INTERNATIONAL TRADE: THEORY AND EVIDENCE

James R. Markusen

Keith E. Maskus

Department of Economics
University of Colorado, Boulder

October, 2011

Preliminary and Incomplete. This packet constitutes a preliminary submission for evaluation and feed back. It is partial and incomplete, both with respect to the number of chapters and the contents of many of the chapters included.

`

INTERNATIONAL TRADE: THEORY AND EVIDENCE

(Alternative title suggestions most welcome!)

James R. Markusen
Keith E. Maskus

Updated October 15, 2011

ACKNOWLEDGMENT

PREFACE

NOTATION

Part One
TECHNICAL CONCEPTS AND THE GAINS FROM TRADE

1. GLOBALIZATION AND INTERNATIONAL TRADE

1.1 Introduction
1.2 Sources of globalization
1.3 Channels of globalization
1.4 Effects of globalization
1.5 Can Globalization be reversed?
1.6 Summary

References
Endnotes
Tables
Figures

2. PRODUCTION, SUPPLY AND PRODUCTION POSSIBILITIES

2.1 Properties of production functions
2.2 Equilibrium for a single producer
2.3 The production set and the production possibilities frontier
2.4 Competitive equilibrium
2.5 Cost functions
2.6 A note on increasing returns to scale and imperfect competition
2.7 Summary

References
Endnotes
Figures

3. PREFERENCES, DEMAND, AND WELFARE

3.1 Optimization for a single consumer
3.2 A note on homogeneous functions
3.3 Aggregating over households to a "community" utility function
3.4 Interpreting community indifference curves: aggregate demand versus individual welfare
3.5 Summary

References
Endnotes
Figures

4. GENERAL EQUILIBRIUM IN OPEN AND CLOSED ECONOMIES

4.1 General equilibrium in the closed (autarky) economy
4.2 General equilibrium in the open (trading) economy
4.3 The excess demand function
4.4 The shape of the excess demand curve, welfare interpretation
4.5 International general equilibrium
4.6 An introduction to computing solutions to numerical general-equilibrium models
4.7 Summary

References
Endnotes
Figures

5. THE GAINS FROM TRADE

5.1 Gains from trade
5.2 The gains-from-trade theorem
5.3 Limitations of the gains-from-trade theorem
5.4 The distribution of gains between countries
5.5 The distribution of gains within countries I: heterogeneous preferences
5.6 The distribution of gains within countries II: heterogeneous endowments
5.7 The static gains from trade: an example from history
5.8 Summary

References
Endnotes
Figures

Part Two:
CAUSES AND CONSEQUENCES OF TRADE

6. THE CAUSES OF INTERNATIONAL TRADE

6.1 The no-trade model
6.2 Methodology

Figures

7. DIFFERENCES IN TECHNOLOGY: THE RICARDIAN MODEL

7.1 Absolute and comparative advantage
7.2 The production frontier
7.3 Excess demand and international equilibrium
7.4 The role of absolute advantage in wage determination
7.5 The distribution of gains from trade between countries
7.6 Econometric on the Ricardian model
7.7 Summary

References
Endnotes
Tables
Figures

8. DIFFERENCES IN FACTOR ENDOWMENTS I: THE HECKSCHER-OHLIN MODEL

8.1 The Heckscher-Ohlin model: an intuitive approach
8.2 The Heckscher-Ohlin theorem: a more formal approach
8.3 The factor-price-equalization theorem
8.4 The Rybczynski theorem
8.5 The Stolper-Samuelson theorem
8.6 A caveat: factor-intensity reversal
8.7 Empirical evidence on factor endowments and trade
8.8 Summary

References
Endnotes
Tables
Figures

9. DIFFERENCES IN FACTOR ENDOWMENTS II: THE JONES SPECIFIC-FACTORS MODEL

9.1 The Jones specific-factors model
9.2 Analogs to the four theorems of the Heckscher-Ohlin model
9.3 Empirical evidence on preferences for protection or free trade
9.4 Summary

References
Endnotes
Tables
Figures

10. DISTORTIONS AND EXTERNALITIES AS DETERMINANTS OF TRADE

10.1 Departures from our stylized world
10.2 Distinguishing among consumer, producer, and world prices
10.3 Taxes and subsidies as determinants of trade: a small open economy
10.4 Taxes and subsidies as determinants of trade: two identical economies
10.5 Production externalities
10.6 Trade and the gains from trade in the presence of production externalites
10.7 Summary

References
Figures

11. IMPERFECT COMPETITION AND INCREASING RETURNS I: OLIGOPOLY

11.1 General discussion of increasing returns, non-comparative-advantage gains from trade
11.2 Pro-competitive gains: the basics
11.3 Special case I: quasi-linear preferences
11.4 Special case II: Cobb-Douglas preferences
11.5 Summary

References
Endnotes
Figures

12. IMPERFECT COMPETITION AND INCREASING RETURNS II: MONOPOLISTIC COMPETITION

12.1 Trade and the gains from trade through increased product diversity
12.2 A more formal approach to Dixit-Stiglitz and love of variety
12.3 Monopolistic competition in specialized intermediate inputs
12.4 The ideal variety approach to product diversity
12.5 Some useful algebra for Dixit-Stiglitz
12.6 Summary

References
Figures

13. TRADE COSTS, TRADE VOLUMES AND FIRM BEHAVIOR

13.1 Geography and trade costs
13.2 Trade costs and trade volumes in competitive, comparative-advantage models
13.3 Trade costs, prices discrimination, and trade volumes in oligopoly models
13.4 Trade costs, inter and intra-industry trade in monopolistic-competition models
13.5 The core-periphery model
13.6 Heterogeneous firms and firm-level export behavior
13.7 The gravity equation
13.8 Empirical evidence
13.9 Summary

References
Endnotes
Tables
Figures

14. PREFERENCES, PER-CAPITA INCOME AND PRODUCT QUALITY

14.1 Preferences and per-capita income as a determinant of trade
14.2 The Linder hypothesis
14.3 Integrating Linder, monopolistic competition and non-homothetic preferences
14.4 Product quality and willingness to pay
14.5 Empirical evidence on preferences, quality and trade
14.6 Summary

References
Endnotes
Table
Figures

Part Three

FACTOR TRADE, DIRECT FOREIGN INVESTMENT, OFFSHORING

15. TRADE IN FACTORS OF PRODUCTION

15.1 Adding factor trade to goods trade
15.2 Factor trade and goods trade as substitutes
15.3 Factor trade and commodity trade as complements
15.4 Agglomeration: combining monopolistic competition, trade costs, and mobile factors
15.5 Summary

References
Endnotes
Figures

16. DIRECT FOREIGN INVESTMENT AND MULTINATIONAL FIRMS

16.1 Stylized facts, basic concepts
16.2 A basic organizing framework
16.3 A simple monopoly model of location choice
16.4 Monopolistic competition and the choice of exporting versus horizontal production
16.5 The knowledge-capital model
16.6 Outsourcing versus internalization (vertical integration)
16.7 Summary

References
Endnotes
Tables
Figures

17. FRAGMENTATION, OFFSHORING, AND TRADE IN SERVICES

17.1 Stylized facts, basic concepts
17.2 Fragmentation and newly-traded intermediate goods
17.3 Fragmentation and trade in “tasks”
17.4 A gains-from-trade theorem
17.5 Trade and foreign direct investment in services
17.6 Summary

References
Tables
Figures

Part Four

TRADE POLICY

18. TARIFFS AND TRADE SUBSIDIES IN COMPETITIVE TRADE MODELS

18.1 Tariffs, welfare and factor prices in a small economy
18.2 Two key equivalences
18.3 Export subsidies
18.4 Gains from trade with many goods, trade taxes and subsidies
18.5 Monopoly power and the "optimal" tariff
18.6 Tariffs and the theory of the second best
18.7 Effective protection
18.8 Tariffs versus transport and transaction costs, foreign ownership
18.9 Summary

References
Endnotes
Figures

19. QUOTAS AND RELATED BARRIERS

19.1 Quantity and equivalent tariffs in a small economy
19.2 Distribution and dissipation of quota rents
19.3 An algebraic example
19.4 (Non) equivalence of tariffs and quotas, other related policies
19.5 Summary

References
Figures

20. STRATEGIC TRADE POLICY

20.1 Trade policy with increasing returns and imperfect competition
20.2 Export rivalry I: Cournot competition
20.3 Export rivalry II: Bertrand competition
20.4 Cournot with and without entry, adding domestic consumption
20.5 Other issues and further reading
20.6 Summary

References
Endnotes
Figures

21. MULTILATERAL TRADE AGREEMENTS: THE WORLD TRADE ORGANIZATION

21.1 Introduction
21.2 The Logic of Trade Agreements
21.3 The World Trade Organization
21.4 The Theory of Contingent Protection
21.5 Global Trade Policy and Market Externalities
21.6 Empirical Evidence
21.7 Evidence from econometric and CGE studies

Endnotes
References
Tables
Figures

22. PREFERENTIAL TRADE AGREEMENTS

22.1 Introduction
22.2 Welfare Basics: Trade Creation and Trade Diversion
22.3 A General Welfare Theorem
22.4 Endogenous FTA Formation
22.5 Econometric evidence

References
Tables
Figures

INTERNATIONAL TRADE: THEORY AND EVIDENCE

AUTHORS' PREFACE

Our book is intended primarily for a one-semester or one-quarter course in international trade for undergraduate economics majors. It is written specifically for courses and students who have completed a half-year course in intermediate microeconomic theory, a year of university calculus, and hopefully a half-year course in econometrics. The book is not suited to a course which covers both international trade and finance in one semester or quarter, nor is it suited to a course for non-majors. The book also serves as a basic text for master's courses in international economics or business and can in fact be used as a basic reference in higher-level courses before students plunge into journal articles.

Our interest in producing such a book stems from the fact that virtually all alternative texts are not designed for a semester course in trade for economics majors with an intermediate microeconomics prerequisite. Available competing books generally try to serve combined trade-and-finance survey courses, either for non-economics majors or for students without the necessary microeconomic training. As a result, they present watered-down theory with sometimes uneven analytical approaches. Further, such books do not offer deep coverage of important and informative empirical studies in the professional literature.

Thus, our book is aimed at the upper end of the undergraduate and master's markets, using a more specialized and consistent approach. Using tools of geometry and calculus, we try to stay true to the great strength of international trade theory, which is its general-equilibrium approach to all issues. Competing books have adopted a clear trend toward partial-equilibrium representations of inherently general-equilibrium problems. Even worse is the regrettable use of partial-equilibrium "triangles" to try to teach welfare analysis in trade. We also discuss analytically the most significant empirical studies that inform and flesh out the essential trade theory. We think that this combination of consistent theory supplemented by solid and frontier empirical evidence establishes the volume's unique contribution to international trade texts.

Our approach to trade theory and empirical work embodies a philosophy and pedagogy that helps unify the subject and the analysis. We have long maintained that there are many things that cause trade. Furthermore, these underlying causes or "bases" for trade are not competing theories in the usual sense, but rather all of them operate to varying degrees at the same time in all countries.

We believe that the task of theory is to analyze each cause individually and then understand the consequences of changes in the economic or policy environment surrounding each basis for trade. For example, a tariff can have very different effects if the underlying cause of trade is differences in factor endowments in a perfectly competitive environment, versus one in which the cause of trade is scale economies and imperfect competition and countries are virtually identical.

We argue that one role for positive empirical analysis is to examine the importance of alternative bases for trade and to examine which models have assumptions that conform better to real-world data. Armed with both theory and this empirical evidence, applied economists are then better positioned to assess counter-factual questions about policy and to make recommendations with some degree of confidence rather than from a stance of mere ideology. It is all too easy for us to say to politicians and journalists that liberalization is always and everywhere a good thing. But theory tells us that this is clearly not the case, and both theory and

empirical testing of the alternative models and assumptions help us give better advice.

What distinguishes the book is partly its analytical approach, but more importantly the breadth, depth and consistency of its coverage. We have tried to maintain a uniform level of analysis throughout the book, and the same basic "tool kit" is used repeatedly to avoid the costs of developing and learning new analytical constructions for each new topic. One of our first tasks was to develop a standard notation that can be used throughout the book and this notation sheet is attached after the outline. Our perception of standard texts is that they tend to treat one topic on a fairly formal level, such as the Heckscher-Ohlin model, and then resort to chat and anecdote about other topics, such as the industrial-organization approach to trade. We worked hard to develop analyses of quite different topics at approximately the same analytical level and depth of presentation. At the same time, we try to avoid introducing an analytical construction that requires a lot of time to master but is used only once.

The first section of the book, chapters 1-5, first overviews major recent trends in global trade and investment, then concentrates on developing analytical tools and techniques (chapters 2-4). We develop, for example, the “revealed preference” methodology for assessing income and welfare changes that is used repeatedly throughout the book for evaluating the efficiency of production, gains from trade, the effects of distortions, and so forth. Chapter 4 ends with a (optional) section about the methodology for actually solving complex general-equilibrium models, about which we will say more below. Chapter 5 presents a general and thorough analysis of gains from trade. Not only do we analyze the aggregate gains for one country (this is done in all textbooks, though rarely with proper care), but we also discuss the division of total gains *between* countries, and the distribution of one country’s gains among groups *within* the country.

The second part of the book, chapters 6-14, presents the positive theory of trade. Here we present our unified methodology not found elsewhere. As noted above, our view is that many things can cause trade. So we begin in chapter 6 with what we call the “no trade model”. What would have to be true for countries to choose *not* to trade and for there to be no gains from trade? The resulting list of conditions then becomes the outline for the following chapters in which those restrictions are relaxed one at a time. Trade costs have attracted a great deal of attention in the last decade or two, and in appreciation of this we will discuss how trade costs modify the results of each theoretical approach (e.g., factor-price equalization in the Heckscher-Ohlin model). Following the theory in each chapter, we present econometric evidence on that theory and evidence from a few numerical simulation models where appropriate.

The third part, chapters 15-17, continues to focus on the positive theory of trade, turning to trade in factors, foreign direct investment and trade in intermediate goods and services. Classical topics are covered in the chapter 15 on trade in factors of production while chapter 17 focuses on recent work on fragmentation: the expansion of trade at the extensive margin with new intermediate goods and services becoming tradeable. Chapter 16 has a significantly expanded and more analytical treatment of multinational firms and direct foreign investment, topics that have achieved the status of a major theoretical and empirical sub-field.

The fourth part, chapters 18-23, turns to trade policy, with analyses of alternative instruments (tariffs, quotas, subsidies) grounded in the theory and evidence presented in earlier chapters. After considering these policy instruments, the text discusses several critical areas of global policy that are the subjects of extensive ongoing research. These include the formation and impacts of preferential trade areas, the political economy underlying the selection of domestic and international trade policies, and the role of multilateralism in trade agreements.

Appendices to the book are motivated by a widespread frustration with the limitations of analytical techniques for all but extremely simple general-equilibrium models. Many books present analytical theory under the assumption of two goods, two countries, two factors, and zero trade costs. In the case of industrial-organization models, there are often only one factor, a zero income elasticity of demand for the increasing-returns sector, and various partial-equilibrium assumptions. Empirical researchers take these simple approaches and apply their insights to multi-good, multi-country, multi-factor trade data generated in a world with multiple tariffs, transport costs, non-tariff barriers, and a host of domestic distortions. It is not always clear that the simpler theoretical specifications are consistent with such complex data generation.

Because of this tension and especially because of the limitations of analytical theory, we have decided to include some brief appendices outlining the techniques for the numerical simulation of theoretical models, but do not feel that we should extend this into the calibration of models to actual data for policy-analysis experiments. Markusen has had long experience with teaching these techniques, and feels that he can open up a world of possibilities for students in a relatively sparse number of pages. Professors, if they wish, can make assignment to students to use these simple models to run counter-factual experiments (e.g., trade liberalization) and interpret the results in light of theory.

Notation

We have worked to make the notation consistent throughout the book and common to all chapters. This is not an easy task and we went through a number of iterations trying to achieve consistency and avoid conflicts. Often, we had to begin anew when a plan arrived at a possible confusion such as the use of same letter to denote a variable and an index, even though this is not technically an inconsistency. For example, it is common to use "C" (upper case, lower case, subscript, superscript) to denote "consumption", "country" and "cost", something we very much want to avoid. While many letters of the Roman and Greek alphabet are not drafted into service, we worked from the principle that notation should be intuitive if possible or at least not counter-intuitive and consistent if possible with common practice.

In the end and largely in pursuit of generality, we abandon some of the "classic" notation that many professors grew up on if not actually love. Instead of goods X and Y and factors of production K and L with prices r and w, we have chosen to adopt more general notation that permits us to refer to an arbitrary good, factor or price. This is much more convenient when summing over sectors and inputs as in, for example, a gains-from-trade proof and it permits immediate generalization to more than two goods and factors.

The approach we adopt is similar to that taken in more "modern classics" such as Dixit and Norman (1980) and Woodland (1980). When we teach from the book, we use our lectures rather than the book to inject specific examples to add concrete relevance to the general cases of the book, such as translating factors V_1 and V_2 into capital and labor, or skilled and unskilled labor.

We can now list our general approach, and then turn to the specifics. We do not suggest that students read through this now, but we hope that it offers a convenient location to allow a reader to refer back to when he or she forgets the notation later on.

(1) In general, *quantities* are denoted with *upper-case* letters, and *prices* and *costs* are denoted with *lower-case* letters. Taxes and subsidies are also denoted with lower-case letters. In the two-good case, *lower case* letters are also used to refer to *ratios* in order to economize on notation, especially in labeling graphs (this is an element of the "classic" notation).

(2) *Subscripts* can be used to denote or index (a) a specific good or factor of production, (b) a specific country. Double subscripts are common, but when it is perfectly clear, one subscript may be dropped in order to economize on notation.

(3) A *superscript* is used to denote a specific equilibrium value of a variable, such as its autarky or free-trade value.

(4) Italics are used for variables.

(5) Greek letters are used to denote parameters, such as technology parameters.

h, f countries or regions: home (*h*) and foreign (*f*), often used as subscripts on quantity or price variables. Occasionally, subscript r (region) give the general reference, *r = h, f*

X_i goods or services: i = 1,...,n; a specific good is denoted by a number; e.g., X_1, X_2

X_i:	production quantity of X_i
D_i:	consumption (demand) quantity of X_i
M_i:	imports or excess demand (D_i - X_i) for X_i (N.B., M_j is *negative* when good j is exported)
E_i:	exports or excess supply (X_i - D_i) for X_i

sometimes also used with a country/region subscript; e.g., M_{hi} , M_{fi} imports of good i by country/region *h* or *f*.

X, D, M with no commodity subscript, these denote production and consumption vectors or bundles; e.g., $X = \{X_1,..., X_n\}$, $D = \{D_1,...,D_n\}$ (typically used to label graphs)

sometimes used with a country/region subscript; e.g., X_h, X_f

V_j factors of production such as capital (*K*), labor (*L*), other (*S*) such as skilled labor, land or natural resources, subscript is a number: j = 1,...,m.

sometimes used with a specific-good subscript: V_{ij} is the use of factor j in good i

V with no factor subscript, this denotes a vector of all factors of production; e.g., $V = \{V_1,..., V_m\}$.

sometimes used with a country subscript; e.g., V_h, V_f, to denote a country's endowment (total supply) vector. These are general *fixed* (in inelastic supply).

in the two-good case, a lower-case *v* may be used to refer to the ratio of V_1 to V_2, as in the capital-labor ratio of the classic approach; e.g., $v = V_1/V_2$, $v_h = V_{h1}/V_{h2}$.

U utility or welfare, often used with a country subscript superscript; e.g., U_r, utility in country/region r.

e supercript — denotes a specific equilibrium, such as e = a (autarky), e = * (free trade), e = t (tariff- restricted) or e = s (subsidized trade). Examples:

- U_r^a autarky utility in country/region c
- X_r^* vector of production in free trade in country/region c
- V_{ij}^t good i's use of factor j under tariff protection

p_i^* — world price of X_i

p_{ir}^e — domestic price of good X_i in country or region r in equilibrium e; e.g., p_{ih}^a is the autarky price of good i in region h.

p — with no commodity subscript, p is a vector of goods prices: $p = \{p_1,...,p_n\}$

sometimes used with a country subscript and/or a specific equilibrium superscript; e.g.,p_r^a is the vector of autarky commodity prices in region r.

in the two-good case, p will denote the relative price of good 1 in terms of good 2:
e.g., $p^* = p_1^*/p_2^*$, $p_h^a = p_{h1}^a/p_{h2}^a$

q_i, q_{ir} — used to denote consumer price of X_i and price of X_i in country r in situations where taxes or subsidies make consumer and producer prices unequal. In the two-good case, q is the relative price of good 1 in terms of good 2: $q = q_1/q_2$.

w_i, w — price of factor V_i, vector of factor prices.

sometimes used with a country and/or equilibrium subscript: e.g., w_{r1}^a is the autarky price of factor V_1 in country r.

in the two-good case, w will denote the relative price of factor 1 in terms of factor 2:
e.g., $w = w_1/w_2$, $w_h^a = w_{h1}^a/w_{h2}^a$

$c_i(w_1, w_2)$ — cost of producing good i as a function of factor prices.

F(..) denotes a function of the variables and/or parameters within parenthesis; e.g., $X_1 = F(V_{11}, V_{12}, \alpha)$ is read X_1 is a function of the amounts of V_1 and V_2 used in X_1 production and the parameter α. F may carry a sector (industry) subscript.

MRS marginal rate substitution:
in the case of utility, it is the slope of an indifference curve between two goods
in the case of production, it is the slope of an isoquant between two factors

MRT marginal rate of transformation: the slope of the production possibilities frontier in the two-good case.

PPF production possibilities frontier (two-good case)

mr_i marginal revenue of an individual firm in industry i

tc_i total cost of an individual firm in industry i (just c_i in competitive models).

mc_i marginal cost of an individual firm in industry i

fc_i fixed cost of an individual firm in industry i

REFERENCES

Dixit, Avinash K. and Victor Norman (1980), *Theory of International Trade*, Cambridge: Cambridge University Press.

Woodland, Alan, D. (1982), *International Trade and Resource Allocation*, Amsterdam: North Holland.

PART ONE

TECHNICAL CONCEPTS AND THE GAINS FROM TRADE

Chapter 1

GLOBALIZATION AND INTERNATIONAL TRADE

1.1 Introduction

We live in a world that is highly interconnected by a bewildering array of complex economic transactions, social and environmental problems, and international political collaborations and conflicts. Examples from global economics are found in the news everyday. A decision by American policymakers to subsidize the production of ethanol, a form of gasoline containing an additive produced from corn, is seen by many as a key reason that grain prices are high around the world. The spectacular emergence of China as a major exporter of manufactured goods has affected wages in both rich and poor countries. As large corporations, such as Microsoft, Intel, Toyota, General Electric, and Siemens have expanded their investments in affiliates in many nations around the world, they have built global production networks that share technological knowledge across locations to produce increasingly complex goods that could be sold anywhere. Today, a major cultural product, such as a Hollywood movie or a jazz band's latest compact disk, is likely to employ creative personnel from around the world, with various components of the product recorded, mixed or edited in different locations.

The importance of international connections in trade, investment, and skilled services can be illustrated by considering the apparently simple act of making and bringing to market an item of apparel, say a fashionable woolen men's suit. The initial task is to design the suit, a highly creative activity that generally takes place in the headquarters of a major fashion label, such as Armani or Hugo Boss. Beyond that, the firm must locate reliable suppliers of raw wool, which could be farmers in New Zealand, Argentina, Scotland, or elsewhere. The wool needs to be spun into yarn and then woven into finished fabrics, tasks that are likely to be done in low-wage economies with abundant labor, such as Vietnam or Bangladesh, both major centers of fabric manufacture. The fabrics then are shipped to locations where they are combined with such other materials as buttons and zippers into high-quality sewn garments. These locations are most likely to be in somewhat higher-productivity economies, such as China, Malaysia or Mexico and the firms involved typically work as independent sub-contractors to many retailers rather than affiliates of one. The garments are then shipped to brand-name apparel companies, who sell them to high-end department stores and specialty retailers, and to generic trading companies that may ultimately sell them in discount or outlet stores.

This simple story illustrates a number of key factors in global trade and investment. The brand-name firms generally do not engage in actual production. Rather, their role is to design products that will entice consumers to pay for quality and fashion. Indeed, they are rarely involved in managing the international supply chain because their focus is on original design. Thus, there are specialized outsourcing firms that take on this supply management task, becoming an important middle actor in the act of getting suits from raw wool to consumer apparel. One prominent example is Li & Fung Limited, a Hong Kong-based sourcing company with offices in dozens of rich and poor countries and global revenues in 2008 of more than $14 billion.[1] Its business is to work with thousands of sub-contractors across the world in many industries to have products assembled according to the originating firms'

[1]See Gereffi and Memedovic (2003) for a primer on global production chains in apparel, while the Li & Fung company is described at http://www.lifung.com/eng/business/service_chain.php.

specifications. As such, it is an excellent example of a multinational enterprise, even if unfamiliar to most Americans, Canadians and Europeans. Finally, there are several international trade flows described in this example, with wool moving from New Zealand to Vietnam, yarn moving from Vietnam to Mexico, and garments being transported from Mexico to the United States and Canada for final purchase.

This textbook explains the fundamental determinants and impacts of this extensive international organization of economic activities into trade, investment, outsourcing, and the global use of knowledge. To begin this journey, consider how modern-day globalization came about.

1.2 Sources of Globalization

The concept of globalization has numerous definitions, depending on the subject matter being explained. To international economists it has a simple definition, albeit one with powerful implications. Specifically, globalization occurs when the markets of different countries become more integrated and interconnected through economic transactions that cross national borders. These transactions can be in real merchandise, various forms of services, financial instruments, investments in local production facilities by multinational firms (a process called foreign direct investment, or FDI), temporary and permanent labor migration, and technological information. They can involve individuals, trade between unrelated firms, transactions within international enterprises, and governments. What drives these transactions and how they are organized is endlessly fascinating, and the subject of this book.

It is useful to distinguish among the *sources* of globalization, the *channels* through which transactions occur, and the *effects* such integration seems to have on national economies. Consider first the sources: what causes economies to become more integrated over time? Economists generally focus on three major factors, all of which have been important in spurring more interconnected global markets in the last 60 years.

An obvious first factor is that countries have chosen to become more integrated through successively reducing their restrictions on international trade in goods and services and barriers to foreign investment. Beginning in the late 1940s, a small number of richer countries began jointly to bring down their taxes on imports, called tariffs, through negotiation and renegotiations of a treaty called the General Agreement on Tariffs and Trade (GATT).[2] This process continued through nearly five decades until 1994 when the current World Trade Organization (WTO) was founded. Currently the WTO membership comprises nearly every country in the world and each member must commit to limits on its tariffs while engaging in periodic negotiations to reduce trade restrictions. The GATT and WTO also were instrumental in liberalizing various quantitative barriers to trade, such as import quotas and favoritism in government procurement programs.

More recently, WTO member countries have begun opening their markets to international suppliers of particular services, such as transportation, retailing, insurance, and banking. Even more, both developed and many developing nations have greatly liberalized their rules under which international enterprises are permitted to invest in their markets. It is now easy for firms to locate their affiliates in countries that offer the best combinations of labor skills, public infrastructure, natural resources, and other factors that encourage FDI. Indeed, where 30 years ago many developing countries preferred to make it

[2]We discuss trade barriers in chapters 18-20 and the international institutions governing trade and investment in chapters 21-23.

difficult for FDI to come into their markets, or tried to control it through various requirements imposed on the multinational firms, now nations are more likely to offer various inducements and subsidies to attract such flows. The reason is that FDI frequently brings with it new jobs and improved technology.

The WTO is by no means the only forum through which trade and investment restrictions have been reduced. Many countries chose to liberalize their barriers unilaterally, recognizing that doing so could achieve significant economic efficiency gains and improve consumer welfare and prospects for growth. Others did so under pressure from international institutions, such as the International Monetary Fund and the World Bank, which often attach their lending programs to economic liberalization and trade reforms. Yet another major force toward lower trade restrictions is the proliferation in the 1990s and 2000s of regional preferential trade accords, such as the North American Free Trade Agreement involving the United States, Canada and Mexico, and MERCOSUR, involving Brazil, Argentina, Paraguay and Uruguay. These agreements generally eliminate tariffs on trade within the region and also require relaxation of specific investment barriers.

A second important factor is the remarkable reduction in certain transportation costs in international commerce since the 1950s.[3] There are many costs involved in getting products from a factory in one country to customers in another. There are the within-exporter surface transportation costs by railroad or truck to ocean port facilities or airports. These costs have fallen considerably in countries that invested in roads and other transport infrastructure. Similarly, the efficiency of the transit ports, whether by sea or air, matters considerably for shipping costs. The increasing use of large containers for shipping massive quantities packed tightly has increased this efficiency and encouraged more trade by ocean vessel. Similarly, the powerful jet engines on today's large cargo jets make it capable to ship considerable quantities of goods through the air, especially those products where rapid delivery is crucial..

In addition to these physical factors, transport costs depend on the prices charged for freight services by shipping companies and the premiums paid to insurance firms to cover losses if goods are damaged or destroyed in transit. It is interesting to note that the cost savings from containerization have been considerably offset by limited competition and high charges in the shipping industry over the last few decades.[4] As a result, a far larger share of world trade now takes place via air freight than was the case 20 years ago, meaning in turn that the speed at which goods are traded has increased considerably.

Shipping costs also depend on access to loans from banks, called trade credits, which are used to pay the short-term charges and then repaid after exporting concerns are paid for the goods they send to importing firms. In the global credit crunch that emerged in 2008 and 2009, the volume of trade credits fell sharply, diminishing what was already a significant decline in foreign trade. For example, a report in March 2009 found that 47 percent of banks surveyed had decreased the amount of letters of credit (short-term loans) for exports between the end of 2007 and the end of 2008, with a marked increase in fees for originating such loans.[5] A rapid recovery in trade finance is important to avoid a sustained slump in

---

[3]The importance of transportation costs and other trade costs will be discussed in detail in Chapter 13.

[4]See Hummels (2007) for an extensive review of ocean and air freight costs.

[5]International Chamber of Commerce Banking Commission, *Rethinking Trade Finance 2009: an ICC Global Survey*, Paris, March 31, 2009.

global merchandise trade.

Reductions in shipping costs are an example of the third great source of globalization, changes in technologies that bind international economies closer together. Examples surround us every day. Improvements in telecommunications make it possible for customers in Toronto to talk to technical assistants at a call center in Mumbai to answer questions about the functioning of a computer designed in Texas and assembled in Kuala Lumpur. Powerful computer programs permit consumers to locate desirable products and services that are available over the internet from companies down the block and around the world. A family can now purchase airplane tickets to go abroad and reserve hotels in which to stay on line in a matter of minutes at nearly zero transactions costs, a process that used to take much time and involve a costly travel agent. The same is true for firms looking for high-quality industrial inputs and supplies, which use complex business-to-business information networks to locate global sources.

Important technological improvements also arrive from sources far beyond telecommunications and software. Companies continually invest in research and development (R&D) to improve the quality of their products and make them more distinctive. Just as consumers like to choose from a greater range of products made within their home economies, they enjoy more variety from international product differentiation and quality improvements. We need only to think of the dramatic expansion in the number and types of wines that are traded among countries now, with varieties from Europe, the United States, Australia, New Zealand, Chile, South Africa and many other locations available in local retail outlets.

Similarly, manufacturing firms and service providers find their costs reduced, or profits increased, as they take advantage of better and more distinctive input supplies from multiple global sources. Consider, for example, the construction of a Boeing 747 passenger aircraft. To assemble this plane Boeing procures inputs from over 200 suppliers of materials, components, airframe systems, avionics, engines, power systems, and production equipment, with many of these companies headquartered in different nations. All are engaged in R&D programs to develop new products, as is Boeing. This symbiosis between technical change and product variety is a critical element in the growth of world trade, as we will see later in the book.

As the Boeing case indicates, a fundamentally important means of efficiently organizing production is to engage in *outsourcing*, or the procurement of inputs and services from firms outside the final producer. Outsourcing refers to contracting with other firms to provide inputs and a specific form contributing to globalization is *offshoring*, or the use of suppliers and sub-contractors from countries other than the headquarters location of the originator firm. In essence, offshoring is the fragmentation of production of a good into different stages across unrelated firms around the globe.

1.3 Channels of Globalization

Falling trade barriers and increasing access to new forms of technology have expanded the possibilities for international transactions of all kinds. In this section we examine trends in the major forms of transactions: international trade in goods and services, portfolio capital flows, foreign direct investment, contracts for technology and labor migration. These are the fundamental means by which citizens and firms of different nations interact with each other economically. They are, therefore, the basic conduits through which integration of markets ties countries more closely together.

Before that consider recent trends in economic growth in selected countries. In Table 1.1 we show, first, that the period from 1980 to 2007, roughly the most recent era of major globalization, saw

significant increases in nominal per-capita gross national income (GNI). These figures are stated at so-called purchasing power parity (PPP) exchange rates, which adjust each country's currency value relative to the U.S. dollar to reflect underlying costs of consuming a particular basket of goods and services.[6] For example, because of high land costs and wages in western Europe, the costs of living, especially for local services that are not subject to much import competition, tend to be higher than in other nations and the PPP rate adjusts the per-capita GNI downward for those countries. Just the opposite is true in most developing economies, where local services are cheap due to low wages and land prices. In these economies the PPP rate adjusts the per-capita GNI upward.

As noted in the first two data columns of the table, average income in the United States more than tripled in this period, rising from $12,150 in 1980 to $45,840 in 2007. This corresponds to about a 4.8 percent annual increase in nominal incomes and purchasing power. The other developed economies in the table experienced somewhat similar percentage increases also, as did Mexico, Brazil and South Africa. However, there were astonishing income expansions in the Asian developing economies, including Singapore, South Korea, China and India. Singapore and South Korea both saw nine-fold increases in their average gross national incomes, rising to $47,950 in the former case. Indeed, it is no longer sensible to refer to these countries as "developing"; they are both global leaders in finance, innovation and high-technology production. China's PPP-adjusted nominal average income rose by a factor of 22 in this period, surely the largest such increase for a large economy in recorded history over a 27-year period. India's rose by a factor of nearly seven.

While increases in nominal incomes are instructive, we should consider also the expansion of real, inflation-adjusted gross domestic product (GDP), which is a measure of how much domestic production capacity grows over time. In the next column of Table 1.1 we see that output growth averaged 3.1 percent in the United States, highest among the developed economies. Japan average 2.3 percent, though this was the result of rapid growth in the 1980s and quite low growth in the so-called "lost decades" of the 1990s and 2000s. As for the developing economies, again there was a sharp distinction between the slower-growing countries (Mexico, Brazil, and South Africa) and the Asian economies, who grew much more rapidly. Indeed, China has averaged 10 percent real economic growth over 27 years, again a historically high figure. South Korea, Singapore and India were not far behind.

These data raise a number of points worth making here. First, countries vary widely in their growth performances. In general, developing economies may be expected to grow faster than richer economies as they catch up in terms of access to technology and capital and as they educate larger shares of their populations. All of these factors raise the productivity and incomes of the catching-up economies, leading to a *convergence* of incomes over time. However, this convergence can arrive at radically different rates, as the incredible growth of the East Asian economies attests. Second, while such data can not establish any linkage between globalization and economic growth, they do pose a fundamental question. Is it possible that those countries that become more rapidly engaged in international trade and investment tend to grow faster? Our overview of trends in international activity should shed some light on this issue.

---

[6] A crude, but effective, annual attempt at computing PPP exchange rates may be found in *The Economist* magazine, which attempts to standardize the consumer cost of a Big Mac hamburger. The most recent Big Mac Index is in the July 16, 2009 edition. For a technical analysis and review, read Rogoff (1996).

International Trade

Much of this textbook is concerned with explaining theories of how countries and firms trade, what types of goods they exchange, and the effects of those activities. To motivate this analysis, consider some basic data on the foreign trade volumes and patterns for selected countries. Continuing with Table 1.1, the fourth and fifth data columns show merchandise trade, which is the sum of all commodity exports plus imports, divided by GDP. This is a common measure of "openness" to international trade, for it indicates how important international trade is relative to the size of the economy. In general, countries that lower their trade barriers would expect to see this ratio rise over time, though there are many other determinants of the relative size of international trade exposure.

These figures demonstrate that for most countries the last three decades have been a period of substantial growth in the importance of international trade. In the United States, for instance, commodity trade grew from 17 percent to 23 percent of GDP. In turn we can conclude that considerably larger shares of employment, income and consumption are now associated with foreign commerce than was true in 1980. In Canada this ratio rose from 48 to 61 percent. The difference in these two countries is easily explained in terms of *market size*, which is a key factor in explaining trade volumes. Canada is a much smaller economy than is the United States. Smaller economies tend to rely more on exports to support production and imports to provide consumption varieties, while larger economies are more diverse. As might be expected, the United States is Canada's largest trade partner and the sheer size of the U.S. economy supports significant amounts of Canadian trade. This situation holds also for Mexico, which saw its ratio of trade to GDP rise by a factor of 2.7 in this period. Both Canada and Mexico have seen large increases in trade volumes in the period since 1995 when the North American Free Trade Agreement was launched.

Although the UK trade share fell in this period, in Germany it rose from 48 percent to 72 percent. This increase reflects in part the close proximity of that country to its major trading partners in the European Union. Thus, *distance* from markets is another important determinant of international trade prominence. We can see this factor again in Japan, which is a large economy but its significant distance from major markets in North America and Europe imply relatively small trade shares. Australia's trade is also strongly affected by large distances.

An interesting economy in this context is Singapore, where the value of merchandise exports and imports together are over three times the size of GDP. This is possible because Singapore has long been a center of *entrepot* trade, meaning that goods often come there and are quickly transshipped elsewhere after some local processing. Singapore is a major regional port through which goods are shipped from a source country, such as Indonesia or Malaysia, to a destination country such as the United States or Japan. Both sides of these transactions constitute trade flows for Singapore. Hong Kong, a special administrative region of China, is similar and acts as a transshipment point for Chinese trade.

The data for China again are remarkable. China began its economic modernization reforms in 1978 and has continuously opened its markets to trade, most significantly when it joined the World Trade Organization in 2001. In the 30-year period since 1980 China has seen the ratio of merchandise trade to GDP rise from 20 percent to 68 percent. Given the very rapid growth of output, this implies a far faster expansion of trade. India has also seen a marked increase, with its ratio of trade to output more than doubling. Both of these economies have greatly reduced their restrictions on trade in this period and both have seen massive increases in international economic activity. Without doubt the entrance of these two countries, with extremely large labor forces engaged in export production, represents one of the most

important shifts in world competition in recent times.

The final four data columns of Table 1.1 provide the values of exports and imports in 2007, broken down into merchandise, including agriculture, mining, and manufacturing, and services. In economics services are often referred to as *non-traded goods*, reflecting the notion that many of them are locally provided and do not cross borders. However, trade in such services as transport, tourism, business and engineering services, and electronic offshoring to call centers are all becoming increasingly important. Even medical services, where patients in one country may travel to another for a cheaper surgical procedure or specialized therapy, have grown rapidly since the 1990s. Another example is higher education: tuition payments from foreigner students to American universities amount to tens of billions of dollars each year.

In 2007 the United States exported $1.162 trillion in merchandise, while importing $2.020 trillion, leaving a sizeable trade deficit of $858 billion. While trade imbalances like this are widely discussed in the media, economists point out that trade in services is also important to consider. Thus, the United States had a surplus of $120 billion in services, leaving an overall *current account* deficit of $738 billion, still a large figure. In 2007 Germany was the largest single merchandise-exporting country in the world and ran a sizeable trade surplus, though its services trade was nearly balanced. China also exported more goods by value than did the United States, reflecting again its massive labor force and production capacities. Less well-known is that China is a major importer of both goods and services.

To demonstrate that there has been a substantial growth in trade in the first decade of the 21st century, we show the evolution of merchandise exports, imports and the trade balance for the United States, Japan and China in Figures 1A, 1B, and 1C, respectively. China's trade was essentially balanced at around $250 billion of both imports and exports in 2000. By 2008 it was exporting around $1.4 trillion and importing $1.1 trillion.

While aggregate figures are interesting, a key subject for our study is the determination of which products countries tend to export and import. As we shall see, this issue depends essentially on the concept of *comparative advantage*, which arises from the fact that countries are relatively more efficient at producing some goods than others, depending on technology, factor endowments, consumer preferences, and other variables.

The entries in Table 1.2 provide a basic accounting of the major export and import industries for our country sample as a crude measure of comparative advantage as revealed by trade data.[7] For example, the data suggest that the United States tends to specialize its exports in crude materials,

---

[7] In this case revealed comparative advantage (RCA) is calculated as the ratio of a nation's sectoral exports, such as transport equipment, divided by total exports, with that ratio divided by the similar ratio of sectoral imports to total imports. This adjustment essentially neutralizes the fact that if an economy has a large overall trade deficit, for example, that fact will tend to reduce exports relative to imports in all sectors. Thus, a simple comparison across industries of the export-import ratio will be misleading. If the RCA exceeds unity it indicates that in that sector the economy has larger than average net exports, suggesting the economy has an export specialization in the industry. If it is less than unity the sector displays larger than average net imports, or comparative disadvantage. The industry titles in Table 1.2 reflect ratios that are much larger than one as export sectors and ratios much smaller than one as import sectors.

scientific equipment, industrial machinery, and chemicals. The first industry reflects an abundance of natural resources and forest land, while the others reflect significant endowments of capital, engineering and technology. In contrast, the US import bundle emphasizes petroleum, apparel and textiles, and beverages and tobacco. Clearly, high American wages make it uneconomic to produce labor-intensive textiles and clothing and, indeed, much of that industry has migrated to developing economies.

Similar comments apply to the trade patterns of other countries. Canada and Australia are large net exporters of primary commodities, such as raw materials, food, and petroleum. Japan, Germany and the United Kingdom have export bundles much like that of the United States, while all the developed economies import apparel. Japan imports petroleum and food, reflecting its scarcity of minerals and agricultural land. In contrast, Mexico, Brazil, India and China all are significant apparel exporters, while China also specializes in such miscellaneous manufactures as toys and video games. In general, developed economies tend to export machinery, transport equipment, scientific equipment and other capital-intensive and technology-oriented goods, while developing economies export relatively labor-intensive goods, such as apparel and textiles. This fact suggests there are substantial gains from this sector-based trade among countries to move products from lower-cost locations to higher-cost locations.

Readers should recognize, however, that these calculations are made at a high level of industry aggregation. There are many sub-industries within transport equipment or industrial machinery, for example, and countries export some variants of these goods while importing others. Thus, there is substantial two-way, or *intra-industry* trade masked by these figures. We analyze this important phenomenon in Chapters 12 and 13.

Foreign Direct Investment and Technology

While the growth in international trade since 1980 is impressive, the expansion of foreign direct investment through the global operations of multinational enterprises is what really stands out. Consider the figures in Table 1.3, which show the ratios of FDI stocks to GDP in 1980 and 2005.[8] The inward stock refers to investments in affiliates and subsidiaries in a country owned by foreign enterprises, while the outward stock is just the opposite. For example, Honda Motor Company, a Japanese enterprise, owns a number of assembly factories in the United States and the capital stock in those factories is included in the inward calculation. Microsoft, a U.S.-owned enterprise with headquarters in Seattle, owns many research and distribution facilities abroad.

It is clear from Table 1.3 that the period 1980-2005 saw tremendous growth in globalization through FDI. In the United States, for example, the inward FDI stock rose from three percent of GDP to 13 percent, signaling a very large expansion of MNE operations in that nation. Indeed, in 2005 almost 5.7 million people worked for affiliates of foreign enterprises in the United States.[9] On the other side, the

[8] A "stock" means the accumulated value of past investments, with those investments made further back in the past being discounted due to depreciation. For purposes of understanding the impacts of FDI a stock measure is better than the "flow" measure of the current year's investment, since past expenditures affect current production and employment.

[9] Detailed data on both U.S. inward and outward FDI and the operations of MNEs may be found at the Bureau of Economic Analysis website:
http://www.bea.gov/international/di1fdiop.htm

outward FDI stock rose from 7.8 percent of U.S. GDP to 16.4 percent, while direct employment of American affiliates abroad in 2005 amounted to nearly 11 million jobs. The United States was not alone in this growth. Canada's outward FDI stock nearly quadrupled as a proportion of national output, while Mexico's inward relative stock rose by a factor of 7.6, reflecting a massive shift of production from domestic firms to foreign-affiliated firms. Much of these increases in Canada and Mexico may be attributed to the impacts of NAFTA, which has encouraged rationalization of production throughout North America via international investment.

