginal cost cut-off of domestic supply in country i
- $\psi_{iT}^*,\ (i = 1, 2)$ marginal cost cut-off of export in country i
- ψ_{1TA}^* marginal cost cut-off of export in country 1 in the arbitrage equilibrium
- ψ_{2A}^* marginal cost cut-off of domestic supply in country 2 in the arbitrage equilibrium
- $b_i,\ (i = 1, 2)$ heterogeneity level in mean-preserving spread i
- C^M manufacturing aggregate
- C^A consumption of the composite good in sector A
- σ elasticity of substitution
- $P_i,\ (i = t, e)$ price index under CT or ETS policy
- e_i emission input for variety i
- l_i labor input for variety i
- β input share of emission
- k parameter of Pareto distribution, degree of heterogeneity
- F fixed input of production
- p^e cost of per-unit emission
- $\bar{E}$ total emission target of the government
- t marginal tax rate of CT
- f lump-sum tax (subsidy) of CT
- T total tax revenue

- $p_i,\ (i = t, e)$ market price under CT or ETS policy
- $q_i,\ (i = t, e, o)$ firms' output level under CT or ETS policy and optimal allocation
- $N_i,\ (i = t, e, o)$ mass of active firms under CT or ETS policy and optimal allocation
- $M_i,\ (i = t, e, o)$ mass of entrants under CT or ETS policy and optimal allocation
- $\varphi_i,\ (i = t, e, o)$ marginal cost cut-off of production under CT or ETS policy and optimal allocation
- $W_i,\ (i = t, e, o)$ welfare level under CT or ETS policy and optimal allocation
- $\bar{e}$ initial emission allowances in ETS
- s emission price in ETS
- γ preferences for children
- B the upper limitation of preferences for children
- L_A number of unskilled workers in region 1
- L_M number of skilled workers in region 1
- L_A^* number of unskilled workers in region 2
- L_M^* number of skilled workers in region 2
- H total number of skilled workers in two regions
- ϕ trade freeness
- λ share of skilled workers in region 1
- $\bar{\gamma}$ preferences for children cutoff of skilled workers between two regions
- $\bar{n}$ average fertility rate in region 1
- $\bar{n}^*$ average fertility rate in region 2

Chapter 1

Introduction

1.1 Backgrounds and research objects

Spatial Economics mainly investigates the uneven distribution of economic activities across regions. The market in Spatial Economics is featured by imperfect competition, increasing returns to scale, and trade costs across regions. Such framework allows us to observe the variation of spatial structure endogenously in a general equilibrium. This thesis applies some typical models of Spatial Economics in different scales to investigate some widely concerning issues around the world.

First, we start from a global scale and investigate the trade pattern between a developed country and a developing country incorporating a special binary demand and Melitz-type heterogeneity (Melitz 2003; Melitz & Ottaviano 2008). Starting from Krugman (1980), the new trade theory (NTT) has contributed to many interesting findings in trade patterns based on the Dixit-Stiglitz type constant elasticity of substitution (CES) framework (Dixit & Stiglitz 1977). However, such a tractable setting with a constant markup is not widely supported by the empirical facts (e.g., De Loecker 2011; De Loecker and Warzynski 2012; Feenstra and Weinstein 2017). Some research extends this to a more general class of variable elasticity of substitution (VES) preferences and finds new results that the CES framework can not explain (Behrens & Murata 2007, 2012; Chen & Zeng 2018).

This thesis extends this stream of research and tries to explain some new results between a "South-North" trade. In contrast to the homothetic preferences, we adopt a non-homothetic binary preference based on Foellmi *et al.* (2018a). It allows us to investigate the consumption of indivisible goods such as cars, refrigerators, and smartphones, for which most individuals only require one unit. This special 0-1 form concentrates on the extensive margin (mass of available varieties), which can be regarded as the other extreme of the CES framework. With CES preferences, individuals can only make choices along the intensive margin as they have to consume all the varieties in the market (The marginal utility of one variety will become infinite when the consumption is close to zero). Although 0-1 preferences are very stylized, they can capture both the income effect and pro-competitive effect with tractable analysis. The differences between the two polar cases are essential references for other research that considers the more general VES preferences. Moreover, we can analytically investigate the effect of international arbitrage when the price gap across countries gets large enough.

Meanwhile, a new trend of literature considers the firm-level differences in the supply side. It emphasizes the importance of firm heterogeneity (Melitz 2003), which is also called "New" new trade theory (NNTT). The repeated converse between focus on the demand side and supply side accompanies the development of trade theory. Traced back to Adam Smith's absolute advantage model, Ricardian's comparative advantage model, and Heckscher–Ohlin model, these trade models concentrate on the supply side with different factor endowments. Then the NTT moves the focus to the demand side, whereas the NNTT pulls it back again. Furthermore, keeping the firms homogeneous, Foellmi *et al.* (2018a) find that the binary demand can still explain some results of Melitz-type firm heterogeneity, such as the presence of "export zero" during bilateral trade. They state that the demand side plays an important role, and low willingness to pay is associated with low income per capita, leading to a market price gap between rich and poor countries. Therefore, firms from rich countries might not export to a poor country due to the threat of international arbitrage. However, after applying the firm heterogeneity to the demand framework of Foellmi *et al.* (2018a), we find some new results as the firm selection works. Due to the asymmetric selection effect across countries, the market competition differs, and the pro-competitive effect plays an important role as well as the income per capita. This indicates the gap in market prices also depends on the gap of available mass of varieties, so the country with a higher income per capita does not always have a higher market price during trade under certain circumstances. Additionally, we observe a new trade pattern of "export only" apart from the "export zero", whereas the direction is from the poor countries to the rich countries. Our research clarifies that both the demand and supply sides are crucial.

Moreover, this research studies the typical finding of NTT named the home market effect (HME). Generally, the HME is defined from three perspectives within a two-region framework. The first point is in terms of firm share, indicating that the larger region has a higher ratio of firm share than that of the labor market share; the second point is in terms of wage rate, indicating that the larger region always has a higher wage rate; the third point is in terms of trade pattern, indicating that the larger country is the net-exporter of the differentiated sector. Since we consider a one-sector and one-factor model, we cannot investigate the HME in terms of trade patterns due to the trade balance. Under the CES framework, the HMEs in terms of firm share and wage rate are known to be equivalent. However, this equivalence no longer holds after considering general VES preferences (Chen & Zeng 2018). Using the binary preferences, we also check the consistency of the two perspectives and find that the HME in terms of firm share may reverse with the trade costs.

Second, we focus on a national scale and investigate the efficiency of different emission regulation policies within one country. The increasing greenhouse gas (GHG) emissions and frequent extremely hot or cold weather have been a worldwide concern for a long time, especially after the Kyoto Protocol held in 1997 in Kyoto. More countries have started to apply emission regulation methods to reduce GHG discharges and tend to improve air quality. Among the various regulation policies, two of the most commonly adopted are the carbon tax (CT) and emission trading scheme (ETS). These two methods are also regarded as the representatives of price and quantity control, respectively (Weitzman 1974). Namely, the CT can control the price of emissions to affect the production of firms, but it is difficult to limit the quantity of emissions directly; In contrast, the ETS can fix the total emission cap, but the market determines the emission price. Functionally, these two policies are both proposed to promote the incentive of emission reductions (Metcalf 2007; Stavins 2007) and carbon-saving innovation (Milliman and Prince 1989; Fischer, Parry, & Pizer 2003). However, whether they still have the same efficiency under different design elements and special circumstances, such as transaction costs (Stavins 1995; Baudry et al. 2021) and uncertainty (Weitzman 1974; Shinkuma & Sugeta 2016), remains a controversial debate.

In this thesis, we build a general-equilibrium model with one factor (labor) and two sectors (one dirty manufacturing sector and one clean composite goods sector) to analyze the industrial organizations under these two emission regulation policies. Once again, we apply the Melitz-type firm heterogeneity into this model, which is crucial from three aspects. First, the heterogeneity captures both uncertainty and asymmetric information. On the one hand, the entrants can not realize their productivity before sinking the entry costs, and the productivity randomly conforms to a distribution; on the other hand, only the firms can realize their productivity, while the government can only observe the entire distribution. Second, the selection effect is essential for the analysis of market outcomes, and some low-productivity firms should be eliminated from the market. Namely, the mass of entrants and active firms are not equal. This allows us to observe the differences in firms' entry and exit processes under the CT and ETS. Third, only with heterogeneous firms can we observe the emission trading across the firms. In the homogeneous case, firms' production and emission discharge are identical, no firms need to trade the emission allowances with others. Instead, emission trading only happens between the firms and the government in the form of auction.

Unlike previous literature, we do not include the externality of emissions in the utility function. In contrast, we compare the welfare levels of two policies under an identical exogenous emission target. This is because in the real world, the emission target of one country is usually determined by negotiations at international climate conferences such as the Kyoto Protocol (1997) and the Paris Agreement (2015), which the national governments can not easily change. Since the total emission cap is given, our model's price/quantity control is embodied in another perspective. Under the CT, the price control can directly affect firms' revenue level so that the government can control the productivity cutoff to adjust the mass of active firms. Under the ETS, quantity control can determine the allocation of initial emission allowances to adjust the mass of entrants. We find that the two policies have different impacts on the market outcomes and resource allocations across sectors, leading to different efficiency and welfare levels. Meanwhile, the degree of heterogeneity plays an important role, indicating that countries with different development levels should choose different policies.

Finally, we turn to a national inter-regional scale to investigate the relationship between fertility rate, agglomeration, and sorting. This research is closely related to the New Economic Geography (NEG) that was started by Krugman (1991). In NEG, workers are allowed to migrate across regions so that we can observe the process of demographic agglomeration as well as the industry. We adopt the tractable framework of Pflüger (2004) to divide the population into two groups, skilled workers and unskilled workers. The skilled workers are mobile across regions and are used as the fixed input of the firms; the unskilled workers are immobile and are used as the marginal input of the firms.

In this thesis, we try to clarify both the impact of agglomeration and sorting on the regional fertility rate. The decreasing fertility rate is widely observed in metropolitan areas nowadays. However, to explain this phenomenon, we need to consider two aspects. On one hand, the agglomeration during urbanization rises the child-rearing costs, so the residents become reluctant to have children. On the other hand, more households with lower fertility intention are sorted into the large cities, which leads to lower fertility than the small cities. Most literature focuses on the agglomeration effect, while the role of sorting is easy to be neglected. To figure out the roles of these two effects, we assume heterogeneous households to analyze their location choices in the equilibrium. Moreover, we find that due to the existence of child-rearing costs as a dispersion force, the skilled workers will evenly redisperse into two regions during trade liberalization. This process occurs even if the population size of the two regions is asymmetric.

In all three studies of this thesis, heterogeneity plays a crucial role. Firms or individuals have different choices due to the heterogeneity, and the cutoffs between the choices appear. The variation of the cutoff value is an essential mechanism, which is called the selection effect in Chapters 2 and 3, and called the sorting effect in Chapter 4. Regarding the difference, the heterogeneity contributes to our results from several perspectives. Chapter 2 considers a trade issue. The selection effect determines the extensive margin, which affects the market competition in two countries and leads to the price gap. Chapter 3 compares two policies within an autarky model. The selection effect under different policies induces different market outcomes and resource allocation. Chapter 4 depicts the mobility of skilled workers with heterogeneous preferences for children in a NEG model. The sorting effect determines the location choices and further enlarges the fertility gap between regions.

1.2 Overview of our study

This thesis applies the general-model analysis of Spatial Economics to investigate three worldwide concerned issues, international trade, emission regulation, and fertility rate. Moreover, the three issues are investigated from three different scales, which correspond to a global, individual country, and intra-country inter-regional scopes, respectively. The main body includes three parts.

The first part consists of a research paper by Lin and Zeng (2023) titled "International trade with binary demand and heterogeneous productivity." Using a trade model between a developed and a developing country with binary preferences (Foellmi *et al.* 2018a) and heterogeneous productivity (Melitz 2003; Melitz & Ottaviano 2008), this study finds that firm selection brings four new results with the possibility of arbitrage. First, we observe a price reversal, such that the price in the developed (high-income) country can be lower than that in the developing (low-income) country under particular circumstances. Second, we demonstrate the existence of export-only firms in the developing country due to the large price gap; meanwhile, arbitrage enlarges the profits gap between the two markets. As a result, some medium-productivity firms abstain from selling in the developing countries despite having the ability. Third, the selection effect of domestic supply is stronger in the small country when trade is more liberalized. This shows that the results of Felbermayr and Jung (2018) based on the CES framework are not robust and may reverse. Finally, we find that a higher degree of heterogeneity simultaneously enlarges the gains from trade in the developed country and losses from trade in the developing country. We notice that the average productivity also changes with the degree of heterogeneity. To separate the impacts, we adopt two mean-preserving spreads to keep the average productivity unchanged. Moreover, we check various forms of the HME in this special VES setting. Our result shows that the HME in terms of firm share is reversed when trade costs are sufficiently large, and the HME in terms of wages is always observed.

The second part consists of a research paper by Lin, Pan, and Zeng (2023) titled "Carbon Tax vs Emission Trading in a Monopolistically Competitive Market with Heterogeneous Firms." In this study, we build a two-sector (one dirty sector with emission and one clean sector) general-equilibrium model with monopolistic competition and heterogeneous firms to compare the efficiency between CT and ETS. Capturing the selection effect, this model allows us to examine how each policy shapes the market outcomes and analyze the sources of policy inefficiency. We show that an economy with high heterogeneity is better to adopt the ETS, whereas the CT is superior in a low-heterogeneity economy.

As disclosed by Nocco *et al.* (2014) and Behrens *et al.* (2020), the misallocation of resources between sectors is one of the main sources of market distortion. Therefore, we further compare the two policies with the optimal allocation to investigate how the policies can fix the distortion. The results show that the equilibria under both policies fail to reach the optimum. We verify that the CT/price control can achieve an optimal mass of active firms but an insufficient mass of entrants, while the ETS/quantity control performs just the opposite. Meanwhile, both policies allocate excessive labor resources to the non-polluting sector. Moreover, taking the emission regulation as a resource, we find that the input intensity has a sharp impact on output levels and inter-sector resource allocation. These findings emphasize the importance of the production side while investigating market distortion and provide inspiration for future research.

The third part consists of a working paper by Wang, Lin, & Zeng (2021) titled "Agglomeration, Sorting, and Fertility." In this study, We show a consecutive process of agglomeration and the variation of regional fertility rate during this process. This analysis is based on the Pflüger (2004) type quasi-linear utility and households with heterogeneous preferences for children. We explain the low fertility rate in the agglomerated region from both perspectives of agglomeration and sorting. On the one hand, agglomeration increases the child-rearing costs in the larger region; On the other hand, the higher child-rearing costs sort the households with a lower preference for children to the larger region, which further decreases the fertility rate there. Namely, the regional fertility rate shows a negative relationship with the firm share. Additionally, we clarify the child-rearing cost as a dispersion force, inducing an agglomeration-redispersion pattern across regions during trade liberalization.

1.3 Thesis outline

The remainder of this thesis is organized as follows.

Chapter 2 introduces the contributions of binary preferences and the wide applications in the related literature. Furthermore, we analyze various of interesting findings based the trade model with binary preferences and heterogeneous firms.

In Chapter 3, we compare the efficiency between two widely used emission regulation policies: Carbon tax and Emission Trading Scheme from the perspective of market outcomes and resources allocation. We also calculate the optimal allocation to analyze the sources of policy inefficiency for both policies.

Chapter 4 tries to explain the lower fertility rate in the larger city under a NEG framework. Based on the setting of households with heterogeneous preferences for children, we find that both agglomeration and sorting contribute to the low fertility rate. Moreover, we clarify the child-rearing cost as a dispersion force, which induces a agglomeration-redispersion pattern with the increase of trade freeness.

Chapter 5 gives concluding remarks.

Chapter 2

International trade with binary demand and heterogeneous productivity

2.1 Introduction

This study examines trade, incorporating the binary demand of consumers (Foellmi, Hepenstrick, and Zweimüller, 2018a) and heterogeneous firms (Melitz, 2003).

It is a step-by-step process for economists to create models to disclose various mechanisms. The constant elasticity of substitution (CES) function has contributed a lot to economic theory in recent decades because it allows us to build a tractable general-equilibrium framework. However, a CES framework leads to the property of constant markups, which is not supported by empirical facts (e.g., De Loecker, 2011; De Loecker and Warzynski, 2012; Feenstra and Weinstein, 2017; Mukherjee and Chanda, 2021).

Some preferences have been proposed to allow variable elasticity of substitution (VES). Foellmi et al. (2018a) propose a simple setup that considers binary preferences. We have numerous indivisible goods in the real world. With binary preferences, households either purchase one unit of a particular product or do not purchase it at all, which only allows households to make decisions along the extensive margin. Such preferences can be regarded as the other extreme of CES preferences, with which households have to consume all the varieties and only have the choice along the intensive margin (quantity). Foellmi et al.'s (2018a) approach is considered useful and applied to disclose a puzzle in the product cycles between developing and developed countries (Foellmi, Grossmann, and Kohler, 2018b).

Meanwhile, firm heterogeneity is an important tool for clarifying the different production behaviors of firms when trade begins. Heterogeneous productivity enables us to investigate firm selection, i.e., how firms with lower productivity are driven out of the market, which is called the selection effect. Combining binary demand and heterogeneous firms, our study contributes to the literature in four aspects.

First, it discloses the mystery of a price reversal—the market price in a lower-income country exceeds that in a higher-income country. In a CES framework, the market price is proportional to marginal costs (see, for example, Lemma 2.1 of Zeng (2021) for a formal proof), which are generally proportional to the wage rate. Thus, the market price in a high-income country is higher than that in a low-income country. Previous research has emphasized that variable markups are related to per capita income. For example, Simonovska (2015) finds a positive relationship between the prices of tradable goods and per capita income. Markusen (2013) documents that countries with higher per capita income have higher markups and price levels. However, price reversals are widely observed in the real world, as shown in Table 2.1. In 2019, average wages were $\$66,382.5$ in the USA and $\$39,041.1$ in Japan, while the cost-of-living indices[1] were 69.91 and 83.33, respectively.

Table 2.1 Average wages and cost-of-living index of some typical countries

Country	Average wages (US$)	Cost-of-living index
USA	66382.5	69.91
Australia	54020.8	72.08
Germany	54041.4	67.62
Canada	54119.3	65.01
France	47112.4	74.85
UK	47936.6	65.28
Japan	39041.1	83.33

Source: OECD (2019): https://data.oecd.org/earnwage/average-wages.htm;
Numbeo (2019): https://www.numbeo.com/cost-of-living/rankings_by_country.jsp?title=2019

Moreover, Morel et al. (2011) and Lorne and Shah (2015) find that price reversal occurs in the pharmaceutical industry.

If firms are homogeneous, as reported by Foellmi et al. (2018a), their behavior is symmetric. All firms export when arbitrage does not occur. Consequently, the mass of available varieties is identical in the two countries. The market price is completely determined by the demand side (more specifically, by per capita income in the country), which leads to a positive relationship between the wage rate and the market price. However, after incorporating the heterogeneity, the active selection effect results in asymmetric market competition in two countries. Although the large country possesses a higher income per capita, accompanied by a higher willingness to pay, the mass of available varieties in the large country is relatively larger and the local price there might be relatively lower. Moreover, when trade costs are high, more labor is allocated to supply the domestic market. This shift is relatively large in the bigger country, making the labor resources more scarce in the large country and increasing the wage rate. On one hand, exporting to the

[1] A cost-of-living index is a relative indicator of consumer goods prices, including groceries, restaurants, transportation, and utilities. We use the data of 2019 to avoid the impact of COVID-19.

small market results in only a small profit; on the other hand, the increasing wage rate proportionately increases the marginal cost. More firms in the large country abstain from exports. Consequently, the product market competition in the larger country becomes relatively tougher when trade costs are high, which may finally lead to a price reversal.

Second, we show the existence of pure export or, export-only, firms. Arbitrage may occur if the price gap between the two countries is large.[2] Tradeable firms in either country have two options when facing arbitrage: (i) charge a lower trade price and sell in two markets or (ii) abandon the smaller market and sell only in the larger market. Firms in the two countries balance the trade-offs between these two options according to their productivity. Our results show that, in both countries, some firms with intermediate levels of productivity abandon the market of the smaller country, even if they can serve both markets. Another interesting phenomenon occurs in the smaller country—the cut-off of exports exceeds that of domestic supply, as exports make higher profits than domestic sales; therefore, export-only firms appear. This indicates that exporting is not the privilege of firms with high productivity, which contrasts to the prevailing view in the related literature. Such export-only firms have been observed in the real world. Lu, Lu, and Tao (2014) theoretically show the existence of pure export within middle-productivity firms and empirically verify the results using data from Chinese manufacturing firms. Further, de Astarloa, Eaton, Krishna, Roberts, Rodríguez-Clare, and Tybout (2015) find that some exporters in Bangladesh export most of their output abroad—they describe these firms as "born to export." Moreover, Liaqat and Hussain (2020) claim that pure exporters will appear in developing countries and usually have lower sales, export earnings, and import spending than normal exporting firms do with Pakistani data. Qiu and Yan (2017) extend the theoretical research and attribute the existence of export-only firms to a low fixed export cost, an export tax rebate, and a large efficiency gap. Contrary to previous studies, we show that the existence of export-only firms stems from endogenous variation in the price gap between the two countries without any additional conditions.

Third, the assumption of binary demand makes it tractable for us to investigate how the relative intensity of the selection effects in two countries changes with trade costs. Regarding the selection effect in each country, we obtain results similar to those of Melitz (2003): the extensive margin of domestic supply expands while the extensive margin of export shrinks in both countries when trade costs rise. However, their relative intensity varies with the trade costs. Specifically, when the trade costs are low, the selection of domestic supply is more severe in the smaller country, and this is reversed when the trade costs increase. Thus, the larger market may accommodate relatively more firms only

[2] In this research, the price gap is a relative concept used to measure the ratio between the relative market price in two countries and trade costs.

when trade is sufficiently liberalized. More specifically, the larger market may not have a more than proportionate firm share when trade costs are high, which is consistent with the theoretical result of Chen and Zeng (2018) on the home market effect in terms of firm share, based on a framework with additive preferences more general than CES.

Finally, we explore how heterogeneity affects welfare and gains from trade when preference is binary. Based on a model of continuous demand, Kokovin et al. (2022) report the possibility of harmful trade, indicating a welfare loss when trade starts from a high level. Conversely, Foellmi et al. (2018a) find that globalization may hurt a smaller country when trade costs are low. Here we further show that heterogeneous productivity aggravates it, which occurs in the arbitrage equilibrium when the price gap between the two countries is significant. Meanwhile, Melitz and Redding (2015) argue that intra-industry reallocation brought on by the selection effect plays a vital role in measuring gains from trade. By comparing our heterogeneous firm model with the homogeneous firm model of Foellmi et al. (2018a), we find that productivity heterogeneity leads to higher welfare and gains (losses) from trade. Furthermore, we use two mean-preserving spread models to investigate the continuous impact of heterogeneity while maintaining the average productivity constant. We find that either a higher degree of heterogeneity or a higher average productivity will increase welfare in both countries for the given trade costs. Additionally, these two factors can simultaneously enlarge the gains from trade in the large country and the losses from trade in the small country. Accordingly, a higher degree of heterogeneity improves the efficiency of resource allocation within a country, which can be observed in the increasing domestic varieties in both countries. Conversely, in the more heterogeneous case, the large country absorbs more resources from the small country, leading to a growing imbalance across the countries, attributable to the expanded extensive margins of export-only firms in the small country. Our results suggest that a small country should maintain a certain level of tariff to maximize its welfare, which complements the conclusions of Demidova and Rodríguez-Clare (2009) and Demidova (2017).

Moreover, we examine the existence and consistency of various home market effects (HMEs). In the literature, Krugman (1980) shows via two different CES models that a large country has some advantages, such as being a net exporter, having a more-than-proportionate firm share, and having a higher wage rate, which are the HMEs in terms of trade pattern, firm share, and wages, respectively. Takahashi, Takatsuka, and Zeng (2013) show that these three HMEs are equivalent, and all of them are observed in a single model with mobile capital as a second production factor. Surprisingly, keeping the assumption of two production factors, a recent paper by Chen and Zeng (2018) finds that the HME in terms of firm share may fail if the CES utility is replaced by a general additive VES (variable elasticity of substitution) function. By using the binary preferences, this research shows that the HME in terms of firm share may be reversed even when labor is the only production factor. More specifically, when trade costs are high, the increasing wage rate becomes a burden for production. Meanwhile, the export firms in the small country are less affected thanks to the low production costs and the large export market. This indicates that the firms in the large country face a tougher survival environment. The pressure from both costs and prices pushes firms out of the larger country and finally leads to a reversed HME in terms of firm share.

Further, we complement our model in a case where the population and technological advantages are geographically separated between the two countries. We claim that both advantages lead to higher welfare. By investigating the trade structure between two countries with identical total effective labor, we further verify that a country with a technological advantage has a leading position in determining the trade structure.

2.2 Literature review

The "0-1" form binary demand has been applied to explain many economic issues in both Micro- and Macro-economics. Foellmi & Zweimüller (2006; 2017) use binary preferences to investigate the issues of innovation and growth between rich countries and poor countries. They highlight the importance of income per capita and identify that for the innovators, the inequality between countries has a positive relationship with the product price (price effect) and a negative relationship with the market size (market-size effect). This research is further extended to study the product cycles in a dynamic North-South model (Foellmi, Grossmann, & Kohler 2018b). Meanwhile, Foellmi, Wuergler, & Zweimüller (2014) consider the quality of products and extend the binary preferences to the trinary preferences, so that individuals can further choose to consume high-quality or low-quality products based on their income levels. They use this model to explain the process innovations that transform existing luxuries into mass products for the poor. Moreover, there are many other types of non-homothetic preferences are used in the literature to explain various economic activities (Matsuyama 2002; Rojas & Saffie 2022; Hsu, Lu, & Picard 2022).

Additionally, this study highlights the importance of arbitrage during international trade. In the CES framework, arbitrage will never occur due to the constant mark-up. In the MO model (Melitz & Ottaviano 2008), the arbitrage is also ruled out as the price gap between domestic supply and export is insufficient. Under homothetic preferences, arbitrage can not be observed without some extra assumptions. In contrast, the binary preferences can tractably analyze both the equilibria with and without arbitrage as shown in Foellmi et al. (2018a). They also verify that the predictions of the simple 0-1 model still hold under more general preferences if some conditions are met. Meanwhile, the binary preferences provide an analytical general-equilibrium framework for the research of parallel import. Previously, most research considers this issue within a partial equilibrium (Matsuyama 2000) or game theory (Roy & Saggi 2012; Zeng & Zhang 2020).

The consistency of the HME in terms of firm share and wage rate has been investigated by many different models. Considering capital as the second production factor, Takahashi et al. (2013) show that these two HMEs are equivalent under CES preferences. Keeping labor as the only factor and CES preferences, Felbermayr & Jung (2018) find the equivalence still holds with firm heterogeneity. However, Chen & Zeng find these two aspects of the HME are not equivalent anymore in a general VES framework. They verify that the HME in terms of firm share is reversed in the case of one-way trade from the smaller to the larger country. Some empirical research also supports this result. Redding & Venables (2004) and Hanson (2005) prove the robustness of the HME in terms of wage rate, whereas Head and Mayer (2004) point out that the evidence on the HME in terms of firm share is highly mixed. The result of our study is consistent with the general VES framework that the HME in terms of firm share may reverse under certain circumstances.

2.3 Closed economy

We first implement the model in the simple autarky case and use the subscript c to denote the closed economy. We assume that there are L_c identical individuals and that each individual provides θ_c units of labor.

2.3.1 Demand

The following utility function,

$$U = \int_0^n x(j)dj,$$

is assumed to describe consumer preferences, where $x(j)$ is either 0 or 1 for any $j \in [0, n]$. The mass of variety n is denoted as n_c in the closed economy. The budget constraint is

$$\int_0^{n_c} p(j)x(j)dj = y,$$

where y is the individual income. The individual demand for each variety j is obtained by the first-order condition to maximize utility:

$$x(j) = \begin{cases} 1 & \text{if } p(j) \leq 1/\lambda \\ 0 & \text{if } p(j) > 1/\lambda, \end{cases}$$

where λ is the Lagrangian multiplier that can be understood as the marginal utility of income. All products are differentiated, but they are symmetric in quality. Individuals purchase one unit of good j if the price is not higher than their willingness to pay $(1/\lambda)$. Otherwise, they do not purchase the product.

The analysis above reveals the difference between a continuous demand model and a binary demand model when firms are heterogeneous. When the demand for a variety is continuous, the producer can adjust its price/demand to maximize total profit. In contrast, when demand is binary, the producer cannot increase positive demand by lowering its price. Different varieties enter the utility function symmetrically. The willingness to pay is the same for all varieties, although firms are heterogeneous in productivity. As firms cannot set a price different from the willingness to pay, all varieties have the same price.[3] The aggregate demand for one variety under autarky can be written as

$$X_c(j) = \begin{cases} L_c & \text{if } p(j) \leq 1/\lambda, \\ 0 & \text{if } p(j) > 1/\lambda. \end{cases}$$

Therefore, there exists a single price $p_c = 1/\lambda$ in equilibrium.

2.3.2 Production

Labor is the only factor of production. There is a continuum of potential entrants attempting to enter the market. After sinking f_e units of labor as an entry fee, each firm randomly draws its cost level from a uniform distribution $G(\psi)$ on $[0, \bar{\psi}]$.[4] Specifically, $1/\psi$ is the productivity of each firm; thus, a smaller ψ indicates higher productivity. The density function is

$$g(\psi) = \begin{cases} 1/\bar{\psi} & \text{if } \bar{\psi} \geq \psi > 0, \\ 0 & \text{otherwise.} \end{cases}$$

For simplicity, we do not assume a fixed input as in the study by Melitz and Ottaviano (2008).[5] The profit of firms with a cost level ψ can be written as

$$\pi_c(\psi) = (p_c - \psi w_c) L_c.$$

Hereafter, we choose labor as the numeraire, so that $w_c = 1$ holds.

[3] The detailed explanation of price decision under the 0-1 framework can be seen in the study by Foellmi et al. (2018a). In their homogeneous model, the prices of domestic and imported varieties are also equal to consumers' willingness to pay.

[4] We adopt uniform distribution in this study for better tractability. Nevertheless, through simulations, we find that the results are robust if the Pareto distribution is used. The assumption of a 0 lower boundary ensures the existence of tradeable firms in both countries, which makes the model more tractable by preventing the interference of a choke price when the trade costs are sufficiently large.

[5] Our model is similar to that of Melitz (2003) in the sense of one sector. However, Melitz (2003) requires a fixed input as the CES framework has constant markups without finite choke prices. A fixed input is necessary for Melitz (2003) to display the selection effect.

After recognizing their productivity, firms decide whether to produce or exit the market. The cost cut-off ψ_c^* for firms to be active is determined by the zero-profit condition:

$$\pi(\psi_c^*) = (p_c - \psi_c^*)L_c = 0.$$

Then the firms with higher marginal cost than ψ_c^* will exit the market. The distribution of active firms can be depicted by the conditional distribution of $G(\psi)$ on $(0, \psi_c^*]$:

$$\mu_c(\psi) = \begin{cases} \dfrac{g(\psi)}{G(\psi_c^*)} = \dfrac{1}{\psi_c^*} & \text{if } \psi_c^* \geq \psi > 0, \\ 0 & \text{otherwise.} \end{cases}$$

The free-entry condition takes the following form:

$$f_e = \int_0^{\psi_c^*} \pi_c(\psi) g(\psi) d\psi = \int_0^{\psi_c^*} (p_c - \psi) L_c \frac{1}{\bar{\psi}} d\psi.$$

The labor market clearing condition is

$$L_c\theta_c = \int_0^{\psi_c^*} L_c \psi \mu_c(\psi) N_c d\psi + N_{ec} f_e,$$

where N_c is the mass of active firms, and N_{ec} denotes the mass of firms entering the productivity draw.

The budget constraint now can be rewritten as:

$$p_c n_c = w_c \theta_c = 1,$$

where n_c is the mass of available varieties in the country, which equals the mass of firms in autarky.

The above conditions determine an equilibrium in which the important endogenous variables are

$$p_c = \psi_c^* = \left(\frac{2f_e\bar{\psi}}{L_c}\right)^{\frac{1}{2}}, \quad U_c = n_c = N_c = \theta_c \left(\frac{L_c}{2f_e\bar{\psi}}\right)^{\frac{1}{2}}, \quad N_{ec} = \frac{L_c\theta_c}{2f_e}, \tag{2.1}$$

where U_c is the utility level under autarky. The first equality in (2.1) indicates that the equilibrium price decreases with country size L_c. This reveals the well-known pro-competitive effect in the literature. Moreover, a larger $\bar{\psi}$ implies a lower average productivity because of the assumption of uniform distribution. Thus, the mass of active firms, N_c, decreases with $\bar{\psi}$. This less competitive product market leads to a higher market price p_c. Accordingly, the cutoff cost $\psi_c^* = p_c$ increases with $\bar{\psi}$. Welfare U_c is negatively related to, whereas the mass of entrants, N_{ec}, is independent of $\bar{\psi}$. Note that a large $\bar{\psi}$ can also be interpreted as a higher heterogeneity. we will investigate the different role of $\bar{\psi}$ in greater depth using several models in the subsequent Section 2.4.3.

The above results display an advantage of VES studies based on binary preferences—they are relatively tractable even when both the income effect and the pro-competitive effect are captured.

2.4 Impact of globalization

We now consider a world economy in which a large developed country and a small developing country trade with each other. More specifically, country 1 is a developed country with a population of L_1, and country 2 is a developing country with a population of L_2. We introduce labor efficiency to describe the technological differences between the two countries. Each individual in country 1 provides θ_1 units of labor, and each individual in country 2 provides θ_2 units. Our benchmark model assumes that country 1 has both population and technological advantages, i.e., $L_1 \geq L_2$, $\theta_1 \geq \theta_2$. The opposite case of $L_1 < L_2$, $\theta_1 \geq \theta_2$ will be examined in Section 2.7. Trade is costly, and trade costs have the standard iceberg form: for each unit sold to a foreign market, $\tau > 1$ units must be shipped, and $\tau - 1$ units are lost during transportation.