All of the three European economies saw large increases in their FDI ratios as well. While these trends reflect, in part, globalization of production both into and outside of Europe, to a large extent they are the result of reduced barriers to investment within the European Union and the accession of Eastern European countries to that association. Thus, just as the United States and Canada have expanded FDI into Mexico in search of lower costs, so have companies from the UK, Germany, France and Switzerland shifted production toward the lower-cost east. Switzerland's situation is especially interesting because its outward FDI stock amounts to more than its annual GDP. Switzerland is the classic example of a small, wealthy, and highly skill-abundant nation which is home to design-intensive and technology-intensive companies that shift much of their production activities abroad. Singapore has become an intriguing variant of this case. Its inward FDI stock is 59 percent larger than its GDP. While home to many domestic MNEs, this fact largely is due to global companies from the United States, Japan and Europe locating regional headquarters in Singapore, which then becomes a base for further investments in Asia.

Japan is unusual among developed economies in still having a fairly small representation of inward FDI. Indeed, the relatively low penetration of global MNEs into the Japanese market has long been a characteristic of that economy and a source of contention between Japanese and U.S. policymakers.[10] It is not entirely clear what explains this fact, though barriers to international investment and takeovers are high. While the figures are higher in South Korea, its inward stock is also small in relation to its economic characteristics.

In contrast the relative expansion of inward FDI in China and India has been extraordinary in this period. From a period of virtually no international ownership in 1980, both countries have greatly reduced their formal restrictions on ownership, at least in certain industries and regions.[11] In China a substantial share of this incoming investment is from Taiwan, Malaysia, Canada, Australia and other locations where there are significant populations of Chinese ancestry. These people have successfully established networks of production facilities throughout East Asia, centered on Chinese production.

Also presented in Table 1.3 are figures on technology payments and receipts, which are royalties and license fees charged on intellectual property rights, such as patents, trademarks, copyrights and trade secrets. Thus, for example, the Siemens Corporation, a German-based high-technology firm, earns significant international royalties from selling and licensing rights to use technologies in machinery, electronics, and wind power. As may be seen, these flows have greatly expanded in the period as well, reflecting increasing amounts of global technology transfer, both within MNEs and among unaffiliated firms. The United States is a major net recipient of such revenues because it remains the largest source of

---

[10] See Prestowitz (1988). This book, written by a former U.S. government official, was alarmist about Japanese growth, which seems quaint in light of that country's subsequent economic problems.

[11] Bergsten, et al (2008) discuss the case of China.

newly innovated technologies.

International Labor Migration

The period since 1980 has also seen a substantial increase in global labor migration. In Table 1.4 we present figures for four major destination countries on a standard measure of the international integration of people: the share of population that was born elsewhere. We see immediately the effect of significant inward migration in the United States, with the foreign-born population share rising from 6.2 percent to 12.6 percent. These people come from all over the world but in recent decades the largest shares have come from Central America and the Caribbean, Eastern Europe, East Asia and South Asia. They differ in their characteristics, ranging from relatively unskilled workers entering construction, agricultural and retail jobs, to highly trained medical personnel and software engineers. Most enter legally under certain visa categories, while many illegal entrants may cross the border multiple times during a year. Nearly all seek higher incomes than they earn at home, though some migrants are political refugees and some come to join family members.[12]

All countries have rigorous controls on immigration, though the United States and Canada are among the most open in legal terms. Australia has a long tradition of permitting migration, with recent flows entering primarily from Asian developing economies. Thus, its share of foreign-born in the population rose slightly, from 21.1 percent to 23.8 percent. The UK and Germany tend to be more restrictive than the United States, at least as regards legal immigration from outside the EU. However, the UK has long permitted significant immigration from former colonies, while migrants from Turkey and Southern Europe for many years have entered Germany through guest-worker programs. More recently, members of the European Union have agreed to liberalize restrictions on immigration from anywhere within the EU, resulting in large inflows to the wealthier nations. Thus, as noted in Table 1.4, both the UK and Germany registered significantly higher shares of foreign born in 2005 than 1990.

In an important sense Table 1.4 is misleading for it suggests that labor migration happens generally from poor countries to rich countries. In fact, there are massive migration flows among developing nations also, though again because the migrants hope for better pay and living conditions. Thus, there are large numbers of workers from Bangladesh in India, from the Philippines in the Persian Gulf nations, from Zimbabwe and Lesotho in South Africa, and from Central Asian countries in Russia. International labor migration is truly a global phenomenon.

1.4 Effects of Globalization

This major expansion in the exposure of economies to international transactions is highly controversial precisely because it has substantial impacts on both the well-being of countries and the welfare of individuals within nations. It also fundamentally alters the ways in which societies use resources and make decisions. In this section we highlight some of these impacts in broad terms, leaving analytical treatments to later chapters.

Economists profess almost universal support for free trade, or the complete opening of markets to foreign competition through trade and investment. This attitude comes from the basic logic of

[12] The economics of labor migration are discussed in Chapter 18.

competition: free and open trade pushes countries to specialize their resources in those industries and goods where they are relatively most productive. In turn, this specialization generates greater national output and income through trade than would be possible for countries that remain isolated. Just think, for example, how much poorer the United States would be if each individual state were walled off from trading with other states. Florida would have to produce its own wheat and Nebraska its own oranges, or else they would go without. Similarly in Europe: if investors in Italy were prevented from using brokers in Switzerland the financial sector would be fragmented and unproductive rather than concentrated and efficient. Indeed, this simple but powerful concept of specialization is the foundation on which international economists build their essential claim: globalization tends to raise aggregate incomes and overall living standards in all countries. Countries become more productive because they concentrate their production in areas of true advantage.

As we shall see in the text, there are numerous other reasons why we might generally expect globalization to generate aggregate benefits to a country that embraces open trade. First, being open to trade permits consumers to take advantage of the greater variety of goods available than would exist in a closed domestic market. This process permits more choice regarding qualities and prices, while reducing the cost to households of achieving a given level of consumption benefits. Similarly, domestic enterprise gain access to a wider variety and quality range of inputs they can put into their production processes, thereby raising productivity.

Second, globalization expands the size of markets into which domestic producers can sell their goods as exports, giving them more opportunities to benefit. A recent study of the impacts of Vietnam's 1993 decision to eliminate its restrictions on exporting rice found substantial income gains for that country's rice producers, with a significant reduction in rural poverty and reduction in the use of child labor.[13] Third, foreign competition often breaks down inefficient domestic monopolies, bringing prices down closer to the cost of production and making consumers better off.[14] For example, in numerous developing countries the decision in the 1990s to open their telecommunications industries to international entry has dramatically expanded the range of services for domestic consumers. A related outcome is that greater exposure to global competition often forces inefficient domestic firms to reduce their output or even shut down if they cannot invest sufficiently in greater productivity to compete. Economists think of this *rationalization* of production as a benefit, for it releases labor and capital from inefficient use into more productive firms or new investments.[15] It is the globalization equivalent to the familiar concept of *creative destruction* through competition and innovation.

A further form of significant gains from globalization arrives from the information content of imports, FDI and licensing. Imported capital goods may be more efficient than those produced domestically. Multinational enterprises and licensed joint ventures typically bring with them more advanced technologies or superior means of production that often result in higher productivity in domestic firms. These "spillovers" may happen in a variety of ways, including simple copying of

---

[13] See Edmonds and Pavcnik (2005).

[14] This idea finds significant support in a famous study of Turkish trade liberalization by James Levinsohn (1993).

[15] A good example is a recent study of how Tunisia's proposed liberalization of its closed service sectors would generate large welfare gains of this kind. See Konan and Maskus (2005).

technologies and products, the leaking of such information as engineers change jobs, and the sharing of technical standards between the multinational firm and its input suppliers.[16]

Finally, economists often note the large economic gains that are realized through international labor migration. The most obvious winners are the migrants themselves, who typically make far higher incomes in the countries to which they move, such as the United States and Germany, than they could at home in Mexico, Central America, Africa and South Asia.[17] These migrants send billions of dollars or euros back to their home families in the form of *remittances*, which form a major source of income there and help poor households save and accumulate productive assets. For example, a recent study found that global value of workers remittances exceeded $100 billion by 2005, which far exceeded the entire flow of financial aid from rich to poor countries.[18] In five countries in Latin America, these remittances accounted for more than 10 percent of gross national income. Another benefit of migration is that skilled workers bring technical expertise that may be needed in destination countries. In the United States, for example, foreign software engineers are in high demand among firms in the information technology sector.

Benefits of this kind from openness to trade and investment are significant and well documented across a wide range of country experiences. They are sufficiently powerful that a number of economists write spirited defenses of globalization and strongly resist attempts to reduce the momentum toward tariff cuts and investment liberalization.[19]

This positive view of international trade surely cannot be the entire story, however, since there are frequent news items about people losing their jobs to import competition or outsourcing and entire towns being devastated by the closure of manufacturing plants that were primary employers. Many analysts at certain non-governmental organizations (NGOs) argue that because farmers in poor countries cannot compete with subsidized agriculture in Europe and the United States, greater trade exposure forces them to leave their land and raises rural poverty. Others note that the increased economic activity caused by expanding flows of trade and investment places excessive stress on the use of natural resources, generates more pollution and contributes to climate change. Indeed, a number of prominent economists now wonder if a global policy of free trade is causing more harm than good.[20]

---

[16] Keller (2004) provides an extensive review of this question.

[17] For example, Winters (2004) calculates that if governments in the developed countries would increase their quotas on the inward migration of both skilled and unskilled temporary workers up to three percent of their labor forces, the global welfare gains would be more than $150 billion per year. The effects of migration are discussed in detail in Chapter 18.

[18] See Jennings and Clarke (2005).

[19] Two excellent examples are the books *In Defense of Globalization* by Jagdish Bhagwati (2004) and *Why Globalization Works* by Martin Wolf (2004)..

[20] Two important books along this line are *Has Globalization Gone Too Far?* by Dani Rodrik (1997) and *Fair Trade for All: How Trade Can Promote Development* by Joseph E. Stiglitz and Andrew Charlton (2005).

Globalization, therefore, is an extraordinarily powerful and complex phenomenon with multiple sources, carriers, and impacts. Our goal in this text is to provide a consistent and advanced analytical treatment of international trade and factor flows in order to provide a sound framework within which to study this complexity. For example, trade and technological change can have substantial effects on the distribution of income across types of workers, regions within countries, and across countries. In general, openness favors workers, land and capital that are capable of producing high-quality goods for export. It tends to worsen the lot of individuals and regions that produce goods and services competing closely with imports, inward FDI, and immigrants. In many countries in both the developed and developing world these impacts seem to make the distribution of income more unequal over time.

Furthermore, globalization is heavily criticized because greater international integration can interfere with the attainment of deeply held social preferences. A country or village, for example, may find its culture being changed to be more like international practices that appeal to the bulk of consumers but are disliked by some. Thus, France and Canada have specific policies in place to prevent foreign media such as Hollywood movies and American television programs from overwhelming their cultural industries. Peru, India, and other developing nations recently have adopted legislation to protect their plant resources, traditional folklore and cultural traditions from unwanted use by international firms.[21]

Others are concerned because globalization, despite the economic opportunities it provides, brings people in different countries into closer contact and that can spread problems around the world more rapidly. The H1N1 "swine flu" virus is a good example. Perhaps more relevant is the economic contagion that is associated with financial crises. The global downturn of 2008-2009 was launched by the bursting of the housing bubble in the United States but diffused rapidly into the rest of the world through integrated investment markets. Finally, countries may worry that globalization pressures governments to choose policies that are not necessarily perceived to be in the national interest. Such changes may be mandated by the rules of the World Trade Organization or other international institutions, for example. Or they may be adopted by governments competitively in order to attract multinational enterprises looking to reduce production costs. A "race to the bottom" in tax policy and environmental regulations is often alleged to accompany globalization of investment, though evidence of it is elusive.[22]

1.5 Can Globalization Be Reversed?

These are powerful objections to the idea that free trade is the most beneficial policy a country can pursue. They underlie the fact that, even though economists overwhelmingly prefer open markets, other people are reluctant to embrace that idea and sometimes actively oppose it. For instance, consider the following question asked in a *Wall Street Journal* poll late in 2007:

> "Do you think the fact that the U.S. economy has become increasingly global is good because it has opened up new markets for American products and resulted in more jobs, or bad because it has subjected American companies and employees to unfair competition and cheap labor?"

---

[21] *Mulan* is a Disney movie based on an ancient Chinese story – should anyone in China have been able to prevent this movie from being made or to demand compensation?:

[22] See Basinger and Hallerberg (2004) and Dasgupta et al. (2002).

In June of 2007 42 percent of Americans surveyed thought globalization was good while 48 percent answered that it was bad. By December of that year, in the wake of extensive media coverage of job offshoring, manufacturing plants shutting, rising income inequality and emerging uncertainty about the economy, only 28 percent answered that it was good and 58 percent said globalization was harmful.[23] Attitudes can change quickly.

In this context, we might wonder whether recent globalization is sustainable or might actually be reversed as political pressures mount against it. To consider this question it is interesting to delve into a bit of history. The current era of rapid international integration is unprecedented in its scale and scope, primarily because of the role of information technologies and the entry of vast new populations from China, India and other developing economies into global competition. But it is has important historical roots.

Indeed, the period from 1980 to now is sometimes referred to as the Third Wave of globalization.[24] The First Wave happened from 1870 to around 1915 and was in many ways just as remarkable as the current epoch. It largely involved the industrialized economies of Western Europe and the "new world" agricultural economies of the United States, Canada, Australia, and Argentina. In this era vast new tracts of farmland were opened in the new world, with the commodities being exported in great volumes to Europe. One effect was a steady decline in crop prices and farm incomes in Europe, pushing rural workers off the land. At the same time, the United States was becoming an industrial power because its low wages encouraged production of textiles, apparel, footwear, and other labor-intensive goods. Millions of workers in the "old world" of Europe were forced off the land and out of factories in Germany, Italy, Greece, Ireland, Scandinavia and elsewhere. Many of them migrated across the ocean. Further, the abundant land in North America and Australia attracted large capital flows from Europe to finance the construction of railroads, ports, and factories.

These flows were sizeable in every dimension. Thus, for example, by 1910 the ratio of world merchandise exports to global output reached about 35 percent, approximately the same as it is today. The foreign capital stock mounted to nearly nine percent of GDP of the then-developing economies, such as the United States, Canada, and Australia. And the number of immigrants who became legal permanent residents in the United States averaged nearly 1 million persons per year in the decade 1905-1914, peaking at 1,285,349 in 1907.[25] These figures are quite comparable to the approximately 1 million new legal residents per year since 2000, but represented a far-higher proportion of the labor force in the early 20th century.

Thus, the period 1870-1915 was also a remarkable era of globalization, with similar effects: rising incomes overall, rapidly increasing inequality, particularly harming the unskilled workers in Europe, and a shift of manufacturing capacity from high-wage to low-wage locations. This epoch was critical for the industrial development of the United States and Canada.

---

[23] "Americans' Anti-Global Turn May Stir Race for President," *Wall Street Journal* December 20, 2007.

[24] For detailed discussion, see World Bank (2002) and Bordo, Taylor and Williamson (2003).

[25] Immigration data are from Migration Policy Institute.

However, this first wave was rapidly reversed after 1915 and subsequently global integration collapsed. By 1932 world exports had fallen to just five percent of GDP, while legal immigration into the United States collapsed to just 23,068 persons in 1933. The reasons were straightforward. World War I (1914-1918) greatly disrupted normal trade relations and many of the European economies never really recovered in the 1920s. The onset of the Great Depression of the early 1930s induced massive increases in barriers to trade, investment and migration. The most notorious episode was the passage into law in the United States of the Tariff Act of 1930, widely known as the Smoot-Hawley Tariff Act. This bill sharply raised tariffs on 200,000 imported goods as the United States attempted to push rising unemployment onto foreign labor markets. Other major trading partners quickly retaliated with their own tariff increases, a process that some think contributed markedly to bringing on global depression.[26] At the same time, major countries greatly restricted the number of visas available for migration. Protectionism was the watchword of the day, with declining integration the result.

This was a lesson painfully learned. After World War II (1938-1945), several countries in Western Europe joined the now-developed United States, Canada, Australia and other economies to establish a global system for integration that might preclude another episode like it. Thus was established the International Monetary Fund, the World Bank and, most significantly for this course, the General Agreement on Tariffs and Trade (GATT), which became the World Trade Organization in 1995. The GATT engineered several rounds of tariff cuts among the developed countries and reduced some barriers to investment flows. Under this cooperative approach trade and investment again expanded, generating the Second Wave of globalization from 1950 to 1980, although the figures were modest compared to those in the current Third Wave.

Given this history, can modern globalization be reversed? Much depends on how politicians react to the economic downturn of 2008-2010, the impacts of higher unemployment, and popular resentment of income inequality. By 2009 there were unmistakable signs of an increasing tendency toward protectionism, including higher tariffs in many developing economies, government purchasing laws that favored domestic production, and restrictions on the ability of international firms to take over domestic enterprises in particular countries.[27] Perhaps the most worrying sign is the chronic inability of the WTO member countries to conclude the long-standing Doha Round of trade negotiations.

At the same time, there are three major differences in the world of today compared to that of the early 20th century, all of which argue that globalization is sustainable. First, modern globalization is, for the first time, truly global. With the exception of parts of sub-Saharan Africa, most of the developing world has engineered significant trade and investment liberalization. As a result their economies have become more dependent on export markets and imported technologies. This global integration makes it less likely that countries will reverse course. Second, the major growth of global production networks through multinational enterprises and the the use of information technology and telecommunications means that large companies are more global than national in character now. As a result, they are more likely to oppose protectionism than support it, for political attempts to preserve jobs in one country are likely to raise costs of operations in others. Finally, virtually all nations of the world are members of the WTO, which places restrictions on their ability to raise trade and investment barriers unilaterally, as we

---

[26] See Kindleberger (1973) for an insightful history.

[27] These trends are documented by an association called Global Trade Alert, with udpates available at http://www.globaltradealert.org.

discuss in Chapter 24. Unless that institution collapses there is unlikely to be a significant reversal into protectionism.

1.6 Summary

With this background, we begin our exploration of international trade theory and policy. We will illuminate and deepen the understanding students have of the trends discussed in this chapter by relying on a strongly analytical approach, backed up by a review of important empirical evidence. By the end of the textbook students should have a solid grasp of the modern analytics of international trade, investment, and global policymaking.

REFERENCES

Basinger, Scott J. and Mark Hallerberg, 2004, "Remodeling the Competition for Capital: How Domestic Politics Erases the Race to the Bottom," *American Political Science Review*, vol. 98, 261-276.

Bergsten, C. Fred, Charles Freeman, Nicholas R. Lardy and Derek J. Mitchell, 2008, *China's Rise: Challenges and Opportunities*, (Washington DC: Peterson Institute for International Economics).

Bhagwati, Jagdish, 2004, *In Defense of Globalization*, (Oxford: Oxford University Press).

Bordo, Michael D., Alan M. Taylor and Jeffrey G. Williamson, editors, 2003, *Globalization in Historical Perspective*, (Chicago: University of Chicago Press).

Dasgupta, Susmita, Benoit Laplante, Hua Wang and David Wheeler, 2002, "Confronting the Environmental Kuznets Curve," *Journal of Economic Perspectives*, vol. 16, 146-168.

Edmonds, Eric and Nina Pavcnik, 2005, "The Effect of Trade Liberalization on Child Labor," *Journal of International Economics*, vol. 65, 401-441.

Gereffi, Gary and Olga Memedovic, 20003, *The Global Apparel Value Chain* (Vienna: United Nations Industrial Development Organization).

Hummels, David, 2007, "Transportation Costs and International Trade in the Second Era of Globalization," *Journal of Economic Perspectives*, vol. 21, 131-154.

Jennings, Allen and Matthew Clarke, 2005, "The Development Impact of Remittances to Nicaragua," *Development in Practice*, vol. 15, 685-691.

Keller, Wolfgang, 2004, "International Technology Diffusion," *Journal of Economic Literature*, vol. 42, 752-782.

Kindleberger, Charles P., 1973, *The World in Depression, 1929-1939*, (Berkeley: University of California Press).

Konan, Denise and Keith E. Maskus, 2006, "Quantifying the Impact of Services Liberalization in a Developing Country," *Journal of Development Economics*, vol. 81, 142-162.

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

Prestowitz, Clyde, 1988, *Trading Places: How We Are Giving Our Future to Japan and How to Reclaim It*, (New York: Basic Books).

Rogoff, Kenneth, 1996. "The Purchasing Power Parity Puzzle," *Journal of Economic Literature*, vol. 34, 647-668.

Rodrik, Dani, 1997, *Has Globalization Gone Too Far?* (Washington, DC: Institute for International

Economics).

Stiglitz, Joseph E. and Andrew Charlton, 2005, *Fair Trade for All: How Trade Can Promote Development*, (Oxford: Oxford University Press).

Wolf, Martin, 2004, *Why Globalization Works*, (New Haven: Yale University Press).

Winters, L. Alan, 2004, "The Economic Implications of Liberalizing Mode 4 Trade," in Aaditya Mattoo and Antonia Carzaniga, editors, *Moving People to Deliver Services*, World Bank and Oxford University Press.

World Bank, 2002, *Globalization, Growth and Poverty: Building an Inclusive World Economy* (Oxford: Oxford University Press).

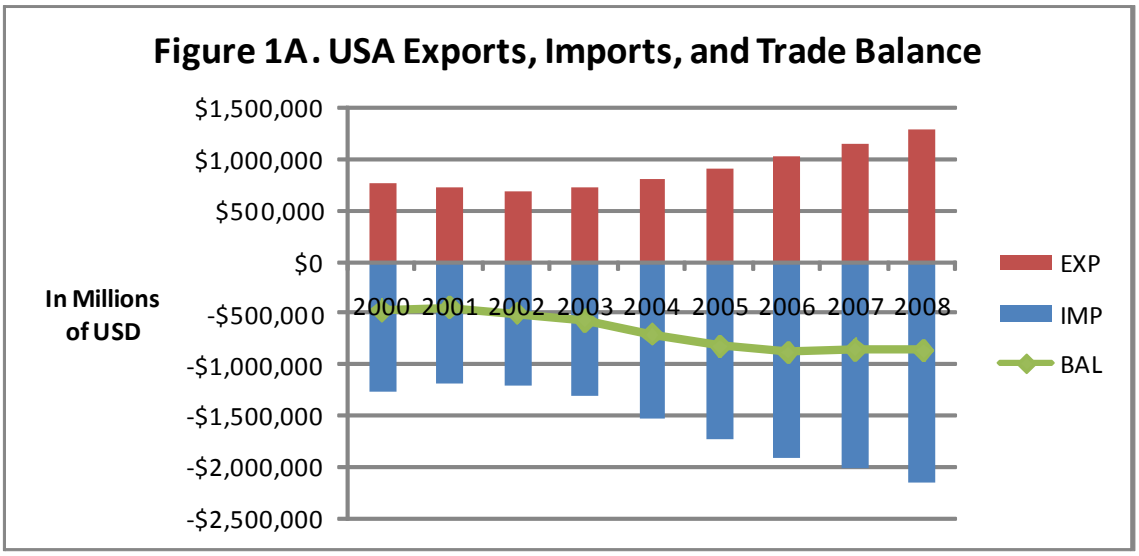


Figure 1A. USA Exports, Imports, and Trade Balance
In Millions of USD
$1,500,000
$1,000,000
$500,000
$0
-$500,000
-$1,000,000
-$1,500,000
-$2,000,000
-$2,500,000
2000 2001 2002 2003 2004 2005 2006 2007 2008
EXP
IMP
BAL

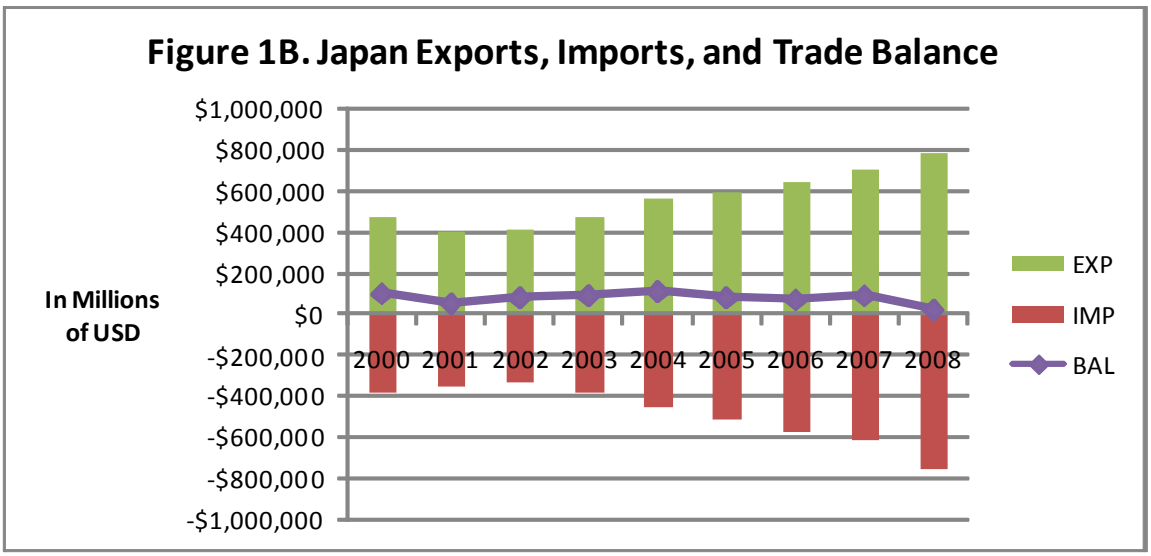


Figure 1B. Japan Exports, Imports, and Trade Balance
In Millions of USD
$1,000,000
$800,000
$600,000
$400,000
$200,000
$0
-$200,000
-$400,000
-$600,000
-$800,000
-$1,000,000
2000 2001 2002 2003 2004 2005 2006 2007 2008
EXP
IMP
BAL

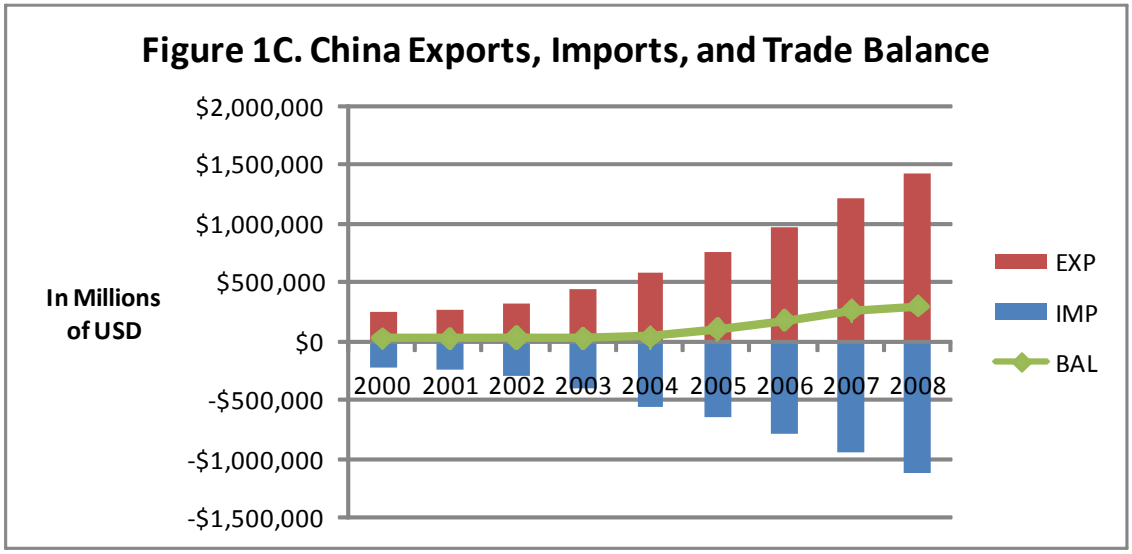


Figure 1C. China Exports, Imports, and Trade Balance
In Millions of USD
$2,000,000
$1,500,000
$1,000,000
$500,000
$0
-$500,000
-$1,000,000
-$1,500,000
2000 2001 2002 2003 2004 2005 2006 2007 2008
EXP
IMP
BAL

Table 1.1 Figures on International Trade									
			Average	Total Merchandise		Exports ($billion)		Imports ($billion)	
	GNI per capita (PPP)		Real GDP	Trade (% of GDP)		Goods	Services	Goods	Services
Country	1980	2007	Growth %	1980	2007	2007	2007	2007	2007
United States	$ 12,150	$ 45,840	3.1	17	23	$ 1,162	$ 456	$ 2,020	$ 336
Canada	$ 10,770	$ 35,500	2.8	48	61	$ 419	$ 61	$ 390	$ 80
Mexico	$ 3,830	$ 13,910	2.6	21	56	$ 272	$ 18	$ 296	$ 24
United Kingdom	$ 8,210	$ 34,250	2.5	42	38	$ 438	$ 273	$ 620	$ 194
Germany	$ 9,870	$ 34,740	2.0	48	72	$ 1,326	$ 206	$ 1,059	$ 250
Australia	$ 8,990	$ 33,400	3.3	30	37	$ 141	$ 40	$ 165	$ 38
Japan	$ 8,920	$ 34,750	2.3	26	30	$ 713	$ 127	$ 621	$ 149
South Korea	$ 2,600	$ 24,840	6.7	62	75	$ 371	$ 62	$ 357	$ 83
Singapore	$ 6,720	$ 47,950	7.0	370	349	$ 299	$ 67	$ 263	$ 70
China	$ 250	$ 5,420	10.0	20	68	$ 1,218	$ 122	$ 956	$ 129
India	$ 420	$ 2,740	6.1	13	31	$ 145	$ 90	$ 217	$ 77
Brazil	$ 3,500	$ 9,270	2.4	19	22	$ 161	$ 23	$ 127	$ 35
South Africa	$ 3,930	$ 9,450	2.4	56	57	$ 70	$ 13	$ 91	$ 16

Sources: World Bank, *World Development Indicators*, World Trade Organization, *Trade Statistics*.

Table 1.2 Indicators of Comparative Advantage, 2007		
Country	Major Net Export Goods	Major Net Import Goods
United States	MATERIALS; SCI EQUIP; IND MACH; CHEMICALS	PETROLEUM; APPAREL; BEVERAGES
Canada	MATERIALS; PETROLEUM; EDIBLE OILS	APPAREL; ELECTRICAL MACHINERY; BEVERAGES
Mexico	BEVERAGES; PETROLEUM; APPAREL; TRANSP	EDIBLE OILS; CHEMICALS; SCI EQUIP; IND MACH
United Kingdom	IND MACH; CHEMICALS; BEVERAGES	APPAREL; FOOD; EDIBLE OILS
Germany	TRANSP; IND MACH; SCI EQUIP	PETROLEUM; EDIBLE OILS; APPAREL
Australia	MATERIALS; FOOD; BEVERAGES	APPAREL; TRANSP; IND MACH; ELEC MACH
Japan	TRANSP; IND MACH; SCI EQUIP	PETROLEUM; APPAREL; FOOD; BEVERAGES
Korea, Rep.	TRANSP; SCI EQUIP; ELEC MACH	EDIBLE OILS; FOOD; MATERIALS; PETROLEUM
Singapore	CHEMICALS; MISC. MFG.	FOOD; TRANSP; APPAREL
China	APPAREL; MISC. MFG; FOOD	MATERIALS; PETROLEUM; EDIBLE OILS
India	APPAREL; MISC. MFG; FOOD; BEVERAGES	EDIBLE OILS; PETROLEUM; IND MACH; SCI EQUIP
Brazil	MATERIALS; FOOD; BEVERAGES; APPAREL	SCI EQUIP; ELEC MACH; CHEMICALS; PETROLEUM
South Africa	MATERIALS; BEVERAGES; FOOD	EDIBLE OILS; ELEC MACH; SCI EQUIP

Source: United Nations Conference on Trade and Development, *Comtrade Country Profiles*, 2-digit SITC Revision 3.
Apparel = textiles, apparel and footwear; Beverages = beverages and tobacco; Chemicals = chemical products and pharmaceuticals; Edible Oils = animal and vegetable fats and oils; Elec Mach = electrical machinery and computers; Food = agricultural and manufactured food products; Ind Mach = industrial machinery; Materials = crude materials; Misc Mft = miscellaneous manufactured products; Petroleum = oil and natural gas; Sci Equip = scientific and photographic equipment; Transp = transportation equipment and motor vehicles.

Table 1.3 Foreign Direct Investment Stocks and Technology Payments and Receipts								
	Inward Stock/GPD		Outward Stock/GDP		Technology Receipts ($m)		Technology Payments ($m)	
Country	1980	2005	1980	2005	1981	2005	1981	2005
United States	3	13	7.8	16.4	$ 7,284	$ 74,826	$ 650	$ 31,851
Canada	20.4	31.6	8.9	35.3	$ 157	$ 2,474	$ 416	$ 1,222
Mexico	3.6	27.3	0.1	3.6	$ 33	$ 180	$ 274	$ 2,094
UK	11.8	37.1	15	56.2	$ 965	$ 30,676	$ 798	$ 14,867
Germany	3.9	18	4.6	34.6	$ 934	$ 34,307	$ 1,479	$ 29,756
Switzerland	7.9	46.9	20	107.4	na	$ 9,799	na	$ 10,900
Australia	7.9	29.8	1.4	22.5	$ 14	$ 2,578	$ 143	$ 3,566
Japan	0.3	2.2	1.8	8.5	$ 794	$ 18,403	$ 1,177	$ 6,385
R of Korea	2.1	8	0.2	4.6	na	$ 1,625	na	$ 4,525
Singapore	52.9	158.6	31.7	94.1	na	na	na	na
China	0.5	14.3	0	2.1	na	na	na	na
India	0.2	5.8	0	1.2	na	na	na	na
Brazil	7.4	25.4	16.4	9	na	na	na	na
South Africa	20.5	29	7.1	16.1	na	$ 45	na	$ 1,071

Sources: United Nations Conference on Trade and Development, *World Investment Report*, 2004 and 2006 editions; Organization for Economic Cooperation and Development, *Technology Balance of Payments*.

Table 1.4 Foreign-Born as Percentage of Population				
Country	1980	1990	2000	2005
United States	6.2	7.9	11	12.6
Germany	9.5	10.1	12.5	12.9
United Kingdom	6.2	6.7	7.9	9.7
Australia	21.1	22.8	23.0	23.8

Source: Organization for Economic Cooperation and Development

Chapter 2

PRODUCTION, SUPPLY AND PRODUCTION POSSIBILITIES

2.1 Properties of production functions

Many of the *causes* of international trade are found in the *differences* between countries in their abilities to produce different goods. These varying abilities are in turn related to underlying aspects of production, such as technologies, factor endowments, competitive conditions, government taxes and subsidies, and returns to scale. However, other aspects of technologies, most notably increasing returns to scale, can create trade and gains from trade for even identical economies. An understanding of these considerations will ultimately help us to understand why the US, for example, exports aircraft and cereal grains, and imports clothing. These same considerations will help us understand the *consequences* of trade, including overall welfare gains and the distribution of those gains between the members of a society.

The purpose of this chapter is to introduce the basic building blocks from production theory that will be used repeatedly in subsequent chapters, so your investment in time now will pay handsomely throughout. These include productivity, returns to scale, factor substitution and factor intensity. We show how these properties at the level of individual industries combine with competition in factor and goods markets to determine the position, curvature, and slope of the economy's production frontier and competitive general equilibrium.

Let X denote a particular good, and let two factors of production, such as capital and labor, be denoted by V_1 and V_2. A basic building block of the supply side of our model will be the production function, as represented by equation 2.1:

$$X = \alpha F(V_1, V_2) \tag{2.1}$$

where α is some scaling parameter. This production function is characterized by a number of properties that will used repeatedly throughout the book. These are as follows:

(A) Total factor productivity. This refers to the amount of output that can be produced from a given supply of inputs. While productivity is of course buried in the function F( ) itself, we will represent total factor productivity in 2.1 by the parameter α: higher values of α indicate higher total factor productivity. Productivity change defined as a change in α is actually a rather special case termed "neutral" or "Hicks neutral" technical change in that it does not affect any factor of production more than any other. More generally, technical change may be "biased" toward one factor or the other leading, for example, to higher productivity for capital or skilled labor. We shall not have too much to say about the latter in this book.

(B) Returns to scale. The term returns to scale indicates how output X responds to an equi-proportionate change in the input of *all* factors together. For example, how does output change if we double the inputs of all factors. If output exactly doubles, this is referred to as constant returns to scale. If output more than doubles, we call this increasing returns to scale. We will focus on these two cases throughout the book.

Consider the following special cases of (2.1) in which X uses only one factor of production, call it labor for example.

$$\text{(i) } X = V^{\gamma},\ \gamma = 1 \quad \text{(ii) } X = V^{\gamma},\ \gamma > 1 \quad \text{(iii) } X = \max[V - FC, 0] \tag{2.2}$$

The first of these three special case exhibits constant returns to scale: output is strictly proportional to input. Case (ii) is characterized by increasing returns to scale. Curves showing output X as a function of V are shown in Figure 2.1.

Figure 2.1

Case (iii) seems conceptually more complicated but it is actually simpler in many ways than case (ii) and is used extensively in trade theory. (iii) exhibits increasing returns to scale due to fixed costs of starting production. (iii) is read "X is equal to the maximum of zero and $(V - FC)$ where FC can be thought of an investment or fixed cost of V required before any actual output is produced. Case (iii) is shown in Figure 2.2. Output is zero until the critical level of input $(V = FC)$ is reached, and thereafter the *marginal* relationship between input and output is constant.

Figure 2.2

(C) Factor substitution and diminishing marginal product. Now return to the case in (2.1) where there are two factors of production. Here are three special cases:

$$\text{(i) } X = V_1 + V_2 \qquad \text{(ii) } X = \min[V_1, V_2] \qquad \text{(iii) } X = V_1^{\beta} V_2^{1-\beta} \quad 0 < \beta < 1 \tag{2.3}$$

In the first of these, case (i), the two factors of production are said to be *perfect substitutes*. An isoquant, showing combinations of V_1 and V_2 that produce one unit of X is shown in Figure 2.3: the isoquant is a straight line. Case (ii) is where V_1 and V_2 *are perfect complements*: they must be used together in equal amounts, and adding additional units of one factor has no added effect on output at all. An isoquant for producing one unit of X is a right-angle curve shown in Figure 2.3.

Figure 2.3

Case (iii) of (2.3) is a special case of the relationship most often shown in microeconomics textbooks. It is a case of smooth substitution and curvature between V_1 and V_2 and this particular special case is called the Cobb-Douglas production function. An isoquant for (iii) is also shown in Figure 2.3 (drawn for $\beta = 1/2$). The relationship between V_1 and V_2 needed to produce a constant amount of X, the slope of an isoquant, is known as the marginal rate of substitution (*MRS*) and can be found by differentiating the production function, holding output X constant.

$$dX = 0 = \beta V_1^{\beta-1} V_2^{1-\beta} dV_1 + (1-\beta) V_1^{\beta} V_2^{-\beta} dV_2 \tag{2.4}$$

$$MRS = \frac{dV_2}{dV_1} = \frac{\beta}{1-\beta} \frac{V_2}{V_1} \tag{2.5}$$

Note that the slope of the isoquant in (2.5) displays the curvature shown in Figure 2.3 and indeed the isoquant never touches the axes in this special case: positive amounts of each factor are necessary for positive output. This curvature is known as a *diminishing marginal rate of substitution*.

Production functions which display a diminishing marginal rate of substitution typically also display *diminishing marginal products* of each factor: increasing the input of one factor holding the other factor constant increases output at a diminishing rate. So, for example, if we increase the input of factor 1 holding factor 2 constant, the first partial derivative (marginal product MP_1) holding V_2 constant and second partial derivative of the Cobb-Douglas function are as follows:

$$MP_1 \equiv \frac{\partial X}{\partial V_1} = \beta V_1^{\beta-1} V_2^{1-\beta} > 0, \qquad \frac{\partial^2 X}{\partial V_1^2} = \beta(\beta-1) V_1^{\beta-2} V_2^{1-\beta} < 0 \tag{2.6}$$

where the second partial is negative since $(\beta - 1) < 0$. A negative second partial derivative is indeed a more formal definition of diminishing marginal product.

An important point is that a production function may exhibit both constant returns to scale and diminishing marginal products (and diminishing marginal rate of substitution). You can check and see that all three production functions in (2.3) have constant returns to scale. You can check this for the Cobb-Douglas case (iii) by multiplying all inputs by the constant of proportionality λ.

$$X = (\lambda V_1)^{\beta} (\lambda V_2)^{1-\beta} = \lambda (V_1^{\beta} V_2^{1-\beta}) \tag{2.7}$$

Doubling both inputs doubles output: constant returns to scale (2.7) occurs with diminishing marginal product (2.6).

2.2 Equilibrium for a single producer

Competition, or more specifically perfect competition, is defined as a situation where an individual firm is sufficiently small relative to the market that it views the prices of its output and inputs as exogenous. Denote the price of a firm's X output at p, and the prices of two inputs as w_1 and w_2. Let partial derivatives of the production function $F( )$ be denoted with a subscript so that, for example, $F_1( )$ is the partial derivative of F with respect to its first argument, V_1, also known as the marginal production of V_1 in producing X. Profits, denoted Π, for the firm are given by revenues minus costs, and the two first-order conditions for determining the optimal use of the two inputs are two equations to determine the two unknowns (the two inputs levels).

$$\Pi = pF(V_1, V_2) - w_1 V_1 - w_2 V_2 \tag{2.8}$$

$$\frac{\partial \Pi}{\partial V_1} = pF_1(V_1, V_2) - w_1 = 0 \qquad F_1 \equiv \frac{\partial F(V_1, V_2)}{\partial V_1} = MP_1 \tag{2.9}$$

$$\frac{\partial \Pi}{\partial V_2} = pF_2(V_1, V_2) - w_2 = 0 \qquad F_2 \equiv \frac{\partial F(V_1, V_2)}{\partial V_2} = MP_2 \tag{2.10}$$

Equations (2.9) and (2.10) are generally referred to as *value-of-marginal-product* conditions. Firms hire a factor up to the point where the *value* of its marginal product, output price multiplied by its physical marginal product, equals the factor price. Dividing (2.10) by (2.9) we have the well-known result that the ratio of the marginal products, the marginal rate of substitution, should equal the factor-price ratio.[1]

$$MRS = \frac{MP_1}{MP_2} = \frac{F_1}{F_2} = \frac{w_1}{w_2} \tag{2.11}$$

Given what we have just learned, this is a convenient place to briefly introduce the term *factor intensity*. This is generally used to make comparisons across industries about the degree to which an industry uses certain factors of production more intensively than others. Specifically, we can define it in our two-factor case as a ranking across industries of the ratio of factor use V_2/V_1 at a common factor-price

ratio w_1/w_2 (common *MRS*). Suppose that we have two industries (X_1, X_2). Let V_{ij} denote the *optimal* amount of factor j needed in the production of good i (when there is a double subscript indicating an industry and a factor, the first subscript denotes the industry, the second the factor). MRS_i be the marginal rate of substitution in industry i. Industry 2 is intensive in factor 2 and industry 1 is intensive in factor 1 if

$$\frac{V_{22}}{V_{21}} > \frac{V_{12}}{V_{11}} \quad \text{at} \quad MRS_2 = MRS_1 = \frac{w_1}{w_2} \Rightarrow \text{industry 2 is factor 2 intensive} \tag{2.12}$$

This is shown in Figure 2.4, where good 2 is intensive in factor 2 and good 1 intensive in factor 1. As a specific example, suppose both goods have Cobb-Douglas technologies as in (2.3) above, but with different values of β and specifically $\beta_1 > \beta_2$. Then using (2.5), the optimal factor intensity ratios are given by

$$\frac{V_{22}}{V_{21}} = \frac{1-\beta_2}{\beta_2}\frac{w_1}{w_2} > \frac{V_{12}}{V_{11}} = \frac{1-\beta_1}{\beta_1}\frac{w_1}{w_2} \tag{2.13}$$

In the Cobb-Douglas case, a relatively higher value of β for an industry implies it is relatively more factor 1 intensive.

Figure 2.4

2.3 The production set and the production possibilities frontier

In many cases, we will not want to deal in detail with factor markets and individual industry production functions. We will want instead to take a short-cut and deal directly with the aggregate production set of the economy and in particular the production possibilities frontier of the economy.

(A) Total factor productivity: absolute and relative. Consider perhaps the simplest possible case of the production function in (2.1), with one factor of production and constant returns to scale in both of two industries. The single factor is in fixed supply $\bar{V}$ and is allocated between sectors 1 and 2 in the amounts V_1 and V_2. The production side of this economy is given by

$$X_1 = \alpha_1 V_1 \qquad X_2 = \alpha_2 V_2 \qquad V_1 + V_2 = \bar{V} \qquad -\frac{dX_1}{dX_2} = \frac{\alpha_2}{\alpha_1} = MRT \tag{2.14}$$

where the last equation follows from the fixed supply of *V* and *MRT* stands for *marginal rate of transformation*, the slope of the production frontier, defined as positive (the slope's absolute value).

The production frontier is shown in Figure 2.5. It is linear, and the end points reflect both the size of the factor endowment *V* and the *absolute* levels of productivity α_1 and α_2. We can think of the distance of the frontier from the origin as the economic size of the economy, a combination of the factor endowment and productivity. It is very important to note for analysis later in the book, however, that the slope of the frontier depends only on the ratio of the productivity parameters, their relative levels α_2/α_1, and not on their absolute levels. To summarize, the economic size of an economy depends on the absolute levels of productivity, and the economy's relative ability to produce goods depends upon the ratio or relative productivities.

Figure 2.5

(B) Returns to scale. Now complicate our example just a little and allow both sectors to have

increasing returns to scale as in (ii) of (2.2) and set the α's equal to one.

$$X_1 = \alpha_1 V_1^{\gamma} \quad X_2 = \alpha_2 V_2^{\gamma} \qquad \gamma > 1, \qquad V_1 + V_2 = \bar{V} \tag{2.15}$$

$$-\frac{dX_2}{dX_1} = \frac{V_2^{\gamma-1}}{V_1^{\gamma-1}}, \qquad V_i^{\gamma-1} = X_i^{\frac{\gamma-1}{\gamma}}, \qquad -\frac{dX_2}{dX_1} = \left[\frac{X_2}{X_1}\right]^{\frac{\gamma-1}{\gamma}} = MRT \tag{2.16}$$

The production frontier for this economy is shown in Figure 2.6; it is a convex curve or alternatively, the production set is non-convex. Intuitively, a point mid-way between the two endpoints, denoted A in the Figure, is not a feasible production point. With increasing returns to scale, taking half of the endowment out of good X_2 and giving it to X_1 generates less than half of the output of each good that would be achieved if all the endowment were allocated to just one of the goods.

Figure 2.6

Now consider the alternative formulation of scale economies: those arising due to fixed costs. Let our simple economy be given by

$$X_1 = \max[V_1 - FC, 0] \qquad X_2 = \max[V_2 - FC, 0] \qquad V_1 + V_2 = \bar{V} \tag{2.17}$$

The production frontier for this economy is shown in Figure 2.7. The maximum outputs of X_1 and X_2 require incurring a single fixed cost. Producing both goods, however, requires two fixed costs, and hence the linear segment of the frontier with both goods produced does not connect with the end points of the frontier. Note the similarity between Figure 2.6 and Figure 2.7. The former has a smooth curvature while the latter is "kinked". But both share the property that the production set is non-convex and the mid-point A of a line connecting the two endpoints of the frontier is not a feasible production point. This property proves important in later analysis.

Figure 2.7

(C) Factor endowments and factor intensities. Return now to the case where both industries have constant returns to scale, but assume that there are two or more factors of production. In this case, the production set must be convex (proof is omitted here) and indeed it is general strictly convex. For many technologies, the production frontier is a smooth concave ("bowed out") function. Let's look at two cases, both of which are widely used in trade models. The first is generally referred to as the specific-factors model (due mostly to Ronald Jones) and the second is the Heckscher-Ohlin model. The specific-factors model has two goods produced with three factors. Only one factor, we'll call it V, is mobile between the two sectors. The other two are fixed in their respective sectors, such as capital goods or resources that have no use in the other sector. We can denote these at K_1 and K_2 and these are treated as fixed. The production side of the economy is given by

$$X_1 = F_1(V_1, \bar{K}_1) \quad X_2 = F_2(V_2, \bar{K}_2) \quad V_1 + V_2 = \bar{V} \qquad -\frac{dX_2}{dX_1} = \frac{F_{21}(V_1, \bar{K}_1)}{F_{11}(V_2, \bar{K}_2)} \tag{2.18}$$

where F_{i1} is the marginal product of V_1 in producing X_i. Assume that the law of diminishing returns holds, so that the marginal product of V in each sector is decreasing in the ratio of V to the specific factor K. The production frontier will be strictly concave, or "bowed out" as shown in Figure 2.8

Figure 2.8

The fact that the production frontier is strictly concave (production set strictly convex) is a fairly general result provided that there are at least as many factors as goods and, of course, the goods must differ in their optimal factor intensities. A second case that forms one of the work-horse models of traditional trade theory, the Heckscher-Ohlin model, has two goods and two factors, with both factors useful in both industries.

$$X_1 = F_1(V_{11}, V_{12}) \quad X_2 = F_2(V_{21}, V_{22}) \quad V_{11} + V_{21} = \bar{V}_1 \quad V_{12} + V_{22} = \bar{V}_2 \tag{2.19}$$

The factor-market side of the economy is typically described via an Edgeworth box, as shown in Figure 2.9. The horizontal axis gives the economy's total supply of V_1 and the vertical dimension its total supply of V_2. The origin for measuring inputs into the X_1 industry is at the lower left or southwest corner, and the origin for measuring inputs into the X_2 industry is at the upper right or northeast corner. Any point in the box is a division or allocation of the total endowment between the two industries.

Figure 2.9

Not all allocations (points in the box) are efficient. The set of efficient or specifically *Pareto efficient* allocations are given by tangencies between X_1 and X_2 isoquant. The set of all such efficient allocations is called the *contract curve*. If X_1 is intensive in V_1 (refer back to Figure 2.4), then the contract curve in Figure 2.9 must lie below the diagonal of the Edgeworth box. That is, the slope of a ray from O_1 to a point on the contract curve must have a flatter slope (lower V_2/V_1) than the slope of a ray from the O_2 to that point. This is the case shown in Figure 2.9.

A somewhat informal proof that the production frontier for this economy is strictly concave ("bowed out") is as follows. Consider the (inefficient) allocations on the diagonal of the Edgeworth box in Figure 2.9. Due to constant returns in both industries, if we plot the values of the X_1 and X_2 isoquants as we move up this diagonal in the output diagram Figure 2.10, we will get a straight line. Point A in Figure 2.9 corresponds to point A in Figure 2.10. However, this is not an efficient allocation. We could move from allocation A in the Edgeworth box to point B and get more X_2 output without reducing X_1 output. Or we could move from A to allocation C, which gives us more X_1 output with the same X_2 output as at inefficient allocation A. Point B and C in the Edgeworth box Figure 2.9 correspond to the efficient production points B and C on the production possibilities curve in Figure 2.10. The production frontier is strictly bowed out as shown in Figure 2.10.

Figure 2.10

2.4 Competitive equilibrium

Although we now know something about the production frontier, (a) we have not shown that the economy operates on this frontier nor (b) that equilibrium will occur at the best point on the frontier. However, for a competitive, undistorted economy this will indeed be the case. This result is know casually as Adam Smith's "invisible hand" and more formally in economic theory as the first theorem of welfare economics. The term distortions refers to anything (in addition to perfect competition) that leads to deviations between prices and marginal costs of production or cause different agents to face different prices for the same good or factor. Taxes and subsidies are common forms of distortions and will be treated in detail later in the book. There are two common approaches to showing the Smithian result, and we have simplified them here, particularly by exploiting the assumption that the total supplies of all factors are fixed.

(A) Approach I: calculus of optimization. First-order conditions equating the value of marginal product of a factor to its price were given in (2.9) and (2.10) above. If we have two industries and two factors, these conditions are given by

$$p_1 F_{11}(V_{11}, V_{12}) = w_1 \qquad p_1 F_{12}(V_{11}, V_{12}) = w_2 \tag{2.20}$$

$$p_2 F_{21}(V_{21}, V_{22}) = w_1 \qquad p_2 F_{22}(V_{21}, V_{22}) = w_2$$

The first implication of these condition is factor market allocation efficiency. If we divide the first equation in each row of (2.20) by the second, we see that

$$MRS_1 = \frac{F_{11}}{F_{12}} = MRS_2 = \frac{F_{21}}{F_{22}} = \frac{w_1^0}{w_2^0} \tag{2.21}$$

The fact that the MRS in each industry is equal tells us that the competitive outcome must be a factor-market allocation on the contract curve in Figure 2.9. This in turn says that the competitive equilibrium lies on the production frontier.

Second, we can differentiate the production function for each good, and replace the physical marginal products F_{ij} in (partial derivatives) with w_j/p_i from (2.20).

$$dX_1 = \sum_j F_{1j} dV_{1j} = \sum_j \left[\frac{w_j}{p_1}\right] dV_{1j} = \frac{1}{p_1}\sum_j (w_j dV_{1j}) \tag{2.22}$$

$$dX_2 = \sum_j F_{2j} dV_{2j} = \sum_j \left[\frac{w_j}{p_2}\right] dV_{2j} = \frac{1}{p_2}\sum_j (w_j dV_{2j})$$

The summations over the factors on the right-hand side are simply one subtracted from the other: an increase in factor j to industry i must mean an equal decrease in supply from the other industry. $dV_{1j} = -dV_{2j}$. Thus when we divide the second equation of (2.22) by the first, we get

$$MRT = -\frac{dX_2}{dX_1} = \frac{p_1}{p_2} \tag{2.23}$$

This result says that competitive equilibrium is a tangency between the slope of the production frontier and the equilibrium price ratio. This is shown in Figure 2.11, where equilibrium relative price p_1/p_2 is denoted p^0. We can think of the price ratio through the production point as essentially a budget line, and competitive equilibrium puts us on the highest budget line possible.

Figure 2.11

(B) Approach II: profit maximization and revealed preference. The approach to proving production efficiency in a competitive economy using the calculus of optimization is useful at many points later and has a fairly clear and intuitive graphical interpretation. Unfortunately it is somewhat limited, particularly with the respect to "corner solutions" in which an economy does not produce all available goods. In this sub-section, we introduce an alternative methodology that we will use quite a number of times throughout the book.

This is an old, long-standing methodology based on the concept of revealed preference. The idea is straightforward: if we see an optimizing agent choosing an action or alternative A when B is a feasible alternative, then A must yield higher profits or utility than B. The crucial word "feasible" depends on the context. In the case of producers, it refers to an alternative output and input vector that is technologically feasible: those alternative inputs are sufficient to produce the alternative outputs. In the case of consumers, it typically refers to cost: if you buy A when B costs no more or strictly less, then you reveal that you like A more than B. If you buy A when B cost more than A, no preference ranking can be

inferred; that is, you may or may not like B more than A.

Now consider profit-maximizing competitive producers, again with fixed aggregate factor supplies. Suppose that we observed that the output vector X^0 , and the input matrix V^0 are chosen at commodity and factor prices p^0, w^0. Suppose that X^1 , V^1 is any alternative *feasible* production plan.

For each industry i, profit maximization implies that

$$p_i^0 X_i^0 - \sum_j w_j^0 V_{ij}^0 \geq p_i^0 X_i^1 - \sum_j w_j^0 V_{ij}^1 \tag{2.24}$$

That is, if competitive producers are optimizing, the actual output and inputs chosen at these prices must yield profits which are greater than or equal to the profits obtained from any other feasible production plan. Sum over all i industries

$$\sum_i p_i^0 X_i^0 - \sum_i \sum_j w_j^0 V_{ij}^0 \geq \sum_i p_i^0 X_i^1 - \sum_i \sum_j w_j^0 V_{ij}^1 \tag{2.25}$$

Suppose that endowments are fixed at $\bar{V}$. For each factor j

$$\sum_i w_j^0 V_{ij}^0 = \sum_i w_j^0 V_{ij}^1 = w_j^0 \bar{V}_j \tag{2.26}$$

Thus the summation terms on the two sides of (2.24) cancel out. Therefore, the outputs X^0 chosen at prices p^0 maximize the value of production at those prices: (2.24) becomes

$$\sum_i p_i^0 X_i^0 \geq \sum_i p_i^0 X_i^1 \tag{2.27}$$

Geometrically, all feasible production points such as X^1 must lie on or below the price hyperplane p^0 through the actual production point X^0. The price plane is "supporting" to the production set.

This is the same result shown in Figure 2.11 using a rather different approach. The revealed-preference approach also covers corner solutions, such as that shown in Figure 2.12. With the price ratio flatter than the slope of the linear production possibilities curve, the equilibrium is to produce only X_2 at price ratio p^0.

Figure 2.12

2.5 Cost functions

Associated with any production function (which embodies the assumption of technical efficiency) is a cost function, which gives the *minimum* cost of producing a given output at given input prices. This cost function, which embodies both technical and economic efficiency, is said to be *dual* to the production function. The word *duality* in economics is used in several different but related senses. It can refer the fact that for a certain maximization problem (maximize profits) there is an equivalent minimization problem (minimize costs), or it can refer to a change of variables or functions that carry the same information, such as switching from a production function defined over input quantities to a cost function defined over input prices. Assume that the production function *F* has constant returns to scale. The dual of the profit maximization problem in (2.8) - (2.10) is a cost minimization problem

$$\min\ w_1 V_1 + w_2 V_2 \quad \text{subject to } F(V_1, V_2) \geq X \quad \text{yields} \quad c(w_1, w_2)X \tag{2.28}$$

Constant returns to scale in production in turn imply that total costs are linear in output. We cannot devote a lot of space here to this issue, and readers are referred to Varian's classic textbook for a better understanding. What we can do is simply illustrate cost functions for the three production functions given in (2.3) above.

Derivation of the cost functions is left as an (very worthwhile!) exercise for the reader. We have added some scaling parameters such that, at factor prices $w_1 = w_2 = 1$, all three cases yield a minimum unit production cost of one.

$$\text{(i)} \quad X = V_1 + V_2 \qquad c(w_1, w_2)X = \min[w_1, w_2]X$$

$$\text{(ii)} \quad X = \min[2V_1, 2V_2] \qquad c(w_1, w_2)X = \frac{(w_1 + w_2)}{2}X$$

$$\text{(iii)} \quad X = \left[\frac{V_1}{\beta}\right]^{\beta}\left[\frac{V_2}{1-\beta}\right]^{1-\beta} \qquad c(w_1, w_2)X = (w_1^{\beta} w_2^{1-\beta})X \tag{2.29}$$

The intuition is as follows. (i) is the case of perfect substitutes in production: the isoquant is linear and therefore the firm will generally use only one input, *either V_1 or V_2*, the one that is cheaper. Therefore the cost function, the minimum cost of producing one unit, is just the minimum of w_1 and w_2. (ii) is the case where the factor inputs are perfect complements. The firm must use *exactly one unit of V_1 and one unit of V_2* to produce one unit of X. Thus the unit cost function is the *sum* of w_1 and w_2. It is interesting to note and important to understand a feature of (i) and (ii). When the production function has a linear isoquant (i), the cost function has as right angle-isocost curve. When the production function has a right-angle isoquant, the isocost curve is linear.

The Cobb-Douglas case in (iii) is something that you should work out as an exercise. Note that if the production function is Cobb-Douglas, so is the cost function; the two look similar except for a scaling parameter (we have put this on the production function so that the cost function looks quite simple). Note finally that constant returns in production also appears in all three cost functions: total cost is linear in output, doubling output exactly doubles costs at constant factor prices.

As a final point in this section, we can use the result in (2.29) to illustrate Shepard's lemma, a special case of the envelope theorem, a result extremely usual for theoretic work, empirical analysis, and computational models. Shephard's lemma gives the result that the derivative of the unit cost function with respect to w_1 , for example, is the optimal amount of V_1 used to produce one unit of good X. Let $a_i(w_1, w_2)$ denote the *optimal* (cost minimizing) amount of V_i needed to produce one unit of X at these prices. Then

$$c(w_1, w_2) = w_1 a_1(w_1, w_2) + w_2 a_2(w_1, w_2) \tag{2.30}$$

$$\frac{dc(w_1, w_2)}{dw_1} = a_1 + \left[w_1\frac{\partial a_1}{\partial w_1} + w_2\frac{\partial a_2}{\partial w_1}\right] = a_1 = \frac{\partial c(w_1, w_2)}{\partial w_1} \tag{2.31}$$

The fact that a_i's are optimally chosen initially means that expression in brackets is zero: small changes in the a's have no effect on cost. (Think of cost as a function of a_i as U-shaped, with optimization meaning that we are at the bottom of the U: a small change in a_i has no effect on cost.) Thus the *total derivative* of unit cost due to a small change in a factor price is just the *partial derivative* or direct effect on cost, which, in turn, is the initially-optimal input level. As a specific example, the optimal amounts of V_1 and V_2 to produce one unit of output in our Cobb-Douglas example are

$$a_1 = \beta\left[\frac{w_2}{w_1}\right]^{1-\beta} \qquad a_2 = (1-\beta)\left[\frac{w_1}{w_2}\right]^{\beta} \tag{2.32}$$

We will see important applications of this tool several times throughout the book.

2.6 A note on increasing returns to scale and imperfect competition

We will conclude this chapter with a brief note on increasing returns technologies and competition, or rather the general incompatibility between increasing returns and perfect competition. Let's use the fixed-cost technology in equation (iii) of (2.2) and (2.17). The total cost function, average cost function (*tc/X*) and the marginal cost function for this technology are given by

$$tc = wX + wFC \qquad ac = w + w\frac{FC}{X} \qquad mc = \frac{dtc}{dX} = w \tag{2.33}$$

The average and marginal cost curves for this technology are shown in Figure 2.13. Average cost is a hyperbole that approaches, but never actually equals, marginal cost. The incompatibility between this technology and a market structure of perfect competition in which firms are assumed to face a fixed prices is straightforward. If the (fixed) price is equal to marginal cost, the firm can never break even, since marginal cost is always less than average cost. If the price is even a little above marginal cost, then there is some output level at which the firm breaks even and the firm has an incentive to expand output to infinity at that fixed price. That cannot be an equilibrium. Thus this technology cannot support perfect competition as an equilibrium market structure. Much more will be said about this in later chapters.

Figure 2.13

2.7 Summary

This chapter has developed tools and ideas that will be used repeatedly throughout the text. You should have an understanding of how productivities, returns to scale, factor endowments and factor intensities contribute to determining the position, curvature, and slope of the production frontier in the two-good case. You should also have an understanding of how the assumption of perfect competition in goods and factor markets (and the absence of other distortions as well) will lead to efficient output at equilibrium prices. Where these prices come from now leads us to turn to the consumer demand side of the market, and then to general-equilibrium in Chapter 4.

Endnotes

1. There is a problem here if the firm's technology has constant returns to scale. In that case, it can be shown that the marginal products of factors and hence the first-order conditions in (2.9) and (2.10)are invariant with respect to a proportional increase or decrease in the use of all factors. This means the *output level* for a single competitive firm is *indeterminate*. Competitive general-equilibrium theory generally ignores this problem by referring to industries and not to individual firms. We shall continue that tradition here, but will also show the problem formally when discussing homogeneous functions in section 3.2.

Figure 2.1

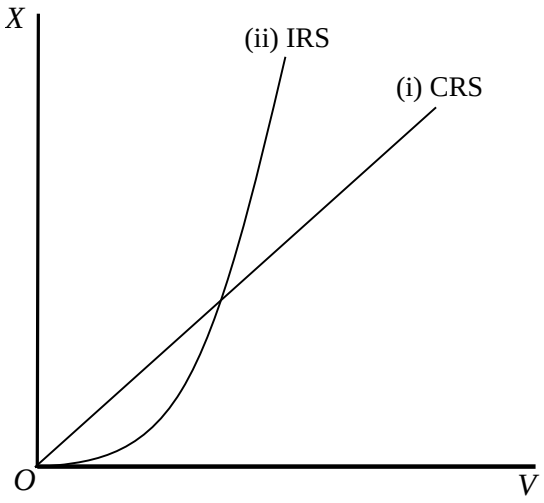


Figure 2.2

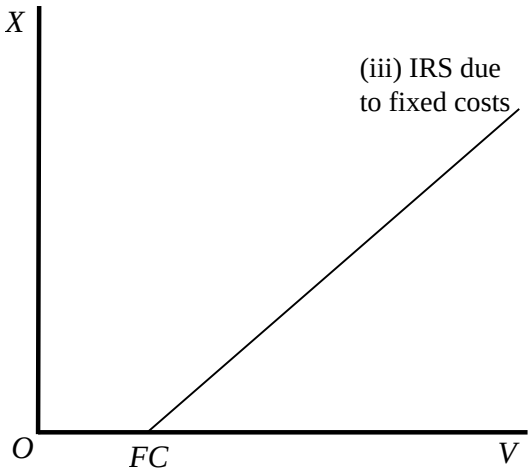


Figure 2.3

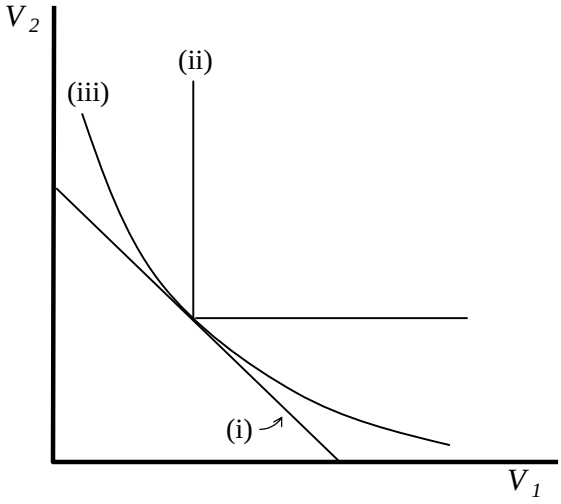


Figure 2.4

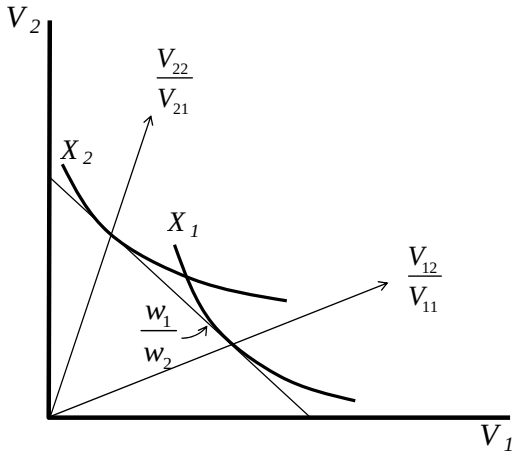


Figure 2.5

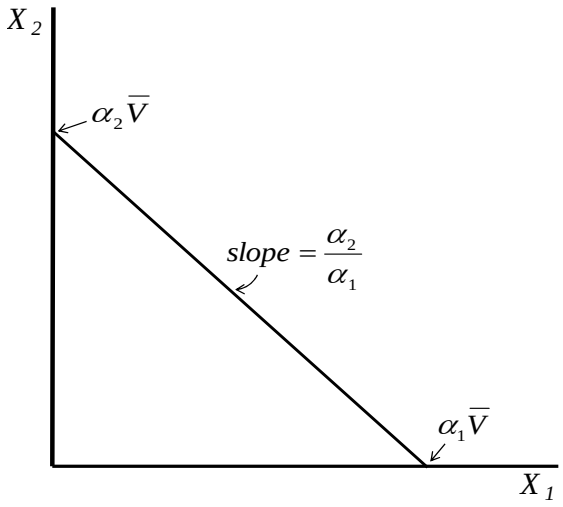


Figure 2.6

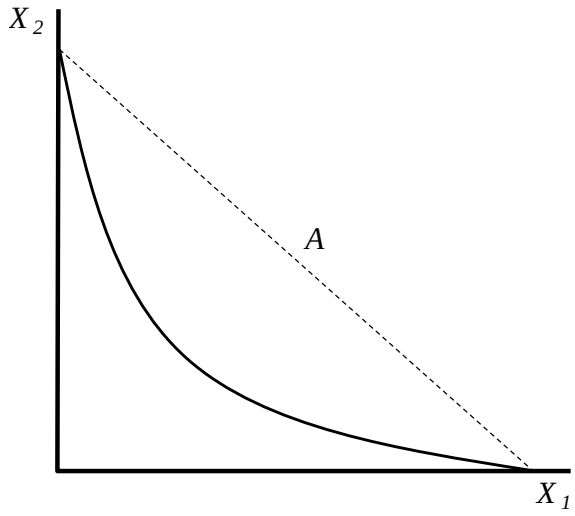


Figure 2.7

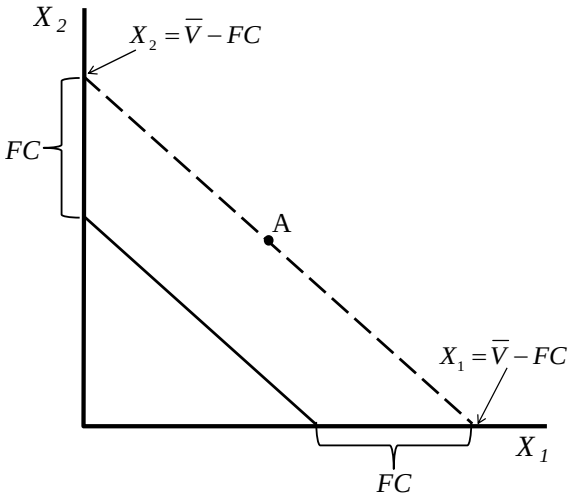


Figure 2.8

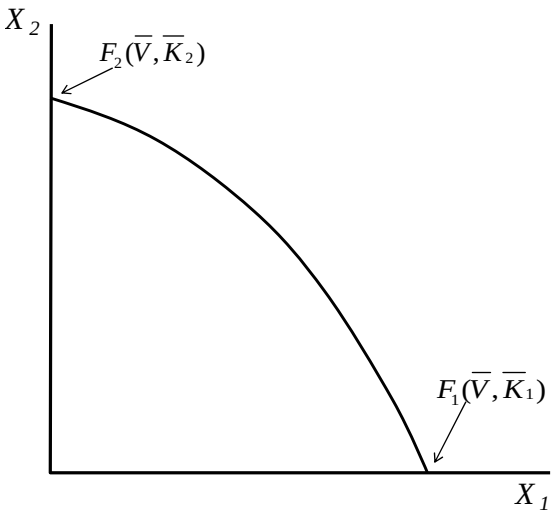


Figure 2.9

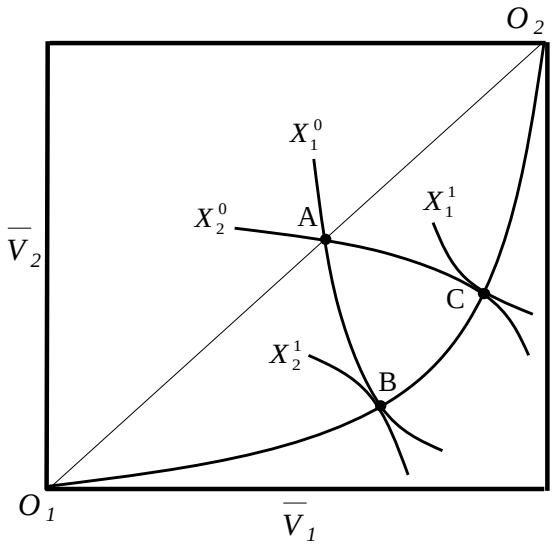


Figure 2.10

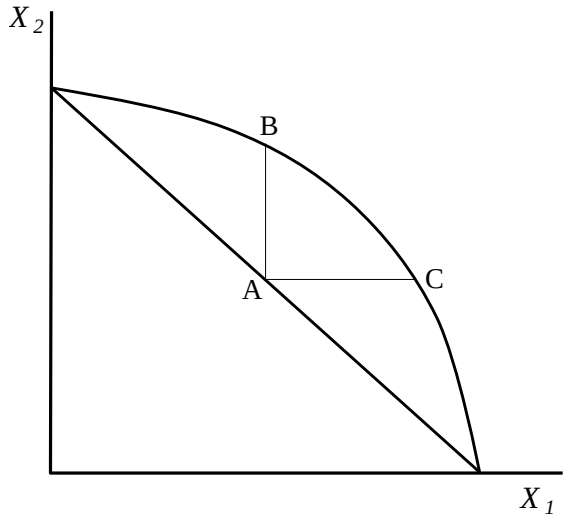


Figure 2.11

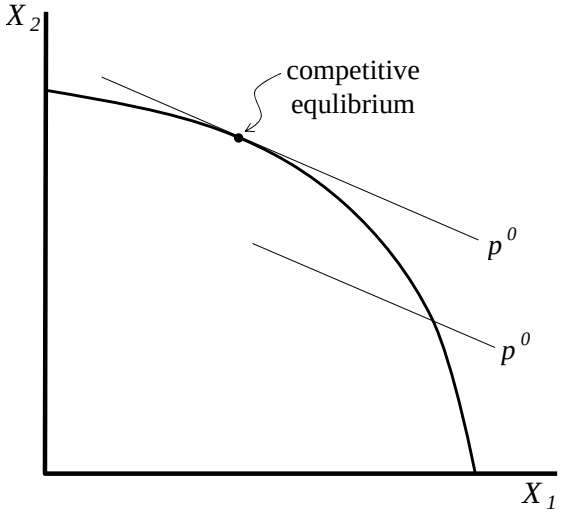


Figure 2.12

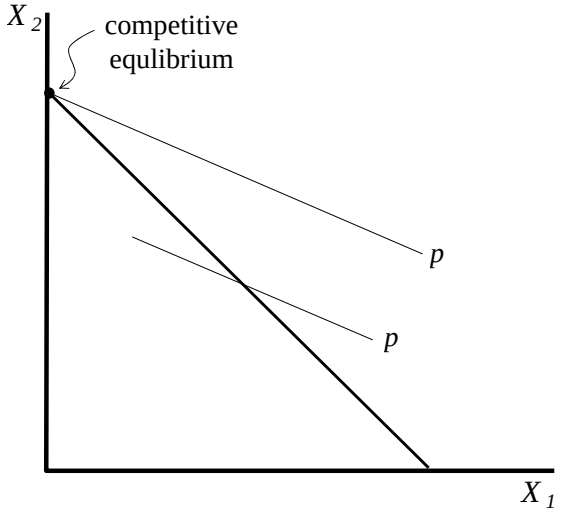
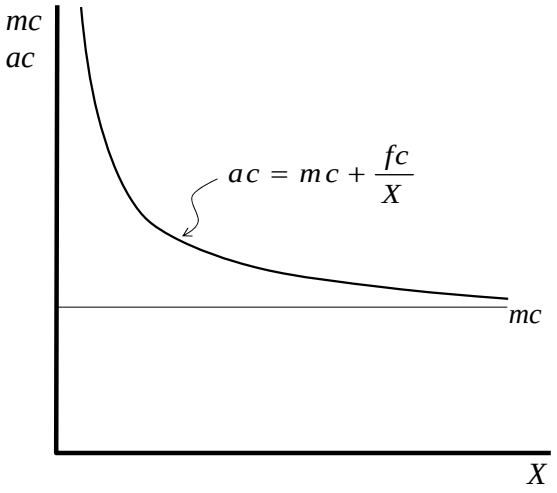


Figure 2.13



Chapter 3

PREFERENCES, DEMAND, AND WELFARE

3.1 Optimization for a single consumer

Students will recall the standard treatment of utility and consumer theory from intermediate microeconomics. A text typically starts with the problem of a single consumer maximizing a utility function *U(X)* subject to a linear budget constraint on total expenditure. Usually, this is set up as a Langrangean function, with a Lagrangean multiplier λ on the income constraint. Letting I denote income, the standard treatment is given by

$$\max U(X_1, X_2) + \lambda(I - p_1X_1 - p_2X_2) \tag{3.1}$$

$$\frac{\partial U}{\partial X_1} - \lambda p_1 = 0 \qquad \frac{\partial U}{\partial X_2} - \lambda p_2 = 0 \tag{3.2}$$

$$I - p_1X_1 - p_2X_2 = 0 \tag{3.3}$$

There are three first-order conditions in (3.2) and (3.3), which can be solved for the optimal quantities of the two goods. The utility-maximizing demand for a good is a function of prices and income, and this particular way of specifying demand is generally referred to a as Marshallian demand function. As an exercise, derive the Marshallian demand function for the Cobb-Douglas function in Chapter 2, ((2.29) iii), with the two goods replacing the two factors as the function's arguments. Show that the demands are given by

$$X_1 = \frac{\beta I}{p_1} \qquad X_2 = \frac{(1-\beta)I}{p_2} \tag{3.4}$$

The well-known graphical representation of this optimization problem is shown in Figure 3.1. The goal is to reach the highest indifference curve, subject to being on the (fixed) budget constraint, and the solution is the tangency at point A.

Figure 3.1

For a lot of theoretical work, especially with numerical simulation models, it turns out to be more convenient to work with the dual problem, that of minimizing the expenditure necessary to reach a given level of utility such as $\bar{U}$. This yields demand functions that are functions of prices and the target utility level, often referred to as Hicksian demand functions.

$$\min p_1X_2 + p_2X_2 + \lambda(U(X_1, X_2) - \bar{U}) \tag{3.5}$$

$$p_1 - \lambda\frac{\partial U}{\partial X_1} = 0 \qquad p_2 - \lambda\frac{\partial U}{\partial X_2} = 0 \tag{3.6}$$

$$\bar{U} - U(X_1, X_2) = 0 \tag{3.7}$$

If you solve the three equation in three uknowns in (3.6) and (3.7), you should derive the Hicksian demand functions for the Cobb-Douglas case of (2.29) iii and show that they are given by

$$X_1 = \beta\left[\frac{p_2}{p_1}\right]^{1-\beta} U \qquad X_2 = (1-\beta)\left[\frac{p_1}{p_2}\right]^{\beta} U \tag{3.8}$$

It is no surprise that these are so similar to the optimal input quantities in (2.29), since the production function there and the utility function here are assumed to be the same. If we then multiply each quantity in (3.8) and add them together, we get the minimum expenditure necessary to buy level of utility U, often referred to the expenditure function.

$$e(p_1, p_2)U = (p_1^{\beta} p_2^{1-\beta})U \tag{3.9}$$

This is naturally the same as the cost function for producing goods from the same Cobb-Douglas function in (2.29). Indeed, the expenditure function is precisely a cost function: the minimum cost necessary at prices *p* to buy one unit of utility. As a final exercise, apply Shepard's lemma to (3.9) to check that you get (3.8) as the optimal input choices.

The Hicksian optimization problem is shown in Figure 3.2. Here the target utility level U is fixed, the problem is achieve this utility level for the minimum expenditure. The solution is the tangency solution at point A, which puts on the lowest possibly budget line to achieve U.

Figure 3.2

3.2 A note on homogeneous functions

Often international trade models focus on the production side of economies and differences in things like technologies and factor endowments as causes of trade. When doing so, we generally assume no differences between countries with respect to demand so that the direction of trade (which countries export and import which goods) is purely determined on the production side. More specifically, it is useful for this reason and for computational reasons as well to assume that the ratio in which consumers demand goods depends only on relative prices on not on income: in the two-good case, the ratio of goods chosen in consumption, X_2/X_1, depends only on the ratio of prices, p_1/p_2. This is a property that characterizes all homogeneous functions. More formally, U is homogeneous of degree i if

$$U(\lambda X^0, \lambda Y^0) = \lambda^i U(X^0, Y^0) \tag{3.10}$$

where the superscript 0 denotes some specific initial values. A homogeneous function with i = 1 is, in economic terminology, constant returns to scale. Or to put it the other way around, a constant-returns-to-scale production or utility function must be homogeneous of degree 1. Differentiating both sides of (3.10), we have

$$U_1(\lambda X_1^0, \lambda X_2^0)\lambda dX_1 = \lambda^i U_1(X_1^0, X_2^0)dX_1$$

$$U_2(\lambda X_1^0, \lambda X_2^0)\lambda dX_2 = \lambda^i U_2(X_1^0, X_2^0)dX_2 \tag{3.11}$$

Diving the second equation of (3.11) by the first, we have

$$\frac{U_2(\lambda X_1^0, \lambda X_2^0)}{U_2(\lambda X_1^0, \lambda X_2^0)} = \frac{U_1(X_1^0, X_2^0)}{U_2(X_1^0, X_2^0)} \tag{3.12}$$

This result says that the marginal rate of substitution *MRS* depends only on X_2/X_1 ratio in consumption and not on the scale. To put it the other way around, holding prices constant, the consumer will choose the same X_2/X_1 ratio independently of income. This result is shown in Figure 3.3: the income expansion path (sometimes called an Engels' curve) is linear from the origin holding prices constant. This is an unrealistic assumption that is probably contradicted by every budget study ever done, but it is a useful one. It will be examined further in a later chapter.

Figure 3.3

Second, note from (3.11) that the marginal utility of a good (or of a factor in production theory) under constant return to scale (i = 1 in (3.10)) is invariant to a proportion change in the use of all goods (factors). That is, the marginal products are homogeneous of degree (i - 1), or homogeneous of degree zero in the case of constant returns to scale. From (3.10), we have

$$U_1(\lambda X^0, \lambda Y^0) = U_1(X^0, Y^0) \qquad U_2(\lambda X^0, \lambda Y^0) = U_2(X^0, Y^0) \tag{3.13}$$

We referred to this earlier in Chapter 2 but postponed discussion of the fact that the first-order conditions for maximizing profits in (2.9) and (2.10) are invariant with respect to proportional changes in inputs, and therefore the supply of a single competitive firm with constant returns is indeterminant.

While on the topic of homogeneous functions, it is interesting to note a couple of properties of the Marshallian and Hicksian functions that we derived above. Note that the Marshallian demand functions in (3.4) are homogeneous of degree zero in prices and income: demands will not change if we double all prices and income (the budget line doesn't move). The Hicksian demand functions in (3.8) are homogeneous of degree zero in prices: if we double all prices, the slope of the budget line doesn't change, so the cost-minimizing choices of X_1 and X_2 don't change. But the minimum expenditure needed to reach the target utility must change as shown in (3.9): the expenditure function is homogeneous of degree 1 in prices if *U* is homogeneous of degree 1. These are in fact very general results and are by no means restricted to the Cobb-Douglas special case.

3.3 Aggregating over households to a "community" utility function

While all of this should sound familiar and relatively straight forward, international trade deals with the whole economy and thus we need to aggregate up from the individual household to the full economy. It unfortunately turns out that this is not at all straight forward.

Suppose that individuals all have preferences of the type described in the preceding sections, which in turn gives rise to demand functions that depend on prices and income. Assume that there are only two goods and that the price ratio is denoted $p = p_1/p_2$. The income of individual *J* is denoted by I_j. If all individuals in the economy face the same price ratio but generally have different incomes, then the total demand for the good *X* can be written as the function *D*:

$$X = D(p, I_1, I_2, ..., I_n) \tag{3.14}$$

where it is assumed that there are *n* individuals in the country. The question of whether or not national or "community" indifference curves exist is nearly the same as the question of whether or not the demand function such as those in equation (3.8) can be written as a function of aggregate income (the sum of the individuals' incomes); that is, whether or not the distribution of income affects total demand.[1] The

intuition here is that when we draw community indifference curves, we are saying that the country has preferences over aggregate bundles of goods that depend only on prices (the slope of the price line) and total national income (the distance the price line is from the origin). Preferences and hence demands are independent of how that aggregate income is distributed. In short can we write total demand as

$$X = D(p, I) \qquad I = \sum_{j=1}^{n} I_j \tag{3.15}$$

Special assumptions are necessary for demand to be independent of the distribution of income and hence for (3.15) to be valid. One problem that arises in aggregation is shown in Figure 3.4, where we have two consumers with identical but nonhomogeneous tastes.[2] In other words, at constant relative prices, the ratio of X_2/X_1 consumed is not independent of income. Specifically, consumers desire more X_1 relative to X_2 as income increases at constant prices.

Figure 3.4

Suppose we have two individuals initially consuming at point *A* in Figure 3.4, but we take some income away from consumer 1 and give it to consumer 2 so that they adjust to consuming at points *B* and *C*, respectively. The chord *AB* is steeper than the chord *AC*, and so the changes in the consumption of the two individuals do not balance. Even though we have not changed either prices or aggregate income, there will now be a higher aggregate demand for X_1 and a lower aggregate demand for X_2. Aggregate demand depends on the distribution of income, and hence community indifference curves do not exist in this situation.

Now consider homogeneous but nonidentical tastes. This situation is shown in Figure 3.5. Consumer 2, who has a relatively strong preference for X_2, is initially at A_2, while consumer 1, who has a strong preference for X_1, is initially at A_1. Now take income away from consumer 2 and give it to consumer 1 holding prices constant. Consumer 2 moves to point B_2, while consumer 1 moves to point B_1 in Figure 3.5. Again, the changes in consumption do not cancel each other out (more X_1 and less X_2 will be demanded) even though prices and total income are constant. Community indifference curves do not exist when preferences differ.

Figure 3.5

This analysis suggests that community indifference curves will exist if all consumers have identical and homogeneous tastes (and, of course, face the same prices) as in Figure 3.3. This is indeed true. However, it is an extremely strong assumption that is nevertheless pervasive in trade theory. A somewhat weaker assumption will do, and it is sufficient simply that all consumers have income-expansion paths (Engel's curves) that are linear and parallel. Many industrial-organization models of trade in fact go to an extreme in order to avoid general-equilibrium income effects and assume that demand in the X_1 sector depends only on prices and not on income (an example of so-called quasi-linear preferences). Much more will be said on this later, but for completeness we show an example in Figure 3.6. If all consumers have these preferences, aggregation is possible since total demand does not depend on the distribution of income.

Figure 3.6

3.4 Interpreting community indifference curves: aggregate demand versus individual welfare

There are two different interpretations of community indifference curves (assuming they exist) and it is very important to distinguish between them. One is what economists call a *positive* interpretation. Under this interpretation, the community indifference curves simply tell us what the country will demand under various price and aggregate income combinations. That is, if we pick income

and prices to determine an aggregate budget line as in equation (3.3), we can find the quantities demanded by the intersection of the highest community indifference curve with the budget constraint. The positive interpretation of community indifference curves does not necessarily attach any welfare significance to the indifference curves.

The *normative* interpretation of community indifference curves does attach a positive welfare significance to moving from a lower indifference curve to a higher indifference curve (or even from moving along one indifference curve) in the same way that we would interpret that move for a single individual. If some trade policy can lead to such a movement, we say that the country is better off (or equally well off).

The pitfalls in the second interpretation are illustrated in Figure 3.7, where A and B are two aggregate commodity bundles and U_a and U_b two community indifference curves. Clearly, under the normative interpretation of community indifference curves, national welfare increases in a move from A to B. However, suppose that the country is composed of two individuals, denoted with subscripts 1 and 2, with identical homogeneous preferences. In the initial situation A, measure the consumption of individual 1 from the origin *O* and measure individual 2's consumption in the opposite direction beginning at *A* as in the Edgeworth box. As drawn in Figure 3.7, individual 1 has initial consumption *OA'* and individual 2 has consumption *A'A*. Suppose that the move from A to B somehow redistributes income from individual 2 to individual 1. This is of no consequence for the community indifference curves. However, the move from A to B clearly has the effect of greatly helping individual 1 but harming individual 2 whose consumption is reduced from *A'A* to *B'B*. The point is that in making normative interpretations of community indifference curves we must remember that *a movement to a higher community indifference curve does not mean that the welfare of all individuals in society has increased*. Similarly, moving along a single community indifference curve does not mean that the welfare of all individuals is being held constant.

Figure 3.7

In later chapters we will present cases where aggregate welfare increases but, due to income re-distribution via factor-price changes, some individuals are worse off. Indeed, this is a relatively common problem with trade policy and it is one of the points of departure in what is called the political-economy approach to trade policy. Throughout the book, we will often make normative interpretations of community indifference curves, but we must keep the caveat just mentioned in mind. One traditional way of avoiding difficulties is to say that, if preferences are indeed identical and homogeneous, then there will always exist some domestic redistribution of income or compensation that will make all individuals better off when the economy moves to a higher indifference curve. The existence of such possible compensations implies that all individuals are *potentially* better off. This is not very satisfactory, however, since in the absence of actual redistributions the individuals who are made worse off are not at all happy with the potential-improvement argument. We will explicitly deal with the income-distribution problem at a number of points in the book, but in many others we will simply attach normative significance to community indifference curves.

3.5 Summary

This chapter has developed tools for the consumption or demand side of the economy. We begin with the standard Marshallian representation of the consumer's optimization problem: maximizing utility subject to a budget constraint gives us demand as a function of prices and income. Then we analyze the Hicksian approach: minimizing expenditure subject to a target utility level gives us demand as a function of prices and utility. Then we turn to the question of aggregating individual demands into aggregate demand and note that extremely restrictive assumptions are needed in order to make this theoretically valid. But having made note of this, we will then follow the tradition of trade theory and more or less ignore this problem in what follows. We will not however, ignore the fact that a rise in welfare as indicated by the aggregate or community indifference curve generally does not imply that all individuals or households within the economy have their welfare increased in the same proportion. We will see

repeatedly that liberalizations or its opposite, protectionism, and changes in world prices exogenous to our own economy not only affect aggregate welfare but also redistributes income within the society. Income redistribution through changes in factor prices or a redistribution between profit income and factor income are just two examples.