For convenience, we use

$$l \equiv \frac{L_1}{L_2}(\geq 1), \quad \theta \equiv \frac{\theta_1}{\theta_2}(\geq 1)$$

to denote the relative population and technology between two countries, respectively. Let $\tilde{\tau}$ be the root of

$$\mathcal{A}(\tau) \equiv l\tau^6 + \tau^3 - \theta^2 l\tau - \theta^2, \tag{2.2}$$

which is the threshold of trade cost for arbitrage to be possible, falling in $[1, \theta)$. Lemma A.2 in Appendix A shows the existence and the uniqueness of $\tilde{\tau}$.

Subsequently, we conduct the equilibrium analysis with a large τ (i.e., $\tau \geq \tilde{\tau}$) and a small τ (i.e., $\tau < \tilde{\tau}$). However, $\tau < \tilde{\tau}$ is impossible when $\theta = 1$ because $\tilde{\tau}|_{\theta=1} = 1$. Therefore, we neglect the special case of $\theta = 1$ in which there is no technological advantage.

2.4.1 No-arbitrage equilibrium

This section considers the case of $\tau \geq \tilde{\tau}$. We first assume inequalities,

$$\frac{1}{\tau} < \frac{p_1}{p_2} \leq \tau, \tag{2.3}$$

where p_1 and p_2 are the equilibrium prices in countries $i = 1, 2$, respectively. Lemma A.3 in Appendix A proves that the inequalities in (2.3) are actually true. Under (2.3), goods are traded internationally without arbitrage.

After sinking f_e units of labor as entry costs, firms in either country learn their cost level from the identical distribution $G(\psi)$ as autarky. Subsequently, the gross profit of a representative firm in country i is:

$$\pi_i = (p_i - \psi w_i)L_i + (p_{-i} - \tau\psi w_i)L_{-i},$$

where $-i$ is the country other than i. The first term of π_i is the profit from the domestic market, and the second term is the profit from exports. There is no additional fixed cost for either domestic or foreign markets. As in most heterogeneity models, there are two cut-offs: one each for domestic supply and exports. The following shows that the cut-off for exports is higher than that for production in either country. Therefore, some firms with low productivity sell goods only in the domestic markets.

Specifically, the cost cut-off for the domestic market ψ_i^* in country i is determined by the zero-profit condition of the local profit:

$$(p_i - \psi_i^* w_i)L_i = 0.$$

Further, the cost cut-off for export ψ_{iT}^* is determined by the zero-profit condition of export profit:

$$(p_{-i} - \tau\psi_{iT}^* w_i)L_{-i} = 0.$$

Subsequently, the cut-offs in the two countries have the following relationship:

$$\psi_i^* = \frac{p_i}{w_i} > \psi_{iT}^* = \frac{p_{-i}}{\tau w_i}. \tag{2.4}$$

The profits of two kinds of firms in two countries are

$$\pi_{iN} = (p_i - \psi w_i)L_i, \quad \pi_{iT} = (p_i - \psi w_i)L_i + (p_{-i} - \tau\psi w_i)L_{-i},$$

where π_{iN} denotes the profit of a firm producing a non-traded variety, and π_{iT} is the profit of a firm producing a traded variety in country $i = 1, 2$.

The free-entry condition in country $i = 1, 2$ is

$$f_e w_i = \int_0^{\psi_{iT}^*} \pi_{iT} g(\psi) d\psi + \int_{\psi_{iT}^*}^{\psi_i^*} \pi_{iN} g(\psi) d\psi,$$

which is simplified to

$$2\bar{\psi} f_e w_i^2 = p_i^2 L_i + \frac{p_{-i}^2 L_{-i}}{\tau}. \tag{2.5}$$

The mass of available varieties in country i is

$$n_i = \int_0^{\psi_i^*} N_i \mu_i(\psi) d\psi + \int_0^{\psi_{-iT}^*} N_{-i}\mu_{-i}(\psi) d\psi, \tag{2.6}$$

where N_i denotes the mass of active firms in country i, and $\mu_i(\psi)$ is the probability density of active firms in country i (i.e., $\mu_i = 1/\psi_i^*$). Subsequently, Equation (2.6) is simplified using the budget constraints and (2.4) to

$$\frac{w_i \theta_i}{p_i} = n_i = N_i + N_{-i}\frac{p_i}{\tau p_{-i}}. \tag{2.7}$$

The balance of payments is written as

$$p_2 L_2 N_1 \frac{\psi_{1T}^*}{\psi_1^*} = p_1 L_1 N_2 \frac{\psi_{2T}^*}{\psi_2^*}, \quad \text{i.e.,} \quad \frac{N_1}{N_2} = \frac{L_1 p_1^3}{L_2 p_2^3}. \tag{2.8}$$

Let

$$p = \frac{p_1}{p_2}, \quad w = \frac{w_1}{w_2}$$

be the relative price and wage rate, respectively. Choose labor in country 2 as the numeraire so that $w_2 = 1$ and $w_1 = w$. Combining Equations (2.5), (2.7), and (2.8) and the budget constraints, we obtain the wage equation (the details are available from the authors upon request):

$$\mathcal{F}(w) \equiv \tau w^3 + \theta l w^2 - w - \tau\theta l = 0. \tag{2.9}$$

This implicitly determines the relative wage rate w as a function of τ, θ, and l. The relative price is given by

$$p = \sqrt{\frac{\theta}{w}}. \tag{2.10}$$

Other endogenous variables can also be obtained from w. We observe a negative relationship between the relative price and the wage rate. This is because a higher wage rate results in a higher marginal cost. After considering heterogeneity, the firms in the large country become less competitive in the export market. Therefore, more labor resources are allocated to the domestic supply, inducing a tougher product competition and a lower market price in the large country. Moreover, Lemma A.3 in Appendix A verifies that the equilibrium wage rate of (2.9) falls in $[\theta/\tau^2, \theta\tau^2)$.

Proposition 1. *In the no-arbitrage equilibrium, (a) the equilibrium relative wage rate w is uniquely given by (2.9), and $w > 1$ holds; (b) the equilibrium wage gap increases with trade cost ($dw/d\tau > 0$); (c) the equilibrium relative price decreases with trade cost ($dp/d\tau < 0$); and (d) a price reversal (i.e., $p < 1$) occurs iff $l > \theta^2 (> 1)$ and $\tau > \tau_{pr} \equiv (l\theta^2 - 1)/(l - \theta^2)$.*

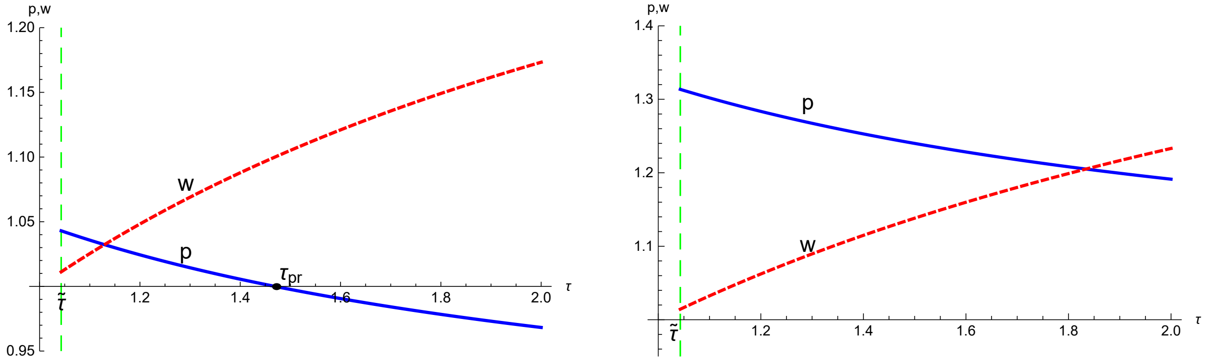
Proof. See Appendix B. □

The results show that the relative price and wage rate have opposite trends when trade costs increase, as observed in (2.10). This is an interesting fact that results from heterogeneity in productivity. According to Foellmi et al. (2018a), the mass of available varieties is identical in the two countries, as all firms can export in a no-arbitrage equilibrium. More precisely, without selection, the market price is only related to per capita income, which equals the wage rate because labor is the only production factor. Therefore, Foellmi et al. (2018a) show a positive relationship between relative prices and wages. Accordingly, a price reversal is not observable when firms are homogeneous.

We use two numerical results to show how the relative wage and price depend on τ, with and without price reversal. They are plotted in Figure 2.1. Parameters of Figure 2.1a are given by[6]

$$\bar{\psi} = 2,\ f_e = 0.2,\ L_1 = 12,\ L_2 = 4,\ \theta_2 = 1,\ \theta_1 = 1.1. \tag{2.11}$$

We only change θ_1 from 1.1 to 1.8 in Figure 2.1b.



(a) With a price reversal

(b) Without a price reversal

Figure 2.1 Relative price (p) and wage (w) in the no-arbitrage equilibrium

In Figure 2.1, $\tilde{\tau}$ (≥ 1) is the lower bound of τ for a no-arbitrage equilibrium. Furthermore, $l > \theta^2$ holds for (2.11). Thus, we are able to observe a price reversal in Figure 2.1a: the price in the high-income country is lower than that in the low-income country when $\tau > \tau_{pr} \approx 1.46927$.

In contrast, Figure 2.1b shows the other case in which $l < \theta^2$ holds. Equation (2.10) tells us that p increases with θ. Thus, the solid curve in Figure 2.1a moves upward when θ increases. When θ is large enough such that $l \leq \theta^2$ holds, we have $p \geq 1$ for any τ. Accordingly, the solid curve does not cross the line of $p = 1$, and price reversal does not occur.

In our setup, a larger θ_i is linked to a higher income per capita and a larger willingness to pay, which are positively related to the market price. Meanwhile, L_i denotes the market size, which is positively related to the degree of market competition and negatively related to market price. They result from the income effect and the pro-competitive effect under binary preferences. Therefore, the occurrence of a price reversal can be attributed to the interplay between these two opposite forces. Additionally, trade costs play an important role in determining the relative intensity of these effects in two countries. Our model captures the competition in both the factor market and the product market, which are observed by Melitz (2003) and Melitz and Ottaviano (2008), respectively. On one hand,

[6] All later simulations in Sections 2.4.1, 2.4.2, and 2.6 use the same parameters of (2.11).

with rising trade costs, labor is reallocated from exports to domestic supply. Due to the higher demand in the larger market, the increasing domestic supply makes the labor resources relatively more scarce in the larger country, which pushes up the relative wage. On the other hand, the increasing marginal costs (both wages and trade costs) make the smaller market less attractive to the exporters in the larger country. As a result, the product-market competition is tougher in the larger country than in the smaller country, leading to a reduction in the relative market price. Finally, when the relative product-market competition dominates the relative factor-market competition, a price reversal occurs. The channel via the product market does not work in Foellmi et al.'s (2018a) because of the firm homogeneity. Without the negative effect of product market competition, the factor-market competition dominates in the larger country, so that the market price is always higher there.

In our model, when τ increases, the extensive margin of export shrinks in either country. However, the proportion of exporting firms in the larger country decreases more sharply than that in the smaller country.[7] This is because exporters in the large country face a higher marginal cost but lower demand compared with those in the small country. In this sense, the selection of exporters is stronger in the larger country when τ increases. This finally leads to a larger mass of varieties in the larger country, lowering the relative price there. We will further analyze the variation of the selection effect in detail in Section 2.6.

Additionally, we provide more analytical support for the fact of price reversal when τ is large by investigating how p_1 and p_2 change with τ. According to (2.5) and (B.1), we derive a price equation for country 1 as follows:

$$\mathcal{H}(p_1) = 2\theta^2\bar{\psi}f_e(L_1p_1^2 - 2\tau\bar{\psi}f_e)^2 - L_2^2p_1^4\left[L_1p_1^2(\tau^2-1) + 2\tau\bar{\psi}f_e\right] = 0. \tag{2.12}$$

The result of $dp_1/d\tau > 0$ is obtained from this implicit function (see Appendix C for details). Furthermore, $dp_2/d\tau = d(p_1/p)/d\tau > 0$ holds according to $dp/d\tau = d(p_1/p_2)/d\tau < 0$ in Proposition 1.

The above results demonstrate that the prices in both countries increase with trade costs, but p_2 increases more. This indicates that the degree of product-market competition decreases in either country, while the small country loses relatively more varieties.

Now we turn to the welfare analysis. As the consumption of each variety is limited to one, welfare in each country equals the mass of available varieties in the market:

$$U_1 = n_1 = \frac{\theta_1 w}{p_1}, \quad U_2 = n_2 = \frac{\theta_2}{p_2}. \tag{2.13}$$

[7] This is shown by $\frac{d}{d\tau}\frac{\psi_{1T}^*}{\psi_1^*}\Big/\frac{d}{d\tau}\frac{\psi_{2T}^*}{\psi_2^*} > 1$, $\frac{d}{d\tau}\frac{\psi_{1T}^*}{\psi_1^*} < 0$, and $\frac{d}{d\tau}\frac{\psi_{2T}^*}{\psi_2^*} < 0$.

Therefore, the trend in available varieties has a direct effect on social welfare. Unlike in the homogeneous firm case,[8] in our model, welfare is always higher in the larger country in the no-arbitrage equilibrium.

Proposition 2. *In the no-arbitrage equilibrium, (a) welfare in the larger country is always higher; (b) an increase in τ reduces welfare in both countries, and its effect is tougher in the smaller country.*

Proof. See Appendix D. □

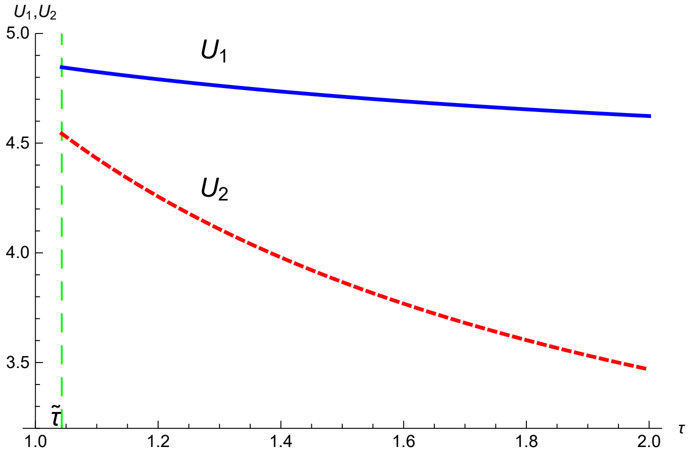


Figure 2.2 Welfare in the no-arbitrage equilibrium

The welfare curves in the two countries are depicted in Figure 2.2, where the parameters are identical to those in (2.11). Both have a downward tendency, while the welfare curve in the smaller country has a sharper slope. This confirms our theoretical results that when trade costs increase, the selection of exporters becomes more severe in the larger country, which further reduces imported varieties in the smaller country. Meanwhile, the smaller country still allocates a relatively large proportion of labor to the export sector, which further enlarges the welfare gap between the two countries. This result is consistent with Proposition 1, which concludes that individuals in the larger country enjoy higher wages and a lower relative price simultaneously when trade costs increase.

2.4.2 Arbitrage equilibrium

We now examine the case of a small τ, i.e., $\tau < \tilde{\tau}$. First, we assume that the inequality

$$\frac{p_1}{p_2} > \tau \tag{2.14}$$

[8] In Foellmi et al. (2018a), all firms can trade when arbitrage does not take place. Thus, the two countries have the same mass of available varieties, and their welfare levels are equal.

holds, where p_1 and p_2 are the equilibrium prices in country $i = 1, 2$, respectively. Lemma A.4 in Appendix A proves that Equation (2.14) is actually true, under which arbitrage occurs. To avoid arbitrage, firms that sell their goods in both countries can only set the price $p_T = \tau p_2$ in country 1, which is lower than p_1. Conversely, firms that sell exclusively in country 1 can charge a higher price, p_1. For convenience, we keep the notation p_1 for the local price of non-traded varieties in country 1 and $p = p_1/p_2$ for the relative price. Therefore, the profits of the non-traded and traded varieties in country 1 are, respectively, as follows:

$$\pi_{1N} = (p_1 - \psi w_1)L_1 \quad \text{and} \quad \pi_{1T} = (\tau p_2 - \psi w_1)L_1 + (p_2 - \tau\psi w_1)L_2.$$

Firms with traded varieties must sacrifice a part of their domestic profits by choosing to serve both markets. Therefore, the cost cut-off ψ^*_{1TA} of exporting firms in country 1 is determined by $\Delta\pi_1(\psi^*_{1TA}) = 0$,[9] where the subscript A refers to arbitrage, and

$$\Delta\pi_1(\psi) \equiv \pi_{1N}(\psi) - \pi_{1T}(\psi) = L_1 p_1 - L_2 p_2 - L_1 p_2 \tau + L_2 w_1 \tau \psi.$$

Thus, we have

$$\psi^*_{1TA} = \psi^*_{1T} - \frac{(p_1 - \tau p_2)l}{\tau w_1} < \psi^*_{1T}, \tag{2.15}$$

where ψ^*_{1T} is the export cut-off in the no-arbitrage equilibrium from (2.4) (for $i = 1$), and the inequality is from (2.14). As $d\Delta\pi_1/d\psi > 0$, only firms with marginal costs lower than ψ^*_{1TA} are willing to export. This inequality suggests that the foreign market is less attractive in the arbitrage equilibrium because some firms abstain from trade to avoid losses in the domestic market, even if they have the ability to trade.

The free-entry condition in country 1 is

$$\int_0^{\psi^*_{1TA}} \pi_{1T} g(\psi) d\psi + \int_{\psi^*_{1TA}}^{\psi^*_1} \pi_{1N} g(\psi) d\psi = f_e w_1,$$

that is,

$$\frac{[L_2 p_2 - L_1(p_1 - \tau p_2)]^2 + \tau L_1 L_2 p_1^2}{2\tau L_2} = \bar{\psi} f_e w_1^2. \tag{2.16}$$

The situation in country 2 is considerably different. The profits of non-traded and traded varieties in country 2 are, respectively, as follows:

$$\pi_{2N} = (p_2 - \psi w_2)L_2 \quad \text{and} \quad \pi_{2T} = (p_2 - \psi w_2)L_2 + (\tau p_2 - \tau\psi w_2)L_1.$$

[9] Note that the determinant of cut-offs here is different from the continuous demand models. If there is no risk of arbitrage, such as in Melitz et al. (2008) or the no-arbitrage equilibrium of Section 2.4.1, the profit of exports is independent of the profit from domestic supply. Therefore, $\Delta\pi(\psi)$ just equals to the export profit in these cases.

However, owing to the large price gap between the two countries, some firms in country 2 make larger profits from country 1 than from the local market. This leads to the presence of export-only firms in country 2, which only serve country 1 through trade. The profits of such firms can be written as

$$\pi_{2O} = (p_1 - \tau\psi w_2)L_1,$$

where the subscript "O" indicates "export-only." The cost cut-off of these export-only firms is derived as follows:

$$\psi_{2T}^* = \frac{p_1}{\tau w_2} > \psi_2^* = \frac{p_2}{w_2}, \tag{2.17}$$

where ψ_2^* and ψ_{2T}^* are identical to (2.4) (for $i = 2$) in the no-arbitrage equilibrium. This inequality indicates that the foreign market is more attractive than the local market in country 2 because of arbitrage. Consequently, some firms in country 2 reconsider whether they can increase their profits by abstaining from the local market. Therefore, the cost cut-off ψ_{2A}^* for a firm to serve the domestic market is determined by $\Delta\pi_2(\psi_{2A}^*) = 0$, where

$$\Delta\pi_2(\psi) \equiv \pi_{2O}(\psi) - \pi_{2T}(\psi) = (p_1 - \tau p_2)L_1 - p_2 L_2 + \psi w_2 L_2.$$

Thus, we have

$$\psi_{2A}^* = \psi_2^* - \frac{(p_1 - \tau p_2)l}{w_2} < \psi_2^*, \tag{2.18}$$

where inequality is obtained from Equation (2.14). As $d\Delta\pi_2(\psi)/d\psi = w_2 L_2 > 0$, only varieties with marginal costs lower than ψ_{2A}^* choose to serve both markets. The inequality also shows that the extensive margin of local varieties in the small country shrinks further because of arbitrage. Some firms are attracted to the larger market and give up the local market even if they have the ability to supply domestically.

The free-entry condition in country 2 is

$$\int_0^{\psi_{2A}^*} \pi_{2T} g(\psi) d\psi + \int_{\psi_{2A}^*}^{\psi_{2T}^*} \pi_{2O} g(\psi) d\psi = f_e w_2,$$

that is,

$$\frac{\tau[L_2 p_2 - L_1(p_1 - \tau p_2)]^2 + L_1 L_2 p_1^2}{2\tau L_2} = \bar{\psi} f_e w_2^2. \tag{2.19}$$

Proposition 3. *In the arbitrage equilibrium, (a) the extensive margin of exports in the larger country shrinks compared with the no-arbitrage equilibrium, excluding some firms with export capability (i.e., $\psi_{1TA}^* < \psi_{1T}^*$). (b) In the smaller country, the selection of domestic supply is more severe than that of exports (i.e., $\psi_{2T}^* > \psi_2^*$). (c) Some firms in the smaller country prefer to become "export only" even if they can serve the domestic market (i.e., $\psi_{2A}^* < \psi_2^*$).*

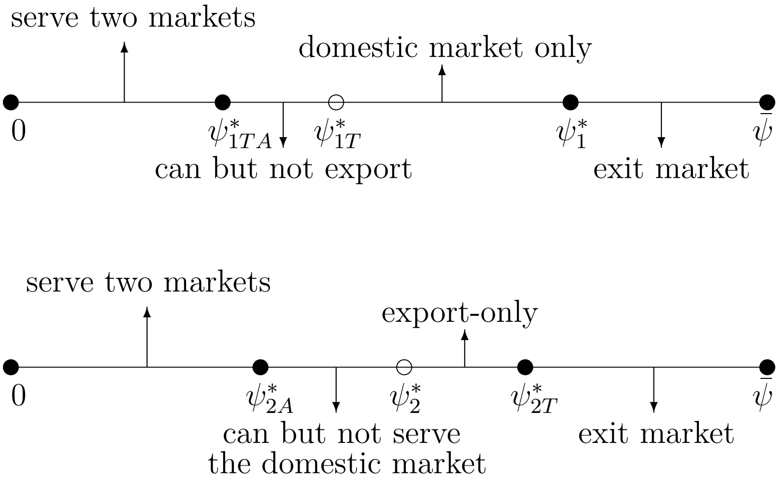


Figure 2.3 The cost cut-offs of two countries in the arbitrage equilibrium

Proof. Results (a), (b), and (c) are obtained from (2.15), (2.17), and (2.18), respectively. □

Figure 2.3 illustrates the results of Proposition 3. In the larger country 1, firms with marginal costs in $[\psi^*_{1TA}, \psi^*_1]$ exclusively serve the domestic market, although firms in $[\psi^*_{1TA}, \psi^*_{1T}]$ can export. In contrast, firms in the smaller country 2 with marginal costs in $[\psi^*_{2A}, \psi^*_{2T}]$ specialize in exports, although firms in $[\psi^*_{2A}, \psi^*_2]$ can serve the domestic market. Moreover, we theoretically prove the existence of export-only behavior among the low-productivity firms in the developing country, which has been empirically reported by Lu et al. (2014) and Liaqat et al. (2020). Unlike Qiu and Yan (2017), who observe the existence of export-only firms after extra assumptions such as asymmetric fixed export costs and export tax rebate rate, our result is obtained directly with the endogenous variation of the price gap between two countries.

The mass of available varieties in country 1 is

$$n_1 = N_1 + N_2, \tag{2.20}$$

while that in country 2 is[10]

$$\begin{aligned} n_2 =& \int_0^{\psi^*_{1TA}} \mu_1(\psi) N_1 d\psi + \int_0^{\psi^*_{2A}} \mu_2(\psi) N_2 d\psi \\ =& \frac{p_2 L_2 - p_1 L_1 + \tau p_2 L_1}{\tau p_1 L_2} N_1 + \frac{\tau(p_2 L_2 - p_1 L_1 + \tau p_2 L_1)}{p_1 L_2} N_2. \end{aligned} \tag{2.21}$$

[10] As the cut-off of exports exceeds that of domestic supply in country 2 when arbitrage appears, the density function of active firms in country 2 becomes $\mu_2(\psi) = 1/\psi^*_{2T}$.

The budget constraint in country 1 is

$$\tau p_2 n_2 + p_1(n_1 - n_2) = w_1\theta_1, \tag{2.22}$$

while that in country 2 is

$$n_2 p_2 = w_2\theta_2. \tag{2.23}$$

The balance of payments is written as

$$\tau p_2 L_1 \int_0^{\psi^*_{2A}} \mu_2(\psi) N_2 d\psi + p_1 L_1 \int_{\psi^*_{2A}}^{\psi^*_{2T}} \mu_2(\psi) N_2 d\psi = p_2 L_2 \int_0^{\psi^*_{1TA}} \mu_1(\psi) N_1 d\psi,$$

i.e.,

$$\begin{aligned} \tau^2 p_2 L_1 N_2 \frac{(p_2L_2 - p_1L_1 + \tau p_2 L_1)}{p_1 L_2} + p_1 L_1 N_2 \Big[1 - \frac{\tau(p_2L_2 - p_1L_1 + \tau p_2L_1)}{p_1L_2}\Big] \\ = p_2 N_1 \frac{p_2L_2 - p_1L_1 + \tau p_2 L_1}{\tau p_1}. \end{aligned} \tag{2.24}$$

Similar to the no-arbitrage equilibrium, Equations (2.16), (2.19), (2.20), (2.21), (2.22), (2.23), and (2.24) lead to the following equations (the details are available from authors upon request):

$$\begin{aligned} \mathcal{K}(p) &\equiv \mathcal{K}_1(p)^2 - \mathcal{K}_2(p) = 0 \qquad (2.25) \\ \mathcal{K}_1(p) &\equiv \frac{\theta(1 - pl + \tau l)}{l\tau(p-\tau)^2 + p^2 - p\tau + \tau^2}, \\ \mathcal{K}_2(p) &\equiv \frac{(1 - lp + \tau l)^2 + \tau l p^2}{\tau(1 - lp + \tau l)^2 + lp^2}, \end{aligned}$$

which implicitly determine the equilibrium relative price. Furthermore, the equilibrium relative wage rate is given by

$$w = \mathcal{K}_1(p)(= \sqrt{\mathcal{K}_2(p)}). \tag{2.26}$$

Appendix E provides the solution of $p(\tau)$ at $\tau = 1, \tilde{\tau}$ and prove that it increases at these two endpoints. Unfortunately, we have no closed form of $p(\tau)$ for general τ from (2.25). Numerical simulations show that $p(\tau)$ and $w(\tau)$ increase with $\tau \in (1, \tilde{\tau})$, as plotted in Figure 2.4, using the same parameters specified in (2.11).

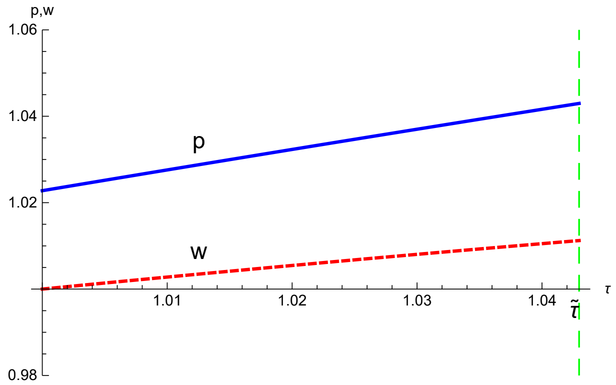


Figure 2.4 Relative price (p) and wage (w) in the arbitrage equilibrium

Moreover, we provide the following results.

Proposition 4. *In the arbitrage equilibrium, (a) the wage rate in the larger country is always higher* ($w > 1$)*; (b) the gap between the relative price and trade costs increases with trade liberalization (i.e.,* $d(p/\tau)/d\tau < 0$*).*

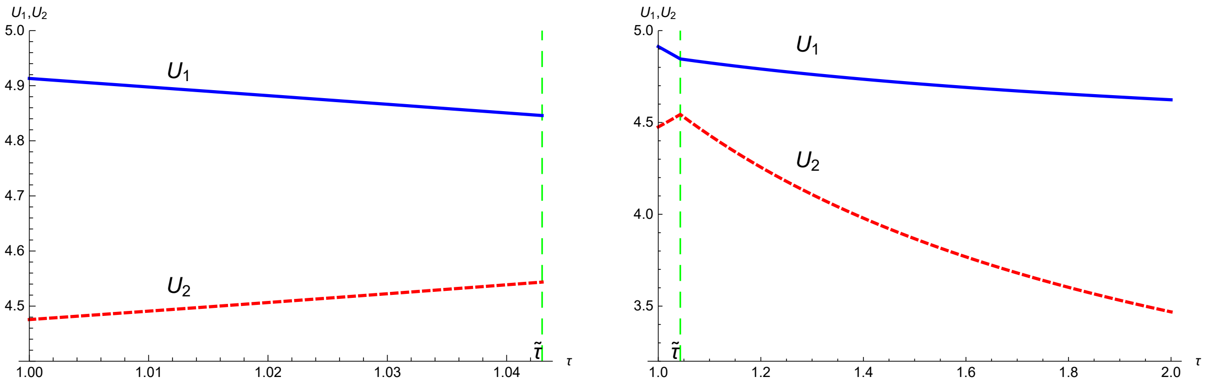
Proof. See Appendix E. □

In summary, we find that trade costs affect the wage rate in the same way as in the no-arbitrage equilibrium, while the relative price shows the opposite trend. This difference can be attributed to the presence of export-only firms, which results in more labor resources for the larger country. When trade costs decrease, more varieties enter the larger country because of the decreasing entrance threshold and the tightened arbitrage condition.

Note that the price reversal observed in the no-arbitrage equilibrium does not occur here. The relative price is always larger than 1, although it decreases with falling trade costs. This is because the selection of exporting firms in the larger country is weakened when trade costs fall, which offsets part of the gap in available varieties between the two countries.

Now, we consider welfare in the arbitrage equilibrium. Using Equations (2.22) and (2.23), the welfare levels in the two countries are as follows:

$$U_1 = n_1 = \frac{w\theta_1 - \tau\theta_2}{p_1} + \frac{\theta_2}{p_2} \quad \text{and} \quad U_2 = n_2 = \frac{\theta_2}{p_2}. \tag{2.27}$$



(a) Welfare in the arbitrage equilibrium

(b) Welfare encompasses both equilibria

Figure 2.5 Equilibrium welfare

The arbitrage equilibrium is not analytically tractable. Therefore, we provide a numerical example. Figure 2.5a depicts the simulation results of (2.27), where the parameters are identical to (2.11). It exhibits the following three facts: (a) welfare in the larger country is always higher ($U_1 > U_2$); (b) welfare in the larger country decreases with trade cost; (c) welfare in the smaller country increases with trade cost. These results are first obtained for the homogeneous case by Foellmi et al. (2018a), who conclude that only the larger country benefits from trade liberalization when the trade costs are low. This is because trade liberalization tightens the arbitrage and expands both the intervals of "export zero" in the larger country and "export only" in the smaller country simultaneously in our model. Note that in the arbitrage equilibrium, export-only firms contribute only directly to the decrease in domestic prices in the larger country, while the price of traded varieties is hardly affected. This means that country 2 allocates numerous resources to country 1, while the rewards from trade are small. For convenience, we show the whole picture of welfare during trade liberalization in Figure 2.5b, which combines the two equilibria of Figures 2.2 and 2.5a.

2.4.3 Average productivity, heterogeneity, and arbitrage

Surprisingly, the smaller country is harmed by trade liberalization in the arbitrage equilibrium. This is first reported by Foellmi et al. (2018a), who assume homogeneous firms. Now, we examine how heterogeneity brings new impacts on the gains from trade.

Noting that $[0, \bar{\psi}]$ is the support set of the random variable ψ, we first investigate the role of $\bar{\psi}$. We rewrite the arbitrage equilibrium conditions as five equations (F.1)–(F.5) for five variables ($\psi_1^*,\ \psi_{1TA}^*,\ N_1,\ N_2,\ w$) in Appendix F. It is straightforward to obtain the following result.

Proposition 5. *In the arbitrage equilibrium, a larger $\bar{\psi}$ lowers the welfare level in both countries and mitigates the losses and benefits of welfare in either country when trade opens.*

Proof. See Appendix F. □

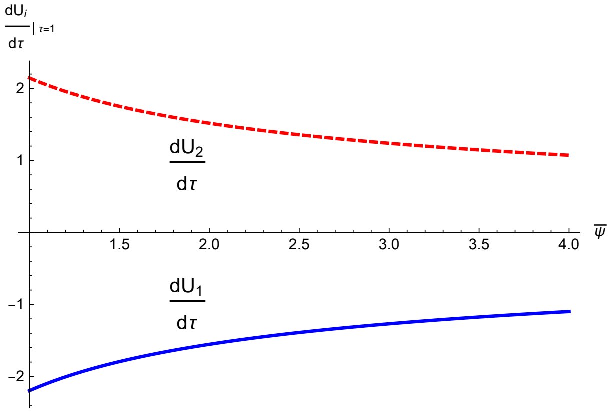


Figure 2.6 Impact of $\bar{\psi}$ on gains from trade

Figure 2.6 shows a numerical example of Proposition 5 with parameters

$$f_e = 0.2,\ L_1 = 12,\ L_2 = 4,\ \theta_1 = 1.1,\ \theta_2 = 1, \tau = 1.$$

The result of Proposition 5 contrasts with the conclusion of Melitz and Redding (2015), who find that heterogeneity increases both welfare and gains from trade. To disclose the different mechanisms, we note that the average productivity also changes with $\bar{\psi}$. The effects of heterogeneity and average productivity on welfare are in opposite directions in our model, while they are in the same direction as in the study of Melitz and Redding (2015).[11] The result of $d(\psi_1^*/\bar{\psi})/d\bar{\psi} < 0$ in Proposition 5 integrates the interaction of these two opposing forces in our model. To investigate the effect of heterogeneity, we use two approaches to remove the effect of changing productivity.