REFERENCES

Deaton, Angus and John Muellbauer (1980), *Economics and Consumer Behavior*, Cambridge: Cambridge University Press.

Green, H. A. John (1976), *Consumer Theory*, London: Macmillan.

ENDNOTES

1. This is not precisely theoretically correct; that is, the existence of community indifference curves and the ability to write demand as a function of aggregate income are not quite the same thing. In particular, there are cases in which the latter is true but the former is not (in particular when heterogeneous consumers each have a fixed share of income). For the purposes of this book, we have decided to avoid a lengthy discussion of the fine points. The analysis which follows develops the intuition and presents restrictions on preferences which guarantee both the existence of community indifference curves and aggregate demand functions. Relatively simple discussions of aggregation can be found in Green (1976) or Deaton and Muellbauer (1980).

2. In other texts, you may see the assumption that preferences are homothetic rather than homogeneous. The assumption that a function is homothetic is a somewhat weaker assumption than assuming it is homogeneous. Specifically, any monotonic transformation of a homogenous function is homothetic. Since for preferences we only require tastes to have the property that more is preferred to less, without really caring by how much more it is preferred, the weaker and more general homothetic property is sufficient for most results in consumer theory. But then again this is not really important for our purposes, so we will use the term homogeneous throughout.

Figure 3.1

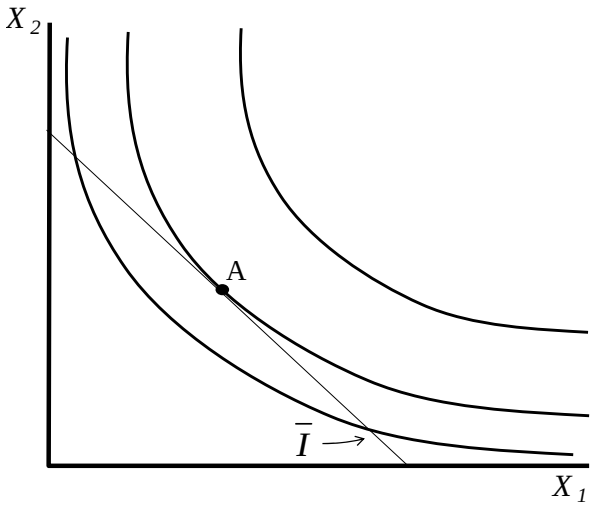


Figure 3.2

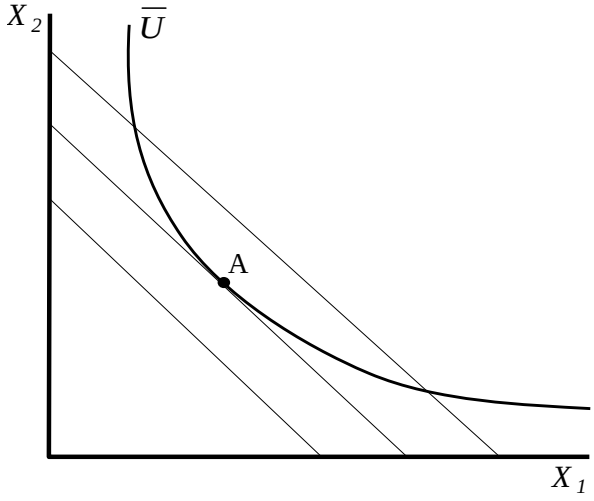


Figure 3.3

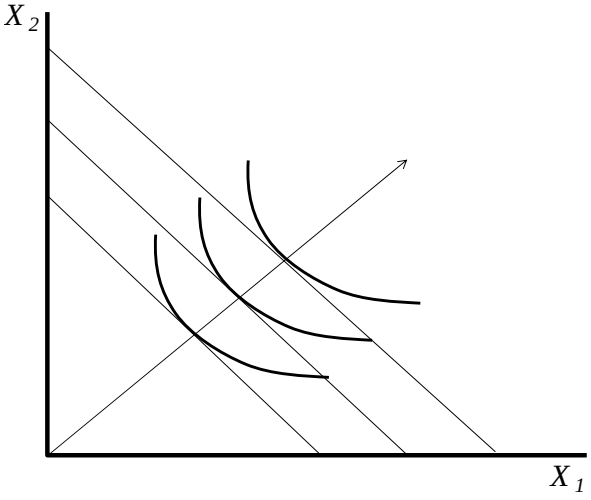


Figure 3.4

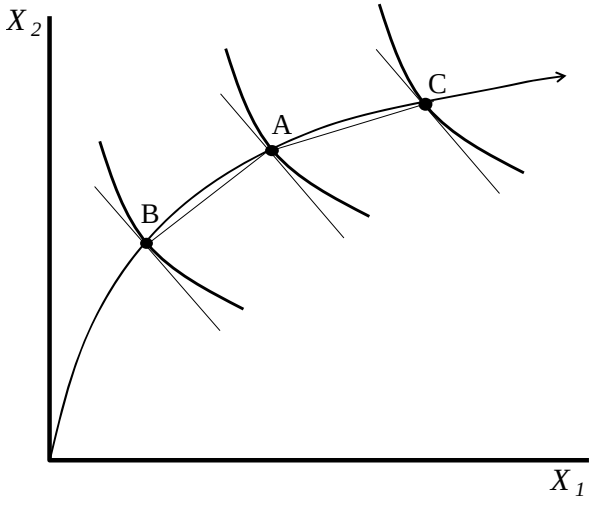


Figure 3.5

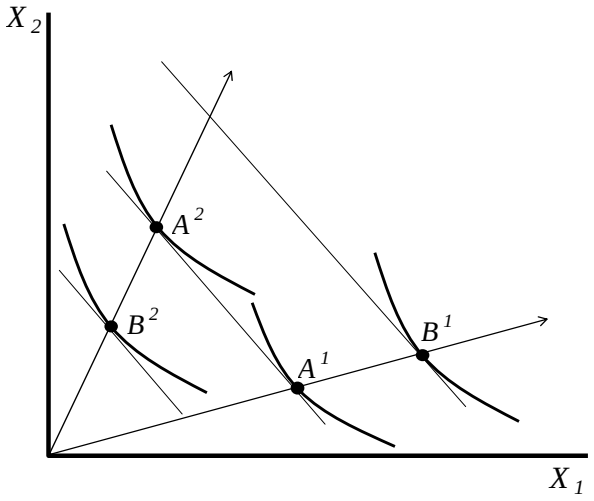


Figure 3.6

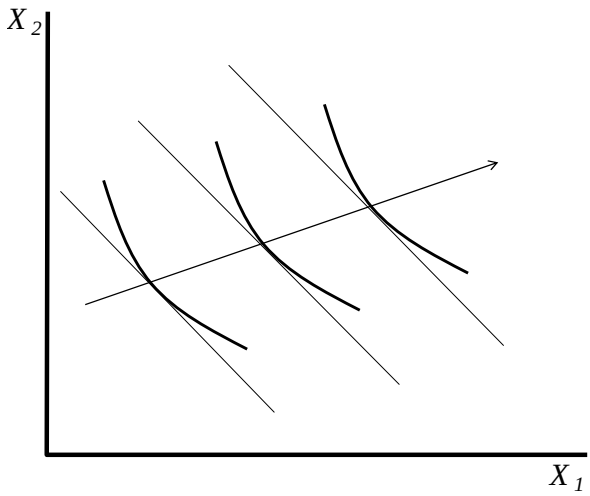
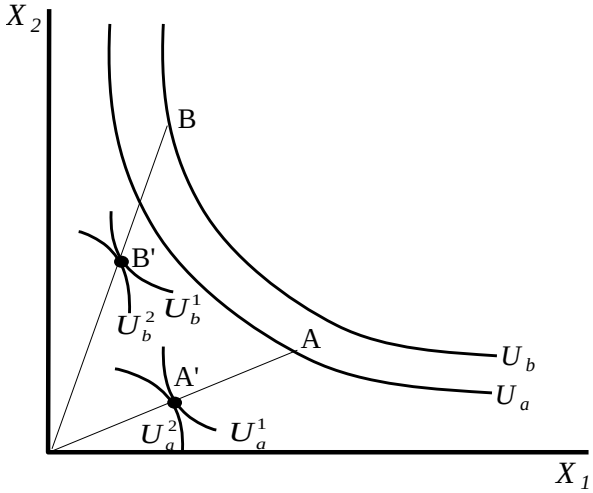


Figure 3.7



Chapter 4

GENERAL EQUILIBRIUM IN OPEN AND CLOSED ECONOMIES

4.1 General equilibrium in the closed (autarky) economy

The two previous chapters developed the tools of production and consumption theory. The purpose of this chapter is to combine the production and demand sides of the economy to arrive at an overall or *general-equilibrium* analysis. This section considers general-equilibrium in a closed economy; that is, an economy that is self sufficient and does not trade. Such an economy is said to be in *autarky*.

We will begin with an informal, graphical representation of general equilibrium, and move later in the chapter to a more rigorous formulation. Three sets conditions determine general-equilibrium in a closed economy. The first of these are *optimization* conditions for producers and consumers. Throughout this chapter producers and consumers are assumed to be competitive. In later chapters we will consider many cases of imperfect competition and other distortions. (1) Competitive, profit-maximizing producers pick outputs such that, at given commodity prices, the marginal rate of transformation is equal to the producer price ratio. This condition was given in equation (2.23) as $p_1/p_2 = MRT$. (2) Consumers pick commodities such that, at given commodity prices, their marginal rate of substitution in consumption is equal to the consumer price ratio. Assuming that consumer and producer prices are the same, this condition is given by $p_1/p_2 = MRS$. The next set of conditions are *market-clearing* conditions: (3) the supply and demand for each commodity must be equal. Let X specifically denote the production of a commodity and D denote consumption (demand) of a commodity. Our three conditions for a graphical representation of general-equilibrium are summarized by

$$\frac{p_1}{p_2} = MRT \qquad \text{producer optimization}$$

$$\frac{p_1}{p_2} = MRS \qquad \text{consumer optimization}$$

$$X_1 = D_1 \qquad X_2 = D_2 \qquad \text{market clearing} \tag{4.1}$$

Figure 4.1 shows an equilibrium for a closed economy that satisfies these three conditions. Producers produce optimally at point A where the slope of the production frontier is tangent to the price ratio $p^a = p_1^a/p_2^a$ ("a" for autarky: superscipts on variables generally denote particular values of those variables). Similarly, consumers consume optimally at point A where the slope of their indifference curve is tangent to the price ratio p^a. And finally, markets clear because the production and consumption points are the same. Note also that the equilibrium at A is *optimal* in the sense that the economy consumes on the highest possible community indifference curve, subject to the constraint that production is feasible.

Figure 4.1

Note for future reference that the equilibrium at point A in turn determines a factor allocation in the Edgeworth box in Figure 2.8 (if indeed the two-factor model is the underlying production structure). Thus factor prices are also determined in general-equilibrium. To the extent that consumers have different factor endowments, the factor prices in turn determine the distribution of income among consumers.

4.2 General equilibrium in the open (trading) economy

Now assume that an economy can engage in trade at a fixed world price ratio, which we will denote by $p^* = p_1^*/p_2^*$. The first two optimization conditions mentioned in the previous section remain unchanged. The only difference is that the values of world prices will in general be different from the values of the prices determined in autarky.

The difference between the closed and open economy equilibria lies in the third condition: market clearing. With international trade, an economy is no longer constrained to consume only what it can produce itself. The loosening of this constraint is indeed the source of gains from trade as we shall see. A trading economy is able to sell some of one good at world prices and use the proceeds to buy the other commodity. Instead of market clearing, we have what we call a *trade balance* condition: the value of what a country sells on world markets must be equal to what it buys. We can define the *excess demand* for goods X_1 and X_2 as $(D_1 - X_1)$ and $(D_2 - X_2)$ respectively. If excess demand is positive, we are consuming more than we are producing, so positive excess demand corresponds to an import good. If excess demand is negative, we are consuming less than we are producing, so this is an export good.

The trade-balance constraint requires that the value of our imports must be equal to the value of our exports. An alternative way of saying this is that the sum of the value of the country's excess demands must equal zero: the positive excess demand for the import good must equal the negative excess demand for the export good.[1] The trade-balance condition is given by

$$p_1^*(D_1 - X_1) + p_2^*(D_2 - X_2) = 0 \tag{4.2}$$

Note that this condition is completely general and does not depend on which good happens to be the import good and which happens to be the export good.

We can rearrange the terms in (4.2) and rewrite the equation in a different way.

$$p_1^* X_1 + p_2^* X_2 = p_1^* D_1 + p_2^* D_2 \tag{4.3}$$

The left-hand side of this equation is the value of production at world prices while the right-hand side is the value of the consumption at world prices. Thus a condition that is exactly equivalent to the trade-balance condition is the requirement that *the value of production equals the value of consumption*.

We can in turn think of the value of production as the income of the country. By placing a line with the slope of the world price ratio p^* through the production point, we then have the "national budget line". This budget line defines national income by evaluating domestic output at world prices. Consumers are then free to choose any point on this budget line, since the value of consumption will be equal to the value of production. This is shown in Figure 4.2, where the fixed world price ratio is given by p^*. Producers optimize by choosing to produce at point X. Consumers optimize by choosing consumption at point D. In the particular case shown, the country imports X_1 ($D_1 > X_1$) and exports X_2 ($D_2 < X_2$). Trade balances insofar as the value of production at world prices equals the value of consumption. To summarize, the conditions for general-equilibrium in an open economy are given as follows.

$$\frac{p_1^*}{p_2^*} = MRT \qquad \text{producer optimization}$$

$$\frac{p_1^*}{p_2^*} = MRS \qquad \text{consumer optimization}$$

$$p_1^*(D_1 - X_1) + p_2^*(D_2 - X_2) = 0 \qquad \text{market clearing} \qquad (4.4)$$

Note finally that the autarky market-clearing condition is a special case of trade balance. The former satisfies the trade-balance condition in that both terms in parentheses are zero. If world prices just happened to be the same as the country's autarky prices, then the trading equilibrium would be identical to the autarky equilibrium.

Figure 4.2

4.3 The excess demand function

We now turn to the larger question of the determination of world prices and an international general equilibrium (our world will consist of two countries). Consider Figure 4.3. The autarky price ratio p^a is shown for reference. At the price ratio $p^{*1} < p^a$, the country produces at X^{*1} and consumes at D^{*1}. Excess demand for good X_1 is positive; i.e., X_1 is imported (recalling $p = p_1/p_2$) . If the relative price of X_1 is lower on the world markets than on the domestic market then the country should buy from the low cost source and hence import the good. Similarly, if X_2 is relatively more valuable on the world market than at home, then exports of X_2 are in order. At the price ratio $p^{*2} > p^a$ in Figure 4.3, producers pick point X^{*2} and consumption at D^{*2}. With the price ratio greater than the autarky price ratio, the home country exports X_1 (the relatively valuable good on the world markets) and imports X_2 (the relatively cheap good on the world market).

Figure 4.3

This is indeed a general result. If the world price ratio exceeds the domestic price ratio ($p^* > p^a$) then X_1 is exported and there is a negative excess demand for X_1. If the world price ratio is less than the autarky price ratio ($p^* < p^a$) then X_1 is imported and there is a positive excess demand for X_1.

In Figure 4.4, we construct an excess demand curve for good X_1 for the country, where excess demand is given by $M_i = D_i - X_i$ ('*M*' for imports). At the autarky price ratio p^a, there is zero excess demand. Price ratios p^{*1} and p^{*2} in Figure 4.4 correspond to the similarly labeled price ratios in Figure 4.3. The excess demand curve for X_1 is labeled M_1 in Figure 4.4. The excess demand curve is thus much like a conventional demand curve, except that the quantity demanded may be either positive or negative. A negative excess demand is simply a desire to supply (export) the good to the world market at that price.

Figure 4.4

4.4 The shape of the excess demand curve, welfare interpretation

What are the factors that lead to the specific shape that an excess demand curve assumes? As in the case of the standard consumer-choice problem, there are both substitution and income effects. The substitution effect here is always negative: an increase in the price of a good decreases demand and increases production, so $M_i = D_i - X_i$ is negatively related to the relative price of p_i. However, the income effect is more complicated than in the standard consumer problem (even under the assumption that all goods are normal). If a good is imported, then an increase in its price generates a negative income effect. But if it is exported, a rise in its price increases income and the income effect is positive. A potential

subtlety thus exists in constructing the excess demand curve comes when $p^* > p^a$. In this case, the curve may bend backward (its slope becomes positive) in the exporting section (negative excess demand) of the curve. Intuitively, as the price of the export good continues to increase, the country gets richer from the sales of that good. This leads consumers to want to devote some of their additional income to purchases of that good. Thus at some point, this "income effect", which leads consumers to demand more of the export good, may outweigh the substitution effect which leads consumers to want less of a good when its price rises.

Finally, movement along the excess demand curve away from p^a in either direction is welfare improving because any change in price leads to an increase in the consumption choices for consumers. As prices fall from p^a, for instance, the previous combination of X_1 imports and X_2 exports now generates a trade surplus, so consumers can buy more imports for less exports (retaining the previous exports as consumption). Similarly, as the price of an export good rises, consumers could maintain their consumption levels of the export, but have the choice of buying more imports as an alternative. In short, the further a country can trade from its autarky price ratio the better off it will be relative to autarky.

4.5 International general equilibrium

Now introduce a second country referring to it as country *f* (foreign), and to the original country as *h* (home). We follow a "typical" treatment of this by writing the axis in reverse order for country *f*; that is, the horizontal axis for country *f* maps excess *supply*, production minus consumption. Figure 4.5 shows an excess supply curve for country *f*, M_{f1}, placed below above the excess demand curve for country *h*, M_{h1}. The autarky price in country *f* is $p_f^a = p_{f1}^a / p_{f2}^a$, less than country *h*'s autarky price ratio, p_h^a.

General-equilibrium in the world economy is then determined at an international price ratio where the excess demands of the two countries are equal and opposite. In Figure 4.5, this occurs at price ratio p^*. At that price, the positive excess demand (imports) of the home country are equal to the negative excess demand (exports) of the foreign country. The market for X_1 clears, which is a condition for international equilibrium: $M_{h1} + M_{f1} = 0$.[2] Note for future reference that the equilibrium price lies *between* the autarky prices of the two countries.[3] This is a general result, at least in competitive models, and it makes economic sense. If, for example, the world relative price of X_1 was greater than the autarky prices of both countries, then both countries would want to export X_1, which cannot be an equilibrium.

Figure 4.5

The *direction of trade* (which countries import and export which goods) at this equilibrium in Figure 4.5 makes economic sense. With the relative price of X_1 higher in country *f* in autarky, *f* will export X_1 and *h* will import X_1 in international equilibrium. We will see many times in the chapters that the direction of trade follows the differences in autarky prices. A major topic of Part II of the book is determining how underlying characteristics of economies, such as technologies and factor endowments, lead to differences in autarky prices.

What about the market for X_2? When we have only two goods and impose a trade-balance condition, we need only examine one market to find international equilibrium. Suppose that both countries are satisfying their trade balance conditions and that the market for X_1 clears. These three equations must then imply that the market for X_2 also clears. The need to find equilibrium in only one of two markets is known as Walras' Law in economics (more generally, if n-1 markets are clearing where n is the number of markets, the nth market must clear as well). If

$$p_1^* M_{h1} = -p_2^* M_{h2} \qquad p_1^* M_{f1} = -p_2^* M_{f2} \qquad M_{h2} = -M_{f2} \text{ then } M_{2h} = -M_{2f} \tag{4.5}$$

4.6 An introduction to computing solutions to numerical general-equilibrium models (may be skipped without loss of continuity)[4]

The treatment of general-equilibrium to this point has been informal and intended to develop a simple intuition about trade and gains from trade. But it is a long way from a methodology that could be implemented with numbers (whether from real data or a straight simulation analysis). The principal problem is that it assumes that we know a lot about the production set and an equation for the production frontier in particular. However, it unfortunately turns out that even for many relatively simple production functions, it is impossible to solve for an equation for the production frontier. This in turn means that it is impossible, for example, to solve for the relationship between product and factor prices.

One approach to solving for general-equilibrium that often occurs to students is to treat it as an optimization problem, essentially exploiting what we called the first theorem of welfare economics earlier in the book: competitive equilibria are Pareto optimal. This is indeed a possible approach, but it is really only viable for a single economy with a single "representative" household on the demand side. If there are multiple household types, characterized by different endowments and/or preferences, or if there is more than one country, the optimization approach breaks down. It is not clear what to optimize subject to what constraints. There are infinitely many Pareto optimal allocations with two household types, but generally only one of those is a competitive equilibrium (e.g., a Pareto allocation in which each household's consumption is equal to its factor income).

The modern approach to general-equilibrium is to formulate a system of equations and inequalities, embedding the optimization at the level of the industry and household, creating a system of n weak inequalities in n unknowns. General equilibrium is then a solution to a square system, not a solution to a single constrained optimization problem. It is actually a bit more complicated than that. In order to accommodate corner solutions in which, for example, not all goods are produced (quantities and prices must be non-negative), it is necessary to associate each equation/inequality with a particular non-negative variable. If the relationship holds as a strict equality, the associated variable is positive. If, in equilibrium, the relationship holds as a strict inequality, the associated variable is zero. A problem of this type is referred to as a (non-linear) complementarity problem in mathematical programming terminology: each (weak) inequality has an associated complementary variable. Economic theory provides the correct association of inequalities and unknowns and the correct direction of the inequality (e.g., cost can exceed price in equilibrium, the good is unprofitable and is not produced, but price cannot exceed cost in equilibrium).

There are several possible approaches to formulating general-equilibrium as a complementarity problem, but one that seems very robust to the inclusion of lots of added features is the Mathieson-Rutherford approach, which formulates a model as composed of four blocks of inequalities and unknowns. (1) *zero-profit or optimization* conditions for all production and trade activities. (2) *market-clearing* or supply-equal-demand relationships for all commodities and markets. (3) *income or budget balance* conditions for all agents (consumers, governments, and so forth). (4) *auxiliary equations* such as equations that determine things like endogenous tax rates, markups and so forth. The last of these will not be treated in this chapter.

Let's see how this works in practice with our two-good, two-factor, closed economy with a single representative consumer. The first task is to solve for cost functions for each sector, and the consumer's expenditure function. Alternatively, Marshallian demand functions may be used. Either way, optimization is embedded at the sectoral and household level. Here is a specification using Marshallian demands.

weak inequality	description	complementary variable	
$c_1(w_1, w_2) \geq p_1$	zero profits in X_1	X_1	(4.6)
$c_2(w_1, w_2) \geq p_2$	zero profits in X_2	X_2	(4.7)
$X_1 \geq D_1(p_1, p_2, I)$	market clearing, X_1	p_1	(4.8)
$X_2 \geq D_2(p_1, p_2, I)$	market clearing, X_2	p_2	(4.9)
$\bar{V}_1 \geq c_{11}(w_1, w_2)X_1 + c_{21}(w_1, w_2)X_2$	market clearing, V_1	w_1	(4.10)
$\bar{V}_2 \geq c_{12}(w_1, w_2)X_1 + c_{22}(w_1, w_2)X_2$	market clearing, V_2	w_2	(4.11)
$I = w_1\bar{V}_1 + w_2\bar{V}_2$	budget balance	I	(4.12)

The first two inequalities (4.6)-(4.7) involved prices and costs and have quantities as complementary variables: if, in equilibrium, marginal cost is strictly greater than price, production is unprofitable and output is zero. All of the market-clearing equations (4.8)-(4.11) are in terms of quantities, and have prices as complementary variables: if, in equilibrium, supply exceed demand then the good/factor is free, its price is zero. The last equation is income-expenditure or budget balance. Note the use of Shepard's lemma in the market-clearing inequalities for factors to get the factor demands on the right hand side of (4.10) and (4.11).

There is one extra equation in the system, another manifestation of Walras' Law: if n-1 markets clear, then the nth does as well. This is generally interpreted as an indeterminacy of the price level. Note that if we find a set of prices and income that solves (4.6)-(4.12), then any multiple scaling these prices and income proportionately up or down also solves the model. This problem is usually dealt with by choosing a numeraire, typically setting one price equal to one and dropping the corresponding equation from the model. This is essentially what we were doing in our graphical presentation above: the solution only determines the relative prices of goods 1 and 2; any multiple of their absolute prices is also a solution. Showing that every relationship in (4.6)-(4.12) remains unchanged with a doubling of prices and income is left to you as an exercise.

While the above system seems simple and intuitive, in many applied problems it turns out to be more useful to work with Hicksian demand and to model utility as though it was a produced good: utility (U) is *produced* with inputs of X_1 and X_2, and it has a *price* associated with it, p_u, which in equilibrium must be the cost of producing one unit of utility, the unit expenditure function. One advantage of this approach is that welfare and the real consumer price index p_u (the price of a unit of utility) are automatically computed as part of the solution. The same model as above can be written as follows.

weak inequality	description	complementary variable	
$c_1(w_1, w_2) \geq p_1$	zero profits in X_1	X_1	(4.13)
$c_2(w_1, w_2) \geq p_2$	zero profits in X_2	X_2	(4.14)
$e(p_1, p_2) \geq p_u$	zero profits in U	U	(4.15)
$X_1 \geq e_1(p_1, p_2)U$	market clearing, X_1	p_1	(4.16)
$X_2 \geq e_2(p_1, p_2,)U$	market clearing, X_2	p_2	(4.17)
$U = I/p_u$	market clearing for U	p_u	(4.18)
$\bar{V}_1 \geq c_{11}(w_1, w_2)X_1 + c_{21}(w_1, w_2)X_2$	market clearing, V_1	w_1	(4.19)
$\bar{V}_2 \geq c_{12}(w_1, w_2)X_1 + c_{22}(w_1, w_2)X_2$	market clearing, V_2	w_2	(4.20)
$I = w_1\bar{V}_1 + w_2\bar{V}_2$	budget balance	I	(4.21)

As in the case of using Shepard's lemma to get factor demands, Shepard's lemma is used in (4.16) and (4.17) to get commodity demands. In this square system, p_u, the real consumer price index is often chosen as numeraire. This has the advantage that all other prices, especially factor prices, are "real" values: a ten-percent increase in a factor price is relative to the cost of the consumer's consumption bundle.

The above indicates that there are often several ways to formulate general equilibrium as a complementarity problem. The second approach has two more equations than unknowns and so seems more complicated (not to mention getting your head around the notion of utility as something that is produced and carries a price, since neither are observed in data). However, it is clear that even the first approach is going to defy analytical solutions, except in extremely simple cases due to the number of dimensions of the model and to the fact that the solution itself determines which weak inequalities hold strictly and which hold with equality (e.g., which goods are actually produced). Thus modelers typically turn to numerical solutions, picking functional forms and parameters for the general model. The computer doesn't care how many dimensions the model has, so the formulation of a model for numerical solution depends on the flexibility it carries and easy of interpretation. Often, more equations and unknowns are preferred.

With this thought in mind, let us now consider an open economy model, and in particular a "small" economy, defined as one that faces fixed world prices. There are several ways to do this, some of which result in smaller-dimension models than the closed economy case: think of our graphical treatment in which two market-clearing equations are replaced by a single balance-of-trade equation. Intuitively, the open economy has fewer *constraints*. However, the other way to think about it is that the open economy offers more *opportunities*. Trade effectively allows a country to transform one good into another through international exchange. Thus our above models can be modified by adding additional activities, import and export activities, with associated quantity variables of imports and exports. This generally turns out to be the more productive approach.

Let ρ_1^* and ρ_2^* be *fixed parameters* for the small economy, giving world prices of X_1 and X_2

respectively. (In a simulation below, we choose X_2 as numeraire and so $\rho_2^* = 1$.) Let E_i and M_i denote the number of units of X_i exported and imported respectively. To allow for either direction of trade depending on world prices, we add four trading activities and four unknowns: E_i and M_i for each good. Our model is extended to the open economy by adding four inequalities (four trading activities, although only two will be strict equalities in equilibrium) and revising the market-clearing equations for X_1 and X_2: production minus exports plus imports equals consumption. The trading activities are represented by zero-profit conditions as in the case of production activities, with the complementary variables being quantities of imports and exports.[5] Here are the added and revised inequalities and variables.

weak inequality	description	complementary variable	
$p_1 \geq \rho_1^*$	zero profits, exports of X_1	E_1	(4.22)
$\rho_1^* \geq p_1$	zero profits, imports of X_1	M_1	(4.23)
$p_2 \geq \rho_2^*$	zero profits, exports of X_2	E_2	(4.24)
$\rho_2^* \geq p_2$	zero profits, exports of M_2	M_2	(4.25)
$X_1 - E_1 + M_1 \geq e_1(p_1, p_2)U$	market clearing, X_1	p_1	(4.16 revised)
$X_2 - E_2 + M_2 \geq e_2(p_1, p_2,)U$	market clearing, X_2	p_2	(4.17 revised)

This surely seems a lot to take in, but analytical solutions are not possible beyond extremely restrictive cases. Once one goes to simulation, the number of inequalities and unknowns is immaterial, and the researcher/student or whomever can analyze much more complex situations than is possible with strictly analytical methods.

In an appendix to this book, we present computer code for this model. In order to catch your interest in this, Figures 4.6 and 4.7 present some simple simulation results for changing the terms of trade, the (fixed) relative price of X_1 in terms of X_2. Figure 4.6 plots the excess demand curve for X_1; it becomes vertical in the export section once specialization is reached but does not bend back (this is a property of Cobb-Douglas preferences assumed in the model). Figure 4.7 plots welfare as a function of the terms of trade. The horizontal axis of Figure 4.7 is the same as the vertical axis of Figure 4.6: by convention, prices are put on the vertical axis of excess demand diagrams, even when they are the exogenous variable as is the case here. Welfare gains from trade are 30 percent of autarky welfare in Figure 4.7 when the relative price of X_1 is 2.5 (X_1 exported in Figure 4.6) or its reciprocal 0.4 (X_1 imported in Figure 4.6).

Figure 4.6 **Figure 4.7**

In the appendix, we will generalize this model to include tariffs and transport costs. Learning to build and compute simulation models requires quite an investment of your time to master as is true of any technique and software, but it is immensely educational to actually build a workable model rather than simply look at graphs and qualitative properties of comparative statics exercises; indeed the latter are only solvable for extremely special cases. In addition to being educational, it is just plain good fun to see theorems verified numerically or see the optimal tariff idea in action, for example.

4.7 Summary

The chapter begins with a simple and intuitive contrast between the conditions for equilibrium between a closed (autarkic) economy and an open (trading) economy. Trade can be thought of as *removing a constraint*: the requirement that anything that is consumed must be produced domestically. Trade can be thought of as *creating an opportunity*: the ability to essentially transform one good into another by exporting the first and importing the second. We showed how an economy may be graphically summarize by an excess demand function, discussed its shape and its welfare interpretation. In the two-good case, Walras' law allows us to find equilibrium in one market only, then the other market must clear as well.

The world prices (or just price ratio in the two-good case) is then determined as the price ratio at which the sum of all country's excess demands are zero. In the two-country case, the import demand of one country must balance the exports (negative excess demand) of the other country. The world price ratio must lie between the two autarky price ratios, and each country will export the good it produces relatively cheaply in autarky and import the good which is relatively costly in autarky.

The graphical approach and its analytical counterpart have great appeal. It is, unfortunately, the case that it is not very useful in practice and, even in very simple models past two goods and two factors, it does not permit a full characterization of equilibrium and of the response of the economies outputs and factor prices to changes in world prices, for example. This is largely due to the fact that we generally cannot characterize much about the production frontier other than concavity from knowing sectoral production functions. Even in the two-good, two-factor case, the graphical and analytical approaches can say little quantitatively; e.g., what are the aggregate welfare and distributional effects of at 20% import tariff?

On account of these difficulties, we have sketched the outline of a methodology used to solve for general equilibrium in more complex situations. This methodology embeds the optimization problems at the sectoral level, that of industries and households. These industry/household results are then used to build square models of n weak inequalities in n unknowns. These are roughly characterized in terms of "blocks": (1) optimization conditions, with outputs and demands as complementary variables, (2) market clearing inequalities with prices as complementary variables, and (3) income balance conditions. A forth set, auxiliary equations and variables such as determining markups, endogenous tax rates and so forth are left to more advanced treatments.

Endnotes

1. In a dynamic model with many time periods, trade need not balance in any one time period. A country can consume more than it producers by selling *assets* which can be thought of as claims to future consumption. The United States has been in a position of *trade deficit* for years, indicating that foreigners are accumulating US assets and hence have a claim on future US production. This is an important topic in international finance, but we will also discuss it in the last chapter of the book.

2. Because of the possibility of backward bending sections of excess demand curves in the exporting region (i.e., negative excess demand) mentioned in the previous section, there is a possibility of multiple equilibria. This will be ignored in this book.

3. Later in the book we show that with imperfect competition, the free-trade price for a good need not lie between the autarky prices of the two countries. Additional competition induced by trade may lead to a fall in the price of the good in both countries.

4. Professors may wish to skip this section, and the numerical appendices in the book. We have tried to make it possible to do this without loss of continuity. The general set up for a method of actually creating a solvable model is instructive however, and all of the tools needed to understand this section (cost and expenditure functions, Shepard's lemma) have already been introduced. There will be a number of general references to this section later in the book but again, we have tried to write subsequent material so that this section is not crucial to things later on.

6. An astute reader might note below that both the export and import equation can hold with equality for a given good. It is possible that a good could be both imported and exported, with only net trade being determined in equilibrium (there are infinitely many gross trade patterns consistent with a given net trade). This indeterminacy is broken in numerical models by imposing a very small trade cost (0.01% will do). For example, equation (4.25) below could be written $p_1(1.0001) \geq \rho_1^* p_{fx}$. We have ignored this problem here, but it is important in actual computable models to add this small trade cost.

Figure 4.1

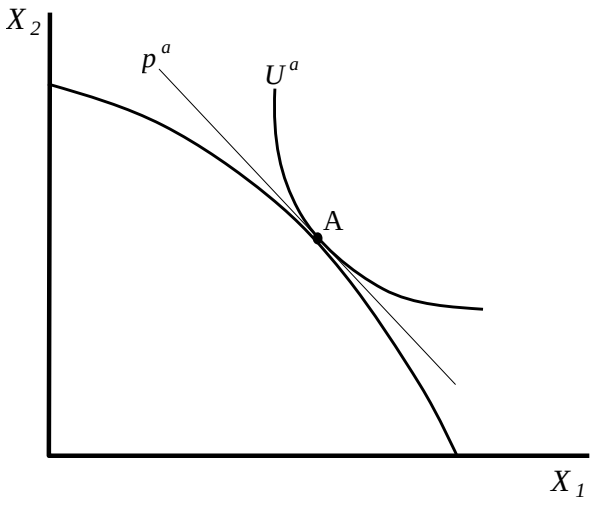


Figure 4.2

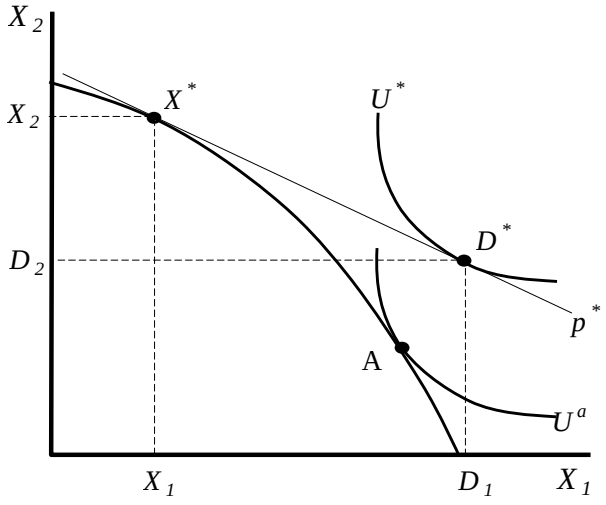


Figure 4.3

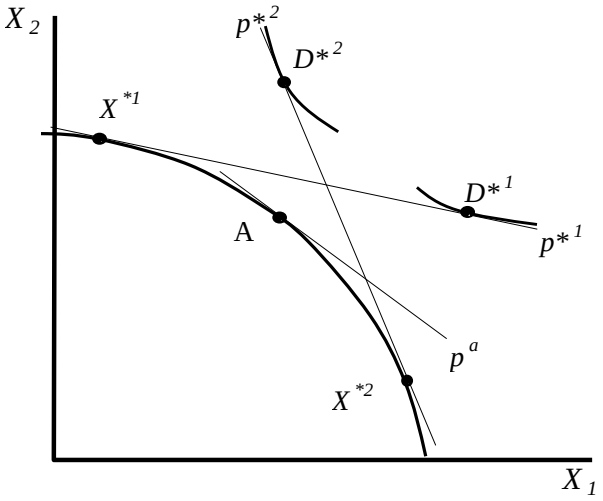


Figure 4.4

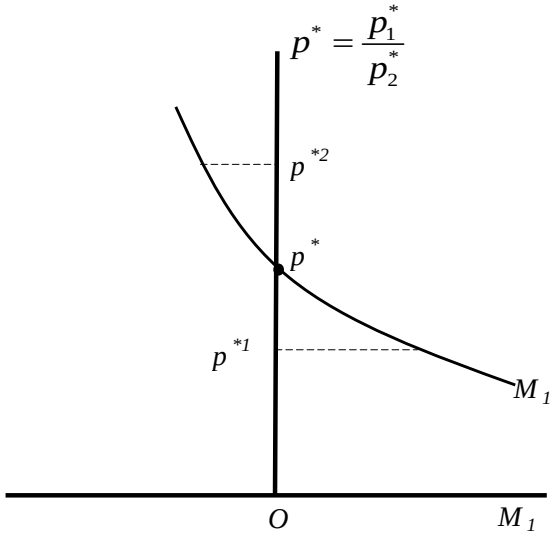


Figure 4.5

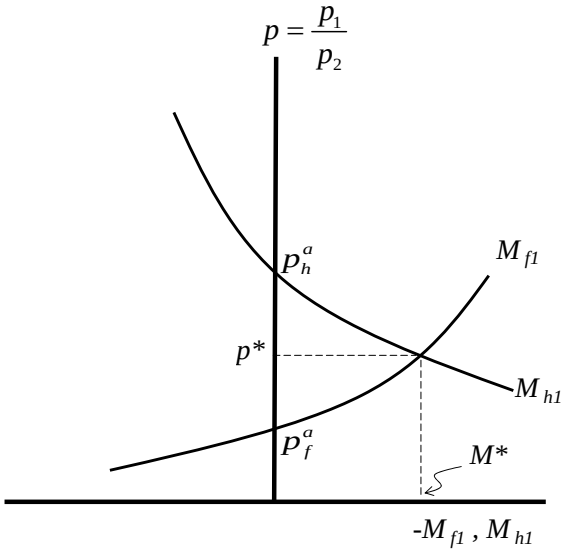


Figure 4.6

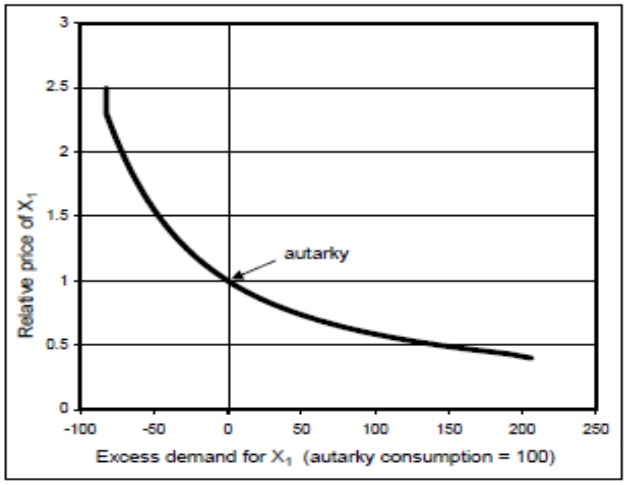


Figure 4.6: Simulated excess demand curve

Figure 4.7

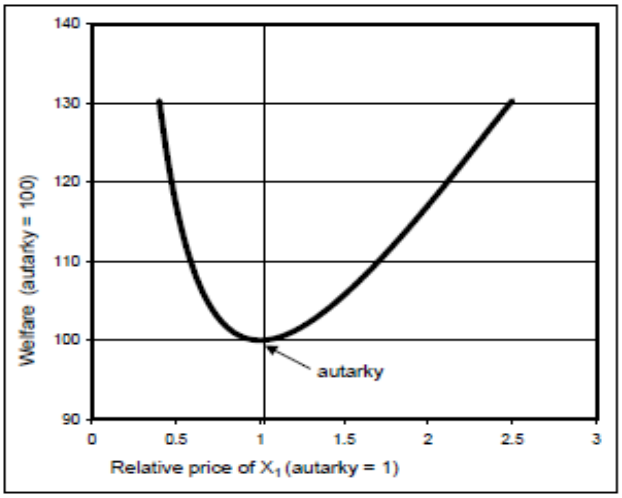


Figure 4.7: Welfare and the terms of trade

Chapter 5

THE GAINS FROM TRADE

5.1 Gains from trade

We are now in a position to address one of the most fundamental issues in the study of international trade: the gains from trade. We will be able to show that, under certain circumstances, a country's overall welfare is in some sense improved by international trade, which should thus be viewed as desirable. We will show that, under a wide range of circumstances, all countries mutually gain from trade; that is, trade is a "positive-sum game".

However, we will also show that not all individuals within a country will necessarily benefit from trade. In other words, while a country's total income is increased by trade, these gains may be very unevenly distributed to the point where some individuals or groups are worse off. A correct understanding of the gains from trade is thus important not only from an academic point of view, but also because of the practical need to evaluate various anti-trade arguments put forward by business, labor, and government groups.

Figure 5.1 shows the production frontier and indifference curves for a single country. Autarky equilibrium occurs at point A, with the economy reaching utility level U^a. Figure 5.1 also shows two alternative world trading price ratios, p^{*1} and p^{*2}. We have constructed the diagram quite deliberately such that these two world price ratios both lead to the same free-trade utility level, U^*. Figure 5.1 is not a formal proof, but it illustrates a result that can be proved more rigorously: *the ability to trade at any price ratio other than the country's autarky prices must make the country better off.*

Figure 5.1

Another thing to note from Figure 5.1 is that the *direction* of trade is of no particular welfare significance. The utility level U^* can be achieved either though the export of X_2 in the case of world price ratio p^{*2} or through the export of X_1 in the case of world price ratio p^{*1}. The point is only that the world price ratio should differ from the domestic autarky price ratio. Given any such difference, the country gains by exporting what is more valuable on world markets than at home and by importing from the rest of the world what is more costly to produce at home than abroad. This is an important point insofar as there are many arguments that attach particular significance (e.g., prestige motives) to what goods a country imports or exports.

Returning to a point made in the first paragraph of this section, it also follows that two countries *mutually* enjoy gains from trade. In Figure 4.5 of the previous chapter, we depict a situation in which two countries have different autarky price ratios and the equilibrium world price ratio is (weakly) between these two autarky price ratios. It thus follows from our analysis of Figure 5.1 that *both countries are (weakly) better off with trade*; the only exception to being *strictly* better off is when the world price ratio is equal to the autarky price ratio of one country, in which case that country is no better off but no worse off with trade. Trade is Pareto improving for both countries.

Figure 5.2 illustrates the point about mutual gains by constructing a special case in which the countries are symmetric, essentially mirror images of one another. The two countries have identical preferences, but different production frontiers. The production frontiers for Home and Foreign are given by $X_{h2}X_{h1}$ and $X_{f2}X_{f1}$ respectively. Home is relatively good at producing X_1 while Foreign is relatively good at prodcing X_2. This difference is then reflected in their autarky price ratios, Home consuming at A_h and Foreign at A_f in Figure 5.2.

Under free trade, both countries specialize in producing only one good in the case shown in Figure 5.2. Free trade allows both of them to reach the same consumption point D at price ratio p^*, with Home producing at X_{h1} and Foreign producing at X_{f2} (each country exports half of the output of its good for half of the output of the other country's good in this symmetric case).

Figure 5.2

We emphasize that this is a very special case; in general there is no presumption that two countries will reach the same utility level through trade or that the gains from trade will be shared equally. The latter point will be made many times throughout the book. The points that (A) both countries do gain and (B) the direction of trade is not necessarily of any significance are very general results.

Before moving on, Figure 5.2 can also be used to illustrate a decomposition that is sometimes useful in analyzing trade distortions. This is the decomposition of total gains into (1) gains from exchange and (2) gains from specialization. Suppose our two countries in autarky in Figure 5.2 magically discover each other's existence after production is fixed for the year. They can still trade from their fixed outputs at A_h and A_f to reach the common point D_e ("e" for exchange). This is often referred to as pure gains from exchange: trade from fixed endowments with no change in production. However, they see that there are additional gains from specializing and so next year they produce at X_{h1} (home) and at X_{f2} (foreign). Then they can trade to reach the mutual consumption point D in Figure 5.2. The distance between D and D_e then measures the additional gains from specialization. Of course, there will be no gains from specialization unless exchange is possible.

5.2 The gains-from-trade theorem

We will now present a somewhat more formal treatment of the gains from trade. In particular, we present a simple proof of what is called the gains-from-trade theorem. This helps make clear the assumptions necessary to ensure that a country gains from trade. In Chapter 2, we showed that a competitive, undistorted economy maximizes the value of final production at world prices. Let superscript '*' denote quantities produced in free trade and superscript 'a' denote the quantities produced in autarky. Subscript i again indexes goods. From our result in (2.25), we have the following inequality.

$$\sum_i p_i^* X_i^* \geq \sum_i p_i^* X_i^a \tag{5.1}$$

Add in the balance-of-trade condition (4.2) or (4.3) and the autarky market-clearing conditions (4.1).

$$\sum_i p_i^* X_i^* = \sum_i p_i^* D_i^* \qquad X_i^a = D_i^a \tag{5.2}$$

Substitute the left-hand equation of (5.2) into the left-hand side of (5.1) the right-hand equations of (5.2) into the right-hand side of (5.1).

$$\sum_i p_i^* D_i^* \geq \sum_i p_i^* D_i^a \tag{5.3}$$

Free-trade consumption is revealed preferred to autarky consumption: this is the gains-from-trade theorem. At free-trade prices, the autarky consumption bundle could have been purchased for the same or less money but wasn't, so the free-trade bundle must be preferred. Another way of stating the theorem is that the assumptions of competition in all markets and the absence of other distortions are *sufficient conditions* to ensure that free trade is preferred to autarky. We will return to this point below.

The gains-from-trade theorem is appealing in its simplicity (once 5.1 is proved), but how robust is it to added complications? The following is a list of modifications to the basic model under which the

theorem remains true.

(A) Robust to the existence of trade costs and tariffs: costly trade versus autarky. Trade costs don't matter: we just interpret the prices p^* as the CIF prices (cost, insurance, freight) at which the country can trade, which are not the prices the foreign partner pays or receives. Tariffs require a more complicated analysis that is postponed until Chapter 18. Trade costs are treated in a number of chapters. The basic result is that increased trade costs and tariffs move the country back toward autarky price and therefore autarky welfare, but cannot make the country worse off than in autarky. This is not true for trade subsidies, discussed in the next section.

(B) Robust to changes in the number of goods produced in free trade versus autarky; robust to the inclusion of intermediate goods and services. Trade may change the set of goods produced relative to autarky. The country may specialize more, producing fewer goods, or new goods and services may be produced as trade creates a demand for things that had no market previously (think of stories about the off-shoring of services in recent years). We noted in Figure 2.12 that the production efficient condition underlying gains from trade certainly covers cases where fewer goods are produced. But it certainly covers the case of more goods as well. Think of the set of i goods in (5.1) as including the set of all possible goods, some of which might not be produced in trade or in autarky. The proof of production efficient in Chapter 2 is valid. For a good produced in autarky but not under trade, the left-hand side of (2.22) is zero, no production, and the right-hand side is negative (production is unprofitable at free-trade prices). For a good produced in free trade but not autarky, the left-hand side of (2.24) is zero, profits are zero, and the right-hand side is negative. Discussion of traded intermediate goods and services is postponed until Chapter 15.

(C) Robust to changes in the number of goods consumed in free trade versus autarky. We will note several times that trade allows for new goods and services to be consumed. The gains-from-trade theorem is not invalidated by this. Consider an economy that is endowed with a single good (or a small number of goods, no variation in production possible). With trade, it is feasible for consumers to choose not to trade and continue their autarky consumption. If they choose to trade some of their endowment for new goods, this must be welfare superior since they are not choosing the alternative, feasible option of autarky.

(D) For small countries, free trade is superior to any level of trade restrictions and/or subsidies. For an undistorted small economy facing fixed prices, it is possible to show that completely free trade is superior to any level of trade taxes and/or subsidies. We will show this in Chapter 18.

5.3 Limitations of the gains-from-trade theorem

Unfortunately, there are many other generalizations of the basic underlying model that invalidate the theorem. Here is a list, all of which are dealt with in more detail later in the book.

(A) Not robust to trade subsidies: subsidized trade versus autarky. Trade restrictions cannot make the country worse off than in autarky, but subsidies to trade can. We will show this in Chapter 18.

(B) Not robust to domestic distortions: "free" trade versus autarky. Domestic distortions within the economy such as taxes, subsidies, imperfect competition and externalities generally invalidate the theorem, and we will show this at a number of points in the book. However, we must emphasize that the basic gains-from-trade theorem states that competition and the absence of other distortions is a *sufficient* condition for gains, not a *necessary* condition. We will show at several points that imperfect competition and increasing returns to scale offer important sources of gains from trade, so gains may be larger when these features are present even though the theorem itself cannot be proven.

In order to get a flavor of the problem, consider imperfect competition in industry i (a monopolist will do) and refer back to (2.24). This inequality is implied by profit maximization for competitive, price-taking firms, but not necessarily for a monopolist. As we know from basic microeconomics, a

monopolist reduces quantity in order to raise price relative to a competitive outcome. Suppose that the output X_i^0 and inputs V_{ij}^0 are observed to be chosen by the monopolist with resulting prices p^0 and w^0. It will generally be true that there is an alterative feasible output/input combination X_i^1 and V_{ij}^1 that generates higher profits *holding those prices constant*. However, the monopolist knows if the firm *actually chooses* output X_i^1 and inputs V_{ij}^1 then the *actual equilibrium* prices will *not* be p^0 and w^0, but will be some other values p^1 and w_j^1; e.g., a higher output of $X_i^1 > X_i^0$ will depress its price $p_i^1 < p_i^0$ sufficiently that profits fall. Inequality (2.24) might be replaced by

$$\left[p_i^0 X_i^1 - \sum_j w_j^0 V_{ij}^1\right] > \left[p_i^0 X_i^0 - \sum_j w_j^0 V_{ij}^0\right] > \left[p_i^1 X_i^1 - \sum_j w_j^1 V_{ij}^1\right] \tag{5.4}$$

Following through the same steps as in (2.24) to (2.27) on the first two inequalities assuming that the other industries are competitive, we would then arrive at:

$$\sum_i p_i^0 X_i^1 > \sum_i p_i^0 X_i^0 \tag{5.5}$$

The monopoly equilibrium does not maximize the value of output at equilibrium prices. Similar problems arise with domestic taxes and subsidies, and externalities.

(C) Not robust to trade barriers for large economies: restricted trade versus free trade. The comparison of free or restricted trade to autarky is important for theory, but it is rarely the actual choice faced by policy makers. Generally, they are choosing between more and less restricted trade. As we noted above, completely free trade is optimal for small economies, but large economies may have some monopoly power in trade, the ability to alter world prices. Some level of trade restrictions might improve prices such that the country is better off. An analysis of this point is postponed to Chapter 18.

5.4 The distribution of gains between countries

If the assumptions of the gains-from-trade theorem hold for all countries, then all countries are assured of mutual gains, or at least of being no worse off than in autarky. However, the theorem says nothing about the size of gains nor, in particular, the distribution of total gains between countries. Trade makes the welfare pie bigger, but how is that added benefit divided?

The situation is captured in a simple way in Figure 5.3, where the Edgeworth box gives the endowment point as E, indicating that country h is endowed with the total world supply of good X_1 and country f is endowed with the total world supply of good X_2. The two indifference curves through the endowment point give the autarky utility of the two countries, U_h^a and U_f^a, with autarky price ratios given by the slopes of the indifference curves through these points, p_h^a and p_f^a. Trading to any point within the "lens" formed by these two indifference curves is Pareto superior to autarky, and the locus BAC of tangency points is the set of Pareto optimal outcomes. However, there are infinitely many Pareto optimal outcomes and they differ in the distribution of the total gains between the two countries. Point A has a relatively even distribution. Point B awards all gains to country f, with country h indifferent between trade and autarky. Point C awards all gains to country h.

Figure 5.3

In a situation where the two countries are two individuals, an economist would first think of this as a bargaining problem between the two individuals: they try to negotiate over the division of the total

surplus. In a situation where there are many buyers and sellers in each country, this indeterminancy is resolved anonymously via market supplies and demands. We can think of the slope of a ray from E through B, A, or C as the world relative price ratio $p = p_1/p_2$. As we pointed out in the first section of this chapter, a country gets larger gains the further the world price ratio is from its autarky price ratio, the slope of its indifference curve at the autarky point. We will say more about this in Chapter 7, when we point out the role of country size in determining this distribution of gains between countries.

5.5 The distribution of gains within countries I: heterogeneous preferences

The preceding sections have shown that a country will gain from international trade in the sense that the country can potentially consume more of both goods. The gains were illustrated with the use of community indifference curves, although such indifference curves are not necessary for the main argument. While trade may result in aggregate consumption gains, it is important in understanding certain trade policy questions to emphasize that the gains from trade are not necessarily distributed evenly among members of a society. Indeed, it is possible that certain groups will actually be worse off in a situation of free trade than in an autarky or a restricted trade situation.

One possibility occurs when individuals in a society have very different tastes. Suppose that all individuals in the society have identical factor endowments, so they have identical incomes and budget lines. Suppose that the world price ratio exceeds the price ratio that would prevail in autarky ($p^* > p^a$) so that the country exports X_1 and imports X_2 (as in Figure 5.1 with $p^* = p^{*1}$). Consider now two individuals with different tastes (but identical incomes). Let AA' be the identical autarky budget line for each of these two individuals in Figure 5.4. Individual 1 has a high preference for X_1 and so chooses his autarky consumption bundle A_1. Individual 2 has a high preference for X_2 and therefore chooses bundle A_2. Their utility levels are given by U_1^a and U_2^a, respectively.

Figure 5.4

As shown in Figure 5.1 ($p^{*1} > p^a$) and Figure 5.4, trade has the effect of raising the relative price of X_1, which we illustrate in Figure 5.4 by rotating the budget line to TT'. Individual 1 increases his consumption from A_1 to T_1 and experiences an increase in welfare from U_1^a to U_1^*. But the increased relative price for X_1 has affected individual 2 so adversely that his consumption falls from A_2 to T_2 and his welfare falls from U_2^a to U_2^*. Thus when individuals have heterogeneous tastes, the gains from trade will be distributed unevenly and some groups may indeed become worse off.

5.6 The distribution of gains within countries II: heterogeneous endowments

A second example of uneven distribution of gains from trade occurs when individuals differ widely in their endowments of goods and/or factors. To keep it simple here, suppose that workers can either produce a unit of X_1 or X_2 but not both. Suppose that country h has lots of X_1 workers few X_2 workers and vice versa for country f. The situation is shown in Figure 5.5, where X_h and X_f are the fixed production points of countries h and f respectively, and both enjoy the same level of autarky utility U^a.

Figure 5.5

Now suppose the countries can trade and that the market outcome is that they trade to the mid-point D between X_h and X_f (production cannot change by assumption). Both countries share equally in gains from trade. However, relative prices change, with trading price ratio p^* in between the autarky price ratios p_h^a and p_f^a.

This price change has important income-distribution consequences. Figure 5.6 shows the effect on the budget line and welfare of "minority" X_2 producers in country h, the country with lots of X_1 producers. The X_2 producer's budget line is anchored on the X_2 axis by his/her endowment at point A in Figure 5.6. The effect of trade as shown in Figure 5.6 is to steepen or rotate the budget line through A in Figure 5.6. The minority X_1 producer in country h is worse off by what is an adverse price change for that individual. Figure 5.7 shows the same outcome for a minority X_1 producer in country f. That producer's budget line is anchored by A in Figure 5.7, and flattens out as a consequence of trade. The minority X_1 producer is worse off by an adverse price change even though the price change benefits the majority X_2 producers.

Figure 5.6 **Figure 5.7**

5.7 The Static Gains from Trade: An Example from History

The analysis in Sections 5.1 and 5.2 refer to the most basic forms in which a country in the aggregate enjoys welfare benefits in moving from autarky to free trade. Opening up to trade offers an opportunity to trade at international prices rather than domestic prices. This opportunity in itself offers a gain from exchange, as consumers can buy cheaper imported goods and producers can export goods at higher foreign prices. Further, there is a gain from specialization as the new prices established in free trade encourage industries to reallocate production from goods that the closed economy was producing at relatively high cost to goods that it was producing at relatively low cost. Thus, the basic allocative, or static, gains from trade arise from shifting the mix of outputs toward goods of comparative advantage (holding fixed the economy's technology and endowments so its PPF remains static) while permitting consumers to take advantage of the new price vector.

While this proposition is straightforward to demonstrate in theory it is far more difficult to marshal real-world data to support it. The reason is deceptively simple: the gains-from-trade theorem is, in essence, a comparison of a consumption bundle chosen at free trade with a consumption bundle chosen at autarky. The latter is virtually never observable since so few countries have existed in a state of autarky during a period in which consumption data would have been recorded. Rather, nearly all countries have engaged in some trade for centuries. There is a further problem. To compute the static gains we must be sure that the opening to trade generates a sufficiently rapid transformation of output and consumption that the data are not seriously affected by the dynamic impacts we might anticipate. These include adopting new technologies from abroad and investing in capital and skills that would shift the PPF.

There is one case from economic history that has been studied closely by Daniel Bernhofen and John Brown (2005) in order to calculate the static gains from opening to trade. Japan in the 1850s remained an essentially closed economy after two centuries of feudal rule by the Tokugawa Shogunate. Those rulers had limited trade to one Dutch ship per year and a few Chinese junks, with all goods traded by agents of the shogun's treasury. By the 1840s Japanese exports were just 1.2 cents per capita and imports just 0.4 cents per capita.[1] This seclusion cut off Japan from the economic and technological changes occurring in Europe, North America and elsewhere. However, an American naval squadron appeared in Tokyo harbor in 1853 to prompt the Japanese government to consider trading with the Western powers. In 1858 Japan signed a treaty that ended the autarky system in July 1859. These foreign powers demanded rapid elimination of quotas and tariffs. By 1866 "...the country went from nearly complete autarky to virtually free trade." (Bernhofen and Brown, p. 209). This situation continued after the Meiji Emperor's restoration in 1868.

Bernhofen and Brown argue that the rapid opening after 1859 makes this episode a unique "natural experiment" for calculating the ensuing allocative impacts for three reasons. First, Japan was a

[1]In those times East Asian trade was measured in units of the Mexican silver dollar, which was slightly more expensive than the U.S. dollar.

small country with relatively competitive product markets so that it would be a price taker on global markets. Second, the free-trade period arrived so quickly that new consumption and output data existed prior to much importation of foreign technology. Third, there was a dramatic change in relative goods prices due to the elimination of tariffs. A fourth convenient element is that Japanese history offers data on production of many tradable goods in both periods.

To estimate the gains that Japan experienced in this rapid transformation the authors use the expenditure function introduced in Chapter 3. Here we define it in terms of consumption bundles rather than general utility. Thus, the expenditure function *e(p,D)* is the minimum expenditure the economy must make at price vector p to purchase the consumption bundle *D* (demand). In this context we can define Japan's static welfare gains from trade as follows:

$$\Delta W^a = e(p^a, D^f) - e(p^a, D^a) \tag{5.6}$$

Here, the superscript a refers to the autarky period of the 1850s, just before opening up. The superscript f refers to the free-trade period of the late 1860s and early 1870s. Under revealed preference the first bundle (free-trade consumption) would not have been affordable at autarky prices. Thus, this *equivalent variation* measure of welfare change asks the question, "How much more would income have to be in autarky to permit Japan to afford the consumption bundle it chose in free trade?" This offers a computation of the implicit income (consumption) gains Japan experienced in the opening, valued at autarky prices. In essence it asks how much better off consumers would have been in autarky had they been given this additional income.

We illustrate this measure with a constant-costs PPF and two goods in Figure 5.8. Autarky consumption and production are at point A for Japan under autarky price ratio p^a ($p = p_1/p_2$). With free trade a new set price ratio p^f arrives and we assume that Japan has a comparative advantage in good X_1. With constant opportunity costs the country specializes in that good, achieving production at X^f and consumption at D^f. The level of real income in autarky, measured in units of good X_1, is the distance OX^f while in free trade it is OS. Thus, the equivalent variation measure in (5.6) can be represented as the distance X^fS. This is the amount of additional income, in units of X_1, that would have permitted the free-trade consumption bundle at autarky prices.

It is impossible to calculate the first term on the right side of (5.6) because it is a fictitious *counterfactual scenario*: what would consumption have been at autarky prices given free-trade income? However, Bernhofen and Brown note that the distance X^fS can equally be represented in terms of the amounts of international trade. Specifically, $p_1^a(SR - X^fR)$ is the added income from trade measured in terms of good X_1. In Figure 5.7, line segment X_1^fR is the quantity of exports E_1^f and the vertical segment RC^f is the quantity of imports of M_2^f. This latter amount in units of X_2 may be expressed in units of X_1, measured at autarky prices, by line segment SR: $RS = (p_2^a/p_1^a)M^f$ *or* $p_1^aSR = p_2^aM^f$ Thus, the welfare measure may be written equally as:[2]

$$\Delta W^a = p_2^aM_i^f - p_1^aE_1^f = \sum_i p_i^a(M_i^f - E_i^f) \tag{5.7}$$

Dividing this calculation by an estimate of real GDP in the 1850s provides an indication of the proportionate gains from trade.

On first glance the measure in (5.7) is not helpful since it also relies on counterfactual amounts of

[2]If the PPF is concave in the normal fashion, this expression is not strictly correct but it does provide an upper bound on the welfare change.

what the quantities of imports and exports would be in free trade if the economy faced autarky prices. These trade figures are unobservable in the data. However, Bernhofen and Brown argue that it is reasonable to use figures on trade quantities in the 1860s as approximations since the tariff cuts were the only significant policy change during the period that should affect international commerce. This is surely a strong assumption and the technique likely misstates the actual static gains from trade.

This assumption is convenient since by that time data on imports and exports were collected by a number of institutions, including the Customs Authority of Japan. Moreover, most of the goods traded could be matched closely to specific goods for which there were recorded prices before the opening up. In autarky Japan's domestic prices for cloth, yarn, iron and foodstuffs – its primary imports after liberalization – were generally more than double or triple world prices, which were cheaper due to significant technological improvements abroad. Most importantly, these prices were quite high in relation to those of Japan's primary export commodities, which were silk and tea. For example, the price of iron bars relative to tea was around nine times the corresponding ratio in London in the 1850s (Huber, 1971).

In the face of such differences, the rapid reduction of trade barriers forced a remarkably rapid adjustment of the Japanese economy. Imports per capita rose from 0.4 to 7.2 cents per capita in 1860, the first year of open trade, and eventually increased 100-fold by the 1870s. Exports per capita rose from 1.2 to 17 cents per capita by 1860 and also increased much more by the 1870s after a period of time in which more mulberry trees (to feed silkworms) and tea bushes could be made productive. These changes occurred as open trade engineered an increase in Japan's terms of trade (its index of export prices relative to import prices) by as much as 340 percent (Huber, 1971).

With such rapid structural changes we might expect large allocative gains from trade. Indeed, the calculations made by Bernhofen and Brown of equation (5.7) suggest a gain of around nine percent of GDP per year after 1865. Put more precisely, the economy's real income would have had to be nine percent higher during the late autarky period to afford the same consumption level it could have obtained if it were engaged in international trade at that time.[3]

Is a welfare gain of nine percent of real GDP a lot or a little? It may seem disappointingly small, given the insistence of trade theory on the existence of overall gains from trade. Keep in mind, however, that these are strictly the benefits of reallocating production and consumption in response to exposing an isolated economy to global prices. The calculations do not account for further potential impacts of access to superior global technologies. Indeed, in an earlier study, Huber (1971) speculated that the gains were as much as 65 percent of GDP, including the effects of technological change after trade liberalization.

Even on its own, however, a change of nine percent of GDP is not a small figure. It implies a permanently higher level of national income by at least this amount, an impact that other policy changes rarely achieve when simulated in static computational models.

5.8 Summary

In this chapter we argue that trade is a positive-sum game, not a zero-sum game. One country does not gain at the expense of another, there are mutual gains from trade. Yet conditions have to be met to guarantee that gains are captured in practice. We show that the assumptions that an economy was competitive and undistorted are needed to prove a simple gains-from-trade theorem. The ability to trade at any prices other than the country's autarky prices makes an economy strictly better off.

Yet while trade increases a country's aggregate welfare or, in the worst case, leaves it unchanged, there are still important questions about the distribution of gains from trade both between trading partners

[3]For reasons noted in the discussion of Figure 5.8, the authors regard this as an upper bound, though with a small error. However, as we suggest their use of actual trade flows to approximate counterfactual flows may understate this gain somewhat. .

and within individual countries. Between countries, there are many possible trades that are Pareto optimal, but have very different distributions of the total gains. In market economies, this is essentially determined by market forces and in Chapter 7 we will see how country size influences this issue. Within countries, the gains can be very unevenly distributed and indeed some households can certainly lose. We noted how this can occur if households have different tastes or different endowments of goods and/or factors. This will be very important later on in understanding the politics of trade policy, and why some groups will oppose liberalization or lobby in favor of protection from foreign competition despite the fact that free trade maximizes overall national income and welfare.

REFERENCES

Bernhofen, Daniel M. and John C. Brown (2005), “An Empirical Assessment of the Comparative Advantage Gains from Trade: Evidence from Japan,” *American Economic Review* 95, 208-225.

Huber, J. Richard (1971), “The Effect on Prices of Japan’s Entry into World Commerce after 1858,” *Journal of Political Economy* 79, 614-628.

Samuelson, P. A. (1939). "The Gains from International Trade. "*Canadian Journal of Economics and Political Science* 5, 195-205.

Samuelson, P. A., (1962). "The Gains from International Trade Once Again. "*Economic Journal* 72, 820-829.

Figure 5.1

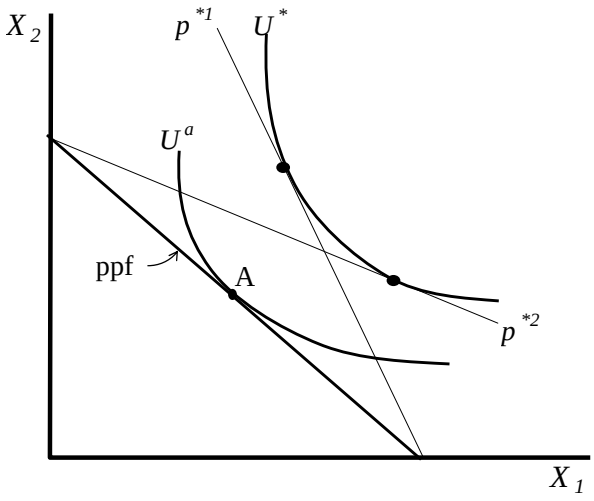


Figure 5.2

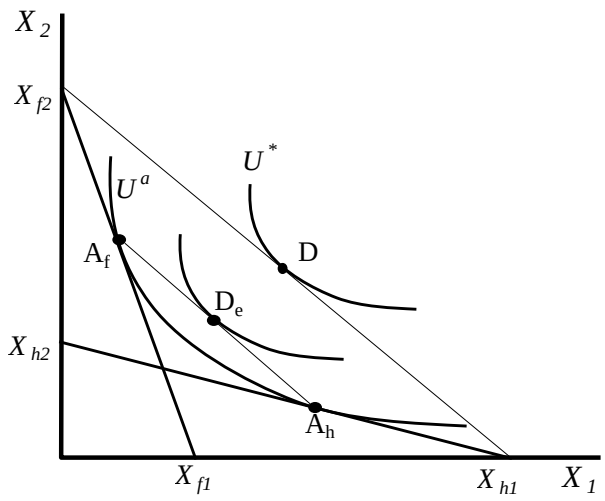


Figure 5.3

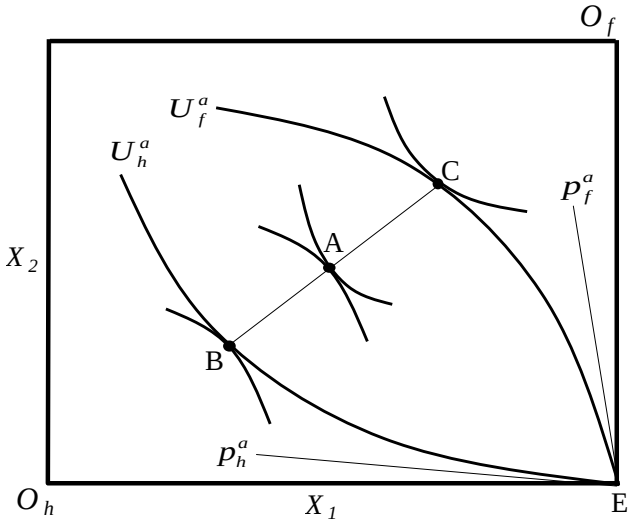


Figure 5.4

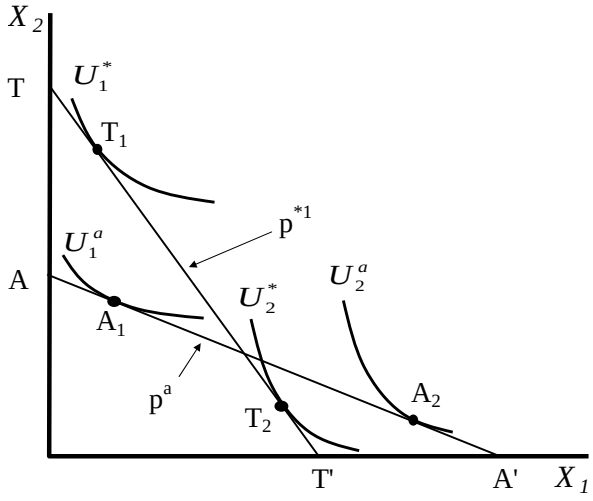


Figure 5.5

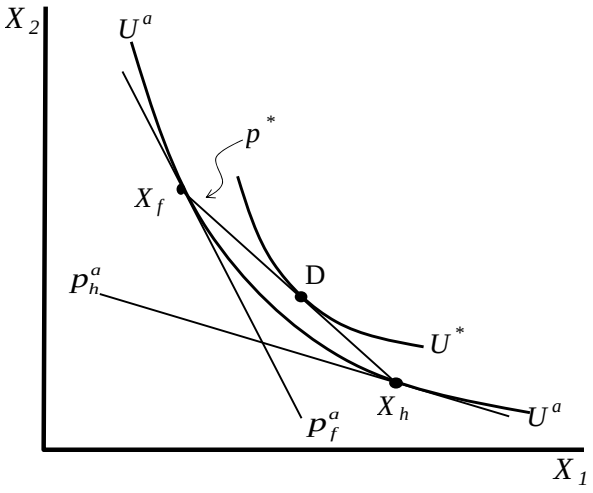


Figure 5.6

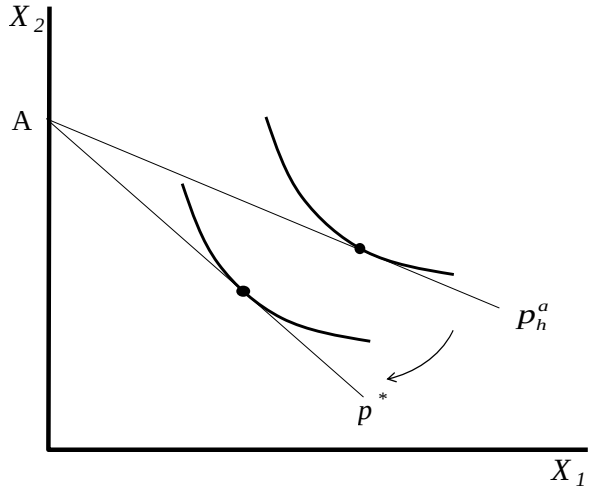
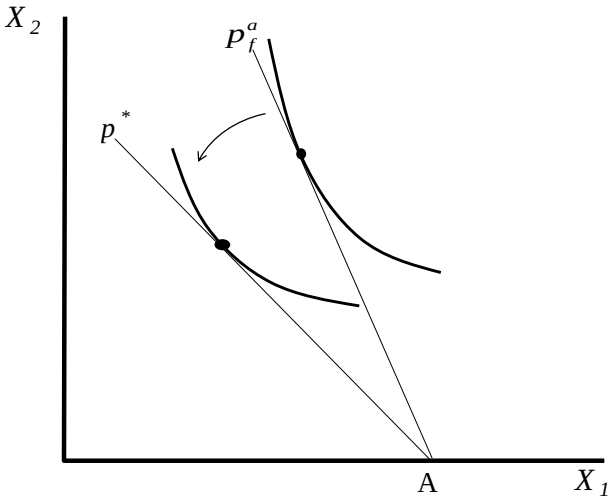


Figure 5.7



PART TWO

CAUSES AND CONSEQUENCES OF TRADE

Chapter 6

THE CAUSES OF INTERNATIONAL TRADE

6.1 The no-trade model

In the previous chapter, we emphasized that countries gain from trade by importing what is relatively costly to produce at home and by exporting what is produced relatively cheaply (efficiently) at home. What are the underlying characteristics of an economy that determine its pattern of comparative advantage? Part II of this book, comprising Chapters 6-14, turns to this question. Chapters 7-10 examine alternative sources of *differences* between countries that can give rise to trade. Chapters 11 and 12 focus on scale economies and imperfect competition, which can give rise to trade even between identical economies. Part II of the book is thus concerned with the determinants of the *direction of trade*: what determines the pattern of imports and exports by particular countries.

In fact, the trade of any country is a complex outcome of many causes all operating at the same time. There is, in general, no single cause of trade, but in order to understand the overall picture, we need to study how each possible cause of trade operates. A convenient method of examining the causes of trade is to begin by imagining a world in which there is no trade. In terms of our simple model, this would be true if all autarky price ratios are identical, and are were no scale economies. Thus we begin by imagining a situation in which all countries have identical, convex production sets, and where the same set of community indifference curves prevails in all countries. We are thus assuming that all countries can be represented by the situation of Figure 6. 1.

Figure 6.1

What assumptions are sufficient to ensure that the demand and supply situations in all countries are identical? On the demand side, it is sufficient to assume that identical and homogeneous tastes exist throughout the world. On the production side we found that three things determine the position and shape of the production possibility curve - technology, returns to scale, and factor endowments. Thus to achieve identical production possibility curves in all countries, it is clearly sufficient to assume that all countries have the same constant-returns production functions and that all countries have the same factor endowments.

These assumptions will give the same aggregate demand and supply relationships in all countries, but there is one further restriction that we must impose. We are seeking conditions such that commodity price ratios will be the same in all countries, and this will only be the case if commodity prices are determined by aggregate demand and supply. We must, in other words, ensure that equilibrium prices are determined by the tangency between the highest community indifference curve and the production possibility curve as shown in Figure 6.1, and to ensure this we assume there are no distortions, in the model. Distortions include taxes, subsidies, and imperfect competition. We can thus write down a set of five conditions that are jointly sufficient to guarantee the no-trade situation. These are:

1. Identical production functions among countries
2. The same relative endowments in all countries
3. Constant returns to scale
4. Identical and homogeneous tastes in all countries
5. No distortions (e.g., taxes, subsidies, imperfect competition)

While these five conditions are sufficient to imply that there will be no trade, there are obviously many other models that could be invented in which autarky prices would be identical so that no trade would take place. In other words, while these assumptions are sufficient for no trade, they are by no means necessary. This is illustrated in Figure 6.2, where HH′ and FF′ are the production frontiers for countries *h* and *f*, respectively. Production conditions are clearly different in the two countries, with *h*

producing relatively more X_2 and *f* relatively more X_1 at any common price ratio. Demand conditions also differ, however, and in the situation shown, the differences in demand offset the production conditions, leaving autarky prices identical.

Figure 6.2

These five conditions are important not simply because they describe a world in which there will be no trade, for such a situation is clearly not of much interest, but because they summarize the various conditions that can *cause* trade. If four of the five conditions hold, in other words, then the relaxation of the fifth will give rise to a situation in which trade will be possible. These five conditions can thus be thought of as the five broadly defined determinants of, or *bases* for trade.

6.2 Methodology

The relaxation of any one of the five assumptions listed could give rise to a situation in which international trade could take place. The approach that will be employed will be to relax each of the assumptions in turn, maintaining all four of the others, and to examine the implications for international trade. For example, to examine the implications of factor-endowment differences, it will be assumed that while preferences are identical and homogeneous across countries, all countries have the same production functions, that these production functions exhibit constant returns to scale, that endowments are proportional in all countries, and that there are no distortions in the model.

At this stage of the analysis, then, the question of whether the model is "realistic" is not a relevant one, for no claim has been made about its predictive powers. In each of the models developed in subsequent chapters, the strict assumptions made are necessary in order to isolate the effects of the particular determinant being examined. The assumptions of no distortions, of identical production functions, and so on, are made, not as descriptions of the real world, but simply as a methodological device to allow individual determinants to be considered in isolation.

While the question of realism is not relevant for the kind of *theoretical* experiments that we have just described, it is of course the principal focus of *empirical* analysis. If one is interested in empirical tests of trade models, one will be faced with the question of what assumptions are appropriate for a model used to explain real-world trade flows. If the implications of the various determinants of trade models are different, then ideally we would wish to include any variable which could cause trade. In practice, of course, some simplification will be necessary, and it will be very much up to the investigator to decide which of the variables he thinks are important and which are not, and how the model is to be constructed.

To make this last point a bit more strongly, it is reasonable to believe that *none* of conditions 1-5 hold between any two countries in the world (although in some cases, a condition would be "close" to holding). In examining the characteristics of North American versus the European Union versus Africa, for example, we would find that the US and Canada had a higher ratio of land endowment to labor endowment relative to Europe. North America and Europe have higher endowments of physical and human capital relative to unskilled labor when compared to Africa, and have superior technology as well. We would find that many important industries such as aircraft, autos, and chemicals had strong scale economies. We would find that tastes differ across countries, and that they are far from homogeneous in any country (e.g., the share of income spent on food declines steadily with per capita income). Countries have tax systems that differ significantly from one another, and many industries are characterized by small numbers of firms and significant imperfect competition (generally those industries with strong scale economies).

The assumption that two countries have only one basis for trade (only one of conditions 1-5 fails to hold) is made for the purposes of understanding that basis' individual contributions to determining trade. It is the goal of empirical analysis to determine the quantitative importance of the five bases for trade.

As a final point, we should emphasize that theories of trade based the relaxation of one of the five factors identified above are not conflicting theories. It is not the case, as it might be in physics for example, that if one theory is true, then the others must be false. All of these factors surely contribute in some positive measure to actual trade. The empirical question is thus not which are true and which are false, the question is how much does each basis contribute to total trade.

Figure 6.1

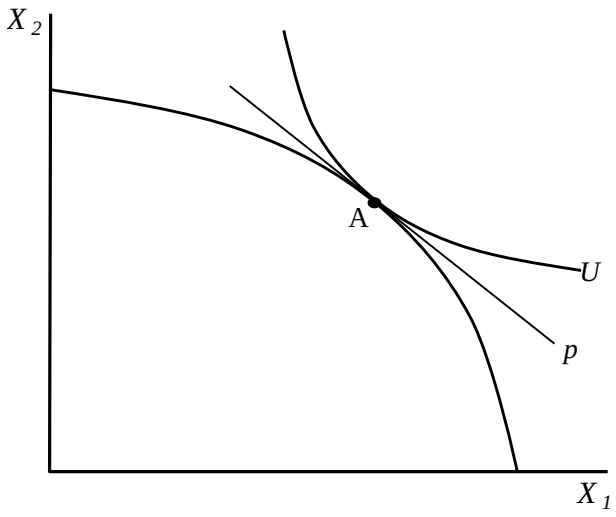
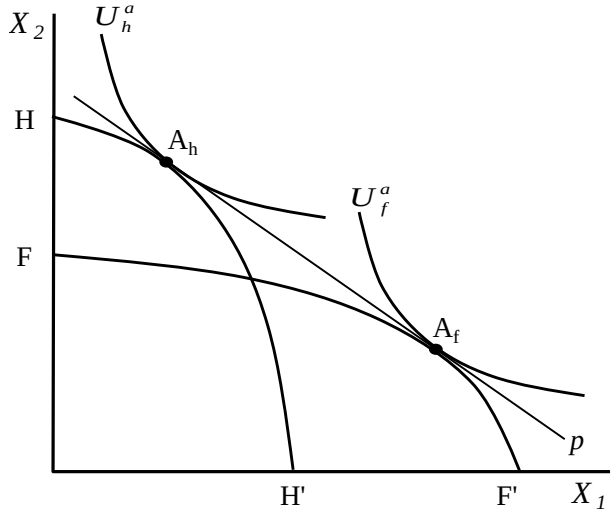


Figure 6.2



Chapter 7

DIFFERENCES IN TECHNOLOGY: THE RICARDIAN MODEL

7.1. Absolute and Comparative Advantage

The determinants-of-trade question will be analyzed Part II by relaxing, in turn, each of the five assumptions from Chapter 6 and examining the implications for international trade. The first model we consider is one in which production functions (technology) differ across countries. This model is often associated with 19th century British economist David Ricardo. In order to keep the model simple and the focus as clear as possible, we will assume that there is only one factor of production, labor, which we denote by L in this Chapter. By differences in technology, we mean that the amount of output that can be obtained from one unit of labor differs across countries. The one-factor model can be thought of as a special case of condition 2 of the previous chapter: with one factor, the issue of differences in relative endowments does not arise.[1]

Constant returns to scale are assumed. In terms of the production functions of Chapter 2, constant returns and one factor together imply that the production possibility frontier will be linear. As we will see later, this assumption significantly simplifies the analysis. We also impose the remaining conditions of Chapter 6: there are no distortions such as imperfect competition or taxes, and tastes are identical and homogeneous (in fact, the last assumption is not needed for any of the principal results).

The Ricardian model, then, assumes that labor is the only constraint on the production process. Assuming that two goods, X_1 and X_2, are produced with *constant returns*, the production functions and the labor constraint can be written in a very simple form

$$X_1 = F_1(L_1) \quad \text{specifically} \quad X_1 = \alpha_1 L_1 \tag{7.1}$$

$$X_2 = F_2(L_2) \quad \text{specifically} \quad X_2 = \alpha_2 L_2 \tag{7.2}$$

$$\bar{L} = L_1 + L_2 \tag{7.3}$$

where α_1 and α_2 are some positive constants. Note that α_1 and α_2 are the *marginal products of labor* in industries X_1 and X_2 respectively: α_1 and α_2 give the additional outputs obtained from one unit of labor.

Now introduce two countries, *h* and *f* as before. The assumption that production functions differ between countries implies that the values of α_1 and α_2 will be different in the two countries, and so we will give the α's country subscripts as well. Two pieces of terminology and the difference between them are quite important. Absolute advantage refers to the comparison of the α's for a given industry across countries. Thus we have the following example:

$$\alpha_{h2} > \alpha_{f2} \quad \text{defines country } h \text{ as having an } \textit{absolute advantage} \text{ in good } X_2 \tag{7.4}$$

The term comparative advantage refers to the relative productivity in the two industries across countries. For example:

$$\frac{\alpha_{h2}}{\alpha_{h1}} > \frac{\alpha_{f2}}{\alpha_{f1}} \quad \text{defines country } h \text{ as having a } \textit{comparative advantage} \text{ in good } X_2 \tag{7.5}$$

An important contribution of Ricardo (as we understand it), was to point out that a sufficient

condition for the existence of gains from trade is that there exists comparative advantage; that is, a strict inequality in (7.5) (whether greater than or less than) is sufficient for gains from trade. One country could have an absolute productivity advantage in both goods, but that is not relevant for trade. It is relevant to other questions such as wage comparisons between countries, and that will be discussed later.

The proof of the argument is relatively straightforward. Assume that (7.5) holds. Beginning in a situation in which both countries are producing both goods, let us move some labor in country *h* out of X_1 and into X_2 and some labor out of X_2 and into X_1 in country *f*. That is, labor is reallocated in each country toward the comparative-advantage industry.

$$dL_{h2} = -dL_{h1} > 0 \qquad dL_{f1} = -dL_{f2} > 0 \tag{7.6}$$

Then the changes in the *total world output* of the two goods will be

$$dX_1 = -\alpha_{h1}dL_{h2} - \alpha_{f1}dL_{f2} \qquad dX_2 = \alpha_{h2}dL_{h2} + \alpha_{f2}dL_{f2} \tag{7.7}$$

Set the first equation to zero, reallocating labor within each country to hold world X_1 output constant, and solve for

$$dL_{f2} = -\frac{\alpha_{h1}}{\alpha_{f1}}dL_{h2} \quad \Leftrightarrow \quad dX_1 = dX_{h1} + dX_{f1} = 0 \tag{7.8}$$

Substitute (7.8) into the right-hand equation of (7.7), replacing dL_{f2} with (7.8).

$$dX_2 = \left[\alpha_{h2} - \frac{\alpha_{h1}}{\alpha_{f1}}\alpha_{f2}\right]dL_{h2} = \alpha_{h1}\left[\frac{\alpha_{h2}}{\alpha_{h1}} - \frac{\alpha_{f2}}{\alpha_{f1}}\right]dL_{h2} > 0 \tag{7.9}$$

which follows from (7.5) and (7.6). This show that, holding the world output of one good constant, it is possible to reallocate labor in each country toward the sector of comparative advantage so that the world output of the other good rises. Note that the sign of (7.9) depends only on comparative advantage and is not invalidated if one country is absolutely more productive in both industries.

Finally, note that a pattern of comparative advantage is both a necessary and sufficient condition to be able to increase the world output of one good without decreasing the output of the other. If there is no comparative advantage, that is if (7.5) is a strict equality, then (7.9) is zero (the term in brackets is zero).

7.2 The production frontier

Figure 7.1 shows the production frontier for a single economy. The "economic size" of the country, the distance of the production frontier from the origin, is determined by a combination of its labor endowment and its absolute productivity levels. Total production capacity can be high due either to a large population or high productivity: for total GDP it will not matter which is the case, but it will certainly matter for wages discussed below. The slope of the production frontier is determined only by its comparative-advantage ratio and indeed is equal to the ratios used in (7.5).

Figure 7.1

Figure 7.2 shows production frontiers for two countries on a per-worker basis (or for countries with identical labor forces L) which satisfy the pattern of comparative advantage in (7.5). H_2H_1 is the production frontier for country *h* and F_2F_1 is the production frontier for country *f* under the assumption

that not only is (7.5) satisfied, but also each country has an absolute advantage in their comparative-advantage sector. The two production frontier cross one another. Holding country *f*'s technology constant, $H_2'H_1'$ shows the frontier for country *h* when it has an absolute advantage in both goods: *h*'s production frontier per worker will then lie entirely outside *f*'s frontier.

Figure 7.2

In a symmetric case (countries and mirror images of one another, and half of income is spent on each good) , the gains from specialization and trade are as illustrated in Figure 5.2 of the previous chapter. As an exercise, try to draw a situation where one country is absolutely economically bigger as in the case of F_2F_1 and $H_2'H_1'$ in Figure 7.2.

7.3 Excess demand and international equilibrium

The fact that countries *can* potentially gain from trade does not necessarily ensure that they *will* in fact capture these gains. If we assume that the economies are competitive and undistorted, then the gains-from-trade theorem applies and the worst outcome for one country is that it captures zero gains.

Figure 7.3 is essentially a repeat of Figure 5.1 and can be used to construct the excess demand curve M_h for one country, arbitrarily chose to be country h, in Figure 7.4. At the autarky price ratio, the slope of the production frontier H_2H_1, the country consumes at point A in Figure 7.3 but is actually indifferent to producing at any point on the production frontier and trading to point A. Thus there is a flat section of the excess demand curve for X_1 in Figure 7.4 at the autarky price ratio. Point H_1 in Figure 7.3 corresponds to being specialized in and exporting X_1 and thus to point H_1 in Figure 7.4 and point H_2 in Figure 7.3 corresponds to being fully specialized in X_2 and hence importing X_1 at point H_2 in Figure 7.4.

Figure 7.3 Figure 7.4

Similarly, price ratios p^{*1} and p^{*2} in Figure 7.3 correspond to p^{*1} and p^{*2} in Figure 7.4, where the country exports X_1 at the former and imports X_1 at the latter.