2.4.3.1 Comparison with the homogeneous model

The first approach compares Foellmi et al.'s (2018a) welfare properties with ours. The marginal cost of production in the homogeneous model is adjusted to the mean value of

[11] Melitz and Redding (2015) adopt the Pareto distribution $g(\varphi) = \kappa\varphi_{\min}^{\kappa}\varphi^{-(\kappa+1)}$, $\varphi \in [\varphi_{\min}, \infty)$, where φ is the productivity and $\kappa > 1$ is the shape parameter. A smaller κ implies both a higher degree of heterogeneity and a higher level of average productivity ($\kappa\varphi_{\min}/(\kappa - 1)$). Conversely, in our model, ψ is the marginal input, a larger $\bar{\psi}$ implies a higher degree of heterogeneity but a lower level of average productivity ($2/\bar{\psi}$).

productivity in the heterogeneous model. Additionally, we equalize the entry cost f_e in the heterogeneous model to the fixed cost F in the homogeneous model. Figures 2.7 and 2.8 show a numerical example, in which the parameters are specified as

$$f_e = F = 0.2, \;\; L_1 = 12, \;\; L_2 = 4, \;\; \theta_1 = 1.1, \;\; \theta_2 = 1, \tau = 1.$$

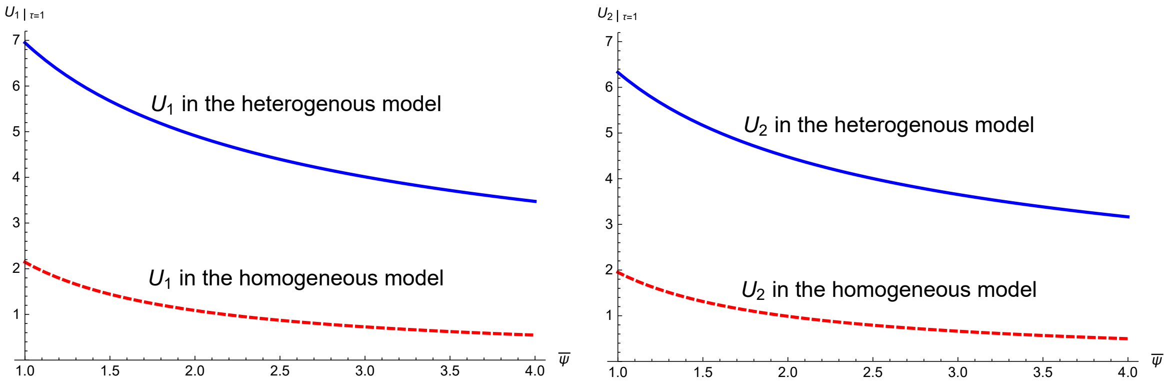


Figure 2.7 Comparison of welfare

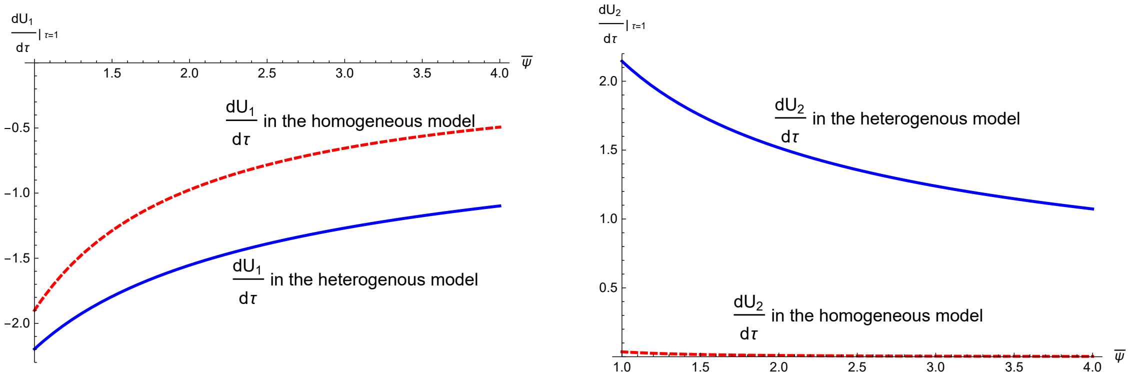


Figure 2.8 Comparison of gains from trade

In Figure 2.7, we plot how welfare is related to $\bar{\psi}$. We observe that the welfare curves in the heterogeneous model in both countries are always higher than those in the homogeneous model. Figure 2.8 shows how $\bar{\psi}$ impacts $dU_i/d\tau$ when $\tau = 1$, which measures the gains or losses in the two countries when trade opens. The curve of the heterogeneous model is more negative in country 1 and more positive in country 2. The results suggest that productivity heterogeneity enlarges the gains and losses in either country when trade opens.

2.4.3.2 Impact of heterogeneity under a mean-preserving spread

The comparison between homogeneous and heterogeneous models is based on a discrete comparative static approach. We now introduce the second approach, building two mean-preserving spread models to examine the effect of productivity heterogeneity based on a continuous comparative static.

Model 1. Uniform distribution with a mean-preserving spread

We modify the interval of the uniform distribution we used above to $[\bar{\psi}/2 - b_1, \bar{\psi}/2 + b_1]$, where $0 \leq b_1 < \bar{\psi}/2$. A larger b_1 value indicates a higher degree of heterogeneity. This setting maintains the average productivity $\bar{\psi}/2$ constant when we investigate the impact of the heterogeneity level by increasing b_1. Details of the amended model are provided in Appendix F.1.

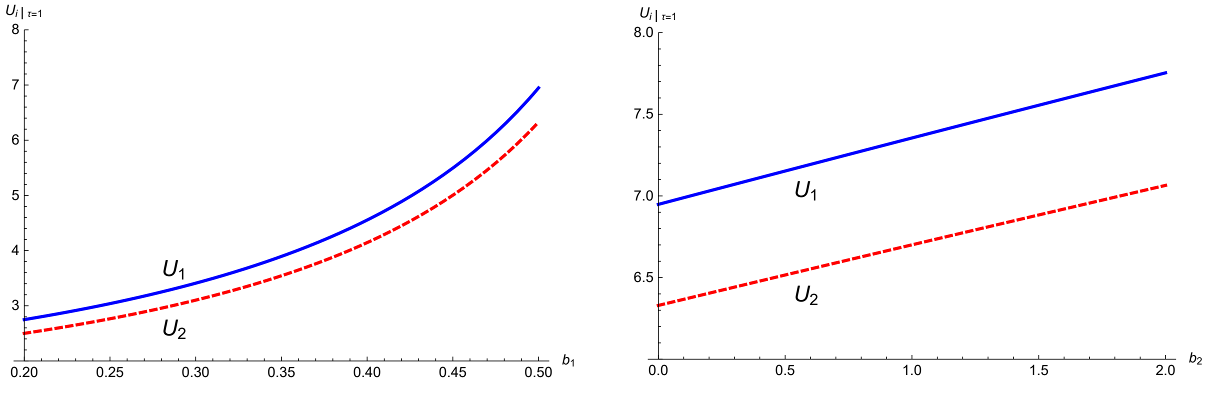
Model 2. Quadratic distribution with a mean-preserving spread

To exclude potential special characteristics caused by the uniform distribution, we apply another quadratic distribution $g(\psi) = b_2\left[(\psi - \bar{\psi}/2)^2 - \bar{\psi}^2/12\right] + 1/\bar{\psi},\ \psi \in (0, \bar{\psi}]$ to recheck the results. In this distribution, a larger b_2 represents a higher degree of heterogeneity, whereas both the average productivity and distribution interval remain unchanged. Details are provided in Appendix F.2.

However, these models are not analytically tractable. Therefore, we use some numerical examples to illustrate the continuous comparative statics of firm heterogeneity. The parameters in the numerical example are specified as follows.

$$\bar{\psi} = 1,\ f_e = 0.2,\ L_1 = 12,\ L_2 = 4,\ \theta_1 = 1.1,\ \theta_2 = 1, \tau = 1.$$

The simulation results are summarized in Figures 2.9 and 2.10 (Mathematica code to draw the figures is available from the authors upon request.)



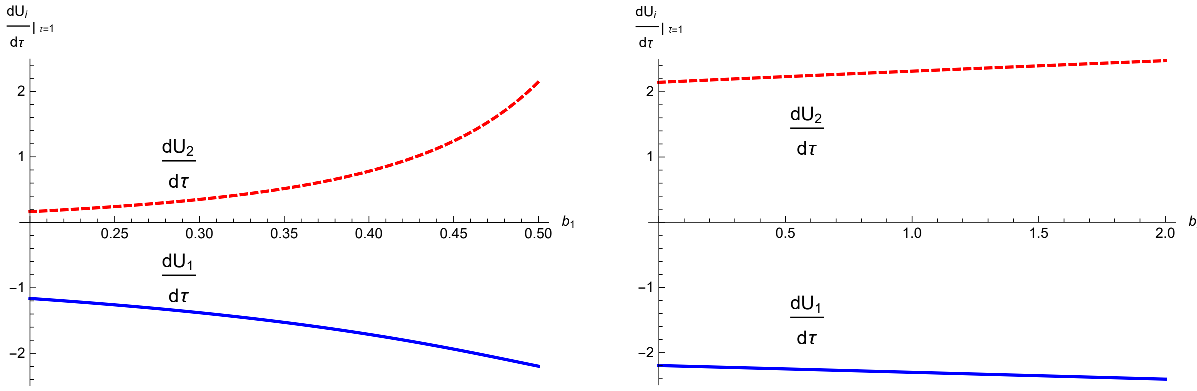
(a) Model 1

(b) Model 2

Figure 2.9 Welfare in mean-preserving spread models

Figure 2.9 shows that in both cases, welfare increases with the degree of heterogeneity. Intuitively, more resources are allocated to high-productivity firms so that fewer resources are wasted and more varieties are available in the market. This can also be observed in the increasing domestic varieties in both countries. They are consistent with the study by Melitz and Redding (2015) for continuous demand.

Figure 2.10 illustrates the level of gains from trade with different heterogeneities. It is obvious that in both cases, either the gains from trade in the large country or the losses from trade in the small country are enlarged by a higher degree of heterogeneity. Consequently, more resources are absorbed from the small country into the large country. This result can be attributed to the new group of export-only firms in our model. More specifically, the mass of export-only firms in the small country increases with the degree of heterogeneity, which wastes cheap labor in the small country and contributes to providing more varieties in the large country.



(a) Model 1

(b) Model 2

Figure 2.10 Gains from trade in mean-preserving spread models

The results of Models 1 and 2 are qualitatively similar. Model 1 can also be connected to the results based on the first approach. In fact, the two countries' welfare and gains from trade gradually converge to the value of the homogeneous case when b_1 decreases to 0. Additionally, when $b_2 = 0$, Model 2 is simplified to a uniform distribution. The results of Sections 2.4.3.1 and 2.4.3.2 are consistent. All of them suggest that increasing heterogeneity increases welfare in both countries and enlarges gains and losses when trade opens up. The trends observed in Figure 2.6 show that the effect of decreasing average productivity dominates that of increasing heterogeneity in our model.

2.5 Two HMEs

Based on CES preferences, Krumgan (1980) demonstrates various HMEs, describing the advantages of a large market when increasing returns, monopolistic competition, and trade costs are involved. Specifically, in an economy of two countries, it is shown that the wage rate in a larger market is higher, which is called the HME in terms of wages. It is also shown that the larger country accommodates a more-than-proportionate share of firms, which is called the HME in terms of firm share. When consumers have binary preferences, we have the following results.

Proposition 6. *The HME in terms of wages appears in both types of equilibria. The HME in terms of firm share will reverse when the price reversal occurs.*

Proof. See Appendix G. □

Figure 2.11 plots the curves of p, w, and the ratio of active firm N_1/N_2 for $\tau \in [1, 2]$, in which two types of equilibria are combined. The blue line in the right panel is the ratio of effective labor $(\theta_1 L_1)/(\theta_2 L_2)$.

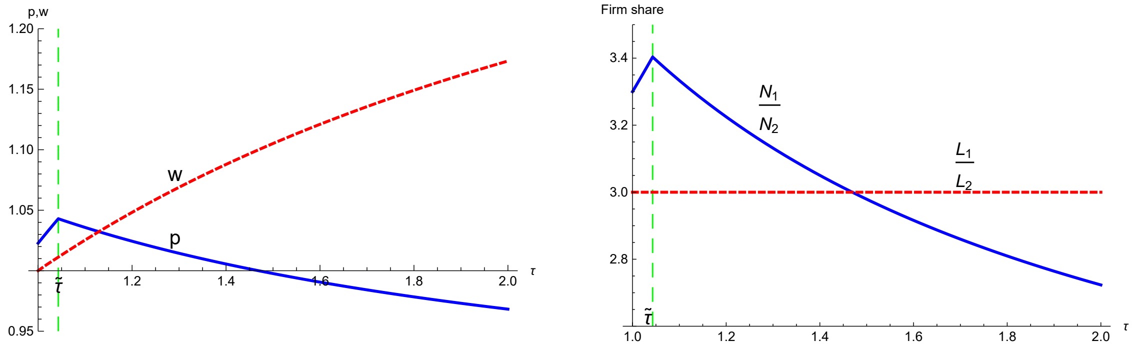


Figure 2.11 Combination of two equilibria

The result on the HME in terms of firm share is related to the decreasing relative price. As in the proof given in Appendix G, $w = p$ holds at the critical point of the firm share reversal. Namely, the firm share reversal starts when two curves of w and p cross in the left panel of Figure 2.11. Specifically, when the trade cost increases, more resources will be reallocated from the exporting to the local production, which induces the increase in active firms in both countries. However, the rising wage rate in the large country becomes a burden in the supply side. On one hand, a higher wage rate leads to a higher production cost in the large country; on the other hand, the increasing wage rate brings a higher consumption power of individuals in the large country, which allows many firms in the small country to remain exporting. This further aggravates the surviving pressure of the firms in the large country. As a result, the increasing speed of active firms is much slower in the large country, and finally leads to a reverse of the firm share. This means that a large market may not have an advantage in firm location as predicted by authors basing their research on CES frameworks.

Under CES preferences, Takahashi et al. (2013) find that the HMEs in terms of wages and in terms of firm share are equivalent. Later, Chen and Zeng (2018) report that the equivalence is not true when the preferences are described by general additive utility functions. Both papers assume mobile capital as one more production factor. The above Proposition 6 reconfirms the invalid equivalence even if there is only one production factor when the CES is replaced by binary preferences.

2.6 Selection effect

Based on the framework of 0-1 preferences, the above results come from varying extensive margins. Adding to the previous work of Foellmi et al. (2018a) with homogeneous firms, we obtain some new results after implanting heterogeneity. This is because the selection effect generates gaps in the extensive margins between traded and non-traded varieties in countries 1 and 2, which vary with trade costs. In this section, we explore how the intensity of the selection effect changes with trade costs in the two types of equilibria.

2.6.1 Selection effect in the no-arbitrage equilibrium

In the no-arbitrage equilibrium, the labor market clearing condition

$$L_i\theta_i = \int_0^{\psi_i^*} \psi L_i N_i \mu_i d\psi + \int_0^{\psi_{iT}^*} \tau\psi L_j N_i \mu_i d\psi + N_{ie} f_e$$

gives the mass of entry firms in country i:

$$N_{ie} = \frac{\bar{\psi}}{\psi_i^*} N_i = \frac{L_i\theta_i}{2f_e}, \quad i = 1, 2. \tag{2.28}$$

Consequently, the mass of entry firms in both countries is independent of trade costs, and that of active firms is only related to the cut-off of production, which is similar to the results of Melitz (2003). However, owing to the 0-1 utility setting, we do not need to consider the change in the intensive margin; therefore, the model is more tractable. We now check the relationship between the cut-off of domestic supply and trade costs:

$$\psi_i^* = \frac{p_i}{w_i} = \frac{\theta_i}{U_i}, \quad i = 1, 2,$$

where the second equality is obtained from Equation (2.13). According to Proposition 2, we know that $d\psi_i^*/d\tau > 0$, $i = 1, 2$.

Next, we check the relationship between the cut-off of exporters and trade costs using (2.5):

$$\psi_{1T}^* = \frac{p_2}{\tau w_1} = \sqrt{\frac{2\bar{\psi} f_e}{\tau(\tau p^2 L_1 + L_2)}}, \quad \psi_{2T}^* = \frac{p_1}{\tau w_2} = \sqrt{\frac{2\bar{\psi} f_e}{\tau(\frac{\tau L_2}{p^2} + L_1)}}.$$

Using the inequality $d(\tau p^2)/d\tau > 0$ in Appendix D, we can easily verify that $d\psi_{1T}^*/d\tau < 0$ and $d\psi_{2T}^*/d\tau < 0$. These results are similar to those of Melitz (2003), who found that increasing trade costs weaken the selection of domestic supply but strengthens that of exports in both countries.

Regarding the relative selection effect, we have

$$\frac{\psi_1^*}{\psi_2^*} = \frac{p}{w}, \tag{2.29}$$

$$\frac{\psi_{1T}^*}{\psi_{2T}^*} = \frac{1}{pw} = \frac{1}{\sqrt{\theta w}} < 1, \tag{2.30}$$

where the inequality holds as $\theta > 1$ and $w > 1$ from Proposition 2. Thus, the selection of export firms is stricter in country 1 in the no-arbitrage equilibrium, whereas the intensity of the selection of domestic supply is correlated with the ratio of relative price and wages.

2.6.2 Selection effect in the arbitrage equilibrium

In the arbitrage equilibrium, the labor market clearing conditions

$$L_1\theta_1 = \int_0^{\psi_1^*} \psi L_1 N_1 \mu_1 d\psi + \int_0^{\psi_{1TA}^*} \tau\psi L_2 N_1 \mu_1 d\psi + N_{1e} f_e,$$

$$L_2\theta_2 = \int_0^{\psi_{2A}^*} \psi L_2 N_2 \mu_2 d\psi + \int_0^{\psi_{2T}^*} \tau\psi L_1 N_2 \mu_2 d\psi + N_{2e} f_e$$

give the mass of firm entrants in country i:

$$N_{1e} = \frac{\bar{\psi}}{\psi_1^*} N_1 = \frac{L_1\theta_1}{2f_e}, \quad N_{2e} = \frac{\bar{\psi}}{\psi_{2T}^*} N_2 = \frac{L_2\theta_2}{2f_e}.$$

Note that the cost cut-off of trade now exceeds that of domestic supply in country 2; thus, the mass of active firms is determined by ψ_{2T}^*. By comparing this with Equation (2.28), we know that the masses of firm entrants are identical in the two types of equilibria. The cut-offs of domestic supply are as follows:

$$\psi_1^* = \frac{p_1}{w_1}, \quad \psi_{2A}^* = \frac{p_2 - (p_1 - \tau p_2)l}{w_2}.$$

The simulations in Figure 2.12 show that when τ increases, the domestic cut-off decreases in country 1 and increases in country 2. The trend in country 1 is different from that in the no-arbitrage equilibrium.

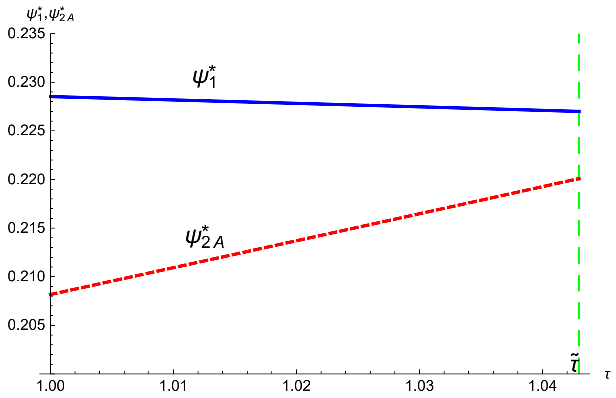


Figure 2.12 Cut-offs of domestic supply in an arbitrage equilibrium

The cut-off of exporters in the two countries is

$$\psi_{1TA}^* = \frac{p_2 - (p_1 - \tau p_2)l}{\tau w_1}, \quad \psi_{2T}^* = \frac{p_1}{\tau w_2}$$

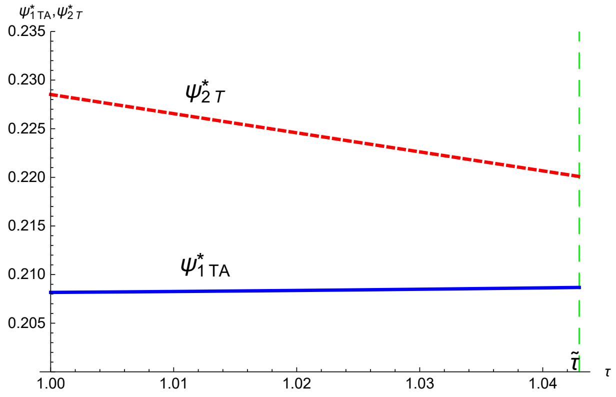


Figure 2.13 Cut-offs of exporters in an arbitrage equilibrium

The simulations in Figure 2.13 show that when τ increases, the export cut-off increases in country 1 ($d\psi_{1TA}^*/d\tau > 0$) and decreases in country 2 ($d\psi_{2T}^*/d\tau < 0$). Again, the trend in country 1 is different from that in the no-arbitrage equilibrium. Combining the results of country 1 in Figures 2.12 and 2.13, we find that more firms export and exit from the market simultaneously when trade costs increase. This is in sharp contrast to the results of typical models under CES and quasi-linear preferences (Melitz 2003; Melitz and Ottaviano 2008). As discussed in Proposition 4, increasing trade costs also increases the cost of arbitrage, which shrinks the gap between the relative price and the trade costs. As domestic profits gradually decrease, more domestic firms with productivity in $[\psi_{1TA}^*, \psi_{1T}^*]$

choose to export. Simultaneously, less productive firms are squeezed out of the market because of the more competitive labor market.

Figures 2.12 and 2.13 also explain the welfare variation in the arbitrage equilibrium from the perspective of the selection effect. The increasing cut-off levels of exporting firms in country 1 and domestic firms in country 2 suggest that there are more available varieties in country 2. Arbitrage is not beneficial to the less developed country, as it allocates too many resources to serve the large country. Specifically, the welfare gap comes from the different choices of firms with productivity in $[\psi_{1TA}^*,\ \psi_{1T}^*]$ and $[\psi_{2A}^*,\ \psi_2^*]$, which we investigated in Section 2.4.3. With rising trade costs, the productivity intervals of these firms shrink as the large market becomes less attractive; thus, more firms are selected to serve country 2.

Subsequently, the relative values of the cut-offs are written as

$$\frac{\psi_1^*}{\psi_{2A}^*} = \frac{p}{w[1-(p-\tau)l]} > 1, \tag{2.31}$$

$$\frac{\psi_{1TA}^*}{\psi_{2T}^*} = \frac{1-(p-\tau)l}{pw} < 1, \tag{2.32}$$

where the inequalities are from $p > \tau > w > 1$ provided by Lemma A.4 and Equation (2.26).[12] While both (2.30) and (2.32) show the same trend regarding the selection of exports when τ changes, the inequality in (2.31) is similar to the trend in (2.29) for a small τ.

The discussions in Section 2.6.1 and 2.6.2 are summarized as follows.

Proposition 7. *(a) The selection effect of exports is always stronger in country 1. (b) The selection effect of production is stronger in country 1 when* $\tau > (l\theta^{\frac{2}{3}} - \theta^{-\frac{2}{3}})/(l-1)$ *and stronger in country 2 when* $1 < \tau < (l\theta^{\frac{2}{3}} - \theta^{-\frac{2}{3}})/(l-1)$.

Proof. Result (a) is obtained from (2.30) and (2.32). Meanwhile, (b) is derived from (2.29) and (2.31). As the relationships are clear in (2.30), (2.31), and (2.32), we only need to verify whether $\psi_1^*/\psi_2^* > 1$ in (2.29) holds with different trade cost values. Note that $w = p$ holds when $\tau = \frac{l\theta^{\frac{2}{3}}-\theta^{-\frac{2}{3}}}{l-1}$ in the no-arbitrage equilibrium, according to (2.9) and (2.10). Moreover, $\frac{l\theta^{\frac{2}{3}}-\theta^{-\frac{2}{3}}}{l-1} > \theta^{\frac{1}{2}} > \tilde{\tau}$ holds due to $\theta > 1, l > 1$, and Lemma A.2. Additionally, we have $dp/d\tau < 0$ and $dw/d\tau > 0$ from Proposition 1. Therefore,

$$\frac{p}{w} \gtrless 1 \ \Leftrightarrow\ \frac{\psi_1^*}{\psi_2^*} \gtrless 1 \ \Leftrightarrow\ \begin{array}{l} \frac{l\theta^{\frac{2}{3}}-\theta^{-\frac{2}{3}}}{l-1} > \tau \geq \tilde{\tau}, \\ \tau > \frac{l\theta^{\frac{2}{3}}-\theta^{-\frac{2}{3}}}{l-1}, \end{array}$$

which results in (b).

□

[12] From the second equality of (2.26), we have $w < w^2 = \tau - \frac{(\tau^2-1)(1-lp+\tau l)^2}{\tau(1-lp+\tau l)^2+lp^2} < \tau$.

These results are consistent with our results for the autarky case. The cut-off of production under autarky is $\psi^* = (2f_e\bar{\psi}/L)^{1/2}$ from Equation (2.1). Thus, a larger country has a stronger selection effect when trade costs are sufficiently high. The results are different from those in the existing literature as we can now capture the impacts of factor-market and product-market competition simultaneously. For example, our propositions show that the results of Felbermayr and Jung (2018), based on the CES framework, are not robust. Using a typical Melitz (2003) model, they find that a larger country always features a lower selection effect. However, noting that the mass of entrants in either country is independent of the trade costs, the ratio of the domestic supply cost cut-offs equals the ratio of firm share, i.e., $\psi_1^*/\psi_2^* = (N_1/L_1\theta_1)/(N_2/L_2\theta_2)$. Thus, the larger country may not accommodate a more than proportionate share of firms (the home market effect in terms of firm share). This result confirms that of Chen and Zeng (2018), who show that this home market effect obtained under the CES framework is not always true if the utility function is non-CES. In fact, many typical results derived from CES frameworks must be revisited using more general frameworks.

2.7 Extension

In the benchmark model, we assume that the developed country 1 has both population and technological advantages. This section examines the case in which these two advantages are geographically separated. More specifically, the developed country 1 is the smaller country: $l \equiv L_1/L_2 < 1$ and $\theta \equiv \theta_1/\theta_2 > 1$. To investigate which advantage plays a major role, we mainly investigate the edge case of $L_1\theta_1 = L_2\theta_2$. Two countries are assumed to have the same amount of effective labor. Numerical results are provided later for more general cases.

2.7.1 $l\theta = 1$ (No-arbitrage)

To spare space, we relegate the detailed calculations to Appendix H. As the trade structure does not change when arbitrage does not appear, the wage rate and relative price are still given by Equations (2.9) and (B.1) when $l < 1$ in the no-arbitrage equilibrium. Subsequently, we directly use $\theta = 1/l$ to simplify the wage equation (2.9):

$$\mathcal{F}_E(w) \equiv \tau w^3 + w^2 - w - \tau = (w-1)\left[\tau(w^w + w + 1) + w\right] = 0. \tag{2.33}$$

It is straightforward to determine that $w = 1$ is a unique solution. Therefore, we can solve all the endogenous variables of the no-arbitrage equilibrium:

$$\begin{aligned} & p = \sqrt{\theta},\ w = 1,\ p_1 = \psi_1^* = \tau\psi_{2T}^* = \left[\frac{2\tau\bar{\psi}f_e}{L_1(1+\tau)}\right]^{\frac{1}{2}}, \\ & p_2 = \psi_2^* = \tau\psi_{1T}^* = \left[\frac{2\tau\bar{\psi}f_e}{L_2(1+\tau)}\right]^{\frac{1}{2}},\ U_1 = n_1 = \frac{\theta_1}{p_1},\ U_2 = n_2 = \frac{\theta_2}{p_2} \\ & N_1 = \theta_1\left[\frac{\tau L_1}{2f_e\bar{\psi}(1+\tau)}\right]^{\frac{1}{2}},\ N_2 = \theta_2\left[\frac{\tau L_2}{2f_e\bar{\psi}(1+\tau)}\right]^{\frac{1}{2}}. \end{aligned} \tag{2.34}$$

We revisit our previous propositions of the no-arbitrage equilibrium in this subsection and discuss it in more detail later with the arbitrage equilibrium. Based on the results of (2.34), we can verify that (a) the wage rate and relative market price are constant and independent of the trade costs, and (b) price reversal does not occur. As θ_i is larger in country 1, income per capita and market price are both higher in country 1. Further, (c) the welfare level is higher in the country with the technological advantage.

2.7.2 $l\theta = 1$ (Arbitrage)

Subsequently, we turn to the arbitrage equilibrium when the trade costs are sufficiently small. As shown in Appendix H.1, the equilibrium relative price $p > 1$ as long as $l < 1$. Accordingly, arbitrage still occurs in country 1; therefore, the trade structure is not affected by population size. As wages and relative prices are constant in the no-arbitrage equilibrium, the threshold trade cost value between the two equilibria can also be explicitly solved.[13]

We use $\theta = 1/l$ to simplify the functions (2.25) and (2.26) of the arbitrage equilibrium. The relative price is implicitly given by

$$\begin{aligned} \mathcal{K}_E(p) &\equiv \mathcal{K}_{1E}(p)^2 - \mathcal{K}_{2E}(p) = 0 \\ \mathcal{K}_{1E}(p) &\equiv \frac{1 - pl + \tau l}{l\left[l\tau(p-\tau)^2 + p^2 - p\tau + \tau^2\right]}, \\ \mathcal{K}_{2E}(p) &\equiv \frac{(1 - lp + \tau l)^2 + \tau l p^2}{\tau(1 - lp + \tau l)^2 + lp^2}, \end{aligned} \tag{2.35}$$

and the equilibrium relative wage rate is given by

$$w = \mathcal{K}_{1E}(p)(= \sqrt{\mathcal{K}_{2E}(p)}). \tag{2.36}$$

Based on the implicit function, we can obtain the following conclusions when arbitrage appears: First, the market price and wage rate are higher in the country with higher

[13] $\tilde{\tau}_E = \sqrt{\theta}$ according to (H.2).

productivity (θ_i), and both the relative price and wage rate are nonmonotonic with trade costs. Second, the welfare level of country 1 is higher.

2.7.3 Welfare and selection effect

We further examine the impact of $\bar{\psi}$ on welfare and gains from trade. The left panel of Figure 2.14 shows that a larger $\bar{\psi}$ harms welfare in either country, which is similar to the message in Figure 2.7 for the case of $l > 1$. Meanwhile, a larger $\bar{\psi}$ diminishes the gains from trade in country 1 and the losses from trade in country 2 according to the right panel, which is similar to Figure 2.8.

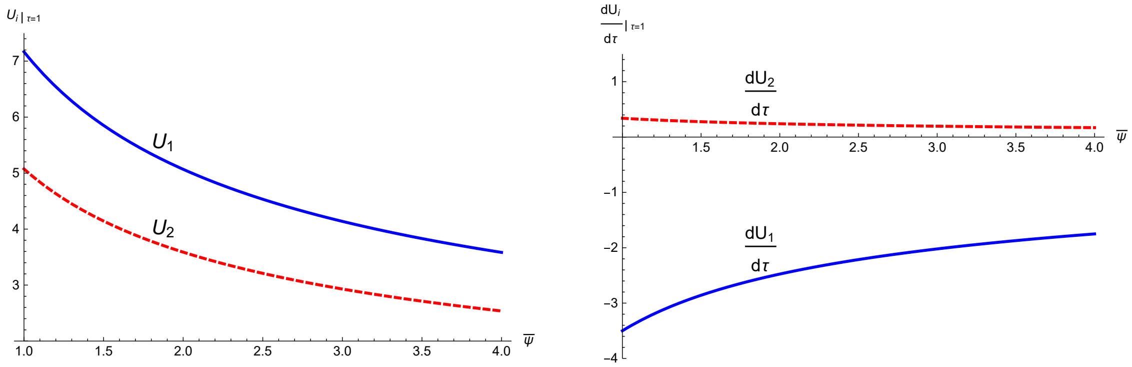


Figure 2.14 Extension: the impact of $\bar{\psi}$ on welfare and gains from trade

Finally, we check the variation of cut-off values to investigate the selection effect. As the no-arbitrage equilibrium is analytically solved in (2.34), we only show numerical examples of arbitrage equilibrium as follows.

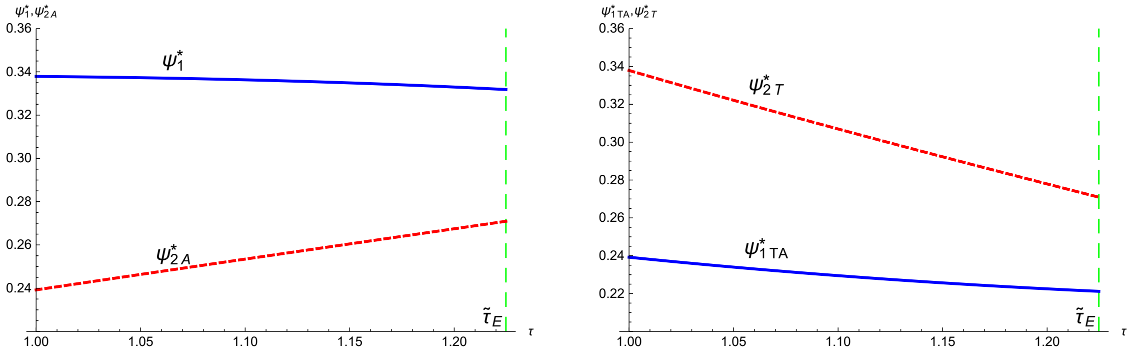


Figure 2.15 Extension: cut-offs in the arbitrage equilibrium

Comparing the left panel of Figure 2.15 with Figure 2.12, we find that the monotonic properties are identical. A rise in trade costs lowers ψ_1^* but increases ψ_{2A}^*. Meanwhile, comparing the right panel of Figure 2.15 with Figure 2.13, we find that ψ_{2T}^* has a similar behavior, but ψ_{1TA}^* behaves differently.

In summary, we conclude our discussion of the extension part as follows.

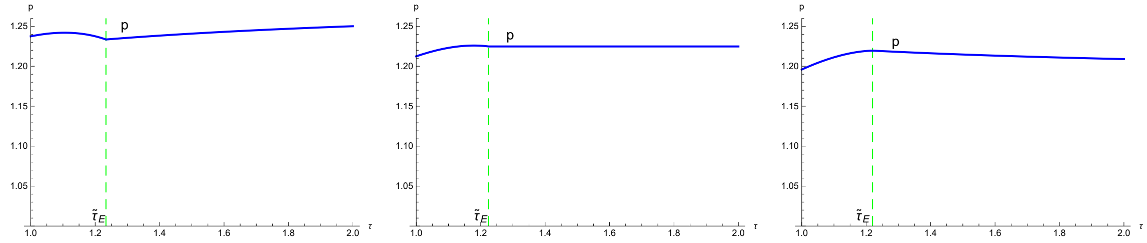
Proposition 8. *In the edge case of $l\theta = 1$ with $\theta_1 > \theta_2$, (a) arbitrage is possible only in country 1, and the market price is always higher in country 1; (b) wages are equalized between the two countries in the no-arbitrage equilibrium but higher in country 1 in the arbitrage equilibrium; (c) the welfare level is always higher in country 1; (d) an increase in $\bar{\psi}$ mitigates the impact of trade liberalization on welfare; (e) the selection effect of exports is always stronger in country 1, whereas that of domestic supply is always stronger in country 2.*

Proof. See Appendix H. □

The results are considerably similar to our previous model when $l > 1$, indicating that technological advantage is more important in determining the trade structure. When the advantages are separated in the two countries, the market price is always higher in the country with the technological advantage, and price reversal does not occur. This is because technological advantage provides a higher income per capita, whereas population advantage represents a tougher pro-competitive effect. Both lead to lower market prices in country 2. However, the lower price in country 2 is insufficient to invert the higher welfare in country 1. In the arbitrage equilibrium, the risk of arbitrage reduces the profits of trade and leads to higher exclusive profits in country 1. Therefore, many resources are absorbed into the developed country, which induces a welfare loss in the developing country. In this sense, the population advantage seems less important when facing arbitrage. Our conclusions suggest that in the trade conducted by two countries with similar total effective labor, the country with higher productivity possesses a leading position. Additionally, the risk of arbitrage is more likely to be a threat to developing countries than developed countries.

2.7.4 General cases

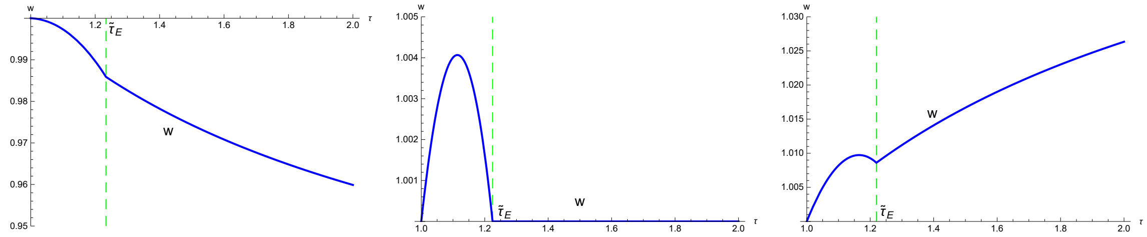
We now use numerical examples to show how the relative price, wage rate, and welfare of the two countries change when the two advantages are not necessarily equalized between the two countries ($\theta l \neq 1$). The parameter values in the examples are given in Appendix H.3. By renaming the countries, we maintain the assumption that $\theta \geq 1$ without loss of generality.



(a) $\theta l < 1$ (b) $\theta l = 1$ (c) $\theta l > 1$

Figure 2.16 General cases: relative market prices.

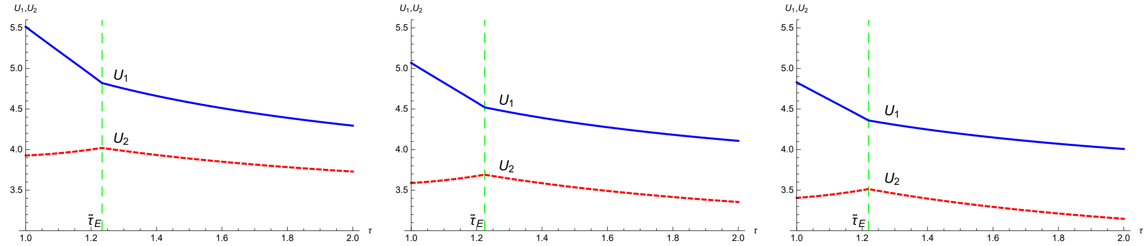
Figure 2.16 shows that $p \geq 1$ always holds when $l < 1,\ \theta \geq 1$. However, the impact of trade costs τ on p depends on the ratio of total effective labor (θl). The three panels show that the trends of relative price p change with the relative endowment of effective labor.



(a) $\theta l < 1$ (b) $\theta l = 1$ (c) $\theta l > 1$

Figure 2.17 General cases: wage rate.

Wages also show a close relationship with the allocation of advantages, as shown in Figure 2.17. The three panels show that the wage gap displays a bell shape with respect to trade costs. However, it may decrease monotonically when $l\theta < 1$.



(a) $\theta l < 1$ (b) $\theta l = 1$ (c) $\theta l > 1$

Figure 2.18 General cases: welfare

The welfare results are plotted in Figure 2.18. The three panels show similar trends, suggesting that both advantages are beneficial to national welfare and that the impact of trade liberalization is mainly determined by technological advantage. Specifically, only the country with a technology advantage benefits from trade after arbitrage occurs, which is independent of the population distribution.

Numerical examples indicate that the advantages of total labor endowment from two dimensions (L_i and θ_i) may have different levels of influence. The edge case of $\theta l = 1$ helps understand their balance.

Chapter 3

Carbon tax vs. emission trading in a monopolistically competitive market with heterogeneous firms

3.1 Introduction

Emission reduction has been one of the most concerning issues in global climate change since the mid-20th century. Among the various regulation approaches, the carbon tax (CT) and emission trading scheme (ETS) are two of the most commonly adopted policy instruments around the world. This research builds a two-sector general-equilibrium model with monopolistic competition and heterogeneous firms to compare the efficiency of these two policies.

In contrast to the long history of the CT, the world's first emission cap-and-trade system,[14] named SO_2 allowance-trading system, was established in the US in 1994-2010 (Stavins 2019). Meanwhile, the first multilateral trading scheme for multi-greenhouse gas emissions, named the EU ETS, was formally launched in Europe in 2005.[15]

There has been a lot of controversy over the efficiency comparison between a CT and an ETS in the literature from various perspectives.[16] One of the important arguments attributes the reasons for non-equivalence to the uncertainty and asymmetric information, which can be traced back to Weitzman (1974). He theoretically explains the difference in efficiency between the CT and ETS if the market has uncertainty and asymmetric information. His model has a single firm, and the equilibrium is considered to be in the short run. Later Spulber (1985) further compares the two policies in the long run with entry/exit of firms and shows that the effluent tax and tradeable permit are equivalent under perfect foresight and certainty. Spulber's result is amended by Shinkuma and Sugeta (2016) by incorporating asymmetric information and uncertainty. They find that in the long-run, with an entry cost, the ETS induces insufficient market entry, whereas the CT can induce either excessive or insufficient market entry. Additionally, they show that when the entry cost is sufficiently low, the ETS is always superior to the CT; otherwise,

[14] The "ETS" and the "cap-and-trade system" are interchangeable in our study to indicate the trade of emission allowances.

[15] See the website of the European Commission (https://reurl.cc/eWvLVx).

[16] Many studies investigate the differences in specific design and operating elements between two policies, such as the government's acting strategy (Ishikawa and Kiyono 2006; Wirl 2012; Kiyono and Ishikawa 2013; Eichner and Pethig 2015), transaction cost (Stavins 1995; Baudry et al. 2021), and footloose capital (Lai 2022).

either policy can be more efficient.

The analysis of Shinkuma and Sugeta (2016) is based on a perfectly competitive market with a homogeneous good. However, in the real world, most dirty goods are heterogeneous and imperfectly substitutable. In addition, emission-intensive industries are generally characterized by increasing return to scale and a large mass of firms (Zeng and Zhao 2009; Batrakova and Davies 2012; Kreickemeier and Richter 2014; Forslid et al. 2018). Thus, we build a two-sector general-equilibrium model with monopolistic competition to compare the policies. Fortunately, the Melitz type heterogeneity (Melitz 2003) is similar to the uncertainty and asymmetric information addressed in Shinkuma and Sugeta (2016). Exploring the selection effect[17] allows us to examine how each policy affects the market allocation and analyze the sources of policy inefficiency. We verify that the CT/price control can achieve an optimal mass of active firms but an insufficient mass of entrants, while the ETS/quantity control performs just the opposite. Meanwhile, the distinction of market outcome under each policy induces different labor allocations across sectors. Specifically, our study leads to the following two results.

First, the degree of productivity heterogeneity is a crucial determinant of the superiority between the two policies. Given the total emission cap, an economy with high heterogeneity does better to adopt the ETS, whereas the CT is superior in a low-heterogeneity economy. Under the CT, the government can adjust the mass of active firms, but not the mass of entrants, by controlling the lump-sum tax (subsidy). Conversely, under the ETS, the government can adjust the mass of entrants rather than the mass of active firms, by controlling the initial permit allocation. These mechanisms lead to different market outcomes, embodied by fewer/more active firms, fewer/more entrants, and a stronger/weaker selection effect. They also induce different resource allocations across sectors, resulting in different market efficiencies.

Second, we find that both the CT and ETS fail to reach the social optimum. After comparing the market allocation of either policy with the optimum, the source of market distortions is disclosed. Our analysis shows that under the CT, the market has a proper mass of active firms but too few entrants, which results in low average productivity. In contrast, the market under the ETS has a proper mass of entrants but too few active firms, which leads to insufficient varieties and a resource loss in the entry costs. Moreover, we verify that under either policy, excessive labor resources are allocated to the non-polluting sector, which also induces excessive emission in the polluting production. Our results

[17] The selection effect means that entrants with lower productivity are driven out of the market (Melitz 2003; Melitz and Ottaviano 2008). In this chapter, a tougher selection effect indicates a higher average productivity of active firms as more low-productivity entrants are eliminated, allowing the mass of entrants and active firms to be endogenously determined.

are closely related to the literature studying the market distortions in an imperfectly competitive market with one production factor (Nocco et al. 2014; Dhingra and Morrow 2019; Behrens et al. 2020). In contrast, taking emission regulation as a resource, our model shows that the resource-allocation share has an impact on equilibrium output.

3.2 Literature review

Many papers provide nice reviews regarding the similarity and distinctions between the two policies (Stavins 2019; Aldy and Stavins 2012). Apart from the uncertainty and asymmetric information we mentioned, many other factors are incorporated in literature from different fields to compare the emission tax and cap-and-trade system. The similar functions of the two policies can be mainly concluded from three aspects: 1). Incentive for emission reductions (Metcalf 2007; Stavins 2007); 2.) Incentive for carbon-saving innovation (Milliman and Prince 1989; Fischer, Parry, and Pizer 2003); 3.) Competitiveness effect in the energy-intensive manufacturing sector (Aldy and Stavins 2012; Goulder and Schein 2013).

Meanwhile, many studies claim the difference between the two policies while considering the specific design elements. Stavins (1995) and Konishi and Tarui (2015) find that the distribution of initial permits might affect the efficiency of ETS, while Baudry, Faure, and Quemin (2021) show that transaction costs of permits also matter. Coria and Kyriakopoulou (2018) investigate the different impacts of emission taxes, uniform emission standards, and performance standards on the size distribution of firms. As the processes of the two policies run differently in reality, the complexity and administration are quite different (Goulder & Schein 2013). Moreover, this will lead to a different level of corruption and market manipulation (Metcalf 2019).

Our research also contributes to the stream of literature that studies the market distortions in an imperfectly competitive market. Assuming one production factor, Nocco, Ottaviano & Salto (2014), Dhingra & Morrow (2019), and Behrens, Mion, Murata & Suedekum (2020) report three margins of potential inefficiency: proper selection of active firms, proper output of each firm, and mis-allocation of labor resources between sectors. To the best of our knowledge, there is no existing related literature to consider this issue with multi factors. Taking emission regulation as a kind of resource, we have two production factors in our model. Our study suggests that there might exist more sources of market distortion in multi-factor production.

3.3 Model

3.3.1 Demand

There are two types of goods in the economy: a continuum of differentiated goods in a polluting manufacturing sector M and a homogeneous good in a clean sector A. All individuals have the same preferences, characterized by the following quasi-linear utility function (Pflüger 2004):[18]

$$U = \alpha \ln C^M + C^A, \quad C^M = \left[\int_0^n x(i)^{\frac{\sigma-1}{\sigma}} di \right]^{\frac{\sigma}{\sigma-1}}, \ \alpha \in (0,1), \ \sigma > 1, \tag{3.1}$$

where C^M is the manufacturing aggregate, C^A is the consumption of the composite good in sector A, n is the mass of available varieties, $x(i)$ is the consumption of variety i, and σ is the substitute elasticity between two varieties. The budget constraint of each individual is written as

$$\int_0^n p(i)x(i)dj + p_A C^A = y,$$

where y is individual income, including wages w and the transfer payment from the government. The transfer payment differs under different policies, which we will introduce in detail later.

The utility maximization yields the demand functions

$$x(i) = \alpha \frac{p(i)^{-\sigma}}{P^{1-\sigma}}, \quad C^M = \frac{\alpha}{P}, \quad C^A = y - \alpha, \tag{3.2}$$

where

$$P \equiv \left[\int_0^n p(i)^{1-\sigma} di \right]^{\frac{1}{1-\sigma}} \tag{3.3}$$

is the price index.

3.3.2 Production

The labor endowment of the economy is L. The composite good C^A is produced with a constant return to scale technology in a competitive market and does not generate emissions. Choosing C^A as the numeraire, we have $p_A = w = 1$. Varieties in M are produced under increasing returns to scale in a monopolistic competition market, and each firm produces one variety. Specifically, after sinking f_e units of labor as an entry cost, each firm randomly draws its marginal input level $\varphi \in (0, \bar{\varphi}]$ from a Pareto distribution $G(\varphi) = (\varphi/\bar{\varphi})^k$ with density function $g(\varphi) = (k\varphi^{k-1})/\bar{\varphi}^k$. The positive constant

[18] We neglect the negative externality of emission in the utility, since we compare the policies to reach the same emission target. In particular, Section 3.4.3 shows that the neglected part leads to the same optimal emission target.

k determines the shape of the marginal input distribution, where a smaller k indicates a higher heterogeneity. To start production, each firm needs a fixed input of F units of labor. Emission is generated when manufacturing goods are produced. Since firms can input labor for emission abatement, we follow Copeland and Taylor (1994) and write the production function as

$$q(e_i, l_i, \varphi_i) = \frac{e_i^{\beta} l_i^{1-\beta}}{\beta^{\beta}(1-\beta)^{1-\beta}\varphi_i},$$

which yields the emission-related cost function

$$c_i(q_i, \varphi_i) = (p^e)^{\beta}\varphi_i q_i,$$

where p^e is the cost of per-unit emission and β denotes the input share of emission. The government has two available policies, a CT and an ETS, under which p^e takes different forms. It equals to the tax rate t under the CT, determined by the government; whereas equals to the emission price s under the ETS, determined by the emission market. Additionally, the emission of firms with marginal cost φ_i is $e(\varphi_i) = \beta(p^e)^{\beta-1}\varphi_i q_i$.

The profit function of a firm with marginal input level φ_i can be written as

$$\pi(\varphi_i) = p(\varphi_i)q(\varphi_i) - \mathcal{T}(q, \varphi_i), \quad \mathcal{T}(q, \varphi_i) = \begin{cases} t^{\beta}\varphi_i q + f + F, & \text{CT} \\ s^{\beta}\varphi_i q - \bar{e}s + F, & \text{ETS} \end{cases}$$

where $\mathcal{T}$ is the total cost consisting of production cost and transfer payment from the government, $t(= p^e)$ is the carbon tax charged for each unit of emission and f is the lump-sum carbon tax. When the government adopts the ETS, no extra tax is charged. $\bar{e}$ is the initial emission quota of allowances allocated to each entrant, and $s(= p^e)$ is the emission price for unit emission allowance in the ETS. Moreover, we assume that the total emission target of the government is $\bar{E}$. We impose the following assumptions for the validity of our model.

Assumption 1. *Parameters are assumed to satisfy the following inequalities:*

$$k + 1 - \sigma > 0, \tag{3.4}$$

$$f_e < \frac{F(\sigma-1)^2}{\sigma(k+1-\sigma)}, \tag{3.5}$$

$$\max\left\{\frac{1+k-\sigma}{k(\sigma-1)^2}, \frac{1}{k(\sigma-1)}\right\} < \beta < 1. \tag{3.6}$$

Inequality (3.4) ensures that there exist sufficient high-productivity firms. Otherwise, no firms produce in the market. Inequality (3.5) excludes the case that entry cost is over-high, in which no potential entrants are willing to enter the market. Inequalities in (3.6) limit the intensity of emissions. If the intensity of emissions is too low in the production, the low demand of emission allowances will lead to a lower carbon price and initial transfer payment $\bar{e}s$. Therefore, the government can not attract sufficient entrants into the market, leading to the failure of the ETS policy. We impose this assumption as we want to compare the efficiency of two policies when both fully work.

3.4 Equilibrium under two policies

3.4.1 The CT policy

Following Shinkuma and Sugeta (2016), the government imposes two kinds of carbon taxes. One is a per unit emission tax t, and the other is a lump-sum tax (subsidy) f, which can be either positive or negative. We use a subscript "t" to denote the CT case. Therefore, the total tax revenue of the economy is written as

$$T = t\bar{E} + fN_t, \tag{3.7}$$

where N_t is the mass of active firms under a CT. The total tax revenue will be evenly redistributed to individuals.

Together with (3.2), the profit maximization yields

$$p_t(\varphi_i) = \frac{\sigma}{\sigma - 1} t^{\beta} \varphi_i, \; q_t(\varphi_i) = \alpha L \frac{p_t(\varphi_i)^{-\sigma}}{P_t^{1-\sigma}}.$$

Defined by (3.3), the price index P_t under the CT policy can be rewritten as

$$P_t = \left[\int_0^{\varphi_t^*} p_t(\varphi_i)^{1-\sigma} N_t \mu_t(\varphi_i) d\varphi_i \right]^{\frac{1}{1-\sigma}} = \frac{\sigma t^{\beta}}{\sigma - 1} \left(\frac{kN_t}{k - \sigma + 1} \right)^{\frac{1}{1-\sigma}} \varphi_t^*,$$

where $k - \sigma + 1$ is assumed to be positive according to (3.4), $\mu_t(\varphi_i)$ is the distribution of active firms when the government imposes tax

$$\mu_t(\varphi_i) = \frac{g(\varphi_i)}{G(\varphi_t^*)} = \frac{k\varphi_i^{k-1}}{\varphi_t^{*k}},$$

and φ_t^* is the marginal cost cutoff of active firms under the CT policy.

The zero cutoff profit condition is

$$\pi_t(\varphi_t^*) = \alpha L \frac{k - \sigma + 1}{\sigma k N_t} - F - f = 0,$$

which yields

$$N_t = \frac{\alpha L(1 + k - \sigma)}{\sigma k(f + F)}. \tag{3.8}$$

The free entry condition indicates that

$$f_e = \int_0^{\varphi_t^*} \pi_t(\varphi_i) g(\varphi_i) d\varphi_i = \left(\frac{\varphi_t^*}{\bar{\varphi}} \right)^k \left(\frac{\alpha L}{\sigma N_t} - F - f \right).$$

Combining this with (3.8), we obtain

$$\varphi_t^* = \bar{\varphi} \left[\frac{f_e(k - \sigma + 1)}{(f + F)(\sigma - 1)} \right]^{\frac{1}{k}}.$$

Additionally, the mass of entrants can be derived as

$$M_t = \frac{N_t}{G(\varphi_t^*)} = \frac{\alpha L(\sigma - 1)}{k\sigma f_e}.$$

Note that the emission amount of the firm with marginal cost φ_i under the CT is $e_t(\varphi_i) = \beta t^{\beta-1}\varphi_i q_t(\varphi_i)$. Therefore, the aggregate emission is written as

$$E = \int_0^{\varphi_t^*} e_t(\varphi_i) N_t \mu_t(\varphi_i) d\varphi_i = \alpha\beta L \frac{\sigma - 1}{t\sigma}. \tag{3.9}$$

The government sets tax rate t to achieve the emission target $\bar{E}$. Equation (3.9) gives

$$t = \frac{\alpha\beta L(\sigma - 1)}{\sigma\bar{E}}. \tag{3.10}$$

Interestingly, the two types of taxes play different roles in the policy. The marginal tax t is used to control the total emission, which does not affect the market outcome, whereas the lump-sum tax f affects the level of selection. Moreover, the CT policy can only adjust the mass of active firms through the selection effect, whereas the mass of entrants is independent of the tax.

The government chooses the optimal f to maximize the utility of a representative resident:

$$\begin{aligned}
W_t(f) =& \alpha \ln \frac{\alpha}{P_t} + 1 - \alpha + \frac{T}{L}, \\
=& \alpha \left(\frac{1}{k} + \frac{1}{1 - \sigma} \right) \ln(F + f) + \frac{\alpha f(k + 1 - \sigma)}{k\sigma(F + f)} + \frac{\alpha\beta(\sigma - 1)}{\sigma} \\
& + 1 - \alpha + \alpha \ln \alpha - \alpha \ln \frac{\bar{\varphi}\sigma}{\sigma - 1} \\
& - \frac{\alpha}{k} \ln \frac{f_e(k + 1 - \sigma)}{\sigma - 1} + \frac{\alpha}{\sigma - 1} \ln \frac{\alpha L}{\sigma} - \alpha\beta \ln \frac{\alpha\beta L(\sigma - 1)}{\sigma\bar{E}},
\end{aligned}$$

where the second equality comes from (3.7), (3.8), and (3.10). The FOC is

$$W_t'(f) = -\frac{\alpha(F + \sigma f)(1 + k - \sigma)}{\sigma k(\sigma - 1)(F + f)^2} = 0 \quad \text{giving} \quad f^* = -\frac{F}{\sigma}.$$

Moreover, we have

$$W_t''(f^*) = -\frac{\alpha(k + 1 - \sigma)\sigma^2}{F^2 k(\sigma - 1)^3} < 0.$$

Therefore, the CT equilibrium is described as[19]

$$\varphi_t^*(f^*) = \bar{\varphi}\Big[\frac{\sigma f_e(k+1-\sigma)}{F(\sigma-1)^2}\Big]^{\frac{1}{k}}, \quad N_t(f^*) = \frac{\alpha L(k+1-\sigma)}{kF(\sigma-1)}.$$

The equilibrium welfare under the optimal CT is

$$\begin{aligned} W_t(f^*) =& \frac{\alpha}{k}\ln\frac{F(\sigma-1)^2}{\sigma f_e(k+1-\sigma)} + \frac{\alpha}{\sigma-1}\ln\frac{\alpha L}{F(\sigma-1)} - \frac{\alpha(k+1-\sigma)}{k\sigma(\sigma-1)} \\ &+ \frac{\alpha\beta(\sigma-1)}{\sigma} + 1 - \alpha + \alpha\ln\frac{\alpha(\sigma-1)}{\sigma\bar{\varphi}} - \alpha\beta\ln\frac{\alpha\beta L(\sigma-1)}{\sigma\bar{E}}. \end{aligned}$$

The results show that the government needs to provide a subsidy to firms to encourage them to produce. Intuitively, taxation limits the emission level and increases firms' production costs simultaneously, resulting in insufficient producers. To fix part of the bias, the government has to transfer some taxation into lump-sum subsidy and encourage more entrants to produce. This approach is common in the real world. For example, Bourgeois et al. (2021) find that subsidy recycling has some advantages.

3.4.2 The ETS policy

Applying the ETS policy, the government can control the mass of entrants by adjusting the initial emission allowances $\bar{e}$.[20] We use a subscript "e" in the notations to indicate the case under an ETS. Since the productivity is private information, the government can only observe the distribution, not the specific productivity level of each firm. Therefore, the initial allowance is evenly allocated to all entrants.[21] The mass of entrants is

$$M_e = \frac{\bar{E}}{\bar{e}}.$$

Using emission price s in the ETS, the profit maximization yields

$$p_e(\varphi_i) = \frac{\sigma}{\sigma-1}s^{\beta}\varphi_i, \quad q_e(\varphi_i) = \alpha L\frac{p_e(\varphi_i)^{-\sigma}}{P_e^{1-\sigma}}.$$

[19] Assumptions (3.4) and (3.5) ensure that $\varphi_t^* < \bar{\varphi}$ and $T > 0$ when $f = f^*$.

[20] Some papers (Shinkuma and Sugeta 2016; Lai 2022) assume that the government will allocate a fraction of initial allowances to firms freely (the initial emission rights owe to the producers) and the rest are auctioned (the initial emission rights owe to the consumers). However, in this research, the equilibrium market outcome and social welfare remain unchanged regardless of the initial allowances allocation among the entrants or auctioned. This indicates our model conforms to the Coase Theorem as there is no transaction costs of carbon trading or auction. We show the proof in Appendix I.

[21] In the case of ETS, the entry cost f_e can be explained as a kind of entry license. The government can control the amount of licenses to determine the mass of entrants. Moreover, the firms can only start production after receiving the allowances, so the government can not utilize the initial permits allocation to directly control the mass of active firms due to the information asymmetry.

The price index defined in (3.3) under an ETS can be rewritten as

$$P_e = \frac{\sigma s^\beta}{\sigma - 1} \left(\frac{kN_e}{k - \sigma + 1} \right)^{\frac{1}{1-\sigma}} \varphi_e^*,$$

where $k - \sigma + 1$ is positive according to (3.4), N_e is the mass of active firms, and φ_e^* is the marginal cost cutoff of active firms in the ETS.

The zero cutoff profit condition becomes

$$0 = \pi(\varphi_e^*) - \bar{e}s = \alpha L \frac{k - \sigma + 1}{\sigma k N_e} - F,$$

which yields

$$N_e = \alpha L \frac{k - \sigma + 1}{\sigma k F}. \tag{3.11}$$

Note that the sales of emission allowances are not included in the operating profit, which does not affect firm's decision on production.

The distribution of active firms when government adopts the ETS is written as

$$\mu_e(\varphi_i) = \frac{g(\varphi_i)}{G(\varphi_e^*)} = \frac{k\varphi_i^{k-1}}{\varphi_e^{*k}}.$$

Moreover, we have

$$N_e = M_e G(\varphi_e^*) = \frac{\bar{E}}{\bar{e}} \left(\frac{\varphi_e^*}{\bar{\varphi}} \right)^k.$$

Combining this with (3.11), we obtain the cutoff

$$\varphi_e^* = \bar{\varphi} \left[\frac{\alpha \bar{e} L (k - \sigma + 1)}{\sigma k F \bar{E}} \right]^{\frac{1}{k}}.$$

The mass of firms, N_e in (3.11), is independent of the initial allowances allocation under the ETS. The government only controls the mass of entrants to adjust the productivity level of active firms.

Note that the emission output of firms with marginal cost φ_i in the ETS is $e_e(\varphi_i) = \beta s^{\beta-1} \varphi_i q_e(\varphi_i)$. Therefore, the emission-clearing condition under the ETS is written as

$$\bar{E} = \int_0^{\varphi_e^*} e_e(\varphi_i) N_e \mu_e(\varphi_i) d\varphi_i = \alpha \beta L \frac{\sigma - 1}{s\sigma},$$

from which we can obtain the emission price in the ETS:

$$s = \frac{\alpha \beta L (\sigma - 1)}{\sigma \bar{E}}.$$

Since the mass of entrants is determined by the government, there is no free entry under the ETS. Therefore, firms may have positive net profits, which are evenly redistributed to the individuals. The total profit is

$$\Pi = \int_0^{\varphi_e^*} \left[\frac{p_e(\varphi_i) q_e(\varphi_i)}{\sigma} - F \right] N_e \mu_e(\varphi_i) d\varphi_i - M_e f_e + \bar{E}s$$

$$=\frac{\alpha L}{\sigma} - FN_e - \frac{\bar{E} f_e}{\bar{e}} + \bar{E}s.$$

The government determines the initial allowances $\bar{e}$ to maximize the utility of a representative resident:

$$\begin{aligned}W_e(\bar{e}) =& \alpha \ln \frac{\alpha}{P_e} + 1 - \alpha + \frac{\Pi}{L}\\ =& 1 - \alpha + \alpha\beta - \frac{f_e \bar{E}}{\bar{e} L} + \frac{\alpha(\sigma - \beta k - 1)}{k\sigma}\\ &- \alpha \ln \left\{ \frac{\beta L \bar{\varphi}}{\bar{E}} \left(\frac{\alpha L}{\sigma F}\right)^{\frac{1}{1-\sigma}} \left[\frac{\alpha L \bar{e}(k-\sigma+1)}{\sigma k F \bar{E}}\right]^{\frac{1}{k}} \left[\frac{\alpha\beta L(\sigma-1)}{\sigma \bar{E}}\right]^{\beta-1} \right\}.\end{aligned}$$

The FOC is

$$W_e'(\bar{e}) = \frac{f_e \bar{E}}{\bar{e}^2 L} - \frac{\alpha}{k\bar{e}} = 0 \quad \text{giving} \quad \bar{e}^* = \frac{k f_e \bar{E}}{\alpha L},$$

and the SOC is

$$W_e''(\bar{e}) = -\frac{L^2 \alpha^3}{\bar{E}^2 f_e^2 k^3} < 0.$$

Thus, the equilibrium of the optimal ETS is solved out[22]:

$$\varphi_e^*(\bar{e}^*) = \bar{\varphi} \left[\frac{f_e(k+1-\sigma)}{\sigma F}\right]^{\frac{1}{k}}, \quad M_e(\bar{e}^*) = \frac{\alpha L}{k f_e}.$$

The equilibrium welfare in the ETS with the optimal initial allocation is

$$\begin{aligned}W_e(\bar{e}^*) =& 1 - \alpha + \alpha\beta - \frac{\alpha(\beta k + 1)}{k\sigma}\\ &- \alpha \ln \left\{ \frac{\beta L \bar{\varphi}}{\bar{E}} \left(\frac{\alpha L}{\sigma F}\right)^{\frac{1}{1-\sigma}} \left[\frac{f_e(k-\sigma+1)}{\sigma F}\right]^{\frac{1}{k}} \left[\frac{\alpha\beta L(\sigma-1)}{\sigma \bar{E}}\right]^{\beta-1} \right\}.\end{aligned}$$

3.4.3 Comparison between the CT and the ETS

First, we compare two market outcomes:

$$\frac{\varphi_t^*}{\varphi_e^*} = \left(\frac{\sigma}{\sigma-1}\right)^{\frac{2}{k}} > 1, \ \frac{N_t}{N_e} = \frac{\sigma}{\sigma-1} > 1, \ \frac{M_t}{M_e} = \frac{\sigma-1}{\sigma} < 1. \tag{3.12}$$

The results are summarized as follows

Proposition 9. *Compared to the CT policy, the economy under the ETS has more entrants, fewer active firms, and a stronger selection effect (higher average productivity).*

Proof. See (3.12). □

[22] Assumptions (3.4) and (3.5) ensure that $\varphi_e^* < \bar{\varphi}$, while (3.6) ensures that $\Pi > 0$ when $\bar{e} = \bar{e}^*$.

This result indicates that two policies can shape the market outcome in different ways. Under the CT policy, the government charges the marginal tax to achieve the emission target and transfers part of this tax revenue as subsidies to firms to reduce the distortion. Although the tax encourages more firms to produce, more low-productivity firms survive, leading to a lower average productivity. In contrast, in the ETS, the government directly allocates all revenue from emissions to the entrants, which increases the expected profits and attracts more potential entrants. However, the ETS fails to encourage more firms to produce, inducing an insufficient mass of active firms.

Next, we investigate how the different market allocations affect the equilibrium welfare:

$$\begin{aligned}\Delta W \equiv & W_t - W_e = \underbrace{\alpha \ln \frac{P_e}{P_t}}_{\text{price index gap}} + \underbrace{\frac{T-\Pi}{L}}_{\text{redistribution gap}} \\ &= \frac{\alpha}{\sigma-1} \ln \frac{N_t}{N_e} + \alpha \ln \frac{\varphi_e^*}{\varphi_t^*} + \frac{T-\Pi}{L} \\ &= \frac{\alpha(k+2-2\sigma)}{k\sigma(\sigma-1)}\left(\sigma \ln \frac{\sigma}{\sigma-1} - 1\right). \end{aligned} \tag{3.13}$$

The welfare gap can be divided into two parts: the price index gap and the redistribution gap. The price index gap results from the differences between the mass of varieties and the average productivity level. As we discussed before, the ETS is superior in the average productivity but inferior in the mass of varieties. The redistribution gap indicates the difference between the total income levels. Intuitively, the government needs to choose whether to use the revenue from emission regulations to encourage manufacturing production or to directly redistribute the revenue to households.