Figure 7.5 presents the excess curve for country *h*, M_h , and the excess supply curve for country *f*, M_f , based on their production frontiers in Figure 7.1. Each curve has a flat section at that country's autarky price. In the situation shown in Figure 7.5, international equilibrium occurs at price ratio p^* at which the import demand of *h* ($M_{h1} > 0$) matches the export supply ($M_{f1} < 0$) of country *f*. As we discussed earlier in the book, the equilibrium world price ratio falls between the autarky price ratios of the two countries. If this was not the case, then both countries would want to either import or export the same good. For example, if $p^* > p_h^a$, then both countries would wish to export X_1 which cannot be an equilibrium.

Figure 7.5

When the free-trade price ratio differs from each country's autarky price ratio as in Figure 7.4, we know from Chapter 5 that both countries must gain from trade. However, the gains from trade are not necessarily distributed "evenly", and the country which trades farthest away from its autarky price ratio is the large gainer. We will return to this point in section 7.5 below.

7.4 The role of absolute advantage in wage determination

To this point, we have shown that in the Ricardian model comparative advantage is determined simply by the relative productivity of labor in producing commodities, or, equivalently, by international differences in production technologies. It may seem surprising that wage rates have not entered the discussion at all. After all, there has been much concern expressed in high-income economies about the

possible effects of competition from low-wage workers in developing countries. In the Ricardian model, wages are an endogenous variable reflecting absolute advantage among other determinants. All workers gains from trade but a more productive country will have higher wages.

Suppose that both countries are specialized as in Figure 7.5, and so the wage rate in each country is determined by the competitive conditions that the value of the marginal product of labor equals the wage rate as discussed in Chapter 2.

$$p_2^* \alpha_{h2} = w_h \qquad p_1^* \alpha_{f1} = w_f \qquad \text{thus} \qquad \frac{w_h}{w_f} = \frac{p_2^* \alpha_{h2}}{p_1^* \alpha_{f1}} \tag{7.10}$$

Second, the world price ratio lies (weakly) between the autarky price ratios of the two countries.

$$\frac{\alpha_{h2}}{\alpha_{h1}} \geq \frac{p_1^*}{p_2^*} \geq \frac{\alpha_{f2}}{\alpha_{f1}} \tag{7.11}$$

Third, assume that country h is has an absolute advantage in both goods in addition to having a comparative advantage in good 2, as in Figure 7.2.

$$\alpha_{h1} > \alpha_{f1} \qquad \text{and thus} \qquad \frac{\alpha_{h2}}{\alpha_{f1}} > \frac{\alpha_{h2}}{\alpha_{h1}} \tag{7.12}$$

We can then add an element to the left-hand side of the change of inequalities in (7.11) using (7.12)

$$\frac{\alpha_{h2}}{\alpha_{f1}} > \frac{\alpha_{h2}}{\alpha_{h1}} \geq \frac{p_1^*}{p_2^*} \geq \frac{\alpha_{f2}}{\alpha_{f1}} \qquad \Rightarrow \qquad \frac{p_2^*}{p_1^*} \frac{\alpha_{h2}}{\alpha_{f1}} > 1 \tag{7.13}$$

where the right-hand inequality comes from multiplying the whole chain in (7.13) through by p_2^*/p_1^*. But the right-hand expression is, from (7.10), the ratio of the wage rates in the two countries.

$$\frac{p_2^*}{p_1^*} \frac{\alpha_{h2}}{\alpha_{f1}} = \frac{w_h}{w_f} > 1 \tag{7.14}$$

This demonstrates that if a country has an absolute advantage in both goods, then it must have a higher wage. The higher wage is a result of higher productivity and it should not be interpreted therefore as a reason for a high-wage country not to trade. Of course, this depends on a competitive market-set wage and would not be a valid proof under institutional or other forms of distortionary wage setting.

7.5 The distribution of gains from trade between countries

The Ricardian model is often used to shed some light on the question that we posed in Chapter 5: how are total gains from trade divided between two economies. One principal determinant of this division is the economic size of countries - measured either in terms of their factor endowments or their productivity levels. In other words, a country is "bigger" economically to the extent that its production frontier is further from the origin. Let us therefore conduct a thought experiment with our two countries

in which we make country *f* bigger by shifting its production frontier in a parallel fashion further out from the origin. This can be accomplished either by increasing *f*'s endowment of labor or by improving its technology: increasing α_1 and α_2 in the same proportion.

The effect of this is shown in Figure 7.6 where F_2F_1 is the initial production frontier and $F_2'F_1'$ is the expanded frontier. This expansion will leave country *f*'s autarky price ratio unchanged, and increase the length of the horizontal segment of its excess demand curve (equal to the country *f*'s maximum output of X_1, F_1'). Suppose that p^{*0} in Figure 7.6 is the initial world equilibrium price ratio, then country *f* specializes in the production of X_1 and consumes at point D^0 in Figure 7.6. Now expand the production frontier to $F_2'F_1'$. If we hold the world price ratio constant at p^{*0}, country *f* will now wish to produce at F_1' and consume at D^1. Provided that country *f* wishes to spend some of its increased income on X_2 (homogeneous demand will do, but it is not necessary), then country *f* will wish to export more X_1 and import more X_2 at the existing equilibrium price ratio p^{*0}.

Figure 7.6

The same argument can be repeated for any price ratio. The expansion of the production frontier will lead country *f* to wish to trade more at any world price ratio. The effect on country *f*'s excess supply curve is shown in Figure 7.7. M_fM_f is *f*'s initial excess (import) demand curve. The new excess supply curve is shown by $M_f'M_f'$. The horizontal segment, corresponding to the distance $0F_1'$ in Figure 7.6, expands and similarly the country *f* now desires to trade more at any price ratio other than its autarky price ratio.

Assuming that nothing has happened in country h, the old equilibrium world price ratio p^{*0} can no longer be an equilibrium. country *h* wants to import the same amount of X_1 that it did previously, but country *f* now wants to export more X_1. There is an excess supply of X_1 and the world price ratio must fall to reestablish equilibrium. This is shown in Figure 7.7 as a fall in the price ratio from p^{*0} to p^{*1}.

Figure 7.7

The change in the world price ratio due to the increase in country *f*'s size (productivity) has important implications for the gains from trade. First, Figure 7.6 could be used to show that country *f* gains less by growth than if the price ratio had stayed constant at p^{*0} (try to redraw Figure 7.6 with the price ratio falling from p^{*0} to p^{*1}). Country *f's terms of trade have deteriorated*; that is, the equilibrating price change involves a fall in the relative price of *f's export good, X_1. For* country *h* on the other hand, the price change due to *f*'s growth is beneficial. Country *h*'s *terms of trade have improved*; that is, the price change involves an increase in the relative price of *f's export good (fall in the relative price of f's* import good).

An extreme case of unequal distribution of the gains from trade between countries is shown in Figure 7.8. In this case, country *f* is sufficiently large relative to *h* that the equilibrium occurs at country *f's autarky price ratio* p_f^a. *The equilibrium world price ratio is at one extreme end of the feasible spectrum between the two autarky prices.* country *h* imports the quantity M_h while country *f* exports M_f of good X_1. In this case, country *f* captures no gains, although it is not worse off with trade. All gains from trade are captured by the smaller country *h.* Country *f* is indifferent to trade.

Figure 7.8

Two things can be learned from this exercise. First, in free trade smaller countries are likely to be major gainers from trade. This result is of considerable importance insofar as some smaller countries worry about their position vis-à-vis large countries.

The second thing to learn from this analysis is that a country may benefit from productivity growth in its trading partner. Trade is not warfare as we have noted, and a productivity growth in a trading partner is very unlike your enemy getting better weapons. Figures 7.6 to 7.8 illustrate that country

h benefits from the productivity growth in country *f* in that this productivity growth is passed on to country *h* in terms of a lower price for its import good.[2] Indeed, it is theoretically possible that the deterioration in the term of trade may be so severe for country *f* that it is worse off and all of the gains are captured by *h* (worse off than before growth, not relative to autarky!). This possibility is referred to as "immiserizing growth".

7.6 Econometric Evidence on the Ricardian Model

The Ricardian model's basic assumptions are that there is a single factor of production (labor), outputs are subject to constant returns to scale, and international variations in labor productivity are exogenously given by technological differences. These are extremely strong claims that we can reject simply by observing the world around us. There are many factors of production, including capital, land, and labor of varying skill levels. We can think of many industries that have increasing returns to scale. And, while technologies may be fixed at a point in time there are incentives to invest in better techniques or higher quality that could change the productivity rankings. Furthermore, we can easily reject a basic prediction of the model, that countries completely specialize in particular products or industries. Extreme specialization may be relevant for countries with dominant natural resources, such as Saudi Arabia and its heavy reliance on oil exports. However, this outcome is more related to resource endowments than to technolgical differences. In truth, the Ricardian model is a pronounced abstraction from reality designed to make important points about comparative advantage and the gains from trade rather than to serve as a basis for explaining real-world trade.

Still, it may be true that actual international trade flows are generated to an important degree by differences in labor productivity, making it worth studying the Ricardian model empirically. The theory makes a central claim that in principle should be testable with data and sufficient care to control for other influences on trade. Specifically, relative differences in industry-level productivity across countries should be strongly correlated with comparative differences in exports or net exports. Equivalently, within a country those industries with the highest relative labor productivities should have the highest comparative levels of exports or net exports.

Before describing one study of this hypothesis let us take a brief look at some recent data. One fundamental point of the technology-differences model worth checking first is the extent to which labor productivity and real wages are correlated. Recall from the model that countries with higher labor productivity should experience higher real wages in both autarky (which is unobservable) and free trade. In Table 7.1 we provide recent figures for several countries on a central measures of productivity in the overall manufacturing sector: value added per hour of work. *Value added* refers to the gross value of a product minus the costs of purchased intermediate inputs such as raw materials and processed components. For example, in automobiles the value added in a car is the price received by a manufacturer minus that firm's costs for the parts needed to assemble it. It is a measure of the net revenues available to pay for labor and capital (including profits) and certain taxes. This is the best measure of output actually generated in a firm or industry. Indeed, GDP is defined as the sum of value added across all producing sectors. The Table also has data on earnings per hour worked in manufacturing and annual earnings per employee.

Table 7.1 shows that in 2004 the United States had the highest levels of hourly productivity, whether the data for other countries are measured in market exchange rates or in purchasing power parity (PPP) exchange rates.[3] It is immediately evident that the richer countries are more productive, with value added per hour ranging from just 64 cents in India to $46.10 in Sweden. Hourly earnings are less than value added per hour because value added also goes to paying for capital, land, profits and some taxes. Still, there is clearly a strong relationship between productivity and the hourly wage, with Indian workers averaging 19 cents per hour, Mexican workers $1.77 per hour, and Swedish workers over $17 per hour. The correlation between the first two columns is 0.94. Average annual earnings also clearly fall as workers become less productive. Although the rankings are slightly different, the same conclusions hold for the PPP-based calculations. The correlation between hourly productivity and hourly earnings using that set of exchange rates is 0.91.

Table 7.1

It is also straightforward to show that there is a strong association between *growth* in productivity and compensation over time. Thus, for example, in a sample of 12 countries over the period 1997-2007, South Korea's value added per hour in manufacturing rose by 8.5 percent per year on average while its average real (inflation-adjusted) hourly compensation rose by over 6.0 percent per year, both the highest figures.[4] The growth rates in productivity and compensation were highly correlated in that sample.

While these data are interesting, no one should be surprised that earnings rise with labor productivity, if only because more productive workers generate more value added that supports their incomes. A considerably more rigorous test of Ricardian theory should relate figures on international trade to underlying labor-productivity coefficients that differ by industry and country. This is the approach taken by Stephen Golub and Chang-Tai Hsieh (2000) in a comparison of the United States with seven other industrialized economies and two emerging economies.[5] Their idea was to study econometrically the relationship between measures of bilateral comparative advantage, based on actual trade flows, and relative labor productivities. Thus, they specified and estimated the following econometric equation:

$$\log(X_{ijk}/M_{ijk}) = \alpha_{jk} + \beta_{jk}\log(a_{ik}/a_{jk})_{-1} + \epsilon_{ijk} \tag{7.15}$$

The variable X_{ijk} refers to exports of good *i* from country *j* to country *k*, while M_{ijk} refers to imports of the same good coming into country *j* from country *k*. Thus, the authors argued for using *net exports*, or the ratio of exports to imports, as the appropriate measure of bilateral comparative advantage.

The coefficients a_{ik} and a_{jk} are measures of inverse labor productivity, in this case employment per dollar of value added in good *i*. Note that because the coefficients of the exporting (importing) country *j* are in the numerator (denominator) of the independent variable, an increase in this ratio implies a rise in the relative productivity of the exporter. The coefficient α_{jk} is the constant term in the bilateral regression and β_{jk} s the slope coefficient, which captures the direct impact of relative productivity variations on net exports. If the Ricardian model is correct this coefficient should be significantly positive for that would imply that the country with more productive labor in a good has a higher ratio of exports to its trading partner than imports from its partner. The final term, ε_{ijk}, is a residual error capturing the unexplained difference between net exports and relative labor productivity for each good.

Four comments should be made at the outset about this specification. First, it is too simple to capture reality, for there are many variables other than labor productivity that affect trade flows. Among these are factor endowments, transport costs, and trade barriers. Thus, the authors assumed that such omitted variables are captured in the error term and need not be considered further, a claim that we discuss below. Second, astute readers may wonder why the independent variable is the ratio of absolute labor coefficients rather than the ratio, in each country, of productivity in a good relative to average manufacturing productivity. The latter is what the Ricardian model would suggest in the presence of multiple goods. In fact, this is not a problem because regression techniques actually estimate the impacts of deviations of particular observations from sample averages.

Third, the authors implicitly assumed that labor productivity coefficients are fixed, or exogenous in econometric terms, rather than being themselves a function of variables that change with international trade flows. In models other than the strict Ricardian case this is not a tenable assumption for productivity would depend on, say, factor intensities and output scales, themselves dependent on factor prices that can change with international trade. In that regard, the regressions may suffer from an *omitted variables bias*, which arises from the endogeneity of labor-value added coefficients. To try to control for this problem Golub and Hsieh simply lagged the right-hand side variable by one year to make it pre-determined. In general this is not a sufficient remedy for the endogeneity problem.

Finally, the equations are estimated using data for 21 manufacturing industries. Note that these sectors are defined quite broadly, including such items as food, beverages and tobacco; chemicals excluding drugs; electrical machinery; and motor vehicles. Thus, in each category there are hundreds of

individual commodities that are aggregated into a single dollar value of exports or imports, raising the problem of *aggregation bias*. In practice it is unlikely that labor productivity in producing baked goods is the same as that in soft drinks, or that they are identical in all kinds of electrical machinery, which covers both industrial machines and household appliances.

To estimate this equation the authors assembled data on bilateral trade between the United States (as one partner) and Japan, Germany, the UK, France, Italy, Canada, Australia, South Korea and Mexico for the 21 industries over the period 1970-1992.[6] They also found figures on value added per worker. Because value added was reported in national currencies, Golub and Hsieh translated those figures into dollars using both annual market exchange rates and PPP rates. They pooled the data for each year and estimated a *seemingly unrelated regression (SUR)* specification. In this approach the assumption is that the error terms in any yearly cross-section are independent across bilateral trading partners, and therefore not correlated, as required by basic ordinarily least squares. However, the errors are likely correlated for any bilateral pair over time because trade patterns do not change rapidly, particularly for highly aggregated industries. The SUR technique in effect estimates the 23 cross-section equations simultaneously but permits an unchanging vector of non-zero time-series correlations. It should be noted that data availability varied by country over time, so the estimation period in effect was different for each country pair and the pooled panel was not balanced.

We present their primary results on the slope coefficients in Table 7.2. Because this is a double-log specification the coefficients may be interpreted as elasticities. Thus, for example, in the slope column under market exchange rates the coefficient 0.41 for US-Germany suggests that a one-percent increase in the ratio of US to German labor productivity should expand US net exports with Germany by 0.41 percent. The main question, however, is whether these coefficients are positive and significant as suggested by the Ricardian model. In general the results are supportive: six of the nine coefficients are significantly positive using market exchange rates while three are significantly negative. These negative results for the UK, France and Korea are surprising for they indicate an inverse relationship between productivity and trade performance. However, using PPP rates seven coefficients are significantly positive, including that for Korea, while those for the UK and France become insignificant. Because PPP rates are superior means of comparing costs and productivity across countries than are the more variable market rates, it seems that the technology-differences model is well supported by the data.

Table 7.2

On this basis economists should be encouraged in thinking that comparative advantage really matters in driving trade. This is an intriguing result and supports studying the Ricardian model for its practical relevance in addition to its theoretical elegance. Still, we should not get carried away because the results can be criticized on a number of grounds. Begin with econometric issues. First, as noted, this simple regression of trade on one variable – relative labor productivities – fails to control for other key determinants of international exchange. It is easy to see from the low R^2 terms that this specification does not explain much of the variation in relative net exports.[7] Second, and more importantly, if those determinants, such as trade barriers, economic infrastructure, imperfect competition and capital and land endowments, are correlated with labor productivity the coefficients may be biased to an unknown degree. This omitted variable bias is likely to be significant in these regressions. Third, the fact that data are missing for a number of countries and years, making the panel unbalanced, raises concerns about the stability and robustness of these estimates, though this problem is unlikely to imply that the true coefficients are negative. The net result of all these shortcomings is that while the authors have succeeded in demonstrating that there is a positive correlation between relative labor productivity and net exports, they have not established a causal relationship.

Ultimately, the economic logic of the test is really what matters here. On this score one basic criticism can be raised. Specifically, the results in Table 7.2 are consistent with trade theories other than the Ricardian model. For example, it is easy to show that in a world where trade is caused by differences in factor endowments but where factor prices are not equalized, the relative productive of labor will be higher in capital-abundant countries for most or all industries.[8] Similarly where industries are characterized by increasing returns to scale labor may be more productive for this reason alone, rather

than having innate productivity differences. Thus, the econometric results may be picking up these kinds of factors and cannot be considered a true test of the Ricardian theory in the sense of accepting it and rejecting others. We conclude that labor productivity is certainly correlated with sectoral export performance but at best offers only a partial explanation for actual trade flows.

7.7 Summary

This chapter is the first of several chapter which offer explanations as to the underlying differences between economies that can lead to differences in autarky prices ratios and hence lead to trade. Here we focus on differences in production functions or technology between two countries. Countries can exploit these differences, with each country specializing in the good in which it has a *comparatively* better technology, and exporting that good in exchange for the good in which is has a comparatively poorer technology. The principal results of the chapter are as follows.

The slope of a country's production frontier reflects its *relative* abilities to produce X_1 and X_2. If these relative abilities differ between two countries, a pattern of *comparative advantage* exists. The existence of comparative advantage and potential gains from trade is quite compatible with a situation in which one country has an *absolute advantage* in producing both goods. Such a country can still benefit from trade. Absolute advantage does, however, play a role in determining real wages which can be different in the two countries. Absolute (productivity) advantage is thus important for international real wage comparisons, but not for the direction of trade.

The distribution of gains from trade between the two countries depends on several factors, but one important factor is the difference in absolute "economic" size between them, by which we mean the distance of a country's production frontier from the origin. A country is better off to the extent it trades further away from its own autarky price ratio and closer to the other country's autarky price ratio. We showed that when a country becomes bigger or more productive, its terms of trade tend to deteriorate (the equilibrium world price ratio moves closer to that country's autarky price ratio). This transfers some of the benefit of the productivity growth to the other country.

An implication of the previous point is that small countries, in general, are likely major gainers from international trade. A second implication is that countries benefit from productivity growth in trading partners, although in more general models this requires that the productivity growth be either neutral or concentrated in export sectors. The price mechanism causes this transfer of benefits insofar as countries benefit from their partner's growth through lower prices for import goods.

REFERENCES

Balassa, Bela (1963), “An Empirical Demonstration of Classical Comparative Cost Theory,” *Review of Economics and Statistics* Vol. 4, 231-238.

Deardorff, Alan V. (1979), “Weak Links in the Chain of Comparative Advantage”, Journal of International Economics 9, 197-209.

Deardorff, Alan V. (1980), “The General Validity of the Law of Comparative Advantage”, Journal of Political Economy 88, 941-957.

Dornbusch, Rudiger., Sanley Fischer, and Paul A. Samuelson (1977), "Comparative Advantage, Trade, and Payments in a Ricardian Model with a Continuum of Goods", *American Economic Review* 65, 297-308.

Golub, Stephen S. and Chang-Tai Hsieh (2000), “Classical Ricardian Theory of Comparative Advantage Revisited,” *Review of International Economics* 8, 221-234.

MacDougall, G.D.A., (1951), “British and American Export: A Study Suggested by the Theory of Comparative Advantage,” *Economic Journal* 61, 697-724.

Ricardo, David. (1817). *On the Principles of Political Economy and Taxation.* London: John Murray.

Ruffin, Roy J. (1988), "The Missing Link: The Ricardian Approach to the Factor Endowments Theory of Trade", *American Economic Review* 78, 759-772.

Stern, Robert M. (1962), “British and American Productivity and Comparative Costs in International Trade,” *Oxford Economic Papers* 14, 275-303.

Endnotes

1. In more formal terms, we could shown that the assumption of one factor is equivalent to a special case of the two-factor model: countries have identical relative factor endowments, and both goods use factors in the same proportions (the goods have identical relative factor intensities). The latter assumption is somewhat stronger than condition 2 of Chapter 6.

2. This is not a completely general result. When production frontiers are concave ("bowed out") as they will be in the next chapter, countries will in general not specialize completely. Thus if a country achieves a productivity improvement in an import-competing industry, it will wish to import less, and the terms of trade will move against its trading partner. In general, productivity improvements in a country's export industries are partially passed on to trading partners through lower prices.

3.PPP exchange rates were explained in Chapter One.

4.These figures (not shown here) were calculated by the authors using data available from the U.S. Department of Commerce, Bureau of Labor Statistics.

5.This article followed much earlier studies by MacDougall (1951), Stern (1962) and Balassa (1963).

6. These data are available from the Organization for Economic Cooperation and Development (OECD) and interested students may access them at http://puck.sourceoecd.org/vl=2724352/cl=19/nw=1/rpsv/home.htm.

7. The R^2 statistic captures the ratio of variation in the dependent variable that is explained by variation in the independent variable(s). An R^2 of 0.02 means that just two percent of the changes in net exports can be attributed to changes in labor productivities.

8. On factor-endowments and factor price equalization see Chapter 8.

Table 7.1 International Comparisons of Productivity and Wages in Manufacturing, 2004

	Market Exchange Rates			PPP Exchange Rates		
Country	VA per hour	Earnings per hour	Average Earnings	VA per hour	Earnings per hour	Average Earnings
US	$ 47.47	$ 16.15	$ 34,263.84	$ 47.47	$ 16.15	$ 34,263.84
Sweden	$ 46.10	$ 17.16	$ 33,459.65	$ 38.36	$ 14.28	$ 27,847.19
Netherlands	$ 42.85	$ 22.66	$ 41,238.26	$ 42.20	$ 22.32	$ 40,613.83
Japan	$ 38.94	$ 14.37	$ 32,506.47	$ 31.46	$ 11.61	$ 26,263.97
Australia	$ 36.94	$ 16.78	$ 33,247.49	$ 40.10	$ 18.22	$ 36,090.21
UK	$ 34.89	$ 19.22	$ 40,972.09	$ 32.34	$ 17.81	$ 37,978.26
France	$ 34.60	$ 20.37	$ 38,985.14	$ 33.62	$ 19.79	$ 37,870.91
Canada	$ 33.38	$ 15.37	$ 30,281.55	$ 36.05	$ 16.59	$ 32,702.79
Spain	$ 30.34	$ 14.91	$ 27,750.56	$ 35.86	$ 17.62	$ 32,800.87
Rep. of Korea	$ 16.40	$ 9.39	$ 23,145.03	$ 23.92	$ 13.70	$ 33,773.86
Mexico	$ 8.76	$ 1.77	$ 4,102.92	$ 12.40	$ 2.50	$ 5,811.97
Costa Rica	$ 8.57	$ 1.75	$ 4,325.54	$ 17.49	$ 3.58	$ 8,827.07
Philippines	$ 3.78	$ 0.48	$ 1,097.95	$ 15.81	$ 1.99	$ 4,588.86
Egypt	$ 3.39	$ 0.47	$ 1,374.00	$ 10.68	$ 1.48	$ 4,321.29
India	$ 0.64	$ 0.19	$ 458.55	$ 3.18	$ 0.95	$ 2,292.76

Sources: calculated by the authors from International Labor Organization, *Laborsta Database*; World Bank, *World Development Indicators*; International Monetary Fund, *International Financial Statistics*; and figures at www.NationMaster.com.

Table 7.2 Primary Results from Regressions of Bilateral Net Exports on Relative Labor Productivities

		Market Exchange Rates		PPP Exchange Rates	
Country Pair	Period	Slope (b)	R^2	Slope (b)	R^2
US-Japan	1984-91	**0.14**	0.09	**0.20**	0.10
US-Germany	1977-90	**0.46**	0.06	**0.83**	0.11
US-UK	1979-90	**-0.08**	0.03	-0.02	0.02
US-France	1978-90	**-0.21**	0.02	0.02	0.02
US-Italy	1979-89	**0.26**	0.11	**0.25**	0.01
US-Canada	1972-89	**0.41**	0.02	**0.73**	0.01
US-Australia	1981-91	**0.72**	0.05	**0.89**	0.10
US-Korea	1972-90	**-0.64**	0.02	**0.93**	0.18
US-Mexico	1980-90	**0.46**	0.14	**0.56**	0.18

Source: Golub and Hsieh (2000). Coefficients in bold are significantly different from zero at the one-percent level (99-percent confidence level), based on standard errors that are consistently estimated in the presence of heteroskedasticity.

Figure 7.1

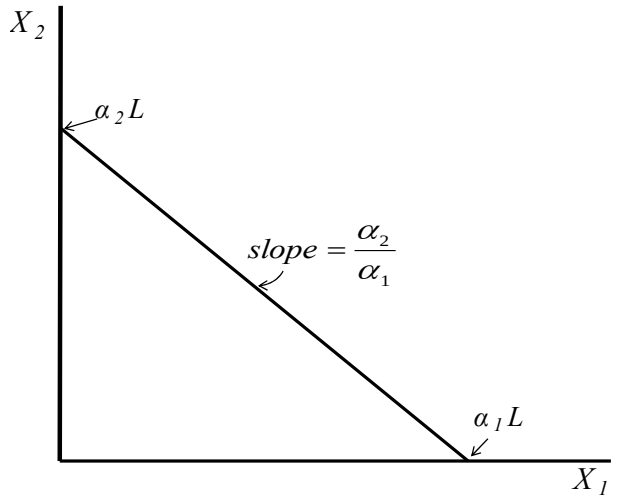


Figure 7.2

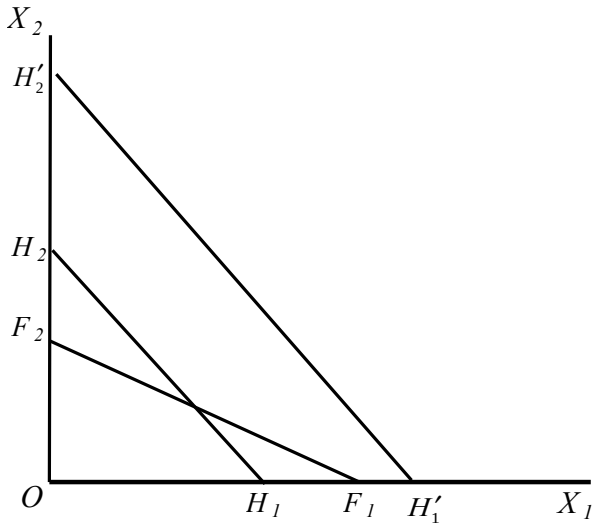


Figure 7.3

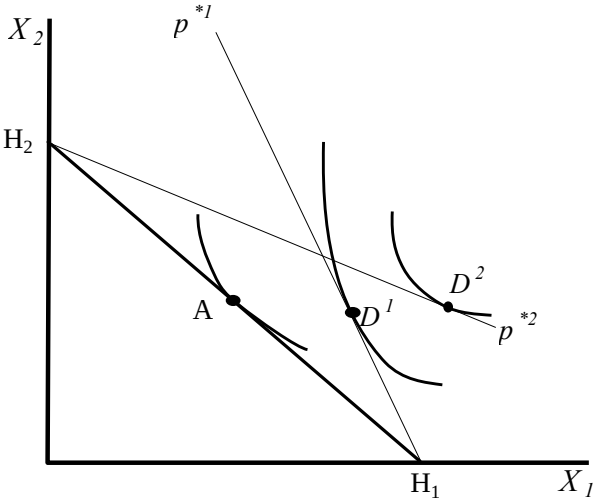


Figure 7.4

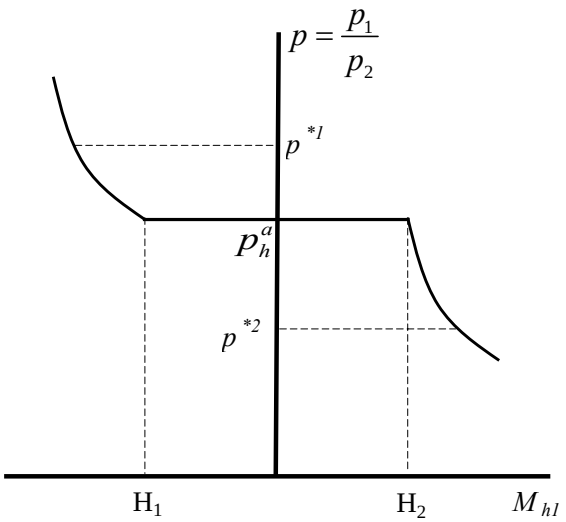


Figure 7.5

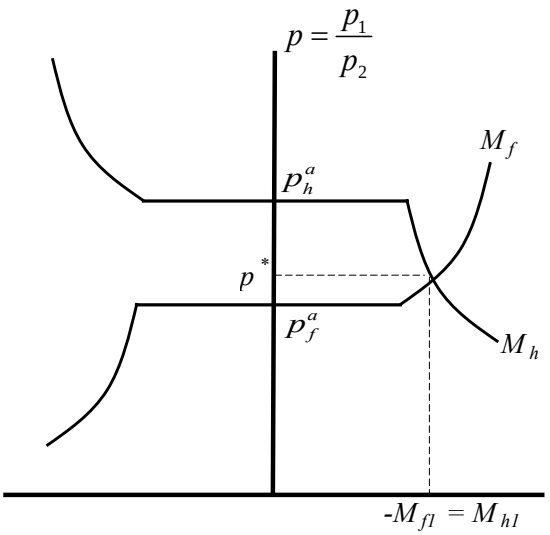


Figure 7.6

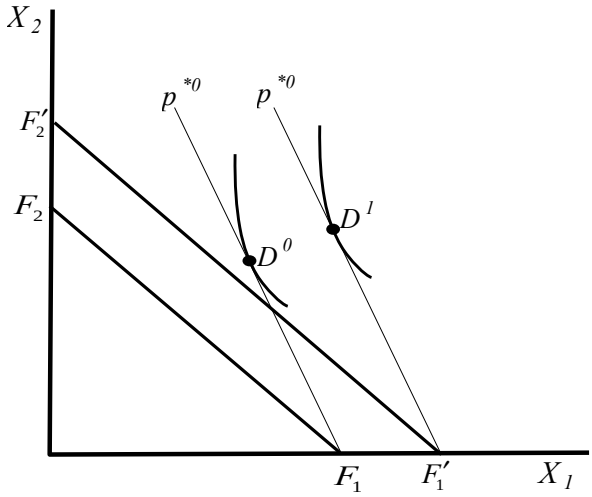


Figure 7.7

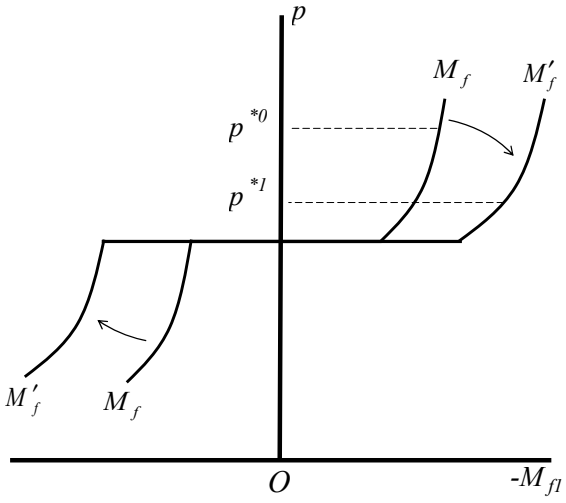
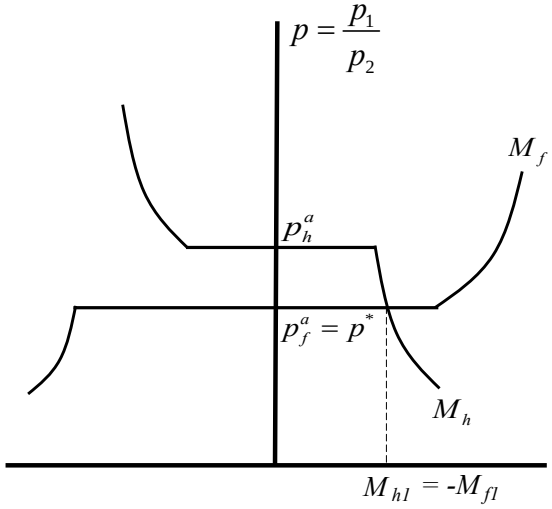


Figure 7.8



Chapter 8

DIFFERENCES IN FACTOR ENDOWMENTS I: THE HECKSCHER-OHLIN MODEL

8.1 The Heckscher-Ohlin model: an intuitive approach

We now turn to a second determinant of trade, differences in factor endowments across countries, combined with differences in factor intensities across goods. In terms of the no-trade model of Chapter 6, we are now going to re-establish the assumption that countries have identical technologies, but relax the assumption that they have identical relative factor endowments. Obviously, this will require that we add a second factor of production. All of the other assumptions of the no-trade model remain in place.

The development of this theory is generally associated with two Swedish economists, Heckscher and Ohlin. Sometime the term Heckscher-Ohlin theory is used to refer to all models in which differences in factor endowments are the driving force of trade. Sometimes it is used in a much more narrow sense to refer specifically to a model with two goods and two factors and in which both factors are mobile between or useful in both industries. We will focus on the two-good, two-factor case here and comment on more general models at the end of the next chapter.

The two-factor Heckscher-Ohlin model is considerably richer than the Ricardian model and allows for more realistic predictions. First, because the production frontier is "bowed out" as discussed in Chapter 2, countries will have much less of a tendency to specialize. In the two-country Ricardian model of the previous chapter, at least one country must be specialized. Second, Heckscher-Ohlin allows for interesting and important income distributional effects from trade. In particular, the owners of the different factors will be in conflict in their views about the desirability of liberal trade versus trade protection. This is an important first step in understanding the political economy of trade protection.

Two goods are produced from two factors which are in fixed (or "inelastic") supply. We adopt the convention of Chapter 2 that when a variable refers to both a good and factor, the first subscript is the industry and the second denotes the factor.

$$X_1 = F_1(V_{11}, V_{12}) \qquad X_2 = F_2(V_{21}, V_{22}) \tag{8.1}$$

$$\bar{V}_1 = V_{11} + V_{21} \qquad \bar{V}_2 = V_{12} + V_{22}$$

Assume that industry X_1 is intensive in the use of factor V_1 when competitive firms choose optimally, and assume that country h is relatively abundant in factor V_1. We are assuming that

$$\frac{V_{11}}{V_{12}} > \frac{V_{21}}{V_{22}} \quad \text{in both countries and} \quad \frac{\bar{V}_{h1}}{\bar{V}_{h2}} > \frac{\bar{V}_{f1}}{\bar{V}_{f2}} \tag{8.2}$$

Figures 8.1 and 8.2 present a very special case in order to provide the intuition behind the main result of the Heckscher-Ohlin model. Points H and F denote the total factor endowments of countries h and f respectively. Isoquants through these endowment points denote the amounts of X_1 and X_2 that would be produced if the country only produced that one good. Thus $\bar{X}_{h2}$ and $\bar{X}_{h1}$ are the endpoints of country h's production frontier, shown in Figure 8.2 and similarly for country f. In this special case, country h can produce absolutely more X_1 and country f can produce absolutely more X_2. Comparative advantage is derived from the intersection between relative factor endowments between countries and

relative factor intensities between industries. Each country has a comparative advantage in the good using intensively its abundant factor.

Figure 8.1 Figure 8.2

In drawing Figure 8.2, we also make symmetry assumption that the countries are mirror imagines of one another and that they spend half their income on each good. In this special case, the autarky equilibria for the two countries are at points A_h and A_f in Figure 8.2 respectively. In autarky equilibrium, each country has the same utility, but each country consumes more of the good in which it has a comparative advantage, the good using intensively its abundant factor. Note that each country will also have a low relative price for the good using intensively its abundant factor. Now permit trade. Each country will export the good using intensively its abundant factor, producing at points T_h and T_f for countries h and f respectively, and both consuming at the same point D in figure 8.2. This is the Heckscher-Ohlin theorem.

> Heckscher-Ohlin theorem: each country exports the good using intensively its relatively abundant factor.

8.2 The Heckscher-Ohlin theorem: a more formal approach

Figures 8.1 and 8.2 show a very special case. The result is much more general than this however and that is what we shall show in this section. Consider a single country and assume for the moment that both industries are producing at world prices p_1 and p_2. One of the requirements for general equilibrium that we discussed in Chapter 4 is that profits are zero. Zero profit conditions give

$$\begin{bmatrix} c_1(w_1, w_2) \\ c_2(w_1, w_2) \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} = \begin{bmatrix} p_1 \\ p_2 \end{bmatrix} \tag{8.3}$$

where again, a_{ij} is the *optimal* amount of factor j used in industry i to produce one unit of good X_i. Using the envelope theorem and its application in Shephard's lemma from Chapter 2 and equation (2.31), we can differentiate the first equation of (8.3) and arrive at the following result.

$$dc_1 = a_{11}dw_1 + a_{12}dw_2 + \left[w_1 da_{11} + w_2 da_{12}\right] = a_{11}dw_1 + a_{12}dw_2 = dp_1 \tag{8.4}$$

The term in brackets is zero: with the a_{ij}'s chosen optimally, small changes in these values have no effect on cost; this is Shepard's lemma. There is a similar equation for industry 2, and (8.4) implies that the right-hand equation of (8.3) also holds in differentials.

$$\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} dw_1 \\ dw_2 \end{bmatrix} = \begin{bmatrix} dp_1 \\ dp_2 \end{bmatrix} \tag{8.5}$$

Invert this two-equation system. By a formula typically known in economics as Cramer's Rule, the inverse mapping is given by

$$\begin{bmatrix} a_{22}/D & -a_{12}/D \\ -a_{21}/D & a_{11}/D \end{bmatrix} \begin{bmatrix} dp_1 \\ dp_2 \end{bmatrix} = \begin{bmatrix} dw_1 \\ dw_2 \end{bmatrix} \tag{8.6}$$

where D is the determinant of the a_{ij} matrix in (8.5). As in Figures 8.1 and 8.2, assume that X_1 is V_1

intensive, so that

$$\frac{a_{11}}{a_{12}} > \frac{a_{21}}{a_{22}} \quad \Rightarrow \quad a_{11}a_{22} > a_{21}a_{12} \qquad a_{11}a_{22} - a_{21}a_{12} \equiv D > 0 \tag{8.7}$$

The determinant D in (8.6) is positive, so the diagonal elements are positive and the off-diagonal elements are negative.

For the rest of the section, let $p_2 = 1$ be numeraire so p_1 is the relative price of good 1 in terms of good 2. (8.6) then implies the following relationship between factor prices and goods prices.

$$\left[\frac{dw_1}{dp_1}\right]_{dp_2=0} > 0 \qquad \left[\frac{dw_2}{dp_1}\right]_{dp_2=0} < 0 \tag{8.8}$$

Intuitively, a rise in the price of good X_1 will lead to an increase in the price of the factor used intensively in good X_1, which is factor V_1 by assumption. This in turn is associated with an optimal movement around an isoquant in each industry, substituting away from the factor whose price has increased and toward the factor whose price has decreased. Thus (8.8) in turn implies

$$\frac{da_{11}}{dp_1} < 0 \quad \frac{da_{12}}{dp_1} > 0 \quad \frac{da_{21}}{dp_1} < 0 \quad \frac{da_{22}}{dp_1} > 0 \qquad (dp_2 = 0) \tag{8.9}$$

We will discuss this in more detail later in the section on the Stolper-Samuelson theorem.

A second requirement for general equilibrium discussed in Chapter 4 is factor-market clearing. Let V_1 and V_2 denote the *exogenous* supplies of factors 1 and 2 (we drop the overbar), and X_1 and X_2 are the *endogenous* production levels. Factor market clearing requires:

$$\begin{bmatrix} a_{11} & a_{21} \\ a_{12} & a_{22} \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} = \begin{bmatrix} V_1 \\ V_2 \end{bmatrix} \tag{8.10}$$

Note that the matrix here is just the transpose $[a_{ij}]'$ of the matrix $[a_{ij}]$ in (8.3) and (8.5). Invert this mapping.

$$\begin{bmatrix} a_{22}/D & -a_{21}/D \\ -a_{12}/D & a_{11}/D \end{bmatrix} \begin{bmatrix} V_1 \\ V_2 \end{bmatrix} = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \tag{8.11}$$

Now divide the first equation by the second, and divide the numerator and the denominator by V_1.

$$\frac{X_1}{X_2} = \frac{a_{22} - a_{21}\dfrac{V_2}{V_1}}{-a_{12} + a_{11}\dfrac{V_2}{V_1}} \tag{8.12}$$

Keeping in mind (8.9), equation (8.12) establishes several crucial things.

(a) the production ratio X_1/X_2 rises with p_1/p_2. This follows from (8.9): the relative V_2 intensity of both industries rise with an increase in the relative price of p_1 (a_{22} and a_{12} rise and a_{21} and a_{11} fall) . The numerator of (8.12) gets larger and the denominator gets smaller. This should be intuitive: the relative supply of good X_1 rises with the relative price of X_1.

(b) the price ratio at which a country just begins to produce X_1 (the numerator of (8.12)) equals zero), is higher in the V_2 abundant country. The higher V_2/V_1, the higher a_{22}/a_{21} must be for the numerator to be zero, which in turn requires a higher relative price of p_1 by (8.9).

(c) the price ratio at which a country stops producing X_2 (the denominator of (8.12) equals zero) is higher in the V_2 abundant country. The higher V_2/V_1, the higher a_{12}/a_{11} must be for the denominator to be zero, which in turn requires a higher relative price of p_1.

These results imply the general-equilibrium supply curves in Figure 8.3, which graphs p_1/p_2 against X_1/X_2. The upper curve is for a relatively V_2 abundant country f and the lower curve is for a relatively V_1 abundant country h.

Figure 8.3

Now add demand, and assume that both countries have identical and homogeneous preferences. Then both countries have the identical demand curve shown by D in Figure 8.3: the ratio in which goods are consumed depends only on relative prices and not on size or income. The autarky equilibria for the two countries are at points F for country f and H for country h in Figure 8.3. Autarky price ratios are given by p_f^a and p_h^a in Figure 8.3.

The final piece of the puzzle is shown in Figure 8.4. With each country having relative low price for the good using intensively the country's relatively abundant factor as in Figure 8.3, the excess demand curves are shown in Figure 8.4 as F and H for countries f and h respectively. Free trade equilibrium is at price p^*. With trade, we then have the Heckscher-Ohlin theorem: each country exports the good using intensively the country's abundant factor.

Figure 8.4

Figures 8.3 and 8.4 are the more general versions of Figures 8.1 and 8.2, or alternatively Figures 8.1 and 8.2 are special cases of 8.3 and 8.4.

8.3 The factor-price-equalization theorem

There are three additional results associated with the Heckscher-Ohlin model. One, the factor-price-equalization theorem, relates to the two-country trade model, while the second and third strictly speaking are just comparative-statics results for a single country, but they have important trade implications. All three depend on the restrictive assumption that a country produces both goods and the factor-price-equalization theorem requires that both countries produce both goods (countries are diversified, or non-specialized). We will begin with the factor-price-equalization theorem, in part because it allows us to assess the plausibility of the non-specialization assumption.

Factor-price equalization (henceforth FPE) is the result that, under very restrictive assumptions, trade will equalize the price of each factor of production across countries even though the factors themselves are not traded: goods trade alone may equalize factor prices. Intuitively, trade in goods indirectly creates competition for factors embodied in the goods trade. The basic idea behind factor-price equalization (henceforth fpe) is seen in equation (8.3). (1) If countries have identical constant-returns

technologies, then their cost functions for goods are the same, and the same two equations in (8.3) apply to both countries. The mapping between two goods prices and two factor prices is identical across countries. Then assume that (2) trade is completely costless so that goods prices are equalized between countries, and (3) both countries produce both goods in free-trade equilibrium. Under these three assumptions, it will be true that the price of each factor is equalized across countries.[1]

Figure 8.5 presents this graphically. Instead of isoquants, it uses unit-value isoquants; that is, combinations of V_1 and V_2 that yield one unit value of output, say \$1 of output. The position of unit-value isoquants thus depends on prices: the higher a good's price, the closer is the unit-value isoquant to the origin, since less physical output is needed to generate \$1 of output. Figure 8.5 shows unit-value isoquants for X_1 and X_2, where the prices that position these isoquants are exogenous for the moment. If assumptions (1) and (2) of the previous paragraph hold, then these two unit-value isoquants will be identical in the two countries. This is step one in the argument. Step two in the proof is that there is one unit-value isocost line, combinations of V_1 and V_2 that cost \$1, tangent to the two unit-value isoquants. This is shown in Figure 8.5 as the straight iso-cost line . At each point on this line, total factor cost is \$1. If both countries are producing both goods, assumption (3) above, then the price of each factor will be identical in both countries.

Figure 8.5

Step three in the argument is that the tangencies between the unit-value isoquants and unit-isocost line establish the optimal V_2/V_1 ratios in the two industries. These are denoted a_{22}/a_{21} and a_{12}/a_{11} in Figure 8.5. The final step is to note that if the countries V_2/V_1 endowment ratio lies inside the cone spanned by the two optimal factor ratios, then it is possible to divide up the endowment so that each industry produces with its optimal V_2/V_1 ratio. The cone is often referred to as the cone of diversification.

Consider point E in Figure 8.5. Point E is the sum of the vectors $0V^2$ and $0V^1$, so the latter two give the factors allocated to industries X_2 and X_1 respectively when the optimal factor ratios are used. When the endowment point lies inside of the cone of diversification, the country will produce both goods. If this is true for both countries so that (1), (2), and (3) are all satisfied, then we have factor-price equalization.

> Factor-price equalization theorem: If (1) two undistorted competitive economies have identical technologies, (2) costless trade equalizes commodity prices between countries and (3) both countries produce both goods in equilibrium, then the price of each factor is equalized across countries.

If the country's endowment is at a point like E^2 in Figure 8.5, it will produce only X_2 with factor prices given by the slope of the X_2 isoquant at that point. If its endowment is at E^1, it will produce only X_1 with factor prices given by the slope of the X_1 isoquant at that point. This establishes the intuition why non-specialization is needed for the theorem.

The limitation of this analysis is that non-specialization is an endogenous outcome of trade, it really should not be imposed as an assumption. In order to treat it appropriately as endogenous, we turn to a world Edgeworth box in Figure 8.6, using a technique made popular by Dixit and Norman (1980). On the vertical axis is the total world endowment of V_2 and the total world endowment of V_1 is on the horizontal axis. The origin for country h is in the lower left or southwest corner and the origin for country f is at the northeast corner.

Figure 8.6

Consider the experiment in which there is only a single country and a single market for each factor, and then solve for the competitive equilibrium, which is referred to as the integrated world equilibrium. Then observe the optimal V_2/V_1 ratios chosen in the two industries and graph these vectors from each origin in Figure 8.6. This will produce the parallelogram shown in Figure 8.6. Now re-

introduce the two separate countries, assuming again (1) and (2) above. If the endowment point E which divides the total world endowment between the two countries lies inside the parallelogram, then the free-trade equilibrium supports factor prices equalization. In term of Figure 8.5, each country's endowment lies inside the cone of diversification.

8.4 The Rybczynski theorem

As noted above, there are two additional results which, strictly speaking are comparative-statics results for an individual country, but they have important trade and policy implications. The Rybczynski theorem concerns the effects of changing endowments on output holding commodity prices constant. It is valid only under the assumption that a country is non-specialized, producing both goods.

The Rybczynski theorem begins with an insight from the FPE theorem and Figures 8.5 and 8.6. If commodity prices are held constant and the country is producing both goods, then factor prices are determined and constant. This in turn determines the optimal V_2/V_1 ratios to use in the two industries as in Figure 8.5. The question addressed by the theorem can be thought of as the experiment of moving the endowment point around in Figure 8.5.

The theorem is illustrated in Figure 8.7, which is an Edgeworth box for a single country. $\bar{V}_2$ and $\bar{V}_1$ are its initial endowments, and A^0 represents its initial equilibrium at fixed commodity prices. Now increase the country's endowment of V_1 by an amount ΔV_1. The new origin for good X_2 is at O_2'. With optimal V_2/V_1 ratios pinned down by the fixed prices, the new allocation in the expanded Edgeworth box must be at point A^1, which preserves the V_2/V_1 ratios in the two industries. Note especially that the amount of factors allocated to the X_2 industry shrinks from O_2A^0 initially to $O_2'A^1$ after more V_1 is added. Factor allocation to X_1, the industry intensive in V_1, expands more than in proportion to the increase V_1 while the allocation to X_2 shrinks. This in turn leads to the Rybczynski theorem.

Figure 8.7

Rybczynski Theorem: Holding commodity prices constant, an increase in the endowment of one factor leads to a more than proportional increase in the output of the good using that factor intensively and a fall in the output of the other good.

The result is illustrated in Figure 8.8, where A^0 corresponds to A^0 in Figure 8.7 and similarly for A^1 in the two Figures. The biased change in the factor endowments leads to an even greater biased change in outputs.

Figure 8.8

A more formal treatment is as follows. Consider (8.10) above. Since the a_{ij} depend only on commodity prices (which fix factor prices), this also holds in differential form. The total derivative of the first equation in (8.10) is given by

$$a_{11}dX_1 + a_{21}dX_2 = dV_1 = \left[\frac{V_{11}}{X_1}\right]dX_1 + \left[\frac{V_{21}}{X_2}\right]dX_2 \qquad (8.13)$$

Dividing through by the total factor endowments V_1 and V_2 (drop the overbar notation), the two equations for the two factors can be written as

$$\left[\frac{V_{11}}{V_1}\right]\frac{dX_1}{X_1} + \left[\frac{V_{21}}{V_1}\right]\frac{dX_2}{X_2} = \frac{dV_1}{V_1} \tag{8.14}$$

$$\left[\frac{V_{12}}{V_2}\right]\frac{dX_1}{X_1} + \left[\frac{V_{22}}{V_2}\right]\frac{dX_2}{X_2} = \frac{dV_2}{V_2} \tag{8.15}$$

The terms in square brackets in (8.14) and (8.15) are shares of each factor used in each industry and hence they are all between zero and one. Using notation and methodology developed in a classic article by Jones (1967), denote the share of factor j which is used in good i with the parameter λ_{ij} and let proportional changes in a variable be denoted by a hat over the variable. Then our two equations become:

$$\begin{bmatrix}\lambda_{11} & \lambda_{21}\\ \lambda_{12} & \lambda_{22}\end{bmatrix}\begin{bmatrix}\hat{X}_1\\ \hat{X}_2\end{bmatrix} = \begin{bmatrix}\hat{V}_1\\ \hat{V}_2\end{bmatrix} \tag{8.16}$$

Now invert the two equation system.

$$\begin{bmatrix}\lambda_{22}/D & -\lambda_{21}/D\\ -\lambda_{12}/D & \lambda_{11}/D\end{bmatrix}\begin{bmatrix}\hat{V}_1\\ \hat{V}_2\end{bmatrix} = \begin{bmatrix}\hat{X}_1\\ \hat{X}_2\end{bmatrix} \tag{8.17}$$

$$D_\lambda = \lambda_{11}\lambda_{22} - \lambda_{12}\lambda_{21} > 0 \quad \textit{where} \quad \frac{\lambda_{22}}{\lambda_{11}\lambda_{22} - \lambda_{12}\lambda_{21}} > 1, \quad 1 > \lambda_{ij} > 0$$

Examining the determinant D and remembering that each λ_{ij} is between zero and one, the magnitudes and signs of the mapping in (8.17) are as follows.

$$\lambda^{-1}\begin{bmatrix}\hat{V}_1\\ \hat{V}_2\end{bmatrix} = \begin{bmatrix}>1 & <0\\ <0 & >1\end{bmatrix}\begin{bmatrix}\hat{V}_1\\ \hat{V}_2\end{bmatrix} = \begin{bmatrix}\hat{X}_1\\ \hat{X}_2\end{bmatrix} \tag{8.18}$$

The Rybczynski theorem gives the effect of increasing each factor individually.

$$\hat{X}_1 > \hat{V}_1 > 0 > \hat{X}_2 \qquad \hat{X}_2 > \hat{V}_2 > 0 > \hat{X}_1 \tag{8.19}$$

Jones referred to this as a “magnification” effect: a change in the endowment of one factor holding prices constant leads to magnified changes in outputs.

While the Rybczynski theorem has nothing directly to do with trade, it does have important implications for a number of empirical and policy issues connected with trade. First of all, it emphasizes that, at least for a fairly small country that has little influence over world prices, large changes in the country’s factor endowment may be absorbed by changing the composition of output with very little effect on factor prices. This is very different from the way that labor economists conceptualize the economy using the partial-equilibrium tools of supply and demand. The general-equilibrium Rybczynski theorem says that at constant world prices, the labor (or capital) demand curve is flat and big changes in

supply may have little effect on the wage (or return to capital).

Second, the Rybczynski theorem has been used to help understand large changes in the composition of output over the last two decades, particularly in East Asia. Some writers have talked about an "Asian miracle" referring to large shifts from traditional agriculture to manufacturing. However, trade economists know that, with sharply falling birth rates and very high savings rates, East Asia changed its factor endowment in away from unskilled labor and toward capital and skilled labor, and the very biased change in the composition of output is perfectly consistent with the Rybczynski theorem. There is no need to appeal to miracles.

8.5 The Stolper-Samuelson theorem

A final theorem connected with the Heckscher-Ohlin model is the Stolper-Samuelson theorem. Technical, it is "dual" to the Rybczynski theorem: the latter is a relationship in quantities, the former a relationship in prices. If one is true, the other must be true. As in the case of the Rybczynski theorem, the Stolper-Samuelson theorem is a magnification effect.

> Stolper-Samuelson Theorem: Holding factor endowments constant, an increase in the price of one good leads to a more than proportional increase in the price of the factor used intensively in producing that good and to a fall in the price of the other factor.

A graphical presentation of the result is given in Figures 8.9 and 8.10. Figure 8.9 is an Edgeworth box for a single economy as in Figure 8.6 above. Let A^0 denote the initial equilibrium in the factor markets. Now assume that the price of good X_1 increases holding p_2 constant. There will now be positive profits to be earned in the X_1 industry at initial factor prices. Holding those prices constant for the moment, the X_1 producers would want to expand along ray a_{12}/a_{11} holding the factor use ratio constant at its optimal value. X_2 would contract along ray a_{22}/a_{21}. But this cannot be an equilibrium. There would be excess demand for factor V_1 (used intensively in X_1) and excess supply of factor V_2. The price of the former must rise to clear the factor market and the price of V_2 must fall. The new equilibrium is at A^1 in Figure 8.9 with factor markets once again clearing.

Figure 8.9

The change in factor prices may be more clear from Figure 8.10, which shows the movement around a single isoquant. A^0 in Figure 8.10 corresponds to A^0 in Figure 8.9 and similarly for point A^1 in the two figures. The increased demand for V_1 and the fall in demand for V_2 at initial factor prices is re-equilibrated by a rise in w_1/w_2 in both figures and a rise in the V_2/V_1 ratios in both industries.

Figure 8.10

This graphical treatment only makes clear that the relative price w_1/w_2 rises when p_1 rises. But the theorem is stronger than that, it says that w_1 must rise more than in proportion to p_1; that is, the real price of w_1 rise and factor V_1 owners are better off no matter what they choose to consume. Refer back to the two-equation system in (8.5). We can transform this mapping into proportion changes that we did with the Rybczynski theorem. For the first equation, we have

$$\left[\frac{V_{11}}{X_1}\right]dw_1 + \left[\frac{V_{12}}{X_1}\right]dw_2 = dp_1 \quad \Rightarrow \quad \left[\frac{w_1V_{11}}{p_1X_1}\right]\frac{dw_1}{w_1} + \left[\frac{w_2V_{12}}{p_1X_1}\right]\frac{dw_2}{w_2} = \frac{dp_1}{p_1} \tag{8.20}$$

The terms in brackets are the shares of each factor's earnings in the total revenue of the industry and all are between zero and 1. Using θ_{ij} for the share of factor j's income in industry i's revenue, these can be written as

$$\begin{bmatrix} \theta_{11} & \theta_{12} \\ \theta_{21} & \theta_{22} \end{bmatrix} \begin{bmatrix} \hat{w}_1 \\ \hat{w}_2 \end{bmatrix} = \begin{bmatrix} \hat{p}_1 \\ \hat{p}_2 \end{bmatrix} \quad \textit{where} \quad D = \left[\theta_{11}\theta_{22} - \theta_{12}\theta_{21}\right] > 0, \quad 1 > \theta_{ij} > 0 \tag{8.21}$$

The sign of the determinant of the share's matrix follows from the assumption that industry 1 is intensive in factor 1. Inverting the equations in (8.21), we have

$$\begin{bmatrix} \theta_{22}/D & -\theta_{12}/D \\ -\theta_{21}/D & \theta_{11}/D \end{bmatrix} \begin{bmatrix} \hat{p}_1 \\ \hat{p}_2 \end{bmatrix} = \begin{bmatrix} \hat{w}_1 \\ \hat{w}_2 \end{bmatrix} \tag{8.22}$$

which has the sign and magnitude pattern given by

$$\theta^{-1} \begin{bmatrix} \hat{p}_1 \\ \hat{p}_2 \end{bmatrix} = \begin{bmatrix} >1 & <0 \\ <0 & >1 \end{bmatrix} \begin{bmatrix} \hat{p}_1 \\ \hat{p}_2 \end{bmatrix} = \begin{bmatrix} \hat{w}_1 \\ \hat{w}_2 \end{bmatrix} \tag{8.23}$$

As noted above, the Stolper-Samuelson theorem is given formally by the magnification relationships

$$\hat{w}_1 > \hat{p}_1 > 0 > \hat{w}_2 \qquad \hat{w}_2 > \hat{p}_2 > 0 > \hat{w}_1 \tag{8.24}$$

A biased change in commodity prices has an even more bias change in factor prices, strongly redistributing income between groups of factor owners. With a rise in p_1, V_1 owners are better off even if they only want to consume X_1 and V_2 owners are worse off even if they only want to consume X_2.

Again, the Stolper-Samuelson theorem has nothing directly to do with international trade, but it does shed light on many external changes to an economy and the effects of domestic trade liberalization or protectionist policies on income distribution. These policies and external shocks change traded commodity prices and hence redistribute income inside the country.

The result is fundamental and often taken as a starting point for analyzing the political economy of trade policy to explain why, for example, some groups will vigorously lobby for protection and against liberalization even though this reduces the aggregate income of the country. A group who might be made worse off by liberalization may understandably have little interest in the argument that it increases national welfare. The latter result does imply that it must be possible to redistribute income following liberalization in a way that makes everyone better off. Unfortunately, how to do this is difficult to calculate in practice and difficult to implement for governments, so the ability of the gainers to compensate the losers is generally not very relevant to practical politics. Much more will be said about this throughout the book.

8.6 A caveat: factor-intensive reversal (may be skipped without loss of continuity)

Unfortunately, there is one complication that can invalidate much of what we said above. The problem is that the zero-profit conditions in (8.3) may not have a unique solution. Suppose that the unit-value isoquant discussed in Figure 8.3 looks like those in figure 8.11. There are two crossings, and hence there are two cones of diversification with different factor prices. This can occur when the elasticities of substitution (curvature of the isoquants) are very different in the two industries: in the case shown, there is a zero elasticity of substitution in industry X_1 (production function (ii) in equation (2.3) of Chapter 2). If country f has its endowment in the upper cone at E_f and country h is at E_h, then both countries will

produce both goods but they will have different factor prices and use different factor intensities even if commodity prices are equalized between countries. This is referred to as factor-intensity reverse: note that for country f, good X_2 is intensive in factor V_2, but for country h, X_1 is intensive in V_2. We will not discuss this further here, but simply note that an additional assumption is needed in order to rule this out for the above theorems to be valid. Ask your professor is you have further questions.

Figure 8.11

8.7 Empirical Evidence on Factor Endowments and Trade

The Heckscher-Ohlin model makes a series of strong assumptions in order to isolate the effects of different relative factor endowments on trade between two countries. These assumptions included identical technologies with constant returns to scale, perfect competition, no factor-intensity reversals, identical and homogeneous preferences, the absence of international factor migration, and no impediments to trade. Again, it is obvious that this set of assumptions does not hold in reality and that the predictions of the model cannot be expected to hold literally. The task is really to assess how significant endowment differences are in explaining trade patterns. This complicated subject occupies our discussion in this section.

Data on International Factor Intensities and Factor Endowments

We begin by looking at straightforward measures in 2000 and 2005 of factor use by industry and factor supplies by country. In Table 8.1 we show U.S. data for 17 manufacturing industries, ranked by capital expenditures (investment) per production worker in 2005. Production labor refers to workers who are engaged in essentially blue-collar activities, such as working on an assembly line. Non-production labor refers to employees who work in essentially white-collar activities, such as clerical work and management. While this is a crude breakdown, many economists think of production workers as less skilled on average and non-production workers as more skilled on average. It is worth noting that the number of production laborers declined in nearly all industries between 2000 and 2005, reflecting the rapid decline in manufacturing jobs in the United States. At the same time, value added generally rose or fell slightly, indicating that remaining workers became more productive.

Table 8.1

Looking at the 2005 figures it is clear that there are very large differences in capital expenditures per production worker, ranging from a low of $2,882 in apparel to almost $170,000 in petroleum and coal products. Investment depends clearly on the state of demand and the oil refining sector spent heavily in 2005 because of high oil prices. It may be more informative to think of chemical products as highly capital-intensive, with a per-worker investment of almost $40,000. It is also evident that the ratio of non-production workers to production workers varies widely across sectors, ranging from a low of 0.23 in textile products to a high of 1.16 in computers and electronic products. The final column provides a rough verbal translation of these data into statements about whether the industry is capital-intensive, skill-intensive, or labor-intensive.

In Table 8.2 we offer data on three types of factor endowments for a selection of countries. The first is a measure of capital stock divided by the labor force. To compute a capital stock requires adding up aggregate investment expenditures in each economy over a period of years and depreciating capital as it ages and we use a standard procedure for this purpose. The capital figures are expressed in constant US dollars (base year 2000), with local currencies converted at PPP rates. Note that it is more appropriate to use stock data than annual investment data since the former captures all capital available for production rather than new capital. Also shown in the table are arable land in hectares per member of the labor force and an estimate of scientists engaged in private and public R&D per 1,000 members of the population. Thus, we present here figures on relative capital, land, and high-skill factor supplies.

Table 8.2

It is immediately obvious that Singapore is both the most capital-abundant and land-scarce nation

in the sample. It also has invested heavily in development of R&D skills. Japan and Finland have similar endowment profiles. We can characterize the United States, Germany, the UK and South Korea as similarly abundant in capital and skills while seeming scarce in land. Canada and Australia have both high capital endowments and arable land. Finally, the developing countries in the sample, including Mexico, Brazil, South Africa, China and India are more labor-abundant.

From such simple measures it appears likely that factor intensities and endowments interact to provide an important explanation for international trade. We now consider what various important empirical studies have concluded about this fundamental question.

The HO Theory in Multiple Dimensions

Before getting to the empirical analysis, we need to undertake some additional theory. In particular, it is obvious that there are more than two goods, two factors and two countries in the world. What can we say about the HO model when there are multiple dimensions? The basic model assumes two factors and countries, so ratios between usages of factors by industry and endowments of factors by country are sufficient to define factor intensities and factor abundances, respectively. But if there are four factors, say capital, land, skilled labor, and unskilled labor, which of these four should be in the denominator when calculating intensities? In Table 8.1 we listed the ratios of capital to production labor and non-production labor to production labor, but these choices were arbitrary. And if there are many countries, does it make sense to compare the labor abundance of, say, China, France, and the United States to that of Mexico? Fortunately we do have theoretical answers to such questions.

Begin by assuming there are two goods and two factors but many nations. Then we can define world factor endowments as the sum of individual country endowments:

$$\overline{V}_{w1} = \sum \overline{V}_{i1} \quad \text{and} \quad \overline{V}_{w2} = \sum \overline{V}_{i2} \tag{8.25}$$

Here the subscript *i* refers to countries, *i = 1...n*. It is then natural to define factor abundance in terms of each country's share of world endowments. Thus, we say that country *i* is abundant in factor 1 and scarce in factor 2 if its share of world supply of the former exceeds its share of the world supply of the latter.

$$\overline{V}_{i1} / \overline{V}_{w1} > \overline{V_{i2}} / \overline{V}_{w2}. \tag{8.26}$$

As noted in equation (8.2) we can write the full-employment or factor-market clearing conditions for both country *i* and the entire world as:

$$A^i X^i = V^i \quad \text{and} \quad A^w X^w = V^w \tag{8.27}$$

Here the *A's* are two-by-two matrixes with factor-output ratios, the *X's* are two-by-one vectors of endogenous outputs and the *V's* are two-by-one vectors of factor endowments and we have omitted the overbars. Let us assume that factor-price equalization holds, so that the *A* matrixes are the same in all countries and the world. Since they are two-by-two they can be inverted just as was done in Section 8.2. Thus, we have

$$X^i = A^{-1}V^i \quad \text{and} \quad X^w = A^{-1}V^w \tag{8.28}$$

Next, consistent with HO theory, assume that preferences are identical and homogeneous in all countries and therefore in the world also. This means that the share of any good consumed by country *i* equals a constant proportion of world consumption, with that proportion given by its share of world output or GDP:

$$C^i = s^i C^w \tag{8.29}$$

By definition, $X^w = C^w$ since the world as a whole consumes what it produces. It follows that $C^i = s^i X^w$. Now define the country's net export (exports minus imports) vector as

$$T^i = X^i - C^i = A^{-1}V^i - s^i A^{-1}V^w \tag{8.30}$$

This can be rewritten in the following fundamental form:

$$T^i = A^{-1}(V^i - s^i V^w) \tag{8.31}$$

It says that trade in goods depends on both factor intensities and the country's *excess factor endowments*, which we define as the difference between the country's own factor endowments and its consumption of the world's endowments. We now have a general definition of factor abundance: a country is abundant in a factor if its own endowment exceeds its consumption of the world's supply of that factor. In the two-factor HO model the first element of this vector is $V_{i1} - s^i V_{w1}$ and the second element is $V_{i2} - s^i V_{w2}$. Thus, we would say that country *i* is abundant in factor 1 if the first element is positive and scarce in factor 2 if the second element is negative. With just two factors and balanced trade it must be true that one excess endowment is positive and one is negative. It follows that $V_{i1} / V_{iw} > s^i > V_{i2} / V_{iw}$ and our definition is equivalent to that in equation (8.26).

Heckscher-Ohlin and Heckscher-Ohlin-Vanek Theorems

Recall from equation (8.6) that if we assume that good 1 intensively uses factor 1 then the sign pattern of the A^{-1} matrix is positive along the diagonal and negative for the two off-diagonal entries. Post-multiplying that matrix by the (+, -) sign pattern of the excess endowments vector in (8.31) we see that the net exports vector T^i has the same (+, -) pattern. That is, the economy exports the good that intensively uses the abundant factor and imports the good that intensively uses the scarce factor. This is a neat demonstration of the HO theorem in two goods and two factors.

But what if there are more goods and factors? It turns out that the simple HO model breaks down and it is easy to see why. Consider first a case where there are equal numbers of factors and goods but these are both greater than two. In that case we can continue to define excess factor endowments as positive or negative entries in the last part of (8.31). We can also invert the *A* matrix and calculate net trade vectors. However, with three or more columns and rows this inverse does not in general preserve the sign of the excess endowments vector. We cannot conclude that positive and negative excess factor endowments map directly into exports and imports.

This result implies that it is impossible to predict the pattern of trade for each commodity even if we know the technology matrix and the vector of excess factor endowments. It is possible that the capital-abundant country may not export the most capital-intensive good, for example. But in an important sense this does not matter. As the FPE theorem suggests, trade in goods is really just a veil for underlying implicit trade in factors. If an economy has an abundance of capital and land and a scarcity of labor we would expect its overall trade in goods to embody this difference, even if we cannot predict that good 1, 2, or 3 is exported or imported. Instead, we should be interested in the *factor content* of trade.

And there is a simple result that is meaningful. The theory was worked out by Jaroslav Vanek (1968) and the resulting factor content theorem is sometimes called the Heckscher-Ohlin-Vanek (HOV) theorem or Factor-Content theorem of trade. Assume there are identical and homogeneous preferences and identical, constant returns technologies and suppose that free trade in goods causes factor prices to be equalized. Then each country exports the services of its abundant factors and imports the services of its scarce factors. The proof is straightforward using the analysis above. If we had not inverted the *A* matrix in equation (8.31) we would directly have:

$$AT^i = (V^i - s^i V^w) \tag{8.32}$$

The left-hand side of this expression is called the factor content of trade since it accounts for all the capital and labor that goes into making exports less what goes into making imports for a country. With two factors, capital and labor for example, the first element of this matrix multiplication measures the total capital used to produce net exports and the second element measures the total labor used. But the right-hand side measures the excess endowments of capital and labor, respectively. Thus, the sign pattern of excess endowments is preserved in the factor content of trade.

This result, however, is true for any number of factors. Note, therefore, that (8.32) gives us a natural definition of factor abundance and factor scarcity. Specifically, for any country *i* and factor *j*

$$\text{if} \quad \frac{V_j^i}{V_j^w} > (<) s^i \quad \text{then country } i \text{ is abundant (scarce) in factor } j. \tag{8.33}$$

This expression states that if a country's share of the global endowment of a factor exceeds (is less than) its proportion of world GDP, it is abundant (scarce) in that factor. Put differently, a capital-abundant country will tend to have a higher share of the world's capital stock than it will consume of world output.[2] We will use this definition in discussing empirical evidence below.

A final question about this theory is what must be true for FPE to hold. Borrowing from our earlier analysis in the chapter, the *A* matrix can be common among countries only if the nations are incompletely specialized. With multiple dimensions this means there must be at least as many goods as factors and that these goods are produced by all countries. Then there are enough traded commodity prices to solve for a unique vector of factor prices. However, if there are more factors than goods , this solution is impossible and FPE will not hold. The specific-factors model in the following chapter is an example of this possibility.

In this context it is worth restating the FPE theorem with many dimensions. Assume that all countries share the same homogeneous preferences and all production functions are identical and display constant returns to scale. Suppose there are an arbitrary number of countries (*n*), goods (*m*) and factors (*v*) but there are at least as many goods as factors. Then in free trade all countries that produce at least *v* goods in common will have the same real factor prices.

Countries that experience common factor prices are said to be in the same *cone of diversification* (though this cone would involve *m* factor prices rather than two). Note that it is possible for different groups of countries to lie in different cones and therefore have identical factor prices within their cone but different factor prices across them.

With this background we can now consider some empirical work on the HOV theorem.

The Leontief Paradox

Our first analysis above says with just two factors and two or more goods, the capital-abundant country will find that its bundle of exports is more capital-intensive than its bundle of imports. This intuitive statement of the factor content of trade was first examined by Wassily Leontief (1953), in one of the most famous empirical study in economics. Leontief had greatly aided the American planning efforts in World War II by developing a technique for accounting for all the inputs required in the production of GNP. This technique, called input-output analysis, recognizes that the production of, say, an automobile requires *primary* inputs, such as capital and labor, in addition to *intermediate* inputs, such as steel, paint, glass, and the like. The prior production of these latter inputs also required capital and labor in addition to other intermediate inputs, which, in turn required capital and labor and so on. Leontief developed a method for assembling these various inputs into an input-output table, which could then be used to compute the total labor and capital embodied in production of any bundle of goods. In terms of our theory the *A* matrix would account for both the direct inputs of capital and labor in making a car and the indirect capital and labor used in intermediate inputs.

An obvious application of input-output analysis was to discover how much capital and labor were required to produce US exports in comparison with US imports. Note that an immediate methodological

problem arises. While it is sensible to use an American input-output table to compute the factor contents of US exports, a computation of the factor contents of US imports would require detailed and consistent data on production techniques in all foreign trading partners. This was infeasible for Leontief, so he calculated the capital and labor required to produce US goods that are similar to (or compete with) American imports. This procedure is valid theoretically under either of two conditions. First, if the factor-endowments model were true and international factor prices were equalized, each country would share the same techniques of production and using the US table would not bias the import computations. Second, if production functions exhibited fixed coefficients, or a constant ratio of capital to labor regardless of the factor-price ratio, use of the US techniques would similarly capture foreign production methods adequately. In fact, the latter justification was taken by Leontief.

Thus, Leontief calculated the capital and labor requirements in the production of a representative bundle of $1 million worth of both exports and import-competing goods in 1947. Unquestionably in that year the United States was the most capital-abundant nation in the world since Japan and Western Europe had been decimated by World War II.. Thus, the United States surely was capital-abundant and labor-scarce relative to the rest of the world, suggesting that exports should be capital-intensive and imports labor-intensive. Nevertheless, Leontief discovered that the capital-labor ratio in US import-competing goods was around $18,200, while that in US exports was perhaps $14,000. Thus, the capital intensity of imports exceeded that of exports by some 23 percent. This unexpected outcome has been termed the *Leontief Paradox.*

Leontief's famous result occasioned great surprise among economists schooled in the Heckscher-Ohlin tradition. Leontief himself was puzzled by the finding and asserted that the issue was really one of measurement. In particular, his belief was that, because of superior education and training in conjunction with better management techniques, American labor was perhaps three times more productive than foreign labor. Thus, in effective labor units the United States was really labor-abundant. While this view anticipated later thinking about labor skills (see below) it was ad hoc and unconvincing. The three-to-one ratio suggested by Leontief was not predicated on a careful evaluation of labor efficiency throughout the world, but was simply the ratio required to get the "expected" result. Moreover, to the extent it was due to better American management or entrepreneurship, we would expect the relative productivity of capital to be enhanced as well.

Numerous attempts have been made to verify Leontief's results, with mixed results. Because the United States may have been an unusual country, similar computations have been made for other countries with all manner of endowments, incomes, and market structures. In some cases the results of these studies were consistent with expectations under the endowment model. For example, Peter Heller (1976) demonstrated that Japan's international trade in the 1960s followed an interesting dual structure. Japan's exports to less-developed nations were capital-intensive and its imports were labor-intensive. However, Japan's exports to more-developed nations were relatively labor-intensive in comparison with imports. In many other studies, however, the results seemed to contradict the Heckscher-Ohlin theorem. Geographically, then, the Leontief paradox is not an isolated event.

A more compelling objection was that 1947 was not a very appropriate year for testing the endowment theory. The HO model relies on the specification of a long-run equilibrium without market distortions. It can hardly be argued that the economies of Europe and Japan were in an equilibrium in 1947, rather they were beginning a process of rapid dynamic adjustment in production and factor supplies. Robert Baldwin (1971) recalculated Leontief's ratios for the United States using the 1958 input-output table and 1962 international trade data, with the result that the paradox was still strongly in evidence. However, the analysis by Robert Stern and Keith Maskus (1981) showed that by 1972 American exports had become capital-intensive relative to imports. Thus, the paradox may have been reversed by the 1970s.

Alternative Explanations for the Paradox

A more revealing line of objections to Leontief's finding has come from noting that the assumptions of the Heckscher-Ohlin model are too strict to be believed. Indeed, the most enduring and valuable outcome from the debate over the Leontief paradox is that it stimulated trade theorists to think

more fully about the implications of departures from those assumptions.

For example, it is possible that the international structure of trade barriers could partially explain Leontief's result. Our analysis of the endowment model showed that free trade would lower the real incomes of each country's scarce factor, providing an incentive for that factor to lobby for import protection. Thus, the United States might be expected to have high trade barriers to labor-intensive imports and some foreign countries might erect restrictions on capital-intensive imports. These policies could reduce the levels of trade below those expected from endowment-based comparative advantage. While this possibility is important, it has proved to be most difficult to test conclusively because tariffs have complicated effects across sectors in general equilibrium.

Another possibility is that preferences differ across countries rather than being identical and homogeneous. Certainly if there are significant *taste biases* in the sense that some countries have strong preferences to consume the goods in which they would otherwise have a comparative advantage, the pattern of trade could be reversed. An additional important possibility is that countries do not share access to identical technologies. This observation is at the core of the so-called product-cycle model and related ideas about trade in a more dynamic context.[i] The issue here is that trade data for a particular year may not reflect a long-run, static equilibrium so much as a short-run, dynamic transition under varying technologies. In the product cycle, for example, American exports of new goods may seem to be labor-intensive when, in reality, they make relatively heavy use of new technological information through the employment of highly skilled, technical labor inputs such as engineers and scientists.

The most substantive practical objection to Leontief's procedure is simply that it is inadequate to suppose that there are only two primary factors of production, capital and labor, in the world. Various forms of land and natural resources also serve as sources of comparative advantage. Indeed, these may be the most relevant factors for the Heckscher-Ohlin model because they are internationally immobile. Further, physical capital and labor exist in different forms. We should not expect the average worker in higher-income countries to share identical productivity characteristics with the average worker in lower-income countries. Rather, national labor forces consist of different endowments of laborers of various skills, with skills being higher in countries that invest more in education and training. Because acquiring new skills often involves a lengthy investment process, it is reasonable to suppose that laborers of widely varying skills do constitute different endowments. For this reason, laborers are often distinguished by their *human capital*, or accumulated investments in education and training. Indeed, trade economists agree that a fundamental determinant of US comparative advantage is a relatively abundant supply of highly-skilled labor.

Recent Studies of Factor Endowments and Trade

A full re-examination of Leontief's procedure was launched by Edward Leamer (1980), who worked out much of the theory above and the definition of factor abundance in equation (8.26). He noted that with more than three factors of production, or in the presence of unbalanced trade, a simple comparison of capital-labor ratios in exports versus imports was incorrect in revealing abundance. Furthermore, to compute the factor contents of $1 million worth of exports and import-competing goods, as Leontief did, essentially imposed balanced trade on data that were not generated in a two-factor, balanced-trade world. In fact, the United States was a net exporter of both capital and labor services in 1947, largely because that country ran a significant trade surplus.

As Leamer noted, condition (8.26) is equivalent to having the ratio of net exports of factor 1 to net exports of factor 2 exceed the ratio of those factor contents in national consumption. Thus, he used Leontief's data to calculate the capital and labor contents of net exports (exports minus imports) in total, rather than for $1 million of each. He found that there were net capital exports of $23.4 billion and net labor exports of 2 million person-years in 1947. Taking the ratio, Leamer calculated a capital-labor ratio in net exports of $11,783 per person-year. In comparison, the capital-labor ratio in overall U.S. consumption was just $6,737 per person-year. In his reckoning, this meant there never was a Leontief Paradox!

Unfortunately, complex matters are rarely that simple to resolve. Look again at condition (8.33),

which states that for a country to be revealed abundant in a factor its share in the global endowment must exceed its share in world GDP. Equivalently, labor could be an abundant factor only if American consumption per worker was less than global consumption per worker. As Richard Brecher and Ehsan Choudhri (1982) pointed out, in 1947 this condition clearly did not hold, for expenditure per worker in western Europe was approximately 50 percent of that in the United States. A different version of the Paradox was in the data after all.

This confusion led economists to work out how to test the HOV model itself rather than simply attempting to "reveal" factor abundance through the factor content of trade. Go back to equation (8.32), which is the HOV theorem. We can label the left-hand side as the actual factor content of net exports, for it is calculated by pre-multiplying the vector of net exports of goods by the matrix of factor intensities. This can be computed for any country, given the *A* matrix and the trade pattern. The right-hand side can be called the predicted factor content of net exports. It simply states that if we have measures of a nation's factor endowments, its share in global GDP, and the world's factor endowments we can calculate what its net exports of factor services must be. Thus, to test HOV we can estimate just how well equation (8.32) fits international data.

Before looking at available evidence consider the data in Table 8.3, where we list for a selection of countries their shares, in year 2000, of world GDP, world capital stock, world arable land, and world supplies of four types of labor broken down by education attainment.[ii] Thus, these data directly measure the comparison in equation (8.33) and are, therefore, theoretically correct estimates of national factor abundance. The first column lists each nation's proportion of global GDP, measured in constant dollars at PPP exchange rates. Thus, in 2000 the United States produced 27.1% of world output, Japan produced 8.9%, China produced 8.3% and so on.

Table 8.3

According to (8.33) if a country's proportion of a world endowment exceeds (is less than) its share of global GDP it is abundant (scarce) in that factor. The United States, for example, has lower shares of almost all endowments than its GDP share, with the exception of post-secondary school graduates and R&D scientists. Thus, it seems to be scarce in capital, land and lesser-educated workers while abundant in higher-educated workers. Canada and Australia are abundant in land while Japan is scarce in land and abundant in capital and research scientists. Brazil is abundant in land while India is scarce in capital and scientists while abundant in lower-educated workers.

A closer look reveals a puzzling anomaly, however. Germany and the UK seem to be scarce in all factors, while the United States, Canada and Australia are scarce in nearly all factors. On the other hand, China is abundant in all factors while India is abundant in all but two. In general, poor countries have disproportionately large numbers of abundant factors and rich countries have disproportionately large numbers of scarce factors. This data feature was first noticed by Daniel Trefler (1995) using figures from 33 countries in 1983. He called it the "endowments paradox" and it is an enduring characteristic of actual data that must be explained. One obvious hypothesis is that factors in poor countries are not used very productively, in which case they generate less output (and a smaller share of GDP) than expected. In contrast, factors in rich countries are used quite productively and generate high levels of output. We therefore have one hint at how to think about relaxing HOV: underlying technologies in different country groups are not the same.

Let us return to testing the theorem. We could start with a weak version of testing condition (8.32), called the *sign test*: the sign pattern of the actual factor content of net exports should be the same as the sign pattern of excess factor endowments. That is, if capital (unskilled labor) were abundant (scarce) in an economy its excess endowment would be positive (negative). Under HOV we should observe net exports of capital services (positive sign) and net imports of unskilled-labor services (negative sign). This test was introduced into the literature by Maskus (1985) and Bowen, Leamer and Sveikauskas (BLS; 1987). It is worth noting some results from the BLS paper. Those authors used the 1967 U.S. input-output table to measure factor inputs (the *A* matrix in our theory), arguing that they should be common across all countries. They also developed measures of 12 types of factor endowments in 1966 and international trade for 27 countries in 1967 in order to calculate the factor contents of net exports and

excess endowments for many nations. They then computed the signs of each factor content and excess endowment. Under HOV, these signs should match completely. However, for the 12 factors the sign matches ranged from a low of 22 percent (managerial labor) to a high of 78 percent (professional and technical workers), with most being around 50 to 60 percent. For the 27 countries the sign matches ranged from 25 percent (Norway and France) to 100 percent (Hong Kong), again with most lying between 40 and 60 percent. These results meant that simple data applied to the HOV theorem produced correct matches only half the time, which is no better than simply flipping a coin!

A further puzzle about these data was noticed by Trefler (1995). In nearly all countries and factors, the factor contents calculated on the left-hand side of condition (8.32) were near zero, implying that actual trade in factor services is much less than its factor-endowments prediction. The author referred to this outcome as the mystery of "missing trade". It appeared from these various studies that the factor-proportions model really does not offer a predictable guide to understanding how factors actually get traded in the world. The endowment paradox, missing trade, and the failure of the sign tests were yet additional surprising findings that spurred economists to think about other explanations of international trade.

Fundamentally, however, this situation was dissatisfying for it failed to explain *why* available data did not meet the basic HOV specification or what accounts for missing trade. Trefler made an attempt to answer these questions but we review here the most informative study. The key insight in this literature, as noted by Donald Davis and David Weinstein (1998), is to recognize that successive relaxations of the basic assumptions underlying the HOV theorem in (8.32) should improve the statistical fit between factor contents of trade and (adjusted) excess endowments.
Thus, one could start an econometric investigation simply by regressing data on actual factor contents of trade against the factor contents predicted by excess endowments vector. Thus, the baseline specification would be:

$$\text{HOV:} \quad AT^i = (V^i - s^i V^w) \qquad \text{(H1)}$$

The observations in this regression would comprise however many factors are in the *A* (input-output) matrix, multiplied by the number of countries in the data set. Recall that the entries in this matrix are the ratios of input used per dollar of output in each industry. Davis and Weinstein employed two factors (labor and capital) and 11 countries, made up of 10 developed nations and an aggregate of other nations they call the rest of the world (ROW), so there were just 22 data points. Note that if HOV held closely in the data, the slope coefficient estimated on excess endowments should be close to unity in value. However, we already know from earlier sign tests that HOV1 can be rejected so we should not expect this outcome to hold.

A first objection to this baseline specification is that it assumes all countries have the same technology matrix *A*, which is taken from the United States. This is obviously unlikely, as Leontief noted in his work. Thus, Davis and Weinstein assembled input-output tables for each of the 10 developed countries, with the 34 industries in the tables defined consistently.[iii] In principle production functions can differ across countries in many ways. However, to make the problem tractable, a second specification the authors considered was that each country's input-output matrix differed from the U.S. matrix by a Hicks-neutral factor. This simply means that the productivity of all factors in a country vary from U.S. levels by a single scalar parameter for each country. This may be stated as $A^i = \theta^i A$, where if $\theta^i > 1$ then country *i* has higher input-output ratios and is, therefore, less productive than the United States. They estimated these parameters by regressing the 68 entries (2 factors x 34 industries) in each country's matrix on the corresponding U.S. entries, with the intercept term measuring the Hicks-neutral factor. The authors found that these parameters ranged from 1.1 (Netherlands and Germany) to 2.0 (Italy). Note also that if factors are less productive we must adjust their endowments downward to reflect the *effective endowments* available in each country. For example, if Italy's productivity parameter is 2, we need to define its effective factor supplies as $V^E = V / 2$. With this theory, the second specification becomes

HOV-HN: $$\theta^i AT^i = (V^{Ei} - s^i V^{Ew}) \qquad \text{(H2)}$$

Davis and Weinstein next noted that these specifications rely on the rigorous assumption of relative factor-price equalization (FPE).[iv] However, if FPE does not hold and countries are in different cones of specialization, we know from section 8.X that factor-output ratios will differ across countries. Indeed, in more capital-abundant countries the techniques should generally be more capital-intensive. In this case the national input-output matrixes would differ systematically with capital-labor endowments. The authors add this idea to the possibility of a Hicks-Neutral average productivity difference by regressing input-output coefficients for each country on an intercept term (average HN term) and an interaction term between country capital-labor endowments and dummy variables for traded and non-traded goods. They defined traded goods as merchandise and non-traded goods as services. Their estimates suggested that a one-percent increase in a country's capital-labor endowment ratio raised the average traded-goods industry's capital-labor input ratio by 0.8 percent. Thus, non-FPE seems important in explaining actual factor intensities and countries may well lie in different cones of production.

Translating this theory into an HOV-type equation is complex and the details are left to interested readers who may consult the original article. The new specification must account for there being different factor intensities across countries in traded goods (due to non-FPE) and also varying intensities within each country between traded and non-traded goods. Thus, the technology matrixes A^i must be split into two components for each country. In turn, consumption in the economy must be split between traded goods (at common international prices) and non-traded goods (which can have different prices across countries). For that purpose Davis and Weinstein assumed that preferences for traded goods are identical and homogeneous in all countries, as in HOV, so each country's share of world income is its share of consumption of global trade in each good. From this they could determine import demands. Finally, preferences between merchandise and services within each country were assumed to be Cobb-Douglas, permitting them to identify the consumption share of non-traded goods. The next specification is therefore:

HOV-nonFPE: $$A^{iT} X^{iT} - (A^{iT} C_d^{iT} + \sum_{j \neq i} A^{jT} M_{ij}) = V^{iT} - s^i V^{wT} \; . \qquad \text{(H3)}$$

The first term on the left-hand side is the factor contents of traded goods production. The second term is the sum of two components. First is the factor contents of domestic consumption of home-produced traded goods. Second is the factor contents of imports, where each foreign country *j's* technology matrix is used. Thus, in this case the authors employed the different input-output matrixes directly. The right-hand side is the difference between country *i's* and the world's endowments used in producing tradeable goods. This equation is scaled for Hicks-Neutral productivity differences also.