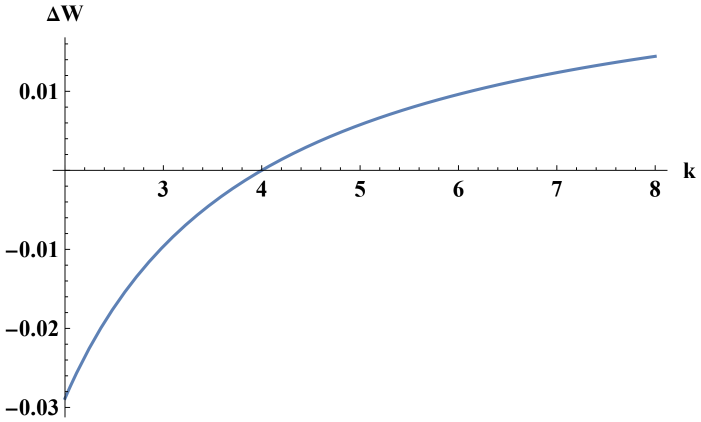


Figure 3.1 The welfare difference with heterogeneity

Interestingly, according to (3.13), we have $d(\Delta W)/d\bar{E} = 0$ (i.e., $dW_t/d\bar{E} = dW_e/d\bar{E}$), which indicates that the welfare gap is independent of the total emission target in our model. As emphasized in footnote 18, this means that the optimal total emission target under either policy is identical, regardless of whether there is an externality of emission in the utility form of (3.1). The degree of firm heterogeneity plays a distinct role in determining the relative efficiency of the two policies. We give a numerical example in Figure 3.1 with parameter value $\alpha = 0.8$, $\sigma = 3$, where the horizontal axis denotes the degree of heterogeneity (a larger k indicates a lower degree of heterogeneity) and the vertical axis denotes the welfare difference shown in (3.13). This result is summarized as follows.

Proposition 10. *The ETS is more efficient than the CT policy when $k \in (\sigma - 1, 2(\sigma - 1))$ and is less efficient when $k > 2(\sigma - 1)$.*

Proof. It is easy to verify that $\sigma \ln \frac{\sigma}{\sigma-1} - 1 > 0$ always holds. Therefore, $\Delta W \lesseqgtr 0$ holds when $k \lesseqgtr 2(\sigma - 1)$. □

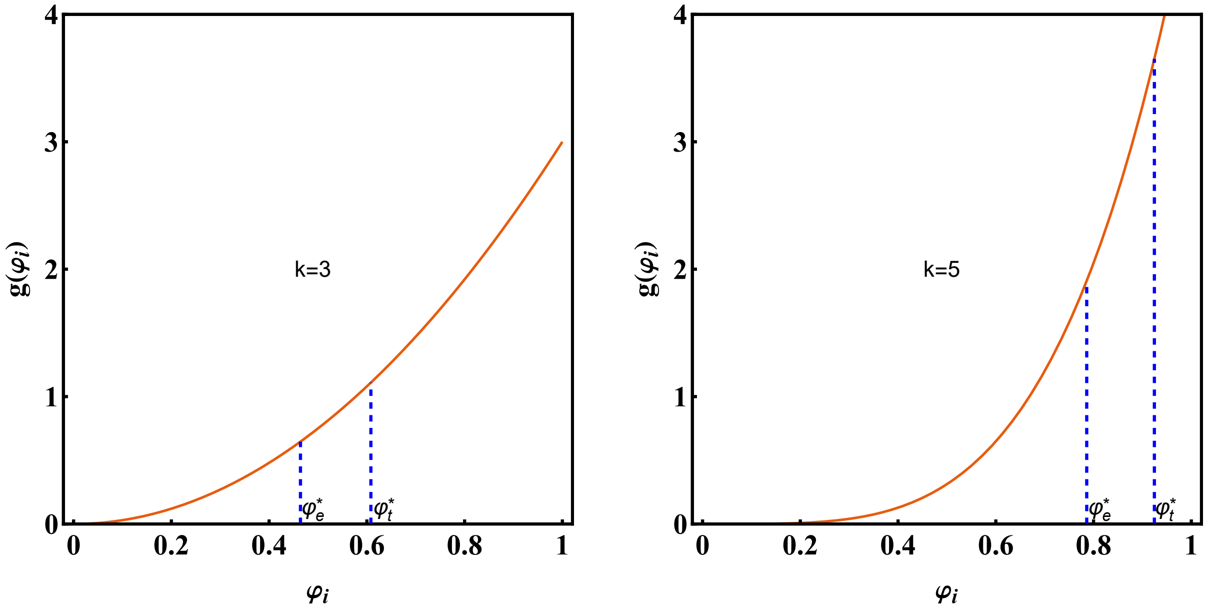
The sharp result of Proposition 10 tells us that the ETS is better if the pollution sector has a high degree of heterogeneity, while the CT policy is better otherwise. As shown in (3.13), the overall welfare gap depends on the shape of the marginal cost distribution, as illustrated by Figure 3.2. When the degree of heterogeneity is sufficiently large,[23] the gap in the productivity level is enlarged, which results in the superiority of the ETS; otherwise, the CT performs better due to a larger mass of active firms. Our result is consistent with Shinkuma and Sugeta (2016), who find that the ETS is more likely to be superior to the CT when the variance of uncertainty increases.

Moreover, we find that when a policy performs better, it may be superior in the price index gap but inferior in the redistribution gap. This indicates that apart from the distinctions in market outcomes, the two policies also lead to different labor allocations between sectors. This is consistent with Behrens et al. (2020), who find that one of the inefficiencies in imperfectly competitive markets comes from the mis-allocation of labor between sectors. We will further examine whether the labor allocation reaches the optimum under the more efficient policy.

3.5 Optimal allocation

Since either of these two policies can be better, the market distortions are not completely removed even in their best equilibria. To understand where the distortions come from, we consider the optimal allocation in this section. We use a subscript "o" in notations

[23] The gap here is a relative value. Note that φ_i denotes the marginal input level. The ratio of productivity levels of ETS to CT is written as φ_t^*/φ_e^*. Figure 3.2 shows that $\varphi_t^*/\varphi_e^*|_{k=3} > \varphi_t^*/\varphi_e^*|_{k=5}$, indicating a larger productivity gap between two policies when heterogeneity increases.



(a) High degree of heterogeneity

(b) Low degree of heterogeneity

Figure 3.2 The cost distribution and cutoffs with different degrees of heterogeneity

to indicate this optimal case. The social planner maximizes the following representative utility with a labor resource constraint:

$$\max_{e_o(\varphi_i),l_o(\varphi_i),M_o,\varphi_o^*} W_o = \frac{\alpha\sigma}{\sigma-1}\left\{\ln\int_0^{\varphi_o^*}\left[\frac{e_o(\varphi_i)^\beta l_o(\varphi_i)^{1-\beta}}{\beta^\beta(1-\beta)^{1-\beta}\varphi_i L}\right]^{\frac{\sigma-1}{\sigma}} M_o dG(\varphi_i)\right\}$$

$$+1-\frac{M_o}{L}\left[\int_0^{\varphi_o^*}\left(l_o(\varphi_i)+F\right)dG(\varphi_i)+f_e\right].$$

$$\text{s.t.}\quad M_o\left[\int_0^{\varphi_o^*} e_o(\varphi_i)dG(\varphi_i)\right]=\bar{E}.$$

The above optimal problem is solved in the Appendix J, giving the following solutions:

$$e_o(\varphi_i)=\varphi_i^{1-\sigma}\frac{\bar{E}F(\sigma-1)}{\alpha L}\left[\frac{f_e\bar{\varphi}^k(k+1-\sigma)}{F(\sigma-1)}\right]^{\frac{\sigma-1}{k}},$$

$$l_o(\varphi_i)=\varphi_i^{1-\sigma}F(1-\beta)(\sigma-1)\left[\frac{f_e\bar{\varphi}^k(k+1-\sigma)}{F(\sigma-1)}\right]^{\frac{\sigma-1}{k}},$$

$$q_o(\varphi_i)=\varphi_i^{-\sigma}\frac{F\bar{E}^\beta(\sigma-1)}{(\alpha\beta)^\beta L^\beta}\left[\frac{f_e\bar{\varphi}^k(k+1-\sigma)}{F(\sigma-1)}\right]^{\frac{\sigma-1}{k}},$$

$$M_o=\frac{\alpha L}{kf_e},\quad \varphi_o^*=\bar{\varphi}\left[\frac{f_e(k+1-\sigma)}{F(\sigma-1)}\right]^{\frac{1}{k}},\quad N_o=\frac{\alpha L(k+1-\sigma)}{kF(\sigma-1)}.$$

We first compare the market outcomes under the two policies to the optimum:

$$N_o:N_t:N_e=1:1:\frac{\sigma-1}{\sigma}\ (<1), \tag{3.14}$$

$$\varphi_o^* : \varphi_t^* : \varphi_e^* = 1 : \Big(\frac{\sigma}{\sigma - 1}\Big)^{\frac{1}{k}} \ (> 1) : \Big(\frac{\sigma - 1}{\sigma}\Big)^{\frac{1}{k}} \ (< 1), \tag{3.15}$$

$$M_o : M_t : M_e = 1 : \frac{\sigma - 1}{\sigma} \ (< 1) : 1. \tag{3.16}$$

Proposition 11. *Compared to the optimal allocation, (i) the economy under the CT has fewer entrants and a weaker selection, while the mass of active firms reaches the optimum; (ii) the economy under the ETS has an optimal mass of entrants but a stronger selection and a smaller mass of firms.*

Proof. See (3.14), (3.15), and (3.16). □

In the CT case, although the mass of active firms is identical to that in the optimal allocation, the market has too few entrants, and the government has to allow more low-productivity firms to produce. Thus, the average productivity level under the CT is lower than the optimum, which becomes a main source of distortion. In contrast, in the ETS case, although the mass of entrants is identical to that in the optimal allocation, the government cannot force more firms to produce. Even if the average productivity level under the ETS is higher than the optimum, a smaller mass of active firms and the too-large sunk costs generate market distortions.

Then we compare the equilibrium input, output, and welfare level with the optimal allocation. The calculations are as follows

$$\frac{e_o(\varphi_i)}{l_o(\varphi_i)} : \frac{e_t(\varphi_i)}{l_t(\varphi_i)} : \frac{e_e(\varphi_i)}{l_e(\varphi_i)} = 1 : \frac{\sigma}{\sigma - 1} : \frac{\sigma}{\sigma - 1}, \tag{3.17}$$

$$q_o(\varphi_i) : q_t(\varphi_i) : q_e(\varphi_i) = 1 : \left(\frac{\sigma - 1}{\sigma}\right)^{1 - \frac{\sigma - 1}{k} - \beta} : \left(\frac{\sigma - 1}{\sigma}\right)^{\frac{\sigma - 1}{k} - \beta}, \tag{3.18}$$

$$\begin{aligned} W_o - W_t =& \alpha(1 + \frac{1}{k} - \beta) \ln \frac{\sigma}{\sigma - 1} - \frac{\alpha\,(1 + k - k\beta)}{k\sigma} \\ =& \frac{\alpha}{k\sigma}\Big[(1 + k - k\beta)\Big(\sigma \ln \frac{\sigma}{\sigma - 1} - 1\Big)\Big] > 0, \end{aligned} \tag{3.19}$$

$$\begin{aligned} W_o - W_e =& \alpha\Big(\frac{1}{\sigma - 1} - \frac{1}{k} + 1 - \beta\Big) \ln \frac{\sigma}{\sigma - 1} \\ & - \alpha\Big[\frac{(1 - \beta)(k\sigma - \sigma + 1) + \beta(k + 1 - \sigma)}{k\sigma(\sigma - 1)}\Big] \\ =& \frac{\alpha[(1 - \beta)(k\sigma - \sigma + 1) + \beta(k + 1 - \sigma)]}{k\sigma(\sigma - 1)}\Big(\sigma \ln \frac{\sigma}{\sigma - 1} - 1\Big) \\ & >0. \end{aligned} \tag{3.20}$$

In summary, we have following proposition:

Proposition 12. *Compared to the optimal allocation, (i) both CT and ETS policies result in more emissions per unit of production; (ii) the two policies allocate less labor in the manufacturing sector, leading to a lower welfare level.*

Proof. (i) See (3.17), (3.19), and (3.20). (ii) We use C_j^A and L_j^A $(j = t, e, o)$ to denote the individual demand for the composite good and the labor allocated to the composite good sector, respectively. They are given by

$$\begin{aligned}
L_t^A &= LC_t^A = (1-\alpha)L + T = L - \alpha L\Big(1 - \beta + \frac{1}{\sigma - 1} - \frac{1-\beta}{\sigma} - \frac{1}{k\sigma}\Big), \\
L_e^A &= LC_e^A = (1-\alpha)L + \Pi = L - \alpha L\Big(1 - \beta + \frac{\beta}{\sigma} + \frac{1}{k\sigma}\Big), \\
L_o^A &= L - M_o\Big[\int_0^{\varphi_o^*}\Big(l_o(\varphi_i) + F\Big)dG(\varphi_i) + f_e\Big] = L - \alpha L\Big(\frac{\sigma}{\sigma - 1} - \beta\Big).
\end{aligned} \tag{3.21}$$

It is straightforward to verify that

$$L_o^A < L_t^A, \quad L_o^A < L_e^A.$$

□

This result shows that labor mis-allocation does occur in the market equilibrium even it is regulated by policies, indicating that both policies allocate too few labor resources in the manufacturing sector. Meanwhile, the bias varies with the degree of productivity heterogeneity.

After taking emissions into account, we find an improper proportion of input factors in the manufacturing production compared to the optimal allocation. The intuition is straightforward. We take the emission regulation as a type of resource that is immobile across sectors. Note that labor is mobile across sectors. As there are not enough labor resources in the polluting sector, firms have to input more emissions into unit production and become more emission intensive. Assuming one production factor, Nocco et al. (2014), Dhingra and Morrow (2019), and Behrens et al. (2020) disclose the sources of market distortion in three parts: the proper selection of active firms, the proper output of each firm, and the mis-allocation of labor resources between sectors. In our model, (3.21) gives a close relationship between emission intensity β and the labor allocation across sectors, showing that resource-allocation parameter β has an impact on the bias of equilibrium output when multiple factors are input. Equation (3.18) also shows that firms might be either over- or under-producing, which is highly dependent on the value of emission input intensity β. In the case of $\beta = 0$, labor becomes the only factor in production, in which all the manufacturing firms under-produce in market equilibria according to (3.18). Thus, our result is consistent with Behrens et al. (2020).

Chapter 4

Agglomeration, sorting, and fertility

4.1 Introduction

Over the past several decades, the world has experienced long-term downtrends in terms of fertility. The median level of the total fertility rate of countries and areas in the world fell by more than half, from 5.2 to 2.4 children per woman between 1950 and 2010. Meanwhile, it is essential to note that the level and pace of the change in fertility vary markedly across spaces. All the countries, where fertility remains above four children per woman are from Asia, Africa, and Oceania. By contrast, very few countries from North America and Europe keep fertility above 2.1. In particular, the fertility rates in East Asia and Southern Europe present a remarkably low level, merely 1.4. Moreover, when we investigate the spatial variation of fertility across regions within a country, the uneven spatial pattern still proceeds. Japanese fertility has entered an era of unprecedented low level in the middle of the 1970s, but the spatial disparity in terms of total fertility rate (TFR) existed as well. Specifically, the fertility rate of Tokyo in 2010, reaches an extremely low level of 1.12, whereas Okinawa and Miyazaki were above 1.60. Such stylized facts remind us that whether regional TFR is featured with spatial variations.

Empirical studies show that human fertility behavior is strongly related to population density. Kondo (2017) empirically clarifies the negative impacts of agglomeration on married couples' decisions to bear children at different life stages in Japan. However, the blocking effect presents an abating trend as married couples' age increases. Employing an extensive data set including individual and household surveys carried out in 44 developing countries, De la Croix and Gobbi (2017) provide empirical evidence over the negative impact of population density on fertility in developing countries. The spatial configuration of human activities seems to have impact on the decline of the number of children. Even if some empirical studies attempt to shed light on the importance of the consequences of spatial agglomeration on fertility, the mechanisms over the impacts of agglomeration on fertility have not been fully uncovered in the empirical literature. More importantly, the sorting effects also work as a potential cause. The lower fertility rate in larger regions may stem from the attraction of people with a lower desire for children. In order to take these different mechanisms into account, we need theoretical studies to involve in addressing the driving forces of the declining fertility rate.

This research develops a spatial economics model with endogenous fertility and economic agglomeration. In the model, there are two different mechanisms to explain the spatial disparity in terms of fertility rate. The first is the agglomeration effect. Economic activities concentrating on limited space results in low fertility choice. The second channel is the sorting effect. People with low fertility intention may ex ante choose to live in larger cities. Integrating these two explanations of fertility variations into an analytical framework is useful to us to comprehend the occurrence of declining TFR and adopt coping strategies.

We investigate a spatial economy of two asymmetric regions. We have in mind that the core region accommodates a larger share of the population such as Tokyo in Japan, or New York in the United States. The two regions have identical preferences, technology, and trade costs other than regional size in terms of immobile workers. The unskilled workers are spatially immobile and may work either in modern manufacturing sectors or in traditional firms. Both manufacturing and traditional goods are differentiated. The skilled workers are freely mobile across regions according to their personal preferences. The production in manufactures shows increasing returns to scale, whereas the traditional goods are produced at constant returns employing unskilled workers.

With our setup, we obtain several sets of main results. First, the skilled workers migrate to the larger region due to demand advantage stemming from market size when the trade costs fall from high to intermediate values. The spatial distribution of skilled workers is biased toward the advantage region in this stage. However, the further reduction of trade costs brings relocation (redispersion) of skilled workers from the core region to the periphery. The results indicate an inverted U-pattern of distribution over economic activities. When childbearing costs are taken into account in the budget constraint, redispersion occurs as the findings in existing studies by considering urban costs (Tabuchi 1998; Ottaviano et al. 2002; Takatsuka and Zeng 2009), housing sector (Suedekum 2006), heterogeneous preference (Tabuchi and Thisse 2002), the immobility of workers with input-output linkages (Venables 1996; Puga 1999), the positive transport costs for traditional goods (Fujita et al. 1999; Picard and Zeng 2005), multiple industries with differentiated labor intensiveness (Zeng 2006), the use of land for production (Pflüger and Tabuchi 2010), and comparative advantage (Picard and Zeng 2010).

Second, we find that the fertility rate in the larger region presents an inverse pattern against the share of skilled workers. We observe that the average fertility rate in the larger region shows U-shape curves with respect to trade freeness. It indicates that the fertility rate in the larger region suffers an initial declining trend and subsequent rise when the two regions are deeply integrated. In contrast, the TFR in the smaller region goes downwards after the initial rising pattern.

This research makes a couple of noteworthy contributions to the literature. Firstly, the impacts of economic agglomeration on TFR in our framework is endogenously initiated by economic integration rather than the exogenous premise as in related studies (Sato and Yamamoto, 2005; Sato, 2007). The fundamental mechanisms explaining agglomeration rely on the interplay among increasing returns at a firm level, mobility costs, and supply and demand linkage. Moreover, when the trade costs fall from large to small values, we can observe that the skilled workers agglomerate or relocate in a smooth way rather than catastrophic changes in traditional models (Krugman, 1991; Ottaviano et al. 2002).

Secondly, our model presents a possible mechanism that explains the variation of fertility across space. The concentration of people in the larger region leads to an increase in an urban scale. Consequently, the childbearing costs in the larger region accordingly go up due to congestion costs. Moreover, agglomeration indicated by the share of skilled workers in our model also corresponds to an increase in manufacturing varieties, which is emphasized in Murayama and Yamamoto (2010). However, the manufacturing variety is exogenously given in Murayama and Yamamoto (2010) and only partial equilibrium is considered. In other words, agglomeration economies and variety expansion to some extent, are both involved in our general equilibrium model. Therefore, the expenditure would be largely occupied by the expansion of endogenous varieties in the budget constraint, which reflects the other explanation for the mechanism behind the recent fertility decrease.

Third, it yields a better theoretical understanding of how agglomeration and sorting interact in determining the fertility rate. Sorting effects imply that people with low reproductive intention prefer to locate in larger regions. Regions accommodating more people with low reproductive intentions also ends up with a lower average fertility rate. Such sorting behavior works as one force in reshaping the lower fertility in denser regions. In turn, this attraction of more individuals and makes the regions become larger, thereby strengthening agglomeration effects.

The remainder of this chapter is organized as follows. In section 4.2, we present evidence on variation in reproductive behavior across space and over time that motivates our theory. The basic model is established in Section 4.3. Section 4.4 examines the possible patterns that how firm location and regional fertility rate evolve with regional integration. Section 4.5 further investigates the impact of market size and human capital stock on the TFR.

4.2 Motivated evidence

In this section, we lay out some novel motivating facts about the occurrence of TFR downtrend in denser areas and the existence of sorting effects. First, we find strong evidence for the association of population agglomeration with lower fertility. That is, TFR in the agglomerated area will net off other characteristics, and shows a lower rate. Moreover, we document the evidence that the low fertility rate in larger cities is in part a reflection of the sorting of the individual with low fertility intention across an urban hierarchy.

4.2.1 Decline in fertility: agglomeration matters

The empirical data we adopt here come from both Japanese prefectures and cities. The employment of stratified spatial units can help us in investigating the mechanism over the impacts of market size on fertility behavior. Specifically, the period for data on prefectures spans from 1975 to 2010, while data at city-level covers the period from 1990 to 2010. The data is provided by the Japanese government and contains the full regional accounts for a balanced panel of 47 prefectures and 500 cities. We also observe the total fertility rate and population, as well as other regional characteristics.

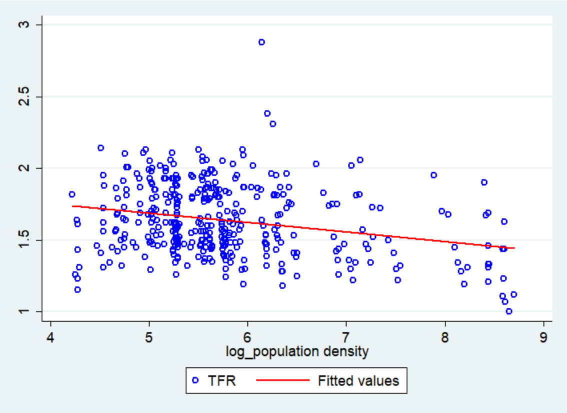


Figure 4.1 Total fertility rate and population density

Figure 4.1 plots the total fertility rate against log population density from 1975 to 2010 for 47 Japanese prefectures. As can be seen from Figure 4.1, locations with higher population densities have lower TFR.

Based on the stylized fact, we also present estimates with Japanese city-level data. Before proceeding to the empirical results, we would like to make a note of the data set. During the span of our investigation with city-level data, large numbers of Japanese peripheral cities experience population loss due to the domestic migration of younger workers to major metropolitan areas. The population densities in these cities are undergoing a declining trend over the studying years. In contrast, only the several core cities, especially Tokyo keep a stable population growth. Meanwhile, we find that the total fertility rate in city-level also present a moderating trend. It indicates that the total fertility rate and population density in most cities are undergoing a common trend over years.

Table 4.1 Agglomeration and fertility rate

	(1)	(2)	(3)	(4)	(5)
	OLS	OLS	OLS	OLS	OLS
Year	1990	1995	2000	2005	2010
$log(empdensity)_{t-5}$	-0.077***	-0.093***	-0.031***	-0.084***	-0.072***
	(0. 0102)	(0. 0101)	(0.0127)	(0.0096)	(0.0108)
$log(area)$	-0.043***	-0.027***	-0.003	0.010	0.0211***
	(0.0077)	(0.0081)	(0.0081)	(0.0072)	(0.0083)
$Infantmortality$	0.001	0.003	0.000	-0.007	-0.0107*
	(0.0025)	(0.0021)	(0.0062)	(0.0048)	(0.0919)
$log\,(income)$	-0.443***	-0.737***	-0.734***	0.753***	0.674***
	(0.0542)	(0.0641)	(0.083)	(0.0686)	(0.0919)
$Marriage\,rate$	-0.000	0.059***	0.027***	0.112***	0.138***
	(0.0015)	(0.0023)	(0.0091)	(0.0082)	(0.0101)
Prefecture	Yes	Yes	Yes	Yes	Yes
Observations	498	502	579	578	566
R-square	0.4536	0.5498	0.3016	0.5096	0.4702

Notes: Superscripts “***” “**” and “*” indicate significant at the 1%, 5% and 10% levels.

Unlike adopting two-way fixed effects cross-prefecture analysis, all specifications are in time-differenced form from 1990 to 2010 with 5 years intervals by focusing on cross-section comparison. As we know, the lower total fertility rate in a city also leads to population implosion which would result in a decline in population density. It accordingly brings about risks in terms of reverse causality. To address these issues, we adopt the lagged value of population density with 5 years in a city to avoid the danger of reverse causality. Moreover, employing a total population with lagged values to measure the market size instead of immobile unskilled workers also keeps pace with the theoretical prediction. The previous total population to a large extent can be regarded as the current permanent inhabitants. The results are shown in Table 4.1.

4.2.2 The role of sorting effects

The previous section studies how much agglomeration matters for spatial variation of fertility rate. We now move to the dual question of whether the low fertility rate may be attributed to regional composition in terms of child number, in addition to local agglomeration externalities. As argued by Behrens and Robert-Nicoud (2015), cross-city differences in size and urban costs may be the most obvious ones, and cities also differ significantly in their composition. Typically, urban economies vary in their industrial structures and urban function from the perspective of industries. Cities also differ in the demographic composition, the set of workers, and the skills they attract. Accordingly, people would make heterogeneous location choices due to different preferences over reproductive behavior. The involvement of migration might alter the average fertility rate in a city. The movers have a different set of personal traits that are associated with lower fertility regardless of whether they migrate. The sorting effects then indicate that heterogeneous workers choose a different location. When the migrants with low fertility intention in a city occupy a certain level, the sorting effects and agglomeration interact to shape the fertility rate of a city.

To debunk the role of sorting effects in lowering regional fertility rates, we adopt the micro-level data to reveal whether the sorting of migrants with different fertility intentions to denser regions has deviated from the results of the previous section. If the sorting effect is prevailing in large cities, strong sorting would thus lead to a right truncation of the distribution of the number of children of residents. We then employ these predictions to address the existence of the sorting effect, which is also considered the potential driver of the fertility difference across cities. Our data are compiled from the Japanese General Social Survey (JGSS), which covers the details of a person's children number and location. Regarding the location, cities are clarified into a three-tiered category of municipalities, large cities (23 wards of Tokyo and ordinance-designated city), other cities and towns or villages. To keep the consistency of municipal classification, this chapter employs the cumulative data set covering the years 2005, 2008 and 2012.

In line with JGSS, we would know that the mean of children number in these three-tiered categories of municipalities are 1.46, 1.65 and 1.85 respectively. To explain this spatial disparity, we have claimed that the impacts of population agglomeration on lowering fertility works as argued above. Differences in fertility across regions would stem from the direct spatial difference in the composition of reproductive preference as well. If the sorting effect works, people would choose their residential location according to reproductive preference. In a general sense, people with lower reproductive intention may sort into denser cities because of stronger preferences for the amenity or wage gains. Those who intend to give more birth would choose small cities to avoid the high costs of child-rearing. If sorting is tough in the larger cities, the distribution of children number is right-truncation relative to smaller cities.

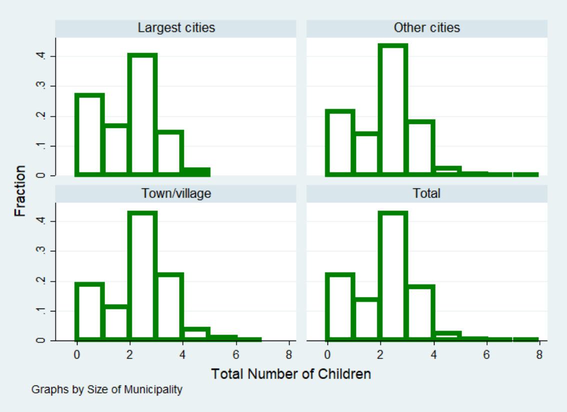


Figure 4.2 Total number of children distribution across space

Figure 4.2 presents the distribution of children number of families in a city across different categories in the JGSS data set. As we can see from Figure 4.2, the distribution of large cities is right-truncation relative to the other two types of cities. Few people with high childbearing intentions would migrate to the large cities. Meanwhile, we find that couples having one child account for a higher fraction in larger cities than the other two groups. Consequently, the accommodation of many residents with lower reproductive intentions leads to a decline in the average level of fertility rate in denser regions. The sorting effect plays a complementary role in the observation of spatial fertility variation. This occurrence offers intuitive evidence that the sorting effect exists.

4.3 Theoretical model

4.3.1 Basic setting

The economy space consists of two regions (1 and 2) and two sectors (manufacturing and composite good). We use an asterisk ($*$) to denote a variable in region 2. The labor force consists of two groups of workers. There are H skilled and perfectly mobile workers. Regions 1 and 2 host L_A and L_A^* immobile unskilled workers, respectively. Let $L = L_A + L_A^*$. Furthermore, we let $l_A = L_A/H$, $l_A^* = L_A^*/H$. The two regions have the same natural endowments, except for an asymmetric population in terms of immobile workers (market size). Without loss of generality, we assume $L_A \geq L_A^*$ hereinafter. Hence, $l_A \geq l_A^*$ as well.

As in Pflüger (2004), preferences are identical across individuals and are described by the following quasi-linear utility function. Specifically, individuals gain utility from A, M, and their number of children, n. Their preferences are characterized by

$$U = \alpha \ln M + \gamma_1 \ln n + A, \tag{4.1}$$

where

$$M = \left(\int_0^N x_i^{\frac{\sigma-1}{\sigma}} di + \int_0^{N^*} x_j^{\frac{\sigma-1}{\sigma}} dj \right)^{\frac{\sigma}{\sigma-1}}$$

is the composite of all varieties produced in the manufacturing sector M. The notation $\sigma > 1$ represents the elasticity of substitution between two manufactured varieties, N (N^*) is the number of varieties produced in region 1 (2), and x_i (x_j) is the demand for a typical traded good i (j) produced in region 1 (2). Individuals are identical except for their heterogeneity in the preference for children. We use a parameter $\gamma_1 \in (0, B)$ to denote the degree of a specific individual's preference for children. Each individual initially knows his/her preference and chooses a location to minimize the raising costs of the children. We assume that γ_1 is a random variable with a uniform distribution in $[0, B]$. When two regions have different child-raising costs, there exists a threshold value $\overline{\gamma}$. People having a γ_1 over the threshold will migrate to region 2, while residents having a γ below the threshold will relocate to region 1. This occurrence of the sorting process is supported by the before mentioned motivating facts.

We follow the literature and assume that the raising costs are highly related to the city size (e.g., Sato and Yamamoto, 2005, Sato, 2007), which is represented by the scale of mobile skilled workers. As in Pflüger (2004), each variety requires a skilled worker as its fixed input. Therefore, the number of skilled workers corresponds to the number of local varieties N_i in this region. The budget constrain of the consumer in region 1 is given by

$$w - bN^c \gamma_2^{1-c} n = PM + A, \tag{4.2}$$

where

$$P = \left[N p_i^{1-\sigma} + N^* (\tau p_j^*)^{1-\sigma} \right]^{\frac{1}{1-\sigma}}, \ \tau > 1 \tag{4.3}$$

is the CES-price index and p_i (p_j^*) denotes the mill price of variety i (j) in region 1 (2). The second term in the LHS of (4.2) represents both the raising child costs related to the number of varieties (Maruyama and Yamamoto, 2010) and the costs related to region congestion (Sato and Yamamoto, 2005). γ_2 denotes an individual's preference for the quality of children. In this sense, there exist two cases that individuals' preference between quantity and quality may be either positive or negative. Intuitively, the positive relationship usually appears in developed regions since affluent individuals won't loose their requirement for children's quality even having more children. In contrast, the negative relationship is more common in developing regions as poor individuals may sometimes pursue the quantity of children blindly. Relevant literature shows that both cases are supported by the real world. Since we use the data of Japan to do the empirical check, we mainly focus on the case that γ_1 and γ_2 are positively related. For simplicity, we assume $\gamma = \gamma_1 = \gamma_2$ in the positive case. What's more, we assume raising children consumes good A.

Standard utility maximization yields the corresponding demand functions:

$$M = \alpha P^{-1}, \quad A = w - \alpha - \gamma, \quad n = \frac{\gamma^c}{bN^c}, \quad x_i = \alpha \frac{p_i^{-\sigma}}{P^{1-\sigma}}, \quad x_j = \alpha \frac{(\tau p_j^*)^{-\sigma}}{P^{1-\sigma}}. \tag{4.4}$$

Turning to the supply side, the traditional sector supplies the homogeneous good under perfect competition employing one unit of L as the only input of a constant-returns technology. Perfect competition in the traditional sector ensures that the equilibrium price equals marginal cost. With free trade in the composite good, the choice of this good as the numéraire implies that in equilibrium, the wage of unskilled labor is equal to one in both regions. In the manufacturing sector, monopolistically competitive firms offer horizontally differentiated goods using both types of workers under increasing return to scale. Specifically, we assume the same technology in both regions. One unit of skilled worker as the fixed requirement and $(\sigma - 1)/\sigma$ units of unskilled workers as the marginal cost are required in production.

In line with the tradition in spatial economics, each variety is assumed to be produced at only one location. Manufactured goods can be traded across space and incur trade costs. We assume the iceberg form of transportation costs between regions: $\tau \geq 1$ units of a manufactured good must be shipped for one unit to reach the other region. In contrast, all shipments of traditional goods are assumed to be costless.

Under foregoing settings, a firm located in region 1 maximizes profits, given by

$$\pi_i = p_i x_i \left(L_A + L_M \right) + p_i^* x_i^* \left(L_A^* + L_M^* \right)$$

$$-\frac{\sigma-1}{\sigma}x(L_A+L_M)-\frac{\sigma-1}{\sigma}\tau x_i^*(L_A^*+\tau L_M^*)-w \tag{4.5}$$

where w stands for the wage prevailing in region 1.

In the classical CES framework, the markup is a constant $\sigma/(\sigma-1)$. Thus, the optimal price of the firms is

$$p_i = p_j^* = 1. \tag{4.6}$$

Having claiming the basic structure, we now solve the model in two steps by short-run and long-run outcomes.

4.3.2 Short-run equilibrium

Following the established traditions of spatial economics, we assume that markets for goods adjust instantaneously, while inter-regional migration of skilled workers is relatively slow. In the short-run equilibrium wages and prices are adjusted to clear the good markets while the share of skilled workers is fixed. Namely, skilled workers are immobile between regions in short-run equilibrium. The skilled workers with heterogeneous preferences are randomly allocated in two regions. Let λ be the share of skilled workers in region 1. Without loss of generality, we assume the skilled worker share is larger in region 1 in the short-run ($\lambda > 1/2$). The numbers of skilled workers in two regions are $L_M = N = \lambda H$ and $L_M^* = N^* = (1-\lambda)H$, respectively.

The zero profit conditions in region 1 and region 2 are given by:

$$\sigma w = \frac{\alpha\left(L_A+L_M\right)}{L_M+\phi L_M^*}+\frac{\phi\alpha\left(L_A^*+L_M^*\right)}{\phi L_M+L_M^*},$$
$$\sigma w^* = \frac{\phi\alpha\left(L_A+L_M\right)}{L_M+\phi L_M^*}+\frac{\alpha\left(L_A^*+L_M^*\right)}{\phi L_M+L_M^*},$$

where $0 \leq \phi = \tau^{1-\sigma} \leq 1$.

For expositional purposes, the above equations are rewritten as

$$\sigma w = \frac{\alpha\left(l_A+\lambda\right)}{\lambda+\phi(1-\lambda)}+\frac{\phi\alpha\left(l_A^*+1-\lambda\right)}{\phi\lambda+1-\lambda},$$
$$\sigma w^* = \frac{\phi\alpha\left(l_A+\lambda\right)}{\lambda+\phi(1-\lambda)}+\frac{\alpha\left(l_A^*+1-\lambda\right)}{\phi\lambda+1-\lambda}.$$

To ensure both sectors A and M are active after trade in two countries, we impose the assumptions of $Nw(\sigma-1) < L_A$ and $N^*w(\sigma-1) < L_A^*$, which lead to a condition $\alpha < \frac{l_A\sigma}{(l_A+l_A^*+1)(\sigma-1)}$.[24]

[24] This assumption is proposed by Pflüger (2004) on p. 568.

4.3.3 Long-run equilibrium

Turn to the long-run equilibrium, skilled workers are freely mobile and choose to live in either region. As argued in basic settings, skilled workers are heterogeneous in reproductive choice. They match their perception of the attributes of a region with their own taste and then decide where to reside. The equilibrium distribution of skilled workers is obtained from the comparison of the utility levels that skilled workers can achieve in both regions.

The indirect utility difference between two regions for individual with γ is written as

$$\begin{aligned}\Delta V(\gamma) \equiv & V(\gamma) - V^*(\gamma) = \alpha \ln \frac{P^*}{P} + (w - w^*) - \gamma c \ln \left(\frac{N}{N^*}\right) \\ = & \frac{\alpha}{1-\sigma} \ln \left[\frac{\lambda\phi + 1 - \lambda}{\lambda + \phi(1-\lambda)}\right] + \frac{\alpha(1-\phi)}{\sigma}\left[\frac{l_A + \lambda}{\lambda + (1-\lambda)\phi} - \frac{l_A^* + 1 - \lambda}{\lambda\phi + 1 - \lambda}\right] \\ & - \gamma c \ln \frac{\lambda}{1-\lambda}. \end{aligned} \tag{4.7}$$

Equation (4.7)indicates that utility difference is determined by three different components. In contrast to Pflüger (2004), the third term acts as additional centrifugal force stemming from child-rearing costs. Since child-rearing costs are related to regional size, agglomeration of skilled workers will lead to the cost rise in raising a child. This burden would render people to relocate from the larger region to the smaller one. Moreover, γ in the third term of (4.7) representing the heterogeneous preference of fertility intention, can be considered as a dispersion force. A skilled worker who prefers more children will move to the smaller region for any intermediate level of the transport costs. The heterogeneity of fertility preference alleviates the HME. Moreover, the heterogeneous preference of fertility choice for mobile workers, allows us to further investigate the composition of regional fertility rate. Accordingly, the average fertility rate in a region would be reshaped when spatial sorting works.

Define $\bar{\gamma}$ as the cut-off value of γ without utility difference ($\Delta V(\bar{\gamma}) = 0$). The utility difference finally sorts the individuals with a higher γ ($\gamma > \bar{\gamma}$) to region 2 and the individuals with a lower γ ($\gamma < \bar{\gamma}$) to region 1.[25] The process of migration will not cease until the critical value $\overline{\gamma}$ equals to the share of skilled labor in region 1, λ. Accordingly, the long-run equilibrium is finally determined by 2 equations with 2 endogenous variables ($\lambda,\ \bar{\gamma}$):

$$\overline{\gamma} = \frac{\dfrac{\alpha}{1-\sigma} \ln \left[\dfrac{\lambda\phi + 1 - \lambda}{\lambda + \phi\,(1-\lambda)}\right] + \dfrac{\alpha(1-\phi)}{\sigma}\left[\dfrac{l_A + \lambda}{\lambda + (1-\lambda)\phi} - \dfrac{l_A^* + 1 - \lambda}{\lambda\phi + 1 - \lambda}\right]}{c \ln \dfrac{\lambda}{1-\lambda}}, \tag{4.8}$$

[25] The sorting direction is obvious from the inequalities $d\Delta V(\gamma)/d\gamma|_{\gamma>\bar{\gamma}} < 0$ and $d\Delta V(\gamma)/d\gamma|_{\gamma<\bar{\gamma}} > 0$.

$$\bar{\gamma} = B\lambda. \tag{4.9}$$

In the equilibrium, equation (4.8) reveals the relationship between the skilled labor share and the critical value of γ while transforming from short-run to long-run. Meanwhile, equation (4.9) determines the destination of migration caused by the unbalance between λ and $\bar{\gamma}$. Intuitively, λ represents the typical agglomeration effect in NEG while $\bar{\gamma}$ indicates the sorting effect brought by the costs of raising children in our model. These two effects are the crucial mechanisms in our model to jointly determine the agglomeration pattern and fertility levels. Individuals must have a choice whether converging to the larger region to enjoy the benefit of agglomeration effect or dispersing to the smaller region to escape from the high child-rearing costs.

Combining (4.8) and (4.9), we depict the equilibrium with one equation:

$$\begin{aligned}\Delta V(\lambda) =& \frac{\alpha}{1-\sigma}\ln\left[\frac{\lambda\phi+1-\lambda}{\lambda+\phi(1-\lambda)}\right]+\frac{\alpha(1-\phi)}{\sigma}\left[\frac{l_A+\lambda}{\lambda+(1-\lambda)\phi}-\frac{l_A^*+1-\lambda}{\lambda\phi+1-\lambda}\right]\\ &-Bc\lambda\ln\frac{\lambda}{1-\lambda}=0\end{aligned} \tag{4.10}$$

Let $n(\gamma)$ be the number of children of a worker of type γ. The average fertility in Region 1 can be expressed as:

$$\bar{n} = \frac{H\int_0^{\bar{\gamma}} n(\gamma)(1/B)d\gamma + L_A\int_0^B n(\gamma)(1/B)d\gamma}{(1/B)\bar{\gamma}H+L_A}.$$

Based on (4.4), (4.8) and the fact of $\lambda = \bar{\gamma}/B$, this $\overline{n}$ can be further written as

$$\overline{n} = \frac{1}{b(1+c)}\left(\frac{B}{\bar{\gamma}H}\right)^c\frac{\bar{\gamma}^{c+1}+l_AB^{c+1}}{\bar{\gamma}+l_AB}. \tag{4.11}$$

Similarly, the average fertility in region 2 can be obtained:

$$\begin{aligned}\bar{n}^* =& \frac{H\int_{\bar{\gamma}}^B n_m^*(1/B)d\gamma + L_A^*\int_0^B n_A^*(1/B)d\gamma}{[(B-\bar{\gamma})/B]H+L_A^*}\\ =& \frac{B^c[(B^{1+c}-\bar{\gamma}^{1+c})+l_A^*B^{1+c}]}{b(1+c)[(B-\bar{\gamma})H]^c[(B-\bar{\gamma})+Bl_A^*]}\end{aligned} \tag{4.12}$$

Starting from the basic setting, following parts are based on the case of $B=1$ ($\bar{\gamma}=\lambda$).

These two equations provide twofold mechanisms for the formation of regional fertility rate. On one hand, the sorting effect operates when skilled workers have a different set of personal behavioral intentions. The skilled workers with lower fertility preference choose to migrate to the larger region 1. Meanwhile, sorting effects also leave mobile workers with higher fertility intention to the smaller region 2. When the sorting mechanism works, the aggregation of workers with the lower reproductive intention in the larger region leads to a lower regional average fertility rate.

On the other hand, more agglomeration of skilled workers in the larger region 1 will bring a negative impact on regional fertility rate. Since one unit of skilled worker is employed in manufacturing production as the fixed input. The number of skilled workers corresponds to the variety amount. A larger share of skilled workers in region 1 indicates that people will face a tighter budget constraint to raise a child according to equation (4.4). The mechanism that the variety expansion reduces the fertility rate is stressed in Murayama and Yamamoto (2010). However, we offer a general equilibrium analysis with mobility of skilled workers, which is different from Murayama and Yamamoto (2010).

Additionally, the concentration of skilled workers in region 1 generates strong congestion diseconomies, which reduce the region's fertility rate. Accommodation of mobile workers in region 1 stands for expansion of urban scale as well. Congestion costs arising from soaring housing price or longer commuting would exhibit a higher living cost and lower fertility rate. Sato (2007) reveals that dis-agglomeration and migration of workers lower the fertility rate by assuming exogenous agglomeration economies. In contrast, the formation of economic agglomeration in our model is endogenously established on increasing return to scale.

4.4 Spatial configuration and regional integration

This section examines the spatial equilibrium by investigating some equations, which are used to determine the equilibrium firm shares and regional fertility rate in both regions. As in most economic geography models (Krugman, 1991; Fujita et al.,1999; Takahashi et al., 2013), we mainly focus on the impacts of economic integration on spatial configuration over the firm location and regional fertility rate.

4.4.1 Regional integration and industrial location

We first investigate how regional economic integration may affect the equilibrium distribution of firms across regions. Although we cannot derive a closed-form solution for the cut-off γ, we can study the impact of regional integration, i.e. the impact of a decrease in transport costs on the cut-off, by an implicit function.

According to (4.7), we can examine the relationship between γ and ϕ.

Lemma 1. *When $L_A \geq L_A^*$, in $[1/2, 1]$, the following equation*

$$\mathcal{F}_1(\gamma) \equiv \frac{\alpha L_A}{\gamma\sigma H} - \frac{\alpha L_A^*}{(1-\gamma)\sigma H} + \left(\frac{\alpha}{\sigma-1} - c\gamma\right) \ln \frac{\gamma}{1-\gamma} = 0 \tag{4.13}$$

has a unique solution

$$\gamma_0 \in \begin{cases} \left[\frac{1}{2}, \frac{L_A}{L_A + L_A^*}\right], & \text{if } \frac{\alpha}{c(\sigma-1)} \leq \frac{1}{2}, \\ \left[\frac{\alpha}{c(\sigma-1)}, 1\right], & \text{if } \frac{1}{2} < \frac{\alpha}{c(\sigma-1)} \leq \frac{L_A}{L_A + L_A^*}, \\ \left[\frac{1}{2}, \min\left\{\frac{\alpha}{c(\sigma-1)}, 1\right\}\right], & \text{if } \frac{\alpha}{c(\sigma-1)} > \frac{L_A}{L_A + L_A^*}. \end{cases}$$

Proof. See Appendix K. □

Depending on the parameters, equation (4.13) may have other roots in $[0, 1/2)$. Nevertheless, when we examine the relationship between γ and ϕ, we focus on the curve $\bar{\gamma}(\phi) \in [1/2, 1]$, which starts at this root when $\phi = 0$.

Mathematical results in Lemma 1 are quite useful because they provide detailed information about the spatial distribution of skilled workers, which leads to the following conclusion:

Proposition 13. *When $L_A \geq L_A^*$, the equilibrium firm share in region has a bell shape or monotonically decreases when the regional trade freeness increases.*

Proof. Let $\bar{\gamma}_0$ and $\bar{\gamma}_1$ denote the equilibrium cutoffs $\bar{\gamma}$ when $\phi = 0, 1$, respectively. Note that $\bar{\gamma}$ is determined implicitly by (4.8) and (4.10). Therefore, $\bar{\gamma}_0$ solves

$$\frac{\alpha}{H(1-\bar{\gamma})\sigma}\left(L_A \frac{1-\bar{\gamma}}{\bar{\gamma}} - L_A^*\right) + \frac{1}{\sigma - 1}[\bar{\gamma}c(\sigma-1) - \alpha] \ln \frac{1-\bar{\gamma}}{\bar{\gamma}}, \tag{4.14}$$

while $\bar{\gamma}_1$ solves

$$\bar{\gamma} \ln \frac{\bar{\gamma}}{1-\bar{\gamma}} = 0. \tag{4.15}$$

We know that $\bar{\gamma}_1 = 1/2$ from (4.15) immediately. Furthermore, the implicit function theorem gives

$$\left.\frac{d\bar{\gamma}}{d\phi}\right|_{\phi=1, \bar{\gamma}=1/2} = -\frac{(L_A - L_A^*)\alpha}{2H\sigma} < 0.$$

The implicit function theorem also gives

$$\left.\frac{d\bar{\gamma}}{d\phi}\right|_{\phi=0, \bar{\gamma}=\bar{\gamma}_0} = \frac{\alpha \mathcal{R}}{\mathcal{S}}, \tag{4.16}$$

where

$$\begin{aligned} \mathcal{R} =& (\sigma-1)[L_A^* \bar{\gamma}_0^2 - L_A(1-\bar{\gamma}_0)^2] - \bar{\gamma}_0 H(2\bar{\gamma}_0 - 1)(1-\bar{\gamma}_0), \\ \mathcal{S} =& (\sigma-1)\alpha[L_A(1-\bar{\gamma}_0)^2 + L_A^* \bar{\gamma}_0^2] + cH\bar{\gamma}_0(1-\bar{\gamma}_0)\sigma[(1-\bar{\gamma}_0)\bar{\gamma}_0(\sigma-1) \ln \frac{\bar{\gamma}_0}{1-\bar{\gamma}_0} \end{aligned}$$

$$- \alpha + \bar{\gamma}_0 c(\sigma - 1)].$$

Depending on parameters (see examples of Figure 4.3), (4.16) can be either positive or negative. Since (4.10) has at most two roots with respect to ϕ, the curve $\gamma(\phi)$ cross any horizontal lines in the $\phi - \bar{\gamma}$ plane at most twice. Thus, the cutoff $\bar{\gamma}$ decreases or shows a bell shape when the regional trade freeness increases. □

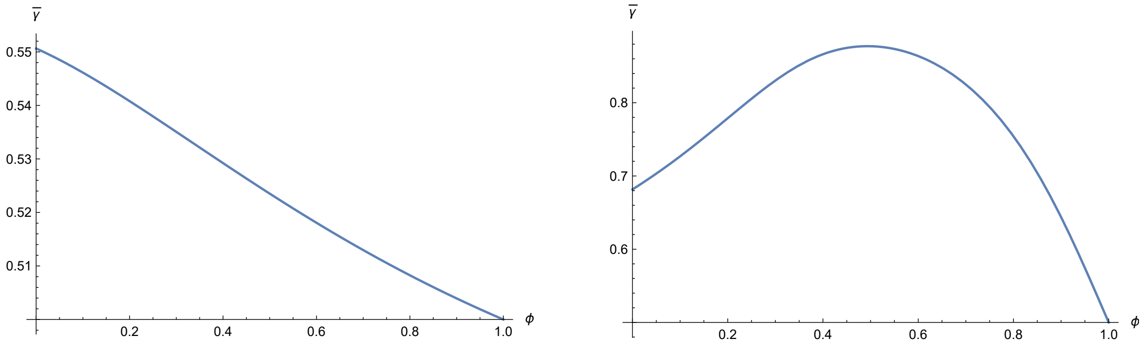


Figure 4.3 The relationship between $\bar{\gamma}$ and ϕ

Figure 4.3 provides two examples with $\alpha = 0.6,\ L_A = 20,\ L_A^* = 10,\ H = 4,\ b = 0.2$, while we choose $\sigma = 20,\ c = 1$ in the left panel and $\sigma = 5,\ c = 0.1$ in the right panel.

We now present the properties of the equilibrium distributions of firms with respect to trade freeness.The results established in this section are intuitively explained in order. Since we assume that $L_A > L_A^*$, the initial larger local demand in region 1 has attracted more skilled workers, which is evident when ϕ is small. Regional integration facilitates this advantage in region 1. Even though centripetal and centrifugal forces interact each other, the agglomeration force still dominates when ϕ is small and results in the home market effect (larger proportionate firm share) in region 1. In line with the tradition of spatial economics, firms move together with their workers. Meanwhile, the quasi-linear upper tier utility alleviates the strength of the demand linkage. We can observe that skilled workers continuously relocate to the larger region 1, when regions integrate further. The absence of catastrophic changes in the location pattern occurs. As trade freeness accelerates, the centrifugal force stemming from the child-rearing costs makes region 1 unattractive. Skilled workers with lower fertility intention tend to disperse to avoid the negative utility in region 1. The redispersion process that workers move to region 2 happens when skilled workers earn less from a better access to consumers than affording cost of child-rearing in the larger region.

It is noteworthy stressing that when $\phi = 1$, λ equals 1/2 in (4.10). In fact, when trade freeness reaches a certain point, the occupation over the share of skilled workers in region 1 starts to decrease. The initial advantage stemming from the local demand of immobile unskilled workers attenuates until the firm share is equally distributed between two regions. It indicates that the HME is muted in our settings.

Proposition 13 contributes to the literature in two respects. The "redispersion" process at small trade costs has been already highlighted in the literature (Tabuchi, 1998; Puga, 1999; Tabuchi and Thisse, 2002; Picard and Zeng 200; Takahashi et al., 2013). The above result provides a new theoretical support to explain the spatial redispersion of economic activities. In addition, the spatial distribution of firm location is determined by the interaction between the HME and the sorting effects.

4.4.2 Trade freeness and fertility

We now turn to the other critical objective of this chapter and ask whether regional integration brings impacts on regional fertility composition. We next analyze how the introduction of heterogeneous fertility preference into a spatial economics model can be employed to investigate the evolution of population structure.

When we examine equations (4.11) and (4.12), although there is no explicit ϕ in the formula of $\bar{n}$ and $\bar{n}^*$, we are still able to unveil the relationship between $\bar{n}$ (or $\bar{n}^*$) and ϕ through $\bar{\gamma}$. According to (4.11), we have

$$\frac{d\bar{n}}{d\phi} = \underbrace{\frac{\partial\bar{n}}{\partial\phi}}_{=0} + \frac{\partial\bar{n}}{\partial\bar{\gamma}}\frac{\partial\bar{\gamma}}{\partial\phi} = \underbrace{-\left(\frac{1}{H\bar{\gamma}}\right)^{1+c}\frac{Hl_A[cl_A+\bar{\gamma}(1+c-\bar{\gamma}^c)]}{b(1+c)(l_A+\bar{\gamma})^2}}_{<0}\frac{\partial\bar{\gamma}}{\partial\phi}, \tag{4.17}$$

$$\frac{d\bar{n}^*}{d\phi} = \underbrace{\frac{[(1+c)(1-\bar{\gamma})+cl_A^*](1-\bar{\gamma}^{1+c}+l_A^*)-\bar{\gamma}^c(1+c)(1-\bar{\gamma})(1-\bar{\gamma}+l_A^*)}{b(c+1)H^c(1-\bar{\gamma})^{1+c}(-\bar{\gamma}+l_A^*+1)^2}}_{>0}\frac{\partial\bar{\gamma}}{\partial\phi}. \tag{4.18}$$

According to the definition of average fertility in two regions (4.11) and (4.12), changes in transportation costs do not have a direct impact on the fertility rate, but rather through changes in the firm share. Namely, we only need to analyze the impact of agglomeration patterns on fertility rates.

In (4.17), $\partial\bar{n}/\partial\bar{\gamma} < 0$ is straightforward as $\bar{\gamma} < 1$ holds. As for $\partial\bar{n}^*/\partial\bar{\gamma}$, the denominator is obviously positive, the numerator

$$\begin{aligned}
&[(1+c)(1-\bar{\gamma})+cl_A^*](1-\bar{\gamma}^{1+c}+l_A^*)-\bar{\gamma}^c(1+c)(1-\bar{\gamma})(1-\bar{\gamma}+l_A^*)\\
>&[(1+c)(1-\bar{\gamma})+cl_A^*](1-\bar{\gamma}+l_A^*)-\bar{\gamma}^c(1+c)(1-\bar{\gamma})(1-\bar{\gamma}+l_A^*)\\
=&[(1-\bar{\gamma}^c)(1+c)(1-\bar{\gamma})+cl_A^*](1-\bar{\gamma}+l_A^*)>0
\end{aligned}$$

Therefore, $\partial\bar{n}^*/\partial\bar{\gamma} > 0$ holds.

The above information and proposition 1 lead to the following conclusion:

Proposition 14. *With the variation of trade freeness, the average fertility rate in the larger region $\bar{n}$ shows the trend opposite to the firm share, while that in the smaller region $\bar{n}^*$ shows the same trend.*

Proof. See (4.17) and (4.18). □

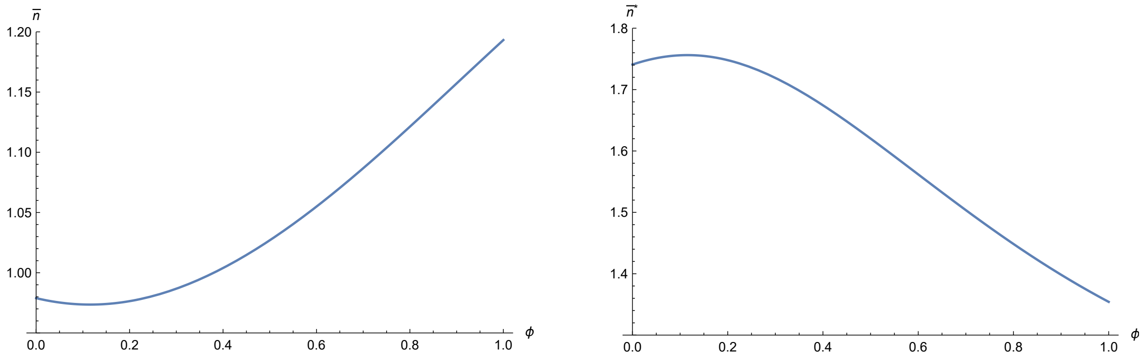


Figure 4.4 The relationship between the average fertility rate and trade freeness

Figure 4.4 provides two examples with identical parameters to the right panel of Figure 4.3.

When agglomeration enhances, more skilled workers migrate from region 2 to region 1. During this process, the sorting effect plays a different role from the agglomeration effect. The new migrants who are sorted to region 1 have higher preferences for children than the original residents. Meanwhile, the average preferences for children further increase in region 2. This indicates that the sorting effect raises the fertility rates in both regions.

In contrast, the child-rearing costs increase in region 1 and decrease in region 2 due to the agglomeration effect. Obviously, the fertility rate in region 2 shows the same trend as in the agglomeration pattern. In terms of region 1, according to (4.17), although these two effects have opposite forces, the agglomeration effect is dominant. Therefore, the fertility rate in region 1 decreases during the process of agglomeration and exhibits the opposite trend to the agglomeration pattern.

4.5 Extension

In this section, we further investigate other factors that might have impacts on the average fertility in our theoretical model.

4.5.1 Market size effects and fertility

Since the setting of our model is asymmetrical, the market size difference plays an important role. Therefore, we first investigate the effect brought by the increase of unskilled workers in region 1.

Proposition 15. *The cutoff $\bar{\gamma}$ increases with L_A.*

Proof. For a given ϕ, (4.8) specifies a relationship between $\bar{\gamma}$ and L_A. Since we focus on the stable equilibrium $\bar{\gamma}(\phi)$ in $[1/2, 1]$, we have

$$\left.\frac{d(V-V^*)}{d\gamma}\right|_{\gamma=\bar{\gamma}} < 0.$$

Therefore, by the implicit function theorem,

$$\frac{d\bar{\gamma}}{dL_A} = -\left.\frac{\frac{d(V-V^*)}{dL_A}}{\frac{d(V-V^*)}{d\gamma}}\right|_{\gamma=\bar{\gamma}} = \left.\frac{\frac{\alpha(1-\phi)}{H\sigma(\gamma+\phi-\gamma\phi)}}{-\frac{d(V-V^*)}{d\gamma}}\right|_{\gamma=\bar{\gamma}} > 0.$$

□

Next, we consider how the average fertility rate in the larger Region 1 depends on L_A and L_A^*. Unfortunately, we have no clear-cut result on this issue. For example, according to (4.11), we have

$$\begin{aligned}\frac{d\bar{n}}{dL_A} =& \underbrace{\frac{\partial\bar{n}}{\partial L_A}}_{>0} + \underbrace{\frac{\partial\bar{n}}{\partial\bar{\gamma}}}_{<0}\underbrace{\frac{\partial\bar{\gamma}}{\partial L_A}}_{>0} \\ =& \frac{H\bar{\gamma}(1-\bar{\gamma}^c)}{b(1+c)\bar{\gamma}^c H^c(\bar{\gamma}H+L_A)^2} - \left(\frac{1}{H\bar{\gamma}}\right)^{1+c}\frac{Hl_A[cl_A+\bar{\gamma}(1+c-\bar{\gamma}^c)]}{b(1+c)(l_A+\bar{\gamma})^2} \\ &\times \frac{\frac{\alpha(1-\Phi)}{H\sigma(\bar{\gamma}+\Phi-\bar{\gamma}\Phi)}}{c\ln\frac{\bar{\gamma}}{1-\bar{\gamma}} + \frac{c}{1-\bar{\gamma}} - \frac{\alpha(1-\Phi)}{\sigma-1}(\frac{1}{\bar{\gamma}\Phi+1-\bar{\gamma}} + \frac{1}{\bar{\gamma}+\Phi-\bar{\gamma}\Phi}) - \frac{\alpha(1-\Phi)}{\sigma H}(\frac{L_A\Phi+\Phi H-L_A}{(\bar{\gamma}+\Phi-\bar{\gamma}\Phi)^2} + \frac{L_A^*\Phi+\Phi H-L_A^*}{(\bar{\gamma}\Phi+1-\bar{\gamma})^2})}\end{aligned}$$

In the last expression, the first term is positive while the second term is negative. Even in the case of $\Phi = 0$, the sign of $d\bar{n}/dL_A$ is indeterminate. We give two examples in Figure 4.5. The parameters are

$$\alpha = 1, L_A^* = 1.1, H = 1, c = 1, \sigma = 20, \Phi = 0$$

in the left panel and

$$\alpha = 1, L_A^* = 1.1, H = 1, c = 1, \sigma = 6, \Phi = 0$$

in the right panel. We can see that the relationship is monotone in the right panel but not monotone in the left panel.

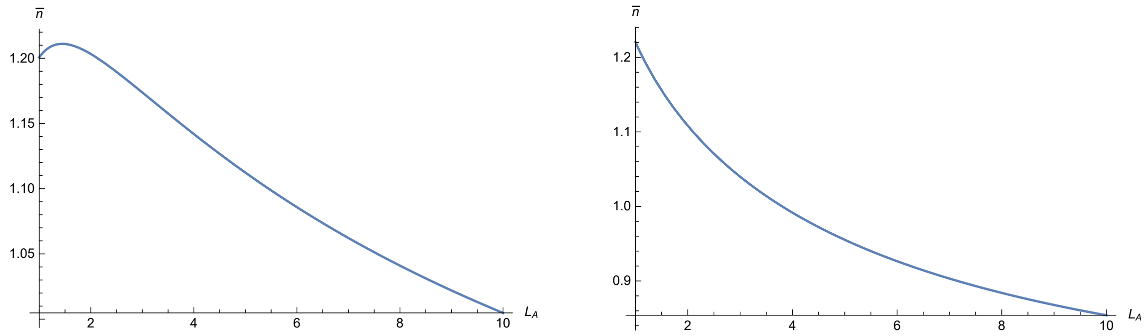


Figure 4.5 The relationship between $\bar{n}$ and L_A

4.5.2 Human capital and fertility

Next, we explore the role of total skilled worker endowment.

Result 1. *The average fertility rate $\bar{n}$ in the larger region 1 decreases with the total amount of skilled workers H while the average fertility rate $\bar{n}^*$ in the smaller region 2 increases.*

Proof.

$$\begin{aligned}\frac{d\bar{n}}{dH} &= \underbrace{\frac{\partial\bar{n}}{\partial H}}_{<0} + \underbrace{\frac{\partial\bar{n}}{\partial\bar{\gamma}}}_{<0}\underbrace{\frac{\partial\bar{\gamma}}{\partial l_A}}_{>0}\underbrace{\frac{\partial l_A}{\partial H}}_{<0} + \underbrace{\frac{\partial\bar{n}}{\partial l_A}}_{>0}\underbrace{\frac{\partial l_A}{\partial H}}_{<0} \\ &= \underbrace{\frac{(1-c)\bar{\gamma}^2 HL_A - c\bar{\gamma}^3H^2 - (c+1)\bar{\gamma}HL_A - cL_A^2}{2bH^{(1+c)}\bar{\gamma}^c(\bar{\gamma}H+L_A)^2}}_{<0} \\ &\quad - \underbrace{\frac{(2-c)\bar{\gamma}^2HL_A + (1-c)\bar{\gamma}^3H^2 - (c+1)\bar{\gamma}HL_A - cL_A^2}{2bH^c\bar{\gamma}^{(1+c)}(\bar{\gamma}H+L_A)^2}}_{unknown} \times \underbrace{\frac{L_A}{H^2}}_{>0} \times \\ &\quad \underbrace{\frac{\frac{\alpha(1-\Phi)}{\sigma(\bar{\gamma}+\Phi-\bar{\gamma}\Phi)}}{c\ln\frac{\bar{\gamma}}{1-\bar{\gamma}} + \frac{c}{1-\bar{\gamma}} - \frac{\alpha(1-\Phi)}{\sigma-1}\left(\frac{1}{\bar{\gamma}\Phi+1-\bar{\gamma}} + \frac{1}{\bar{\gamma}+\Phi-\bar{\gamma}\Phi}\right) - \frac{\alpha(1-\Phi)}{\sigma H}\left(\frac{L_A\Phi+\Phi H-L_A}{(\bar{\gamma}+\Phi-\bar{\gamma}\Phi)^2} + \frac{L_A^*\Phi+\Phi H-L_A^*}{(\bar{\gamma}\Phi+1-\bar{\gamma})^2}\right)}}_{>0} \\ &\quad - \underbrace{\frac{L_A(\bar{\gamma}-\bar{\gamma}^2)}{2b\bar{\gamma}^cH^c(\bar{\gamma}H+L_A)^2}}_{>0}\end{aligned}$$

□

Simulations show that $\bar{n}$ decreases with H. We give a numerical example in Figure 4.6, taking the parameter value as

$$\alpha = 1, L_A = 2, L_A^* = 1, \phi = 0.3, \sigma = 6, c = 1.$$

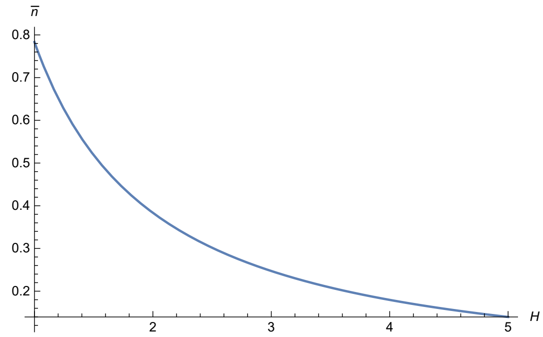


Figure 4.6 The relationship between $\bar{n}$ and H

The increase in H can be comprehended as the advancement of technology or industrial upgrading, which evidently raises the total amount of skilled workers in the economy. However, this progress may reduce the average fertility rate in the larger region and increase the average fertility rate in the smaller region. This conclusion fits the course of human history well.

Chapter 5

Concluding remarks

In Chapter 2, borrowing the binary preferences of Foellmi et al. (2018), this thesis builds a solvable VES trade model to examine the impact of heterogeneous productivity on trade structure. Several interesting findings are observed in this study.

With heterogeneous firms, the determinant of market price is not limited to the demand side, which depends exclusively on per capita income. Thus, we can observe opposite forces when trade costs change, leading to price reversal in the no-arbitrage equilibrium. The larger country may have a lower price even if it provides a higher wage rate. This phenomenon occurs when trade costs are high.

When the price gap is larger than the trade cost, the arbitrage equilibrium replaces the no-arbitrage equilibrium. Subsequently, we find that some export-only firms emerge in the smaller country, aiming to make a large profit from exports. Although such export-only firms are observed in many developing countries, Melitz (2003) excludes their existence by assuming that the fixed costs in foreign markets are higher than those in domestic markets. Qiu and Yan (2017) thus solidify the existence of export-only firms through additional assumptions, such as low export costs, a tax rebate, and a large efficiency gap. By introducing binary demand, we find that such export-only firms generally exist in developing countries.

We confirm an interesting result of Foellmi et al. (2018) regarding the gains from trade—globalization may hurt the smaller country when trade costs are low. This is because the declining trade costs make arbitrage more likely; more firms abstain from exports in the larger country, and more firms choose to only export in the smaller country. Therefore, welfare in the smaller country decreases as fewer varieties become available in the market when trade costs are low. We find that productivity heterogeneity enhances this property. When firms in both countries are more heterogeneous, gains from trade in the large country and losses from trade in the small country increase. This is because resource allocation is more efficient within a country, and the large country absorbs more resources across countries in the heterogeneous case.

Subsequently, we analyze the firm selection with binary demand and find that the selection of domestic supply is more severe in the smaller country when the trade costs are low, and it reverses when the trade costs are sufficiently large.

To investigate the role of population and technological advantage, we assume a large developing country and a small developed country in the benchmark model, and explore the case of a smaller developed country later. The results show that technological advantages determine the trade structure. In other words, the country with a larger θ_i has a leading position in trade.

In Chapter 3, we construct a two-sector (one polluting and one clean) general equilibrium model with heterogeneous firms in a monopolistically competitive market to explore the market allocation and welfare level in both policy equilibria. The two policies result in different market outcomes. The CT/price control can adjust the cutoff of production to reach an optimal mass of active firms, but it leads to low average productivity. In contrast, the ETS/quantity control adjusts the mass of entrants to an optimal level, but it allows too few firms to produce. Our results show that in a country with a low degree of heterogeneity, it is more efficient to charge a carbon tax; otherwise, the ETS is better.

We further compare the policy equilibria with an optimal allocation and find that both policies fail to reach the social optimum. Apart from the biases in the market outcome of the polluting sector, we verify that the mis-allocation of labor between sectors also induces the inefficiency of policy equilibria.

Chapter 4 shows how spatial economics can be embedded into a model of demographic study. The study puts the model to work analyzing fertility variance across space. In a nutshell, this chapter shows how the agglomeration effect interacts with the sorting effect, which determines the average fertility rate. In particular, the analytical results unveil a consecutive process of agglomeration with consideration of labor mobility. Since the costs of raising children act as additional centrifugal forces, we are able to reproduce the bell-shaped evolving process of industrial agglomeration. Moreover, the fertility rate in the larger region presents a U-shape pattern, whereas the smaller region has the opposite one. For ease of exposition, our discussion is based on a simple, two location model. It is straightforward to generalize the model to allow for many heterogeneous locations with a rich transportation cost structure.

References

Aldy, J. E., & Stavins, R. N. (2012), "The promise and problems of pricing carbon: Theory and experience", *The Journal of Environment & Development*, 21(2), 152-180.

Batrakova, S., & Davies, R. B. (2012), "Is there an environmental benefit to being an exporter? Evidence from firm-level data", *Review of World Economics*, 148, 449-474.

Baudry, M., Faure, A., & Quemin, S. (2021), "Emissions trading with transaction costs", *Journal of Environmental Economics and Management*, 108, 102468.

Becker, GS. (1965). "A theory of the allocation of time", *Economic Journal*, 75: 493, 517.

Becker, GS., Murphy K and Tamura, R. (1990), "Human capital, fertility and economic growth", *Journal of Political Economics*, 98(5):12-37.

Behrens, K., Mion, G., Murata, Y., & Suedekum, J. (2020), "Quantifying the gap between equilibrium and optimum under monopolistic competition," *Quarterly Journal of Economics* 135, 2299-2360.

Ben-Porath, Y. (1976), "Fertility Response to Child Mortality: Micro Data from Israel", *Journal of Political Economy*, 84(4): 163-178.

Bourgeois, C., Giraudet, L.G., & Quirion, P. (2021), "Lump-sum vs. energy-efficiency subsidy recycling of carbon tax revenue in the residential sector: A French assessment", *Ecological Economics*, 184, 107006.

Bykadorov, I., Ellero, A., Funari, S., Kokovin, S. and Molchanov, P. (2016) "Painful birth of trade under classical monopolistic competition", available at http://dx.doi.org/10.2139/ssrn.2759872.

Chen, C.-m. and Zeng, D.-Z. (2018), "Mobile capital, variable elasticity of substitution, and trade liberalization", *Journal of Economic Geography*, 18, 461-494.

Copeland, B.R. & Taylor, M.S. (1994), "North-South trade and the environment", *The quarterly journal of Economics* 109(3), 755-787.

Coria, J., & Kyriakopoulou, E. (2018), "Environmental policy, technology adoption and the size distribution of firms", *Energy Economics*, 72, 470-485.

Davies, JB and Zhang, J. (1997), "The Effects of Gender Control on Fertility and Children's Consumption", *Journal of Population Economics*, 10(1): 67-85.

de Astarloa, B.D., Eaton, J., Krishna, K., Roberts, B.A., Rodríguez-Clare, A., and Tybout, J. (2015), "Born to Export: Understanding Export Growth in Bangladesh's Apparel and Textiles Industry", *Working paper*, available at http://146.186.15.14/users/b/x/bxd197/docs/bte_prelim.pdf.

de la Croix D and Gobbi, PE. (2017), "Population density, fertility and demographic convergence in developing countries", *Journal of Development Economics*, 127:13-24.

De Loecker, J. (2011), "Product differentiation, multiproduct firms, and estimating the impact of trade liberalization on productivity". *Econometrica*, 79, 1407-1451.

De Loecker, J. and Warzynski, F. (2012), "Markups and firm-level export status", *American economic review*, 102, 2437-71.

Demidova, S. (2017), "Trade policies, firm heterogeneity, and variable markups", *Journal of International Economics*, 108, 260-273.

Demidova, S. and Rodríguez-Clare, A. (2009), "Trade policy under firm-level heterogeneity in a small economy", *Journal of International Economics*, 78(1), 100-112.

Dhingra, S. & Morrow, J. (2019), "Monopolistic competition and optimum product diversity under firm heterogeneity", *Journal of Political Economy* 127(1), 196-232.

Eichner, T. & Pethig, R. (2015), "Self-enforcing international environmental agreements and trade: taxes versus caps", *Oxford Economic Papers*, 67(4), 897-917.

Feenstra, R.C. and Weinstein, D.E. (2017), "Globalization, markups, and US welfare", *Journal of Political Economy*, 125, 1040-1074.

Felbermayr, G., and Jung, B. (2018). "Market size and TFP in the Melitz model", *Review of International Economics*, 26(4), 869-891.

Fischer, C., Parry, I.W.H., & Pizer, W.A. (2003), "Instrument choice for environmental protection when technological innovation is endogenous", *Journal of Environmental Economics and Management*, 45: 523-545.

Foellmi, R., Hepenstrick, C. and Josef, Z. (2018a). "International arbitrage and the extensive margin of trade between rich and poor countries", *The Review of Economic Studies*, 85, 475-510.

Foellmi, R., Grossmann, S.H., and Kohler, A. (2018b). "A dynamic north-south model of demand-induced product cycles", *Journal of International Economics*, 110, 63-86.

Foellmi, R., Wuergler, T., & Zweimüller, J. (2014), "The macroeconomics of Model T", *Journal of Economic Theory*, 153, 617-647.

Foellmi, R., & Zweimüller, J. (2006), "Income distribution and demand-induced innovations", *The Review of Economic Studies*, 73(4), 941-960.

Foellmi, R., & Zweimüller, J. (2017), "Is inequality harmful for innovation and growth? Price versus market size effects", *Journal of Evolutionary Economics*, 27(2), 359-378.

Forslid, R., Okubo, T., & Ulltveit-Moe, K.H. (2018), "Why are firms that export cleaner? International trade, abatement and environmental emissions", *Journal of Environmental Economics and Management*, 91, 166-183.

Fujita M, Krugman P and Venables AJ, (1999), "The spatial economy: cities, regions and international trade", *MIT Press*, Cambridge.

Goulder, L. H., & Schein, A. R. (2013), "Carbon taxes versus cap and trade: a critical review", *Climate Change Economics*, 4(03), 1350010.

Hanson, G. H. (2005), "Market potential, increasing returns and geographic concentration", *Journal of international economics*, 67(1), 1-24.

Head, K., & Mayer, T. (2004), "The empirics of agglomeration and trade", *In Handbook of regional and urban economics* (Vol. 4, pp. 2609-2669). Elsevier.

Hsu, W. T., Lu, L., & Picard, P. M. (2022), "Income inequality, productivity, and international trade", *Economic Theory*, 1-47.

Ishikawa, J. & Kiyono, K. (2006), "Greenhouse-gas emission controls in an open economy", *International Economic Review*, 47(2), 431-450.

Kiyono, K. & Ishikawa, J. (2013), "Environmental management policy under international carbon leakage", *International Economic Review* , 54(3), 1057-1083.

Kondo, K. (2019), "Does agglomeration discourage fertility? Evidence from the Japanese general social survey 2000–2010", *Journal of Economic Geography*, 19(3), 677-704.

Konishi, Y., & Tarui, N. (2015), "Emissions trading, firm heterogeneity, and intra-industry reallocations in the long run", *Journal of the Association of Environmental and Resource Economists*, 2(1), 1-42.

Kreickemeier, U., & Richter, P. M. (2014), "Trade and the environment: The role of firm heterogeneity", *Review of International Economics*, 22(2), 209-225.

Krugman, P. (1980), "Scale economies, product differentiation, and the pattern of trade", *The American Economic Review*, 70, 950-959.

Lai, Y.B, (2022), "Capital mobility and environmental policy: Taxes versus TEP", *International Tax and Public Finance*, 1-25.

Liaqat, Z., and Hussain, K. (2020), "En route to the world: understanding firms that solely export", *Economics Bulletin*, 40(4), 2872-2886.

Lorne, F., and Sneh, S. H. A. H. (2015), "Price Reversal Pattern of ARV Drugs: A Transaction-Cost Approach Digression", *Expert Journal of Economics*, 3(2).

Lu, J., Lu, Y., and Tao, Z. (2014), "Pure exporter: Theory and evidence from China", *The World Economy*, 37(9), 1219-1236.

Markusen, J.R. (2013). "Putting per-capita income back into trade theory", *Journal of International Economics*, 90(2), 255-265.

Matsuyama, K. (2000), "A ricardian model with a continuum of goods under nonhomothetic preferences: Demand complementarities, income distribution, and north-south trade", *Journal of political Economy*, 108(6), 1093-1120.

Matsuyama, K. (2002), "The rise of mass consumption societies", *Journal of political Economy*, 110(5), 1035-1070.

Melitz, M.J. (2003). "The impact of trade on intra-industry allocation and aggregate industry Productivity", *Econometrica*, 71, 1695-1725.

Melitz, M.J., and Ottaviano, G. (2008), "Market size, trade, and productivity", *Review of Economic Studies*, 75, 295-316.

Melitz, M.J., and Redding, S.J. (2015). "New trade models, new welfare implications", *American Economic Review*, 105(3), 1105-46.

Metcalf, G.E. (2007), "A proposal for a U.S. carbon tax swap: An equitable tax reform to address global climate change", *The Hamilton Project*, Discussion Paper 2007-12. Washington, D.C.: The Brookings Institution.

Metcalf, G.E. (2019), "Paying for pollution: Why a carbon tax is good for America", *Oxford University Press.*

Milliman, S. R., & Prince, R. (1989), "Firm incentives to promote technological change in pollution control", *Journal of Environmental economics and Management*, 17(3), 247-265.

Morel, C. M., McGuire, A., & Mossialos, E. (2011), "The level of income appears to have no consistent bearing on pharmaceutical prices across countries", *Health Affairs*, 30(8), 1545-1552.

Mukherjee, S., and Chanda, R. (2021). "Tariff liberalization and firm-level markups in Indian manufacturing", *Economic Modelling*, 103, 105594.

Murayama, A and Yamamoto, K. (2010), "Variety expansion and fertility rates", *Journal of Population Economics*, 23: 57-71.

Nocco, A., Ottaviano, G.I., & Salto, M. (2014), "Monopolistic competition and optimum product selection", *American Economic Review* 104(5), 304-09.

Ottaviano GI, Tabuchi T, Thisse J-F. (2002), "Agglomeration and trade revisited", *International Economic Review*, 43:409-436.

Pflüger, M. (2004), "A simple, analytically solvable, Chamberlinian agglomeration model", *Regional science and urban economics*, 34(5), 565-573.

Pflüger, M, Tabuchi, T., (2010), "The size of regions with land use for production", *Regional Science and Urban Economics*, 40(6):481, 489.

Puga D. (1999), "The rise and fall of regional inequalities", *European Economic Review*, 43:303-334.

Qiu, B. and Yan, Z. (2017), "Market efficiency, heterogeneous trade costs and export–only firms", *Pacific Economic Review*, 22, 101-122.

Redding, S., & Venables, A. J. (2004), "Economic geography and international inequality", *Journal of international Economics*, 62(1), 53-82.

Rojas, E., & Saffie, F. (2022), "Non-homothetic sudden stops", *Journal of International Economics*, 139, 103680.

Roy, S., & Saggi, K. (2012), "Equilibrium parallel import policies and international market structure", *Journal of International Economics*, 87(2), 262-276.

Sato, Y. (2007), "Economic geography, fertility, and migration", *Journal of Urban Economics*, 61: 372-387.

Sato, Y., Yamammoto, K. (2005), "Population concentration, urbanization, and demographic transition", *Journal of Urban Economics*, 58(1), 45-61.

Schultz, PT. (1985), “Changing World Prices, Women’s Wages, and the Fertility Transition: Sweden, 1860-1910”, *Journal of Political Economics*, 93(6):1126-1154.

Suedekum J. (2006), “Agglomeration and regional costs of living”, *Journal of Regional Science*, 46:529-534.

Shinkuma, T. & Sugeta, H. (2016), “Tax versus emissions trading scheme in the long run”, *Journal of Environmental Economics and Management* 75, 12-24.

Simonovska, I. (2015). “Income differences and prices of tradables: Insights from an online retailer”. *The Review of Economic Studies*, 82(4), 1612-1656.

Spulber, D.F. (1985), “Effluent regulation and long-run optimality”, *Journal of Environmental Economics and Management* 12(2), 103-116.

Stavins, R.N. (1995), “Transaction costs and tradeable permits”, *Journal of Environmental Economics and Management*, 29: 133-146.

Stavins, R.N. (2007), “A U.S. cap-and-trade system to address global climate change”, *The Hamilton Project*, Discussion Paper 2007-13. Washington, D.C.: The Brookings Institution.

Stavins, R.N. (2019), “Carbon taxes vs. cap and trade: theory and practice”, *Cambridge, Mass.: Harvard Project on Climate Agreements.*

Tabuchi T. (1998), “Agglomeration and dispersion: a synthesis of Alonso and Krugman”, *Journal of Urban Economics*, 44:333-351.

Tabuchi T, Thisse JF. (2002), “Taste heterogeneity, labor mobility and economic geography”, *Journal of Development Economics*, 69:155-177.

Takahashi, T., Takatsuka, H. and Zeng, D.-Z. (2013). “Spatial inequality, globalization, and footloose capital”, *Economic Theory*, 53, 213-238.

Takatsuka H and Zeng D-Z, (2009), “Dispersion forms: an interaction of market access, competition, and urban costs”, *Journal of Regional Science*, 49(1):177-204.

Weitzman, M.L. (1974), “Prices vs. quantities”, *Review of Economic Studies*, 41(4): 477-491.

Venables AJ. (1996), “Equilibrium locations of vertically linked industries”, *International Economic Review*, 37:951-958.

Wirl, F. (2012), “Global warming: prices versus quantities from a strategic point of view”, *Journal of Environmental Economics and Management*, 64(2), 217-229.

Yi, JJ and Zhang, JS. (2010), “The effect of house price on fertility: evidence from Hongkong”, *Economic Inquiry*, 48(3): 635-650.

Zeng D-Z. (2006), “Redispersion is different from dispersion: spatial economy of multiple industries”, *The Annals of Regional Science*, 40(2):229-247.

Zeng, D.-Z. (2021). “Spatial economics and nonmanufacturing sectors”, *Interdisciplinary Information Sciences*, 27(1), 57-91.

Zeng, D. Z., & Zhang, B. (2020), “Parallel imports in large developing countries”, *The Annals of Regional Science*, 65(2), 509-525.

Zeng, D.-Z., & Zhao, L. (2009), “Pollution havens and industrial agglomeration”, *Journal of Environmental Economics and Management*, 58(2), 141-153.

Zhang J. (2002), “Urbanization, population transition, and growth”, *Oxford Economic Papers*, 54,91-117.

Appendices

A Trade cost threshold $\tilde{\tau}$

First, we show that $\tilde{\tau}$ is well-defined as the root of (2.2).

Lemma A.2. *Equation $\mathcal{A}(\tau) = 0$ has a unique root $\tilde{\tau}$ in $[1, \theta]$, and $\mathcal{A}(\tau) \gtrless 0$ holds if and only if $\tau \gtrless \tilde{\tau}$ holds.*

Proof. As

$$\mathcal{A}(1) = -(\theta^2 - 1)(l + 1) \leq 0, \quad \mathcal{A}(\theta) = l\theta^3(\theta^3 - 1) + \theta^2(\theta - 1) \geq 0,$$

where the equalities hold only when $\theta = 1$. Subsequently, we know that $\mathcal{A}(\tau) = 0$ has a root $\tilde{\tau}$ in $[1, \theta]$. This root is unique because

$$\mathcal{A}'(\tilde{\tau}) = 6l\tilde{\tau}^5 + 3\tilde{\tau}^2 - l\theta^2 = 5l\tilde{\tau}^5 + 2\tilde{\tau}^2 + \frac{\mathcal{A}(\tilde{\tau}) + \theta^2}{\tilde{\tau}} > 0.$$

Moreover, we have

$$\mathcal{A}'(\tau) = l(6\tau^5 - \theta^2) + 3\tau^2 > l\theta^2(6\theta^3 - 1) \geq 0$$

if $\tau \geq \theta$. Therefore, $\mathcal{A}(\tau) \gtrless 0$ if and only if $\tau \gtrless \tilde{\tau}$ holds. □

Lemma A.3. *The equilibrium wage rate of (2.9) lies in $[\theta/\tau^2, \theta\tau^2)$, and the inequalities in (2.3) hold true.*

Proof. As $\tau \geq \tilde{\tau}$ is assumed, Lemma A.2 indicates that $A(\tau) \geq 0$. Thus, we have

$$\begin{aligned} \mathcal{F}(\theta\tau^2) &= \theta\tau(\theta^2\tau^6 + l\theta^2\tau^3 - \tau - 1) > 0, \\ \mathcal{F}\left(\frac{\theta}{\tau^2}\right) &= -\frac{\theta}{\tau^5}\mathcal{A}(\tau) \leq 0. \end{aligned} \tag{A.1}$$

The above two inequalities indicate that Equation (2.9) has a root $w^* \in [\theta/\tau^2, \theta\tau^2)$. Furthermore, this root is unique because we have

$$\mathcal{F}'(w) = 1 + w^2\tau + \frac{2l\theta\tau}{w} + 2\frac{\mathcal{F}(w)}{w}, \tag{A.2}$$

which implies that $\mathcal{F}'(w^*) > 0$ always holds. Finally, the inequalities in (2.3) are obtained by using (2.10) and $w^* \in [\theta/\tau^2, \theta\tau^2)$. □

Lemma A.4. *For $\tau \in [1, \tilde{\tau})$, the equilibrium relative price falls in $(\tau, \tau + 1/l)$, which gives the inequality in (2.14). Furthermore, $\mathcal{K}'(p) < 0$ holds at the equilibrium.*

Proof. We have

$$\mathcal{K}(\tau) = -\frac{\mathcal{A}(\tau)}{\tau^4(1+l\tau)} > 0,$$
$$\mathcal{K}\left(\tau + \frac{1}{l}\right) = -\tau < 0,$$

where the first inequality follows from $\tau < \tilde{\tau}$. Thus, (2.25) has a root in $(\tau, \tau + 1/l)$. Inequality $\mathcal{K}'(p) < 0$ holds at this root because

$$\frac{d}{dp}\,\frac{\theta(1-pl+\tau l)}{l\tau(p-\tau)^2 + (p^2 - p\tau + \tau^2)} = -\frac{\theta(1+l\tau)[p+(p-\tau)(1+l\tau-lp)]}{[l\tau(p-\tau)^2+(p^2-p\tau+\tau^2)]^2} < 0,$$
$$\frac{d}{dp}\,\frac{(1-lp+\tau l)^2 + \tau lp^2}{\tau(1-lp+\tau l)^2 + lp^2} = \frac{2lp(1-lp+l\tau)(\tau^2-1)}{[l(p-\tau)^2+\tau]^2(1+l\tau)} > 0.$$

□

B Proof of Proposition 1

(a) As $\mathcal{F}(1) = (\theta-1)(1-\tau) < 0$ holds when $\tau > 1$, we know that the equilibrium wage rate satisfies $w^* > 1$ according to (A.1) and (A.2). Lemma A.3 provides the uniqueness of the equilibrium wage.

(b) Using an implicit function theorem to determine the relationship between w and τ,

$$\left.\frac{dw}{d\tau}\right|_{w^*} = -\left.\frac{\frac{d\mathcal{F}(w)}{d\tau}}{\frac{d\mathcal{F}(w)}{dw}}\right|_{\mathcal{F}(w)=0} = \frac{\frac{\theta l w^*(w^*-1)}{\tau}}{2\tau w^{*2} + \theta l w^* + \frac{\tau\theta l}{w^*}} > 0.$$

(c) Let p^* be the equilibrium relative price. Taking the derivative of both sides of Equation (2.10) with respect to τ.

$$\left.\frac{dw}{d\tau}\right|_{w^*} = -\frac{2\theta}{p^3}\left.\frac{dp}{d\tau}\right|_{p^*} \quad \text{so that} \quad \left.\frac{dp}{d\tau}\right|_{p^*} = -\frac{p^3}{2\theta}\left.\frac{dw}{d\tau}\right|_{w^*} < 0.$$

(d) First, note that τ_{pr} is positive when $l > \theta^2$. Second, Equation (2.10) and wage equation (2.9) lead to the following equation, which implicitly determines the equilibrium relative price for a given τ:

$$\mathcal{G}(p,\tau) = \tau lp^6 + p^4 - \theta^2 lp^2 - \theta^2\tau = 0. \tag{B.1}$$

Subsequently, the cut-off of τ for a price reversal is the root of $\mathcal{G}(1,\tau) = 0$, i.e., τ_{pr}. Thus, we have

$$\tau_{pr} = \frac{l\theta^2 - 1}{l - \theta^2} > \frac{l\theta^2 - \theta^4}{l - \theta^2} = \theta^2 > \tilde{\tau},$$

where the last inequality is derived from Lemma A.2. Accordingly, a price reversal can only occur in the no-arbitrage equilibrium, and it occurs iff $\tau > \tau_{pr}$.

C Relationship between p_1 and τ

Using the implicit function theorem to determine the relationship between p_1 and τ,

$$\begin{aligned}\frac{dp_1}{d\tau} &= -\left.\frac{\frac{d\mathcal{H}(p_1)}{d\tau}}{\frac{d\mathcal{H}(p_1)}{dp_1}}\right|_{\mathcal{H}(p_1)=0} \\ &= \frac{4\theta^2\bar{\psi}^2 f_e^2(2\tau\bar{\psi} f_e - L_1p_1^2) - L_2^2p_1^4(\tau L_1p_1^2 + \bar{\psi} f_e)}{3L_2^2L_1p_1^5(\tau^2-1) + 4\tau\bar{\psi} f_e L_2^2p_1^3 + 4\bar{\psi} f_e\theta^2 p_1(2\tau\bar{\psi} f_e - L_1p_1^2)}.\end{aligned}$$

The denominator is positive because $2\tau\bar{\psi} f_e - L_1p_1^2 = \tau L_2p_2^2 > 0$ holds from the free-entry condition of (2.5). Meanwhile, the numerator is as follows:

$$\begin{aligned}&4\theta^2\bar{\psi}^2 f_e^2(2\tau\bar{\psi} f_e - L_1p_1^2) - L_2^2p_1^4(\tau L_1p_1^2 + \bar{\psi} f_e) \\ &\quad= 2\bar{\psi} f_e\theta^2 L_1p_1^2(2\bar{\psi} f_e - \frac{L_1p_1^2}{\tau}) + \bar{\psi} f_e L_2^2p_1^4 - \frac{L_1L_2^2p_1^6}{\tau} \\ &\quad= 2\bar{\psi} f_e\theta^2 L_1L_2p_1^2p_2^2 + \bar{\psi} f_e L_2^2p_1^4 - \frac{L_1L_2^2p_1^6}{\tau} \\ &\quad= L_1L_2p_1^2(2\bar{\psi} f_e\theta^2p_2^2 - \frac{L_2p_1^4}{\tau}) + \bar{\psi} f_e L_2^2p_1^4 \\ &\quad> L_1L_2p_1^2(\frac{\theta^2L_1p_1^2p_2^2}{\tau} - \frac{L_2p_1^4}{\tau}) + \bar{\psi} f_e L_2^2p_1^4 > 0,\end{aligned}$$

where the first equality is from (2.12), and the last inequality is from the fact that $L_1 > L_2$ and $\frac{\theta p_2^2}{p_1^2} = w > 1$ based on (2.10).

Therefore, we have $dp_1/d\tau > 0$.

D Proof of Proposition 2

(a) $U_1 - U_2 = \frac{\theta_1 w}{p_1} - \frac{\theta_2}{p_2} = \frac{\theta_2}{p_2}(w^2p - 1) = \frac{\theta_2}{p_2}(w^{\frac{3}{2}}\theta^{\frac{1}{2}} - 1) > 0.$
(b) As $\frac{dp_2}{d\tau} > 0$ holds, it is easy to derive $\frac{dU_2}{d\tau} = -\frac{\theta_2}{p_2^2}\frac{dp_2}{d\tau} < 0$. Dividing (2.5) by p_1^2 yields

$$U_1^2 = \frac{\theta_1^2}{2\bar{\psi} f_e}(L_1 + \frac{L_2}{A}), \tag{D.1}$$

where $A \equiv \tau p^2$. Rewrite $\mathcal{G}(p,\tau) = 0$ of (B.1) as

$$\mathcal{G}_a(A,\tau) \equiv lA^3 + A^2 - \theta^2 lA\tau - \theta^2\tau^3 = 0.$$

Using the implicit function theorem to determine the relationship between A and τ,

$$\frac{dA}{d\tau} = -\left.\frac{\frac{\partial\mathcal{G}_a(A,\tau)}{\partial\tau}}{\frac{\partial\mathcal{G}_a(A,\tau)}{\partial A}}\right|_{\mathcal{G}_a(A,\tau)=0} = \frac{\theta^2 lA + 3\theta^2\tau^2}{2A^2l + A + \frac{\theta^2\tau^3}{A}} > 0.$$

Differentiating (D.1) with respect to τ obtains

$$2U_1\frac{dU_1}{d\tau} = -\frac{\theta_1^2}{2\bar{\psi} f_e}\frac{L_2}{A^2}\frac{dA}{d\tau} < 0,$$

which leads to $dU_1/d\tau < 0$. Furthermore, based on $dw/d\tau > 0$ and $dp_2/d\tau > dp_1/d\tau > 0$, it is clear that $dU_2/d\tau < dU_1/d\tau < 0$ holds. In other words, welfare in the smaller country faces a sharper decrease when trade costs increase.

E Proof of Proposition 4

Immediately, we have

$$\begin{aligned} p(1) &= \frac{1 + l(2-\theta) + \sqrt{(3 + l\theta)^2 + 4(\theta - l - 3)}}{2(l+1)}, \\ p(\tilde{\tau}) &= \tilde{\tau} \end{aligned}$$

from (2.25). The implicit function also leads to

$$p'(1) = \frac{\mathcal{C}_1\sqrt{[l+\theta+(l-1)(\theta-1)]^2 + 4(l+1)(\theta-1)} + \mathcal{C}_2}{4(l+1)^2(1+l\theta)\sqrt{[l+\theta+(l-1)(\theta-1)]^2 + 4(l+1)(\theta-1)}}, \tag{E.1}$$

$$p'(\tilde{\tau}) = \frac{\mathcal{C}_3}{2(1+l\tilde{\tau})[(\theta+l\theta\tilde{\tau})^2 + l\tilde{\tau}^4(\tilde{\tau}^2-1)]}, \tag{E.2}$$

where

$$\begin{aligned} \mathcal{C}_1 =& 2 + 2l(6 + 4\theta + \theta^2) + l(l-1)(4+\theta)[1 + (2+l)\theta] \\ >& 0, \\ \mathcal{C}_2 =& 2(\theta - 3) - l[12 + 4l + \theta + l(15+8l)\theta + 2\theta^2 + l^2(l+1)\theta^3], \\ \mathcal{C}_3 =& (l-1)\tilde{\tau}^6[\tilde{\tau}(l+1)(1+2l\tilde{\tau}) + 2(\tilde{\tau}^2-1)] + \tilde{\tau}^3(\tilde{\tau}^4-1) + 2\tilde{\tau}^6(\tilde{\tau}^2-1) \\ & - 2(l^2\tilde{\tau}^2 - 1)\mathcal{A}(\tilde{\tau}). \end{aligned}$$

Expression (E.1) is positive because

$$\begin{aligned} & \mathcal{C}_1\sqrt{[l+\theta+(l-1)(\theta-1)]^2 + 4(l+1)(\theta-1)} + \mathcal{C}_2 \\ \geq & \mathcal{C}_1[1 + \theta + (l-1)(\theta-1)] + \mathcal{C}_2 \\ = & 2(1+l)[7(l-1)^2\theta^2 + 2(l-1)^3\theta^2 + 2(\theta^2-1) + (l-1)\theta(7\theta-1)] \\ > & 0. \end{aligned}$$

Meanwhile, (E.2) is positive because $\mathcal{A}(\tilde{\tau}) = 0$, $l > 1$, and $\tilde{\tau} > 1$.

We apply Equation (2.26) to examine the wage rate $w(\tau)$. Noting that $\mathcal{K}_1(p)$ also directly depends on τ, we use the notation $\mathcal{K}_1(p,\tau)$ here. Subsequently, we have

$$\begin{aligned} w(1) =& 1,\\ w'(1) =& \frac{\partial \mathcal{K}_1(p(1),1)}{\partial \tau} + \frac{\partial \mathcal{K}_1(p(1),1)}{\partial p} p'(1)\\ =& \frac{1}{2(l+1)(l+1-lp)(1+l\theta)} \Big\{ \frac{4l(\theta-1)}{\sqrt{4(l+1)(\theta-1)+(l\theta+1)^2}+1+l\theta}\\ & \times \left[3+l+\frac{(l-1)[7+l+(l-1)\theta]}{l\theta+5+\sqrt{4(l+1)(\theta-1)+(l\theta+1)^2}}\right] + l^2 - 1 \Big\}\\ >& 0,\\ w(\tilde{\tau}) =& \frac{\theta}{\tilde{\tau}^2},\\ w'(\tilde{\tau}) =& \frac{\partial \mathcal{K}_1(p(\tilde{\tau}),\tilde{\tau})}{\partial \tau} + \frac{\partial \mathcal{K}_1(p(\tilde{\tau}),\tilde{\tau})}{\partial p} p'(\tilde{\tau})\\ =& \frac{\theta}{2[\theta^2(1+l\tilde{\tau})^2 + l\tilde{\tau}^4(\tilde{\tau}^2-1)]}\\ & \times \{l(\tilde{\tau}^2-1)[(\tilde{\tau}-1)^2+(l-1)(\tilde{\tau}^2+1)] + l^2 - 1\}\\ >& 0. \end{aligned}$$

Now, we turn to prove Proposition 4.

(a) According to Equation (2.25), we can have

$$w^2 = \mathcal{K}_2(p) = \frac{(\tau-1)\left[lp^2-(1-lp+\tau l)^2\right]}{\tau(1-lp+\tau l)^2+lp^2} + 1 \geq 1.$$

The inequality holds because $p \in (\tau, \tau+1/l)$ in the arbitrage equilibrium.

(b) Let $x = p/\tau$. Subsequently, (2.25) can be written as

$$\mathcal{K}(x,\tau) = [\mathcal{K}_3(x,\tau)]^2 - \mathcal{K}_4(x,\tau) = 0, \tag{E.3}$$

where

$$\mathcal{K}_3(x,\tau) \equiv \frac{\left[\frac{1}{\tau} - l(x-1)\right]\frac{\theta}{\tau}}{(1+l\tau)(x-1)^2+x}, \quad \mathcal{K}_4(x,\tau) \equiv \frac{y^2+\tau l x^2}{\tau y^2 + l x^2},$$

and

$$y \equiv \frac{1}{\tau} - (x-1)l,\ x \in [1, 1+\frac{1}{l\tau}],\ y \in [0, \frac{1}{\tau}].$$

Subsequently, we have

$$\frac{\partial \mathcal{K}_4(x,\tau)}{\partial \tau} = \frac{2ylx^2(\tau^2-1)+\tau^2(l^2x^4-y^4)}{\tau^2(lx^2+y^2\tau)^2} > 0,$$

where the inequality holds because of the inequalities $l^2x^4 \geq l^2 > 1 \geq y^2$. Additionally, it is easy to verify $\partial\mathcal{K}_3(x,\tau)/\partial\tau < 0$ from $\partial\{[\frac{1}{\tau} - l(x-1)]\frac{\theta}{\tau}\}/\partial\tau < 0$ and $\partial[(1+l\tau)(x-1)^2 + x]/\partial\tau > 0$. Therefore, Equation (E.3) decreases with τ.

Moreover, the relationship between $\mathcal{K}(x,\tau)$ and x is derived as follows:

$$\begin{aligned}\frac{\partial\mathcal{K}_3(x,\tau)}{\partial x} =& -\frac{\theta(1+l\tau)}{\tau^2[(1+l\tau)(x-1)^2+x]^2}\Big\{2(x-1)+1-\frac{1}{l\tau}+\frac{1-[l\tau(x-1)]^2}{l\tau}\Big\}\\ &<0,\\ \frac{\partial\mathcal{K}_4(x,\tau)}{\partial x} =& \frac{2ylx(y+lx)}{(lx^2+y^2\tau)^2}(\tau^2-1) \geq 0.\end{aligned}$$

The implicit function theorem gives

$$\frac{d}{d\tau}\frac{p}{\tau} = x'(\tau) = -\frac{\partial\mathcal{K}(x,\tau)/\partial\tau}{\partial\mathcal{K}(x,\tau)/\partial x} < 0,$$

leading to the result of (b).

F Proof of Proposition 5

We rewrite the arbitrage equilibrium conditions in the form of cut-off variables. First, we transform Equations (2.4), (2.15), (2.17), (2.18).

$$\begin{aligned}&p_1 = w_1\psi_1^*,\ p_2 = \Big(\frac{\tau w_1\psi_{1TA}^* + p_1 l}{1+\tau l} = \Big)\frac{\tau w_1\psi_{1TA}^* + w_1\psi_1^* l}{1+\tau l},\\ &\psi_{2T}^* = \Big(\frac{p_1}{\tau} = \Big)\frac{w_1\psi_1^*}{\tau},\ \psi_{2A}^* = \Big(p_2 - p_1 l + \tau p_2 l = \Big)\tau w_1\psi_{1TA}^*.\end{aligned}$$

Therefore, we can use w, ψ_1^*, and ψ_{1TA}^* to represent p_1, p_2, ψ_{2T}^*, and ψ_{2A}^*.

Subsequently, the free-entry condition in country 1 (2.16) can be rewritten as

$$L_1\psi_1^{*2} + \tau L_2\psi_{1TA}^{*2} = 2f_e\bar{\psi}, \tag{F.1}$$

and the free-entry condition in country 2 (2.19) can be rewritten as

$$w_1^2(L_1\psi_1^{*2} + L_2\tau^3\psi_{1TA}^{*2}) = 2\tau f_e\bar{\psi}. \tag{F.2}$$

Further, the budget constraint in country 1 (2.22) can be rewritten as

$$(N_1+N_2)\psi_1^* - (N_1+\tau^2N_2)\psi_{1TA}^* + \frac{\tau\psi_{1TA}^*(N_1+\tau^2N_2)(L_1\psi_1^*+\tau L_2\psi_{1TA}^*)}{(\tau L_1+L_2)\psi_1^*} = \theta_1, \tag{F.3}$$

and the budget constraint in country 2 (2.23) can be rewritten as

$$\frac{w_1\psi_{1TA}^*(N_1+\tau^2N_2)(L_1\psi_1^*+\tau L_2\psi_{1TA}^*)}{(\tau L_1+L_2)\psi_1^*} = \theta_2. \tag{F.4}$$

The trade balance (2.24) can be rewritten as

$$L_1 N_2 \psi_1^{2*}(\tau L_1 + L_2) + \tau L_2 \psi_{1TA}^{2*}(\tau^3 L_1 N_2 - L_2 N_1) = L_1 L_2 \psi_1^* \psi_{1TA}^* (N_1 + \tau^2 N_2). \tag{F.5}$$

The five equations above with five endogenous variables (w_1, ψ_1^*, ψ_{1TA}^*, N_1, and N_2) yield the following wage equation:

$$\begin{aligned} \mathcal{M}(w) \equiv & w(1 - \tau w^2) + l(\tau\theta - w^2\theta + \tau w - \tau^3 w) \\ & + [l(\tau w^2 - 1)(\tau - w^2)]^{\frac{1}{2}}(\theta l + \tau w) = 0. \end{aligned} \tag{F.6}$$

This implicitly determines the relative wage rate w as a function of τ, θ, and l. The other variables are given by

$$\begin{aligned} &\psi_1^* = \bar{\psi}^{\frac{1}{2}}\mathcal{D}_1, \quad \psi_{1TA}^* = \bar{\psi}^{\frac{1}{2}}\mathcal{D}_2, \quad N_1 = \bar{\psi}^{-\frac{1}{2}}\frac{\theta_1 L_1 \mathcal{D}_1}{L_1\mathcal{D}_1^2 + \tau L_2 \mathcal{D}_2^2}, \\ &N_2 = \bar{\psi}^{-\frac{1}{2}}\frac{\theta_1 L_2 \mathcal{D}_1 \mathcal{D}_2 (L_1\mathcal{D}_1 + \tau L_2 \mathcal{D}_2)}{(L_1\mathcal{D}_1^2 + \tau L_2\mathcal{D}_2^2)[\mathcal{D}_1^2(\tau L_1 + L_2) - \tau^2 L_2 \mathcal{D}_1\mathcal{D}_2 + \tau^4 L_2 \mathcal{D}_2^2]} \end{aligned}$$

where

$$\mathcal{D}_1 = \left[\frac{2\tau f_e(\tau w^2 - 1)}{L_1 w^2(\tau^2 - 1)}\right]^{\frac{1}{2}}, \quad \mathcal{D}_2 = \left[\frac{2 f_e(\tau - w^2)}{L_2 \tau w^2(\tau^2 - 1)}\right]^{\frac{1}{2}}.$$

Note that $d(\psi_1^*/\bar{\psi})/d\bar{\psi} = -\mathcal{D}_1/(2\bar{\psi}^{\frac{3}{2}}) < 0$ holds, indicating that the cut-off of domestic supply ψ_1^* increases less proportionally with the upper limit $\bar{\psi}$.

Combining the above results with Equations (2.20) and (2.21) yields the following expressions for welfare:

$$U_1 = n_1 = N_1(w) + N_2(w), \quad U_2 = n_2 = \frac{\psi_{1TA}^*}{\bar{\psi}}\frac{L_1\theta_1 + \tau w L_2 \theta_2}{2 f_e},$$

where the second equation comes from the facts of

$$\psi_{2A}^* = \tau w \psi_{1TA}^*, \quad N_{ie} = \frac{\bar{\psi}}{\psi_i^*}N_i = \frac{L_i \theta_i}{2 f_e}, \ i = 1, 2$$

obtained from (2.4), (2.15), (2.18), and (2.28). Thus, we have

$$\frac{dU_1}{d\bar{\psi}} = -\frac{U_1}{2\bar{\psi}}, \quad \frac{dU_2}{d\bar{\psi}} = -\frac{U_2}{2\bar{\psi}}. \tag{F.7}$$

Note that (F.6) is independent of $\bar{\psi}$. Therefore,

$$\begin{aligned} \frac{d}{d\bar{\psi}}\frac{dU_1}{d\tau} &= \frac{d}{d\tau}\frac{dU_1}{d\bar{\psi}} = \frac{d}{d\tau}\left(-\frac{U_1}{2\bar{\psi}}\right) = -\frac{1}{2\bar{\psi}}\frac{dU_1}{d\tau}, \\ \frac{d}{d\bar{\psi}}\frac{dU_2}{d\tau} &= \frac{d}{d\tau}\frac{dU_2}{d\bar{\psi}} = \frac{d}{d\tau}\left(-\frac{U_2}{2\bar{\psi}}\right) = -\frac{1}{2\bar{\psi}}\frac{dU_2}{d\tau}. \end{aligned} \tag{F.8}$$

Accordingly, a higher $\bar{\psi}$ reduces the slopes of both $\frac{dU_1}{d\tau}$ and $\frac{dU_2}{d\tau}$.

F.1 Model 1. Mean-preserving spread model with uniform distribution

In the above setup, the mean productivity decreases with $\bar{\psi}$. Therefore, we need to clarify whether the utility reduction comes from a higher level of heterogeneity or lower mean productivity. To clarify this issue, we assume that the marginal labor input ψ follows a uniform distribution in $[\bar{\psi}/2 - b_1, \bar{\psi}/2 + b_1]$ in both countries, where $\bar{\psi}/2 > b_1 \geq 0$. A larger b_1 represents a higher level of heterogeneity. This setting maintains the average productivity at $\bar{\psi}/2$ when we investigate the impact of the heterogeneity level by increasing b_1.

The equilibrium conditions are rewritten as follows.

The free-entry condition in country 1:

$$\int_{\frac{\bar{\psi}}{2}-b_1}^{\psi_{1TA}^*} [(\tau p_2 - \psi w_1)L_1 + (p_2 - \tau\psi w_1)L_2]g(\psi)d\psi \\ + \int_{\psi_{1TA}^*}^{\psi_1^*} (p_1 - \psi w_1)L_1 g(\psi)d\psi = f_e w_1.$$

The free-entry condition in country 2:

$$\int_{\frac{\bar{\psi}}{2}-b_1}^{\psi_{2A}^*} [(p_2 - \psi)L_2 + (\tau p_2 - \tau\psi)L_1]\, g(\psi)d\psi + \int_{\psi_{2A}^*}^{\psi_{2T}^*} (p_1 - \tau\psi)L_1 g(\psi)d\psi = f_e.$$

The budget constraint in country 1:

$$\tau p_2 n_2 + p_1(n_1 - n_2) = w_1\theta_1.$$

The budget constraint in country 2:

$$n_2 p_2 = \theta_2.$$

where $n_1 = N_1 + N_2$ and $n_2 = \int_{\frac{\bar{\psi}}{2}-b_1}^{\psi_{1TA}^*} N_1\mu_1(\psi)d\psi + \int_{\frac{\bar{\psi}}{2}-b_1}^{\psi_{2A}^*} N_2\mu_2(\psi)d\psi$.

The trade balance:

$$\tau p_2 L_1 \int_{\frac{\bar{\psi}}{2}-b_1}^{\psi_{2A}^*} N_2\mu_2(\psi)d\psi + p_1 L_1 \int_{\psi_{2A}^*}^{\psi_{2T}^*} N_2\mu_2(\psi)d\psi = p_2 L_2 \int_{\frac{\bar{\psi}}{2}-b_1}^{\psi_{1TA}^*} N_1\mu_1(\psi)d\psi.$$

F.2 Model 2. Mean-preserving spread model with quadratic distribution

To avoid the potential influence of the special characteristics of a uniform distribution, the second model assumes a quadratic distribution. The productivity of the entered firms is described by a quadratic distribution:

$$G_q(\psi) = \begin{cases} 0 & \text{if } \psi \leq 0, \\ b_2\left[\frac{(\psi-\frac{\bar{\psi}}{2})^3}{3} - \frac{\bar{\psi}^2}{12}\psi + \frac{\bar{\psi}^3}{24}\right] + \frac{\psi}{\bar{\psi}} & \text{if } \psi \leq \bar{\psi}, \\ 1 & \text{if } \psi > 1. \end{cases}$$

Density of firm productivity is

$$g_q(\psi) = \begin{cases} b_2\left[\left(\psi - \frac{\bar{\psi}}{2}\right)^2 - \frac{\bar{\psi}^2}{12}\right] + \frac{1}{\bar{\psi}}\text{if } \psi \in [0, \bar{\psi}], \\ 0 & \text{if } \psi \notin [0, \bar{\psi}]. \end{cases}$$

Similarly, a larger b_2 indicates a higher degree of heterogeneity, while average firm productivity is always $\bar{\psi}/2$. Subsequently, we rewrite the equilibrium conditions using the new density function as follows.

The free-entry condition in country 1:

$$\int_0^{\psi^*_{1TA}} [(\tau p_2 - \psi w_1)L_1 + (p_2 - \tau\psi w_1)L_2]g_q(\psi)d\psi + \int_{\psi^*_{1TA}}^{\psi^*_1} (p_1 - \psi w_1)L_1 g_q(\psi)d\psi = f_e w_1.$$

The free-entry condition in country 2:

$$\int_0^{\psi^*_{2A}} [(p_2 - \psi)L_2 + (\tau p_2 - \tau\psi)L_1]\, g_q(\psi)d\psi + \int_{\psi^*_{2A}}^{\psi^*_{2T}} (p_1 - \tau\psi)L_1 g_q(\psi)d\psi = f_e.$$

The budget constraint in country 1:

$$\tau p_2 n_2 + p_1(n_1 - n_2) = w_1\theta_1.$$

The budget constraint in country 2:

$$n_2 p_2 = \theta_2.$$

Here, $n_1 = N_1 + N_2$ and $n_2 = \int_0^{\psi^*_{1TA}} N_1\mu_{q1}(\psi)d\psi + \int_0^{\psi^*_{2A}} N_2\mu_{q2}(\psi)d\psi$.

The trade balance:

$$L_1 \int_0^{\psi^*_{2A}} N_2\mu_{q2}(\psi)d\psi + p_1 L_1 \int_{\psi^*_{2A}}^{\psi^*_{2T}} N_2\mu_{q2}(\psi)d\psi = p_2 L_2 \int_0^{\psi^*_{1TA}} N_1\mu_{q1}(\psi)d\psi.$$

G Proof of Proposition 6

At first, the HME in terms of wages (i.e., $w \geq 1$) holds from propositions 1 and 4. Now we examine the HME in terms of firms share.

In the arbitrage equilibrium, we rewrite trade balance equation (2.24) using (2.26) as follows:

$$\frac{N_1}{N_2} = \frac{L_1}{L_2}\frac{\theta\tau}{w}.$$

From the second equality of (2.26),

$$w < w^2 = \tau - \frac{(\tau^2-1)(1-lp+\tau l)^2}{\tau(1-lp+\tau l)^2 + lp^2} < \tau.$$

Accordingly, inequality $N_1/N_2 > L_1/L_2$ holds.

In the no-arbitrage equilibrium, we rewrite trade balance (2.8) using (2.10) as follows:

$$\frac{N_1}{N_2} = \frac{L_1}{L_2}\frac{p\theta}{w} = \frac{L_1}{L_2}p^3.$$

As verified in Proposition 1, $p < 1$ only holds when $l > \theta^2 > 1$ and $\tau > (l\theta^2 - 1)/(l - \theta^2)$. Therefore, the HME in terms of firm share does not hold when the population size gap and trade costs are sufficiently large.

H Analytical proofs for Section 2.7

H.1 Trade structure

Here, we show that $p \geq 1$ holds in the no-arbitrage equilibrium as long as $\theta \geq 1$ and $l < 1$, which is observed in Figure 2.16. Note that the price in the no-arbitrage equilibrium is the root of Equation (B.1). We have

$$\begin{aligned}\mathcal{G}(1,\tau) &= -\left[(\tau-1)(\theta^2-l) + (1+l)(\theta^2-1)\right] \leq 0,\\ \mathcal{G}(\tau,\tau) &= \mathcal{A}(\tau) \geq 0.\end{aligned}$$

The second inequality holds from Lemma A.2 giving $\mathcal{A}(\tau) \geq 0$ when $\tau \geq \tilde{\tau}$.

Moreover, we have

$$\frac{d\mathcal{G}(p,\tau)}{dp} = 6\tau l p^5 + 4p^3 - 2\theta^2 l p = 4\tau l p^5 + 2p^3 + \frac{2\left[\mathcal{G}(p,\tau) + \theta^2\tau\right]}{p}, \tag{H.1}$$

This implies that $G(p,\tau)$ increases at the equilibrium price. Accordingly, Equation (B.1) has a unique solution $p^* \in [1,\tau]$ when $l < 1$.

H.2 Proof of Proposition 8

(a) As verified in Appendix H.1, the relative price in the no-arbitrage equilibrium satisfies $p \geq 1$ when $l < 1$. Therefore, arbitrage is possible only in country 1. As Lemma A.4

still holds when $l < 1$, the relative price falls in $(\tau, \tau + 1/l)$ in the arbitrage equilibrium. Consequently, the market price is higher in country 1.

(b) When $l\theta = 1$, (2.34) yields $w = 1$ and $p = \sqrt{\theta}$ in the no-arbitrage equilibrium. Additionally, we can simplify Equation (2.2) to

$$\mathcal{A}_E(\tau) = \tau^6 + \theta\tau^3 - \theta^2\tau - \theta^3 = (\tau^2 - \theta)(\tau^4 + \theta\tau^2 + \theta\tau + \theta^2). \tag{H.2}$$

Thus, the root of $\mathcal{A}_E(\tau)$, $\tilde{\tau}_E = \sqrt{\theta}$ is the trade cost threshold between the two types of equilibria.

To consider the arbitrage equilibrium, we simplify wage Equation (F.6) using $l = 1/\theta$:

$$\begin{aligned}\mathcal{M}_E(w, \tau) \equiv & w(1 - \tau w^2) + \tau - w^2 + \frac{\tau w}{\theta} - \frac{\tau^3 w}{\theta} \\ & + \left[\frac{(\tau w^2 - 1)(\tau - w^2)}{\theta}\right]^{\frac{1}{2}} (1 + \tau w) = 0.\end{aligned}$$

We have

$$\mathcal{M}_E(1, \tau) = \frac{(\tau^2 - 1)(\sqrt{\theta} - \tau)}{\theta}.$$

Accordingly, equilibrium wage w is 1 iff $\tau = 1$ or $\tilde{\tau}_E$. Moreover, we have

$$w'(1) = -\left.\frac{\frac{d\mathcal{M}_E(w,\tau)}{d\tau}}{\frac{d\mathcal{M}_E(w,\tau)}{dw}}\right|_{w=1,\tau=1} = \frac{\sqrt{\theta} - 1}{2\theta} > 0,$$

$$w'(\tilde{\tau}_E) = -\left.\frac{\frac{d\mathcal{M}_E(w,\tau)}{d\tau}}{\frac{d\mathcal{M}_E(w,\tau)}{dw}}\right|_{w=1,\tau=\tilde{\tau}_E} = -\frac{\sqrt{\theta} - 1}{2\theta} < 0.$$

Therefore, the equilibrium wage always satisfies $w > 1$ when $\tau \in (1, \tilde{\tau}_E)$, indicating that the wage rate in country 1 is always higher in the arbitrage equilibrium.

(c) In the no-arbitrage equilibrium, we have the following from (2.34):

$$\frac{U_1}{U_2} = \left(\frac{\theta_1}{\theta_2}\right)^{\frac{1}{2}} > 1.$$

In the arbitrage equilibrium, we calculate the utility difference according to (2.20) and (2.21).

$$U_1 - U_2 = n_1 - n_2 = \frac{l(p - \tau) + p\tau - 1}{\tau p} N_1 + \frac{(1 + \tau l)(p\tau - 1)}{p} N_2 > 0,$$

where the inequality is due to $p > \tau$ from (a).

(d) According to Appendix F, the signs of (F.7) and (F.8) are independent of l. Therefore, an increase in $\bar{\psi}$ still mitigates the impact of trade liberalization.

(e) In the no-arbitrage equilibrium, it is straightforward to calculate the ratio of cutoffs between the two countries according to (2.34).

$$\frac{\psi_1^*}{\psi_2^*} = \left(\frac{L_2}{L_1}\right)^{\frac{1}{2}} > 1, \quad \frac{\psi_{1T}^*}{\psi_{2T}^*} = \left(\frac{L_1}{L_2}\right)^{\frac{1}{2}} < 1.$$

In the arbitrage equilibrium, inequalities (2.31) and (2.32) still hold when $l < 1$.

$$\begin{aligned} \frac{\psi_1^*}{\psi_{2A}^*} &= \frac{p}{w[1-(p-\tau)l]} > 1, \\ \frac{\psi_{1TA}^*}{\psi_{2T}^*} &= \frac{1-(p-\tau)l}{pw} < 1, \end{aligned}$$

where the inequalities are owing to $p > \tau > w > 1$ from (a) and (2.36).

Therefore, the selection effect of domestic supply is always stronger in country 2, whereas that of exports is always stronger in country 1.

H.3 Simulation parameters

The parameter values to draw Figure 2.14 are

$$\tau = 1, \ f_e = 0.2, \ L_1 = 4, \ L_2 = 6, \ \theta_1 = 1.5, \ \theta_2 = 1.$$

The parameter values to draw Figures 2.16a, 2.17a, and 2.18a are

$$\bar{\psi} = 2, \ f_e = 0.2, \ L_1 = 4, \ L_2 = 8, \ \theta_1 = 1.5, \ \theta_2 = 1.$$

The parameter values to draw Figures 2.15, 2.16b, 2.17b, and 2.18b are

$$\bar{\psi} = 2, \ f_e = 0.2, \ L_1 = 4, \ L_2 = 6, \ \theta_1 = 1.5, \ \theta_2 = 1.$$

The parameter values to draw Figures 2.16c, 2.17c, and 2.18c are

$$\bar{\psi} = 2, \ f_e = 0.2, \ L_1 = 4, \ L_2 = 5, \ \theta_1 = 1.5, \ \theta_2 = 1.$$

I Initial allocation of emission allowances

In this Appendix, we assume only part of the initial allowances are allocated to the firms, while the rest are auctioned by the government. We use ξ to denote the share of allocated initial allowances. Therefore, the mass of entrants becomes

$$M_e = \frac{\xi \bar{E}}{\bar{e}}.$$

The price index is still written as

$$P_e = \frac{\sigma s^\beta}{\sigma - 1}\left(\frac{kN_e}{k-\sigma+1}\right)^{\frac{1}{1-\sigma}}\varphi_e^*,$$

The zero cutoff profit condition remains

$$0 = \pi(\varphi_e^*) - \bar{e}s = \alpha L\frac{k-\sigma+1}{\sigma k N_e} - F,$$

which yields

$$N_e = \alpha L\frac{k-\sigma+1}{\sigma k F}. \tag{I.1}$$

The distribution of active firms is written as

$$\mu_e(\varphi_i) = \frac{g(\varphi_i)}{G(\varphi_e^*)} = \frac{k\varphi_i^{k-1}}{\varphi_e^{*k}}.$$

Moreover, we have

$$N_e = M_e G(\varphi_e^*) = \frac{\xi\bar{E}}{\bar{e}}\left(\frac{\varphi_e^*}{\bar{\varphi}}\right)^k.$$

Combining this with (I.1), we obtain the cutoff

$$\varphi_e^* = \bar{\varphi}\left[\frac{\alpha\bar{e}L(k-\sigma+1)}{\xi\sigma kF\bar{E}}\right]^{\frac{1}{k}}.$$

The emission-clearing condition is rewritten as

$$\bar{E} = \int_0^{\varphi_e^*} e_e(\varphi_i)N_e\mu_e(\varphi_i)d\varphi_i = \alpha\beta L\frac{\sigma-1}{s\sigma},$$

from which we can obtain the emission price in the ETS:

$$s = \frac{\alpha\beta L(\sigma-1)}{\sigma\bar{E}}.$$

Note that the total supply of the emission allowances still equals $\bar{E}$. The price of auctioned allowances should be equal to the ones that are traded in the emission market. The difference is that the auctioned revenue belongs to the government, which is redistributed to the individuals later.

The total profit of firms is

$$\begin{aligned}\Pi =& \int_0^{\varphi_e^*}\left[\frac{p_e(\varphi_i)q_e(\varphi_i)}{\sigma} - F\right]N_e\mu_e(\varphi_i)d\varphi_i - M_e f_e + \xi\bar{E}s\\ =&\frac{\alpha L}{\sigma} - FN_e - \frac{\xi\bar{E}f_e}{\bar{e}} + \xi\bar{E}s.\end{aligned}$$

The government determines the initial allowances $\bar{e}$ to maximize the utility of a representative resident:

$$\begin{aligned}W_e(\bar{e}) =&\alpha\ln\frac{\alpha}{P_e}+1-\alpha+\frac{\Pi}{L}+\frac{(1-\xi)\bar{E}s}{L}\\ =&1-\alpha+\alpha\beta-\frac{f_e\xi\bar{E}}{\bar{e}L}+\frac{\alpha(\sigma-\beta k-1)}{k\sigma}\\ &-\alpha\ln\Big\{\frac{\beta L\bar{\varphi}}{\bar{E}}\Big(\frac{\alpha L}{\sigma F}\Big)^{\frac{1}{1-\sigma}}\Big[\frac{\alpha L\bar{e}(k-\sigma+1)}{\sigma kF\xi\bar{E}}\Big]^{\frac{1}{k}}\Big[\frac{\alpha\beta L(\sigma-1)}{\sigma\bar{E}}\Big]^{\beta-1}\Big\}.\end{aligned}$$

The FOC is

$$W_e'(\bar{e})=\frac{f_e\xi\bar{E}}{\bar{e}^2L}-\frac{\alpha}{k\bar{e}}=0\quad\text{giving}\quad\bar{e}^*=\frac{kf_e\xi\bar{E}}{\alpha L},$$

and the SOC is

$$W_e''(\bar{e})=-\frac{L^2\alpha^3}{\bar{E}^2f_e^2k^3\xi^2}<0.$$

Thus, the equilibrium of the optimal ETS is solved out:

$$\varphi_e^*(\bar{e}^*)=\bar{\varphi}\left[\frac{f_e(k+1-\sigma)}{\sigma F}\right]^{\frac{1}{k}},\quad M_e(\bar{e}^*)=\frac{\alpha L}{kf_e}.$$

Therefore, the market outcome and social welfare are independent of the initial allocate ξ.

J Optimal allocation

The social planner maximizes the following representative utility with a labor resource constraint:

$$\begin{aligned}\max_{\{e_o(\varphi_i),l_o(\varphi_i),M_o,\varphi_o^*\}}W=&\frac{\alpha\sigma}{\sigma-1}\Big\{\ln\int_0^{\varphi_o^*}\Big[\frac{e_o(\varphi_i)^\beta l_o(\varphi_i)^{1-\beta}}{\beta^\beta(1-\beta)^{1-\beta}\varphi_iL}\Big]^{\frac{\sigma-1}{\sigma}}M_odG(\varphi_i)\Big\}\\ &+1-\frac{M_o}{L}\Big[\int_0^{\varphi_o^*}(l_o(\varphi_i)+F)dG(\varphi_i)+f_e\Big]\\ \text{s.t.}\quad &M_o\Big[\int_0^{\varphi_o^*}e_o(\varphi_i)dG(\varphi_i)\Big]=\bar{E}.\end{aligned}\tag{J.1}$$

The planner has no control over the uncertainty in drawing φ_i but knows the underlying distributions $G(\varphi_i)$. Let λ denote the Lagrange multiplier associated with (J.1). The first-order conditions are written as

$$\frac{dW_o}{dl_o(\varphi_i)}=\alpha(1-\beta)M_o\Big[\frac{e_o(\varphi_i)^\beta l_o(\varphi_i)^{\frac{(1-\beta)(\sigma-1)-\sigma}{\sigma-1}}}{\beta^\beta(1-\beta)^{1-\beta}\varphi_iLC^M}\Big]^{\frac{\sigma-1}{\sigma}}-\frac{M_o}{L}=0,\tag{J.2}$$

$$\frac{dW_o}{de_o(\varphi_i)}=\alpha\beta M_o\Big[\frac{e_o(\varphi_i)^{\frac{\beta(\sigma-1)-\sigma}{\sigma-1}}l_o(\varphi_i)^{1-\beta}}{\beta^\beta(1-\beta)^{1-\beta}\varphi_iLC^M}\Big]^{\frac{\sigma-1}{\sigma}}-\lambda M_o=0,\tag{J.3}$$

$$\begin{aligned}\frac{dW_o}{dM_o} =&\frac{\alpha\sigma}{(\sigma-1)M_o} - \frac{1}{L}\Big[\int_0^{\varphi_o^*}\left(l_o(\varphi_i)+F\right)dG(\varphi_i)+f_e\Big] - \lambda\Big[\int_0^{\varphi_o^*} e_o(\varphi_i)dG(\varphi_i)\Big],\\ =&\frac{\alpha\sigma}{(\sigma-1)M_o} - \frac{1}{L}\Big[\frac{\alpha(1-\beta)L}{M_o} + F\Big(\frac{\varphi_o^*}{\bar{\varphi}}\Big)^k + f_e\Big] - \lambda\frac{\bar{E}}{M_o} = 0, \qquad (J.4)\\ \frac{dW_o}{d\varphi_o^*} =&\frac{k\alpha\sigma M_o\varphi_o^{*k-1}}{\bar{\varphi}^k(\sigma-1)}\left[\frac{e_o(\varphi_o^*)^\beta l_o(\varphi_o^*)^{1-\beta}}{\beta^\beta(1-\beta)^{1-\beta}\varphi_o^* LC^M}\right]^{\frac{\sigma-1}{\sigma}}\\ &- M_o\frac{k\varphi_o^{*k-1}}{\bar{\varphi}^k}\frac{[l_o(\varphi^*)+F]}{L} - \lambda M_o\frac{k\varphi_o^{*k-1}}{\bar{\varphi}^k}e_o(\varphi_o^*) \qquad (J.5)\\ =&0\end{aligned}$$

Then from (J.2) and (J.3), we can obtain

$$\left[\frac{l_o(\varphi_i)}{l_o(\varphi_j)}\right]^{\frac{(1-\beta)(\sigma-1)-\sigma}{\sigma}}\left[\frac{\frac{e_o(\varphi_i)^\beta}{\varphi_i}}{\frac{e_o(\varphi_j)^\beta}{\varphi_j}}\right]^{\frac{\sigma-1}{\sigma}} = 1,$$

$$\left[\frac{\frac{l_o(\varphi_i)^{1-\beta}}{\varphi_i}}{\frac{l_o(\varphi_j)^{1-\beta}}{\varphi_j}}\right]^{\frac{\sigma-1}{\sigma}}\left[\frac{e_o(\varphi_i)}{e_o(\varphi_j)}\right]^{\frac{\beta(\sigma-1)}{\sigma}-1} = 1,$$

which can be used to derive

$$\frac{e_o(\varphi_i)}{e_o(\varphi_j)} = \frac{l_o(\varphi_i)}{l_o(\varphi_j)} = \left(\frac{\varphi_i}{\varphi_j}\right)^{1-\sigma}.$$

Substituting this condition back into the emission market clearing (J.1), we have

$$e_o(\varphi_i) = \frac{k+1-\sigma}{k}\,\frac{\varphi_i^{1-\sigma}\bar{\varphi}^k\bar{E}}{M_o\varphi_o^{*k+1-\sigma}}. \qquad (J.6)$$

Multiplying l_i and integrating both sides of (J.2), we derive

$$\alpha(1-\beta) - \frac{1}{L}\int_0^{\varphi_o^*} l_o(\varphi_i)M_o dG(\varphi_i) = 0.$$

Multiplying e_i and integrating both sides of (J.3), we obtain

$$\alpha\beta - \lambda\int_0^{\varphi_o^*} e_o(\varphi_i)M_o dG(\varphi_i) = 0, \quad \text{which gives} \quad \lambda = \frac{\alpha\beta}{\bar{E}}.$$

Substituting λ back into (J.3) and combining that with (J.2), we can get the ratio of optimal labor and emission input for each variety

$$\frac{l_o(\varphi_i)}{e_o(\varphi_i)} = \frac{\alpha L(1-\beta)}{\bar{E}}. \qquad (J.7)$$

Equations (J.6) and (J.7) imply that

$$l_o(\varphi_i) = \alpha L(1-\beta)(k+1-\sigma)\frac{\varphi_i^{1-\sigma}\bar{\varphi}^k}{kM_o\varphi_o^{*k+1-\sigma}}.$$

After substituting $l(\varphi_i)$, $e(\varphi_i)$, and λ into (J.4) and (J.5), we can solve out N_o and φ_o^*. Finally, the optimal value of endogenous variables is rewritten as

$$\begin{aligned}
e_o(\varphi_i) &= \varphi_i^{1-\sigma}\frac{\bar{E}F(\sigma-1)}{\alpha L}\left[\frac{f_e\bar{\varphi}^k(k+1-\sigma)}{F(\sigma-1)}\right]^{\frac{\sigma-1}{k}}, \\
l_o(\varphi_i) &= \varphi_i^{1-\sigma}F(1-\beta)(\sigma-1)\left[\frac{f_e\bar{\varphi}^k(k+1-\sigma)}{F(\sigma-1)}\right]^{\frac{\sigma-1}{k}}, \\
q_o(\varphi_i) &= \varphi_i^{-\sigma}\frac{F\bar{E}^\beta(\sigma-1)}{(\alpha\beta)^\beta L^\beta}\left[\frac{f_e\bar{\varphi}^k(k+1-\sigma)}{F(\sigma-1)}\right]^{\frac{\sigma-1}{k}}, \\
M_o &= \frac{\alpha L}{kf_e},\ \varphi_o^* = \bar{\varphi}\left[\frac{f_e(k+1-\sigma)}{F(\sigma-1)}\right]^{\frac{1}{k}},\ N_o = \frac{\alpha L(k+1-\sigma)}{kF(\sigma-1)}.
\end{aligned}$$

K Proof of Lemma 1

(i) If $\alpha/c(\sigma-1) \leq 1/2$, then $\mathcal{F}_1(\gamma) = 0$ has a solution $\gamma_0 \in [1/2, L_A/(L_A+L_A^*)]$. This is because $\mathcal{F}_1(1/2) = 2\alpha(L_A - L_A^*)/(H\sigma) \geq 0$ and

$$\mathcal{F}_1\Big(\frac{L_A}{L+L_A^*}\Big) = \Big(\frac{\alpha}{c(\sigma-1)} - \frac{L_A}{L_A+L_A^*}\Big)\ln\frac{L_A}{L_A^*} \leq 0.$$

The root γ_0 is unique in $[1/2, 1]$ because

$$\begin{aligned}
\frac{d}{d\gamma}\Big[\frac{\alpha L_A}{\gamma\sigma H} - \frac{\alpha L_A^*}{(1-\gamma)\sigma H}\Big] &= -\frac{\alpha[L_A(1-\gamma)^2 + L_A^*\gamma^2]}{\gamma^2(1-\gamma)^2\sigma H} < 0, \\
\frac{d}{d\gamma}\Big[\Big(\frac{\alpha}{\sigma-1} - c\gamma\Big)\ln\frac{\gamma}{1-\gamma}\Big] &= -\frac{1}{\gamma(1-\gamma)}\Big[c\gamma - \frac{\alpha}{\sigma-1} + c\gamma(1-\gamma)\ln\frac{\gamma}{1-\gamma}\Big] < 0,
\end{aligned}$$

hold in $[1/2, 1]$, where the second inequality comes from the fact of $\alpha/(\sigma-1) \leq 1/2$.

(ii) If $1/2 < \alpha/c(\sigma-1) \leq L_A/(L_A+L_A^*)$, then $\mathcal{F}_1(\gamma) = 0$ has a solution $\gamma_0 \in [\alpha/c(\sigma-1), 1]$ because

$$\mathcal{F}_1\Big(\frac{\alpha}{c(\sigma-1)}\Big) = \frac{[L_A(\sigma-1-\alpha c) - L_A^*\alpha c](\sigma-1)}{(\sigma-1-\alpha c)\sigma cH} \geq 0, \tag{K.1}$$

$$\mathcal{F}_1(1) < 0, \tag{K.2}$$

where the inequality of (K.2) is because

$$\lim_{x\to\infty}\frac{\ln x}{x} = 0.$$

This root γ_0 is unique in $[1/2, 1]$ because $\mathcal{F}_1(\gamma)$ is evidently positive for $\gamma \in [1/2, \alpha/c(\sigma - 1)]$ and $\mathcal{F}_1(\gamma)$ is decreasing in $[\alpha/c(\sigma - 1), 1]$.

(iii) If $\alpha/c(\sigma - 1) > L_A/(L_A + L_A^*)$, then $\mathcal{F}_1(\gamma) = 0$ has a solution $\gamma_0 \in [1/2, \min\{\alpha/c(\sigma - 1), 1\}]$. This is because $\mathcal{F}_1(1/2) > 0$ and (K.1) is negative. This root γ_0 is unique in $[1/2, 1]$ because $\mathcal{F}_1(\gamma) > 0$ holds for $\gamma \in (1/2, \gamma_0)$, $\mathcal{F}_1(\gamma) < 0$ holds for $\gamma \in (\gamma_0, \min\{\alpha/c(\sigma - 1), 1\}$, and $\mathcal{F}_1(\gamma)$ is decreasing in $[\min\{1, \alpha/c(\sigma - 1)\}, 1]$.