In a final specification Davis and Weinstein wanted to account for the possibility that international trade flows (and therefore factor contents) are affected by trade restrictions. For this purpose they estimated a so-called *gravity model* of bilateral trade flows.[v] This model assumes identical and homogeneous preferences so that the demand for imports in country *i* for a product made in country *j* is equal to importers' share in world income times the exporter's production of the good. However, this demand is reduced by trade restrictions between the countries, including the geographical distance between them. In their estimation they included just distance and assumed other influences were captured by a statistical error term. From the first-stage estimation they computed the predicted levels of both domestic consumption and bilateral trade flows in each good. This approach generates the following equation:

HOV-nonFPE-Gravity: $$FCT^{iT} = V^{iT} - (A^{iT} \overline{C}_d^{iT} + \sum_{j \neq i} A^{jT} \overline{M}_{ij}) \qquad \text{(H4)}$$

The left-hand side is the same as that in (H3). The term in parentheses on the right-hand side is the sum of the country' factor contents of its predicted domestic consumption of traded goods and its predicted imports of traded goods. This specification is a considerable departure from the HOV tradition because it

uses the factor contents of predicted commodity flows to measure a country's use of world factors through consumption and imports rather than direct world factor endowments. Note that by employing data constructed in this way the resulting regression fit is almost certain to improve.

It remains to see how well these hypotheses fit the data. In Table 8.4 we provide the basic results reported in the Davis-Weinstein article. The first row of data lists slope coefficients on each regression, which should be close to unity under the HOV theorem, while the second row lists the standard errors of those coefficients. The third row provides the coefficients of determination, or R^2 terms, while the fourth row offers results from sign tests of each specification. In each case the sample size is 22 observations, which comprise two factors for 11 countries.

Table 8.4

It can be seen immediately that the basic HOV equation, using the U.S. technology matrix for all countries, utterly fails to comport with the data. The slope coefficient is essentially zero and insignificant, the equation explains only one percent of variation in the factor contents of net trade, and the sign test is accurate in only 32 percent of the cases. .The simple Hicks-neutral correction in the second data column does not fare much better and the sign test still reveals a coin flip.

The data fit improves considerably when the possibility of non-FPE is introduced through differing production cones in hypothesis H3. The slope coefficient becomes significantly positive, though it is still well below unity, the R^2 rises to 0.96 and the sign test is successful in 86 percent of the cases. The final case, which brings in predicted consumption and trade flows via the gravity equation, raises the slope coefficient to 0.82, close to unity, while the sign test works even better. On this analysis, then, it seems that global data support a depiction of world trade in which countries have broadly Hicks-neutral production functions but have sufficiently different endowments that they produce in different cones under unequal factor prices. Finally, demand patterns and distance do affect the factor contents of trade.

It seems that Davis and Weinstein indeed provided an "account of global factor trade", as they called it. Put differently, there is no longer "missing trade". This is important evidence about the nature of world trade. But it raises a fundamental question: if we must introduce different technologies, variable factor prices, non-traded goods, and trade restrictions into the model, what is left of the HOV theorem? We offer two answers. First, it is obvious from simple observation of the world that factor prices really are not equalized by trade in goods, even if there is a tendency in that direction in the long run. Land prices remain high in Japan and Western Europe, skilled labor commands high wages in the United States, and low-skilled labor remains relatively inexpensive in China and Vietnam. These prices reflect differing productivities and it is not surprising that a simple specification of technology would not capture them.[vi] The HOV theorem is a theoretical abstraction that explains important general-equilibrium concepts and its assumptions cannot be expected to hold in practice. Still, it appears that the extension of the model to a world in which factor prices are not equalized, and capital-abundant (labor-abundant) countries use more capital-intensive (labor-intensive) techniques, fits actual data rather well. Thus, we consider this evidence to be largely consistent with underlying factor-proportions theory.

A second answer, however, is that distance, preferences, and other variables outside the HOV context seem to matter in driving factor trade also. Accordingly, trade theorists have developed alternative theories of product differentiation and firm-level productivity that now stand alongside Heckscher-Ohlin as basic explanations for trade. We consider these theories in detail in later chapters.

8.8 Summary

This chapter has considered a model in which the only important difference between countries is in relative factor endowments. We assume that economies are alike in all other respects, including having equal access to constant-returns-to-scale technologies, sharing homogeneous preferences, and exhibiting no market or policy distortions. Thus, the model differs importantly from the technology-based theory of trade in the prior chapter. With this approach we established several important propositions about economic structure and trade patterns.

A country will export the commodity that uses the abundant factor more intensively, which is the prediction of the Heckscher-Ohlin theorem. Thus, comparative advantage is determined by the structure of factor endowments (or autarky factor prices) in conjunction with relative factor intensities of commodities.

A powerful insight about the effects of trade comes from the factor-price-equalization theorem, in which free trade in goods actually equalizes the relative and absolute prices of homogeneous factors internationally. The essential reason for this result is that trade in goods can substitute for trade in factors. The assumptions under which this theorem holds are severely restrictive and there are numerous reasons why we do not observe such equalization in practice. But there are important tendencies in that direction to the extent that international trade is the result of variations in factor endowments.

The Stolper-Samuelson theorem, which relates changes in commodity prices to changes in real factor prices, provides a fundamental prediction about the effects of trade (or impediments to trade) on the distribution of real incomes between (for example) capital and labor. Because free trade causes exports and imports to rise, it follows that the relatively abundant factor gains real income in each country and the scarce factor loses real income. The gains-from-trade theorem is relevant here, in that the economy overall enjoys a welfare rise in moving from autarky to free trade. However, the theorem will predict political conflicts within a country over trade liberalization versus protection in the absence of an explicit mechanism to redistribute gains.

The Rybczynski theorem, which relates changes in factor endowments to changes in commodity outputs, assuming constant commodity and factor prices, provides the theoretical basis for the Heckscher-Ohlin model. This theorem is important for understanding the effects of factor growth on the evolution of comparative advantage. The theorem is also important for understanding that large changes in factor supplies, especially in a small country with little influence over world prices, can be absorbed by changes in the composition of output with possibly very small changes in factor prices.

REFERENCES

Baldwin, Robert E. (1971). "Determinants of the Commodity Structure of U.S. Trade." *American Economic Review* 61, 126-146.

Bowen, Harry P., Edward E. Leamer and Leo Sveikauskas (1987). "Multicountry, Multifactor Tests of the Factor Abundance Theory," *American Economic Review* 77, 791-809.

Brecher, Richard A. and Ehsan Choudhri (1982). "The Leontief Paradox: Continued," *Journal of Political Economy* 90, 820-823.

Davis, Donald R. and David E. Weinstein (2001). "An Account of Global Factor Trade," *American Economic Review* 91, 1423-1453.

Dixit, Avinash K. and Victor Norman (1980), *Theory of International Trade*, London: Cambridge University Press.

Debaere, Peter (2003), "Relative Factor Abundance and Trade", *Journal of Political Economy* 111, 598-610.

Harrigan, James (1997), "Technology, Factor Supplies, and International Specialization: Estimating the Neoclassical Model, *American Economic Review* 87, 475-494.

Heller, Peter S. (1976). "Factor Endowment Changes and Comparative Advantage: The Case of Japan, 1959-1969." *Review of Economics and Statistics* 58, 283-292.

Jones, Ronald. W. (1965). "The Structure of Simple General Equilibrium Models." *Journal of Political Economy* 73, 557-572.

Maskus, Keith E. and S.Nishioka (2009), "Development-Related Biases in Factor Productivities and the HOV Model of Trade", *Canadian Journal of Economics* 42, 5-9-553.

Ohlin, Bertil (1967), *Interregional and International Trade*, Cambridge: Harvard University Press.

Rybczynski, T. N. (1955). "Factor Endowments and Relative Commodity Prices." *Economica* 22, 336-341.

Samuelson, Paul. A. (1948). "International Trade and the Equalization of Factor Prices." *Economic Journal* 58, 163-184.

Samuelson, Paul. A. (1949). "International Factor Price Equalization Once Again." *Economic Journal* 59, 181-197.

Samuelson, Paul. A. (1953). "Price of Factors and Goods in General Equilibrium." *Review of Economic Studies* 21, 1-20.

Schott, Peter K. (2003), "One Size Fits All? Heckscher-Ohlin Specialization in Global Production", American Economic Review 93, 686-708.

Stern, Robert M. and Keith E. Maskus (1981). "Determinants of the Structure of U.S. Foreign Trade, 1958-76." *Journal of International Economics* 11, 207-224.

Stolper, Wolfgang. F., and Paul. A. Samuelson (1941). "Protection and Real Wages." *Review of Economic Studies* 9, 58-73.

Trefler, Daniel (1995). "The Case of the Missing Trade and Other Mysteries," *American Economic Review* 85, 1029-1046.

Woodland, Alan D. (1982), *International Trade and Resource Allocation*, Amsterdam: North Holland.

Endnotes

1. Unfortunately, these three assumptions are not quite sufficient. It is possible that the two unit-value isoquants in Figure 8.3 cross twice, creating a situation know as factor-intensity reversal. Ruling this out requires an additional assumption. We will discuss this briefly toward the end of the chapter.

2. Strictly speaking, this theorem requires that each country have balanced trade however it is easy to adjust the theorem to account for trade deficits or surpluses.

3. We discuss models of preference differences and product cycles in Chapter 14.

4. For this exercise the "world" consists of 43 countries for which these data exist (a few observations are missing for R&D scientists). This group encapsulates nearly all economies of moderate or larger size and accounts for the great majority of world output.

5. Interested students can find these data from the Organization for Economic Cooperation and Development (OECD), which developed the matrixes.

6. Hicks-neutral technologies are broadly consistent with the HOV assumptions. If all U.S. factors are 50-percent more productive than their counterparts in, say, the United Kingdom, then those factors will all have higher real returns in the United States in free trade. However, the relative prices of factors will be the same in both countries.

7. The gravity model is very important in trade analysis and will be discussed more fully in Chapter 15.

8. A subsequent strand of literature has looked at more general specifications of technology differences, say between developed and developing countries. See Harrigan (1997), Debaere (2003) and Maskus and Nishioka (2009). Another interesting approach is to estimate directly the number of diversification cones that exist among countries, as done by Schott (2003).

Table 8.1 Measures of Factor Intensity for US Manufacturing Industries

	2000					2005				
Industry	Value Added	Production	Capital Exp.	Nonproduction		Value Added	Production	Capital Exp.	Nonproduction	Evident
	($ millions)	Labor (000)	per PL	labor per PL		($ millions)	Labor (000)	per PL	labor per PL	Intensity
Petroleum and coal products	$ 45,748	67	$ 74,624	0.51		$ 117,541	65	$ 169,501	0.58	Capital, Skill
Chemical products	$ 235,614	508	$ 41,112	0.75		$ 328,440	433	$ 38,971	0.76	Capital, Skill
Computer & electronic products	$ 291,125	848	$ 33,227	0.94		$ 226,319	465	$ 33,972	1.16	Capital, Skill
Mineral products	$ 55,722	408	$ 14,820	0.28		$ 64,545	360	$ 14,334	0.29	Capital
Transportation equipment	$ 240,989	1,349	$ 12,529	0.36		$ 254,665	1,104	$ 13,842	0.41	Capital, Skill
Food, beverages & tobacco	$ 255,245	1,244	$ 11,714	0.35		$ 316,389	1,177	$ 13,090	0.34	Capital, Skill
Wood & paper products	$ 114,260	914	$ 12,234	0.24		$ 120,651	765	$ 11,268	0.27	Capital
Miscellaneous products	$ 70,621	501	$ 8,219	0.49		$ 92,974	422	$ 11,044	0.61	Skill
Plastic & rubber products	$ 92,333	862	$ 10,086	0.26		$ 96,348	688	$ 10,127	0.29	Capital
Machinery	$ 148,798	920	$ 10,116	0.52		$ 142,488	683	$ 9,947	0.56	Skill
Printing	$ 63,446	597	$ 7,398	0.39		$ 58,930	457	$ 9,510	0.41	Skill
Metal products	$ 215,545	1,839	$ 8,729	0.30		$ 232,106	1,418	$ 8,545	0.33	Skill
Electrical equipment & appliances	$ 62,991	431	$ 9,069	0.37		$ 54,318	294	$ 6,551	0.43	Skill
Textile products	$ 35,225	475	$ 5,130	0.20		$ 32,395	285	$ 4,633	0.23	Labor
Leather products	$ 4,510	55	$ 2,813	0.25		$ 2,865	29	$ 3,527	0.29	Labor
Furniture & related products	$ 42,267	515	$ 4,011	0.25		$ 46,801	414	$ 3,404	0.29	Labor
Apparel	$ 28,210	423	$ 2,302	0.24		$ 16,319	171	$ 2,882	0.31	Labor

Source: Compiled by authors from US Department of Commerce, *Annual Survey of Manufactures*

Table 8.2 Measures of Relative Factor Endowments								
	2000				2005			
Country	Capital Stock per worker	Arable land per worker (HA)	R&D Scientists per 1000 people		Capital Stock per worker	Arable land per worker (HA)	R&D Scientists per 1000 people	Evident Abundance
Singapore	$ 239,044	0.00	8.08		$ 247,608	0.00	10.45	Capital, R&D
Japan	$ 182,196	0.07	9.55		$ 194,375	0.07	10.55	Capital, R&D
USA	$ 153,689	1.19	8.64		$ 181,856	1.13	8.97	Capital, R&D
Australia	$ 149,347	4.91	6.86		$ 169,374	4.68	6.76	Capital, Land
Germany	$ 160,918	0.29	6.38		$ 162,214	0.29	6.71	Capital, R&D
Canada	$ 142,345	2.82	6.69		$ 156,814	2.55	6.55	Capital, Land
Finland	$ 149,338	0.84	13.42		$ 155,699	0.85	15.00	Capital, R&D
Rep. of Korea	$ 102,235	0.08	4.80		$ 123,959	0.07	7.56	Capital, R&D
UK	$ 102,447	0.20	5.43		$ 117,232	0.19	5.86	R&D
Mexico	$ 48,140	0.64	1.12		$ 50,827	0.58	1.11	Labor
Brazil	$ 39,311	0.70	0.77		$ 37,885	0.63	0.77	Labor
South Africa	$ 31,060	0.95	0.96		$ 30,532	0.86	0.99	Labor
China	$ 13,183	0.18	0.95		$ 20,090	0.18	1.44	Labor
India	$ 7,556	0.42	0.29		$ 9,465	0.37	0.31	Labor

Sources: computed by authors with data available from World Bank, *World Development Indicators*; Food and Agricultural Organization, *FAO-Stat Database*; and Penn World Tables version 6.2.

Table 8.3 Proportions of World GDP and World Factor Endowments, 2000							
				Primary	Secondary	Post-secondary	R&D Research
	GDP	Capital Stock	Arable Land	School	School	School	Scientists
USA	27.10%	23.89%	19.42%	2.25%	11.96%	30.22%	29.20%
Canada	2.45%	2.43%	5.07%	0.80%	0.76%	1.65%	2.49%
Germany	6.18%	6.83%	1.31%	2.15%	4.13%	3.69%	5.89%
UK	4.22%	3.20%	0.65%	2.02%	1.54%	2.43%	3.69%
Australia	1.36%	1.51%	5.24%	0.48%	0.74%	1.19%	1.51%
Japan	8.91%	12.97%	0.50%	3.29%	4.92%	7.63%	14.78%
Rep. of Korea	2.16%	2.44%	0.19%	1.29%	3.10%	3.10%	2.49%
Mexico	2.90%	2.00%	2.78%	2.61%	1.98%	1.77%	1.01%
Brazil	3.39%	3.42%	6.39%	2.87%	1.49%	2.84%	1.46%
China	8.26%	10.11%	14.75%	32.62%	33.33%	9.79%	15.80%
India	4.29%	3.11%	18.02%	16.93%	9.84%	9.00%	2.57%
Countries	43	43	43	43	43	43	36

Sources: computed by authors with data available from World Bank, *World Development Indicators*; Food and Agricultural Organization, *FAO-Stat Database*; and Penn World Tables version 6.2. Figures for GDP are measured with PPP exchange rates at constant 2005 $US.

Table 8.4 Results of Statistical Testing in Davis-Weinstein				
			HOV	HOV
	HOV	HOV-HN	non-FPE	non-FPE & gravity
Statistic	H1	H2	H3	H4
Slope	-0.002	-0.05	0.43	0.82
Standard Error	0.005	0.02	0.02	0.03
R^2	0.01	0.31	0.96	0.98
Sign Test	32%	50%	86%	91%
Observations	22	22	22	22

Source: Davis and Weinstein (2001).

Figure 8.1

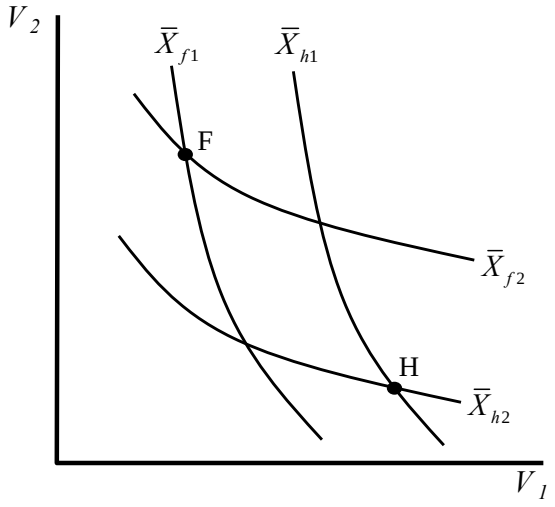


Figure 8.2

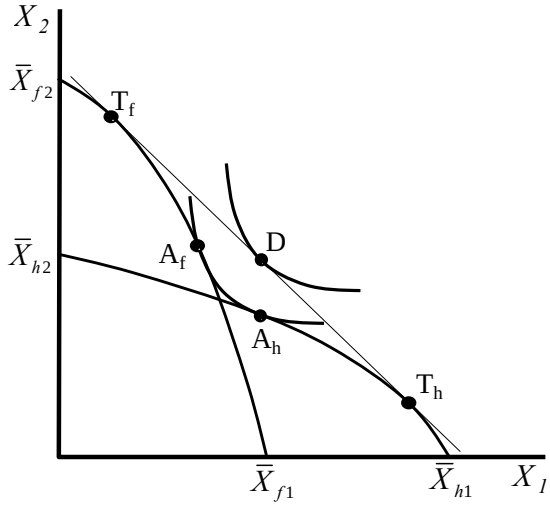


Figure 8.3

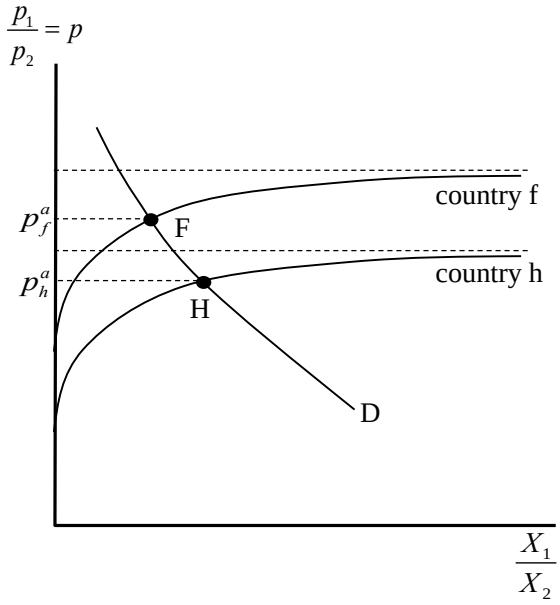


Figure 8.4

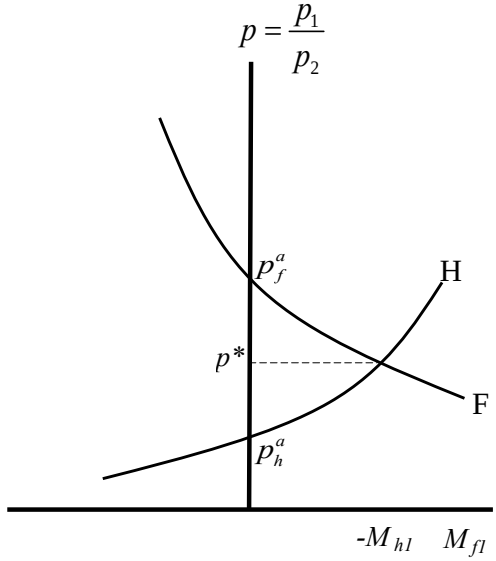


Figure 8.5

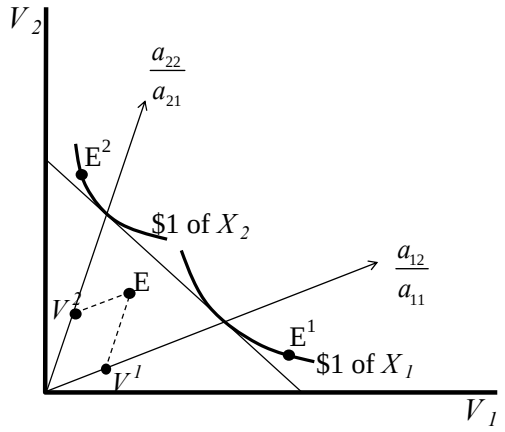


Figure 8.6

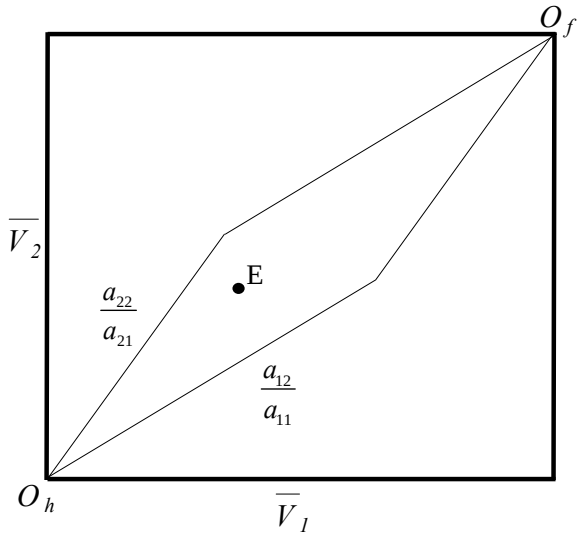


Figure 8.7

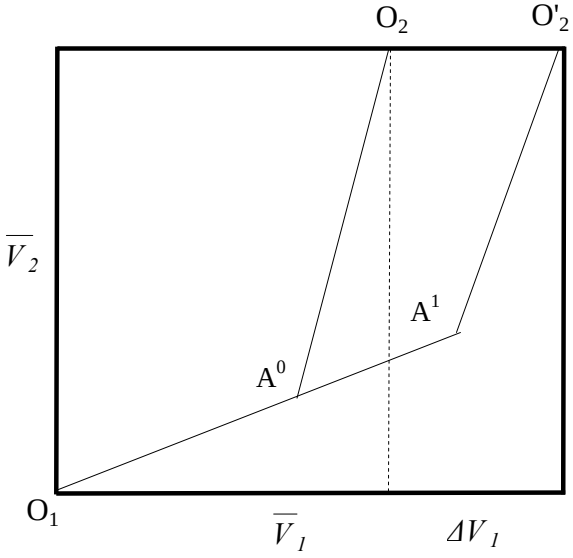


Figure 8.8

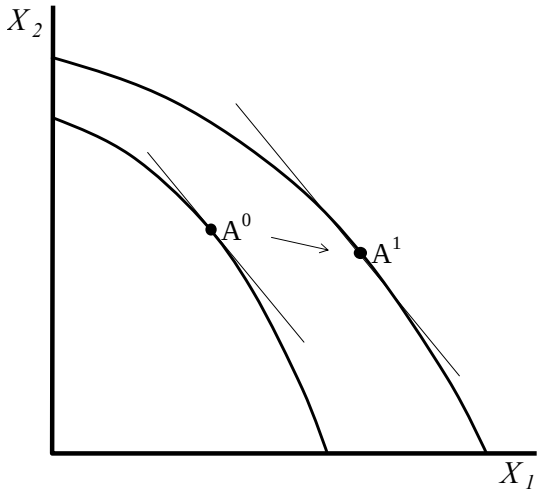


Figure 8.9

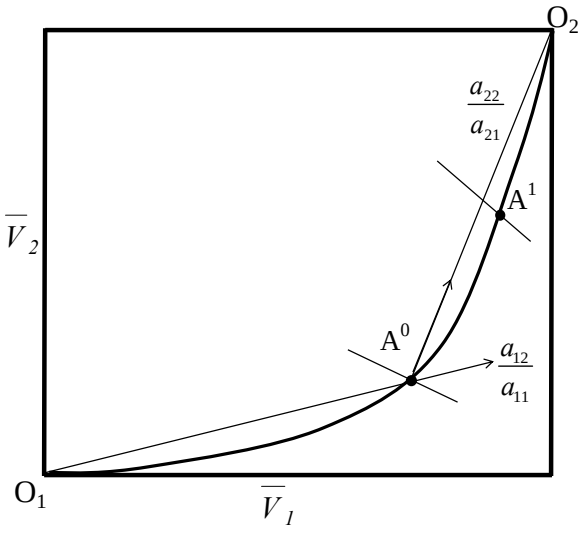


Figure 8.10

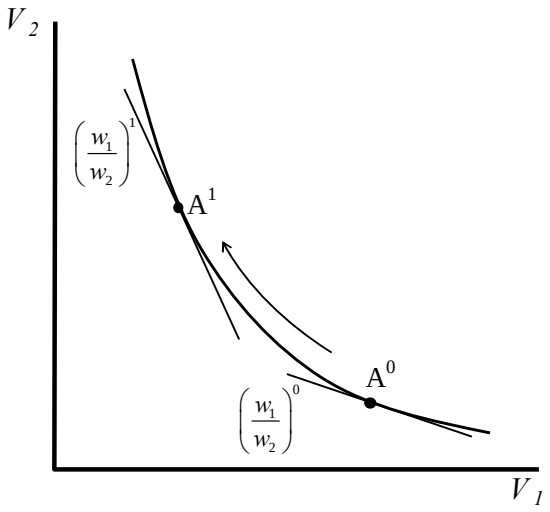
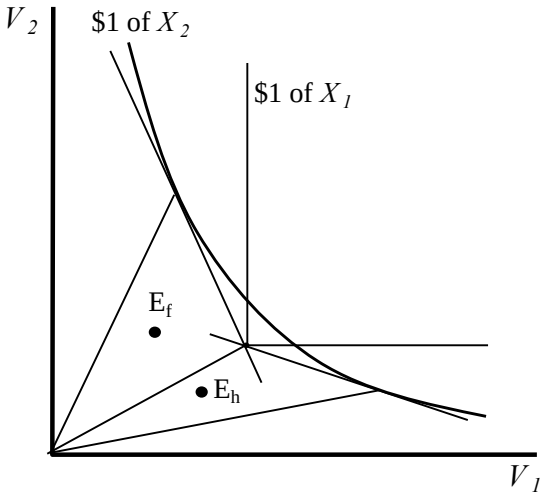


Figure 8.11



Chapter 9

DIFFERENCES IN FACTOR ENDOWMENTS II: THE JONES SPECIFIC-FACTORS MODEL

9.1 The Jones specific-factors model

As noted in the previous chapter, the names Heckscher and Ohlin are often used in a general sense to refer to all models in which differences in factor endowments across countries are the cause of trade. An alternative is to refer to all such models as factor-proportions models. In other cases, the term Heckscher-Ohlin model is used in a much more restricted sense to refer to the two-good, two-factor, two-country case of the previous chapter. In this chapter, we will examine another popular version of a factor-proportions model, which has been developed and popularized by Jones starting in a 1971 paper.

The Jones' formulation retains two goods, but has three factors, making it seemingly more complicated. Two of the factors are "specific" factors, each having a use in only one industry and having no use in the other. The third factor is mobile and can costlessly move between the two sectors as both factors do in the Heckscher-Ohlin model. The fact that only one factor is mobile here makes the model simpler than the HO formulation. As we will see, however, there are offsetting complications. In particular, while it is obvious that each final good is intensive in its specific factor, there is ambiguity about the mobile factor and what happens to outputs if the supply of this factor is increased. Defining the factor intensity of the mobile factor turns out to be complex and there does not emerge a simple, clean theorem about endowments and trade as there does in the Heckscher-Ohlin theorem.

There are several reasons why the specific-factors' formulation is appealing, and perhaps these will become clearer as we proceed. First, the factor-price equalization theorem does not hold and, oddly, trade economist are probably more comfortable with this theorem not holding: it just does not strike us as empirically very plausible. Second and more important, the specific factors' formulation generates results that seem much more plausible from a political-economy point of view. In the Heckscher-Ohlin model, a group of factor owners such as capital or labor are always bound together regardless of what industry they work in. Capital owners should lobby for the capital-intensive industry even if their capital is in the labor-intensive industry. The specific-factors' model instead implies factors are allied with the industry they work in (though again with ambiguity with respect to the mobile factor). Third, the specific-factors' model seems much more plausible for the short and intermediate run, in that it is costly or impossible to move factors between industries over a period of even a few years. Capital, for example, takes on industry-specific forms once installed, and it is not easy to transform shoe-making equipment into agricultural machinery. This is done over the long run by allowing one type to depreciate and channeling all new investment into the other type.

After a number of attempts, we found that the general notation of the previous chapter proved more confusing than helpful in this chapter and so we return to a more traditional notation. We will retain the labeling of goods as X_1 and X_2. However, the mobile factor will be denoted L for labor, with L_1 and L_2 the amounts allocated to the two sectors. The specific factors will be denoted K_1 and K_2. These are in fixed supply and each is only useful in its own sector. Having said this, we don't want to take these labels too literally. The specific factors could be types of human capital that have no use in the other sector. The mobile factor could be land or unskilled labor.

The production side of the economy is given by

$$X_1 = F_1(L_1, K_1) \qquad X_2 = F_2(L_2, K_2) \qquad \bar{L} = L_1 + L_2, \quad K_1 = \bar{K}_1, \quad K_2 = \bar{K}_2 \qquad (9.1)$$

Figure 1 presents an intuitive version of the relationship between factor endowments and

comparative advantage. Suppose that both countries have identical endowments of labor, and suppose that country h has an absolutely larger endowment of K_2 than country f and the latter has an absolutely larger endowment of K_1. Then it is clear that the production frontiers of the two countries must look something like those drawn in Figure 9.1, where HH′ and FF′ are the frontiers of countries h and f respectively. With L endowments equal, each country has a comparative advantage in the good using intensively its abundant factor. That sure sounds like Heckscher-Ohlin, but the caveat about the mobile factor is unfortunately not innocuous as we shall see.

Figure 9.1

As in all competitive models, producers equate the value of the marginal product of a factor to its price. Let w denote the price of labor and r_1 and r_2 denote the prices of the two specific factors. In the specific-factors' model, the following equality express production equilibrium at goods prices p_1 and p_2 (as before F_{i1} is the marginal product of the first factor, labor in this case, in producing good i).

$$p_1 F_{11}(L_1, K_1) = p_2 F_{21}(L_2, K_2) = w_1 \tag{9.2}$$

Differentiate the first equality with respect to the mobile factor L, noting that $dL_1 = - dL_2$ for factor-market clearing. Then divide the second equation by the first to get the marginal rate of transformation, which we know is equal to the price ratio p_1/p_2 in competitive equilibrium.

$$dX_1 = F_{11} dL_1 \qquad dX_2 = F_{21} dL_2 = - F_{21} dL_1 \qquad -\frac{dX_2}{dX_1} = \frac{F_{21}}{F_{11}} = \frac{p_1}{p_2} \tag{9.3}$$

In the specific-factors' model, the slope of the production frontier is just the ratio of the marginal products of labor in the two industries. If the isoquants of the two industries exhibit smooth substitution, then the production frontier of each country must exhibit curvature as in Figure 9.1. Intuitively, begin with only producing good X_2. As we move down the frontier, we are taking labor away from the fixed factor K_2 in the X_2 sector, and so (K_2/L_2) rises, and the marginal product of labor rises. F_{21} in (9.3) rises. As we add labor to the X_1 sector, we are adding labor to a fixed factor, and so (K_1/L_1) falls and the marginal product of labor falls. F_{11} in (9.3) falls. Thus the marginal rate of transformation increases (or the slope becomes increasingly negative) as we move down the production frontier in Figure 9.1

Return now to (9.2). The two sides of this equation, the value of the marginal product of labor *(VMP*) in the two sectors, has a simple graphical representation shown in Figure 9.2. VMP_1 denotes for example, the left-hand side of (9.2). Similar to an Edgeworth box, the horizontal axis is the total supply of labor available to the economy, with the allocation to X_1 graphed from the left and to X_2 graphed from the right. The intersection of the two value-of-marginal-product curves determines the optimal allocation of labor between the two sectors and determines the wage rate. In what follows below, it is important to note that each curve in Figure 9.2 has two parameters: the price of the good and the amount of the specific factor in that sector. Increasing either p_1 or K_1, for example, shifts the curve for X_1 upward: the value of the marginal product of labor increases at any level of L_1.

Figure 9.2

9.2 Analogs to the four theorems of the Heckscher-Ohlin Model

(A) (Non) factor-price equalization. First, we can note directly from Figure 9.2 that factor-price equalization across countries is not going to hold (except by chance). There are two zero-profit equations to determine three factor prices from two goods prices. Factor prices are going to depend on endowments as well even if both countries produce both goods and commodity prices are equalized by trade. Start with two identical countries and commodity prices equalized by trade. If we expand the labor

supply in one country, the horizontal axis of Figure 9.2 lengthens, the two curves move further apart and the wage must fall. Similarly, and beginning again with identical countries, give one country more of one of the specific factors, say K_1, and the VMP_1 curve shifts up. This will increase the wage rate and reduce the price of the specific factor K_1. In summary, even with commodity prices equalized by trade and both countries producing both goods, if the two countries have equal endowment of two factors, the third factor will be cheap where it is abundant and expensive where it is scarce.

(B) Analog to the Stolper-Samuelson theorem. Suppose that we increase the price of good X_1: $dp_1 > 0$, $dp_2 = 0$. Again, this is readily analyzed with Figure 9.2. The curve VMP_1 shifts up as shown in Figure 9.3. Labor is reallocated toward the X_1 sector. With the specific factors fixed, it must be the case that K_1/L_1 falls and K_2/L_2 rises. Recall from Chapter 2 that, with constant returns to scale, the marginal product of a factor depends only on the *ratio* in which factors are used. Use the value-of-the-marginal-product conditions in (9.2) above and add those for the specific factors, and then divide through both sides of each equation by the price of the good; e.g., $p_i F_{il} = w \Rightarrow w/p_{il} = F_{il}$. The changes in the capital-labor ratios in the two industries must imply that the increase in the price of X_1 gives us

$$\begin{aligned} d(w/p_1) &= d(F_{11}) < 0 \qquad d(w/p_2) = d(F_{21}) > 0 \\ d(r_1/p_1) &= d(F_{12}) > 0 \qquad d(r_2/p_2) = d(F_{22}) < 0 \end{aligned} \tag{9.4}$$

Figure 9.3

Recall the algebra we developed in the previous chapter, denoting the proportional change in a variable with a "hat". If a ratio increases, it must mean that the proportional change in the numerator is greater than the proportional change in the denominator. (9.4) gives use a version or analog to the Stolper-Samuelson theorem with a corresponding chain of inequalities for a change in p_2 holding p_1 constant. These chains are given by:

$$\hat{r}_1 > \hat{p}_1 > \hat{w} > 0 > \hat{r}_2 \qquad \hat{r}_2 > \hat{p}_2 > \hat{w} > 0 > \hat{r}_1 \qquad \hat{r}_1 \equiv \frac{dr_1}{r_1} \text{ etc.} \tag{9.5}$$

If the price of good i goes up, then the real income of the specific factor in i rises (the nominal price of the factor must rise by more than the good's price) and the real income of the other specific factor falls (even in terms of the good whose price has not changed). Thus price changes due to trade liberalization, trade protection, or changes abroad create a strong redistribution between the owners of the specific factors. Specific factor owners will lobby hard for their industry and owners of industry-specific physical or human capital will now be enemies of physical or human capital owners in the other industry. This strikes many who study political economy as a much more empirically plausible prediction than that of the Heckscher-Ohlin model's Stolper-Samuelson theorem. As noted above, this is one reason for the specific-factors-model's populatity.

Note from (9.5) however, there is ambiguity about the change in the real income of the mobile factor L. Labor's wage change is caught in between those of the two goods prices. In the first chain of inequalities *($dp_1 > 0$, $dp_2 = 0$)*, the wage rise relative to the (unchanged) price of good 2, but falls in terms of the price of good 1 which has risen. Thus the real income or welfare of labor depends in part on preferences and is more likely to increase if it has a preference for the good whose price has not gone up. In the case of trade liberalization where the relative price of the export good rises for example, this would mean having a preference for the import good makes it more likely that labor gains in welfare terms.

(C) Analog to the Rybczynski theorem. As you recall, the Rybczynski theorem was the effect of an endowment change on the change in outputs, holding commodity prices constant. There is an analog here as in the case of the Stolper-Samuelson theorem, and much of it can understood once again from equation (9.2) and Figures 9.2 and 9.3. First, consider an increase in one of the specific factors, $dK_1 > 0$.

As shown in Figure 9.3, this shifts the VMP_1 curve up and shifts labor toward the X_1 sector to go with the increased K_1. We note from the "X" diagram that w rises with p_1 and p_2 constant by assumption: with prices constant, F_{11} and F_{21} both rise, but this means that both K_1/L_1 and K_2/L_2 rises (in the latter case because L_2 goes down). The X_1 sector pulls in labor, but not in proportion to the rise in K_1. The rise in the capital labor ratio in X_1 means that $dK_1/K_1 > dL_1/L_1$ and therefore that $dK_1/K_1 > dX_1/X_1 > 0$, and $dX_2/X_2 < 0$. Repeating the argument for an increase in K_2 holding K_1, L and goods prices constant, we have our result.

$$\hat{K}_1 > \hat{X}_1 > 0 > \hat{X}_2 \qquad \hat{K}_2 > \hat{X}_2 > 0 > \hat{X}_1 \qquad \hat{K}_1 \equiv \frac{dK_1}{K_1} \text{ etc.} \tag{9.6}$$

Compare this to the result for the Rybczynski theorem in (8.19) of the previous chapter. Now we no longer have the “magnification” effect of an endowment change on outputs. The change in the endowment of a specific factor now has a less-than-proportionate change on the output of its industry. Quantitatively, the results looks much like that shown in Figure 8.8 of the previous chapter, but the biased in the change is quantitatively less in the present case.

The role of the mobile factor L is less straightforward. Figure 9.4 shows the effect of an increasing labor supply, stretching the length of the “X” diagram. Provided that both *VMP* curves are strictly negatively sloped (with respect to their origins), then some added labor is allocated to each sector. We must have

$$\hat{L} > \left[\hat{X}_1, \hat{X}_2\right] > 0 \tag{9.7}$$

Figure 9.4

However, which good has the larger output increase is unclear. It turns out that which good increases relatively more, an issue important for determining the direction of trade, does depend on which one is more labor intensive as measured by factor shares, but it also depends on other properties of the two production functions. In particular it depends on the slopes of the two *VMP* curves in Figure 9.4. Note for example, that if VMP_1 is flat in Figure 9.4 (horizontal), then all of the added labor must go to the X_1 sector. Conversely, if VMP_2 is vertical, all of the added labor must go to sector X_2.

Differentiate the two production functions in (9.1) and then form proportional changes by (1) dividing both sides by the level of output of the good, (2) multiplying and dividing the right-hand side by the price of the good and (3) multiplying and dividing the right-hand side by the labor allocation to the good. We then have

$$\frac{dX_1}{X_1} = \left[\frac{p_1 F_{11} L_1}{p_1 X_1}\right]\frac{dL_1}{L_1} \qquad \frac{dX_2}{X_2} = \left[\frac{p_2 F_{21} L_2}{p_2 X_2}\right]\frac{dL_2}{L_2} \tag{9.8}$$

From our work in Chapter 8 and from the value-of-the-marginal-product conditions in (9.1), we see that the terms in brackets are the shares of labor income in the total value of production in each of the two sectors. Using the same notation as in Chapter 8, we can write these shares as

$$\theta_{iL} = \left[\frac{p_i F_{iL} L_i}{p_i X_i}\right] = \left[\frac{w L_i}{p_i X_i}\right] \tag{9.9}$$

Substituting (9.9) into (9.8), then we get (9.10). Which good increases more depends on the share of labor in each sector, which is a sensible definition of labor intensity.

$$\frac{dX_2/X_2}{dX_1/X_1} = \frac{\theta_{21}}{\theta_{11}} \frac{dL_2/L_2}{dL_1/L_1} \tag{9.10}$$

However, we are unfortunately not done. The changes in outputs also depend on which sector is allocated more of the increase in labor. A good way to conceptualize this is to consider again Figure 9.4 from the point of view of each industry and think about firms in each industry reacting to an increase or fall in the wage rate. The elasticity of demand for labor in an industry can be defined as the proportional increase in demand for labor in response to a proportional increase in the wage rate, holding the price of the output constant. It is typically also defined for holding the prices of other factors constant, but in our case we define it for holding the quantity of the other factor, the sector-specific factor, constant. When defined in this manner and using η_i to denote the elasticity so defined, (9.10) becomes the following.

$$\frac{dX_2/X_2}{dX_1/X_1} = \frac{\theta_{21}}{\theta_{11}} \frac{\eta_2}{\eta_1} \qquad - \left[\frac{dL_i/L_i}{dw/w} \right]_{dK_i = 0} \equiv \eta_i \tag{9.11}$$

With the wage on the vertical axis in Figures 9.1-9.4, a flat *VMP* curve has a high elasticity and a steep curve a low elasticity. As noted in the previous paragraph, an industry responds more to an increase in the labor endowment when there is a higher elasticity of demand for labor (or low in the other industry) as well as when the share of labor in that industry is high.

Now we can summarize our finding to create an analog to the Heckscher-Ohlin theorem about the direction of trade. Unfortunately, it is not as simple and clean as the original.

(D) Specific-factors analog to the Heckscher-Ohlin theorem. (i) for two otherwise identical countries, if one country h has more of specific factor 1 and/or country f has more of specific factor 2, then country h exports good 1. (ii) for two otherwise identical countries, if country h has more labor then country h is more likely to export good 1 if the share of labor in industry 1 is higher than in industry 2 and the elasticity of demand for labor with respect to the wage rate is higher in industry 1.

9.3 Empirical evidence on preferences for protection or free trade

The Jones specific-factors model offers a number of rich extensions of the basic factor-proportions model of international trade. The dimension in which it has attracted the most empirical study is its predictions about the relationships between goods prices and real factor returns. Interest in this question arises because, as noted in section 9.2, we would expect specific factors in the short run to lobby for policies that increase output price in their own sectors and to oppose policies that increase output price in other sectors. As we will demonstrate in a later chapter, protection from import competition through import tariffs or other barriers raises the domestic prices of goods that receive such protection. Thus, specific factors that work in industries facing significant import penetration should favor a protectionist trade policy. However, reductions in trade barriers reduces those same prices, implying a rise in the relative price of goods in other industries, especially of exported products. In turn, specific factors that work in export-oriented industries should favor free trade. The situation is ambiguous with respect to mobile factors, whose real returns depend on consumption preferences.

These predictions may be contrasted with the long-run Stolper-Samuelson (SS) theorem, which predicts that factor owners either favor free trade (if they are abundant) or protectionism (if they are scarce), regardless of the industry in which they work. Thus, in a capital-abundant nation all capital owners should lobby for free trade and all laborers should lobby for tariffs. For example, suppose the United States is capital-abundant. If textiles are labor-intensive and aircraft are capital-intensive in any

long-run equilibrium, we would expect capital owners in both industries to prefer free trade but workers in both industries to want trade protection. The Jones model, in contrast, would predict that in the short run capital owners in textiles would demand protection but capitalists in aircraft would urge free trade. Perhaps the most significant question is whether factor owners tend to lobby based on their short-run (industry-based) or long-run (factor-based) interests.

These predictions are obviously important for policymakers trying to understand how domestic citizens would react to changes in trade policy, such as tariff liberalization or the negotiation of free trade areas. A number of economists have studied this question with varying types of data.

The initial study was performed by Stephen Magee (1980) in what the author considered direct tests of the SS theorem. He set out three testable hypotheses, all of which rejected SS. Here we simply present the first test, which is visually arresting.

Stolper-Samuelson: In the long run all factors are mobile. Thus, labor and capital will oppose each other in their preferences for free trade and protection.

Specific-factors alternative: Suppose that both capital and labor are specific factors in the short run. Then both capital owners and laborers in each sector should share a preference either for protection or free trade.

To consider this hypothesis, Magee read testimony in May and June of 1973 before the Ways and Means Committee of the U.S. House of Representatives, which was considering passage of legislation that became the Trade Reform Act of 1974. He classified testimony by 29 trade associations (representing management, or capital) and 23 labor unions (representing workers) from a variety of manufacturing industries as favoring either protection or free trade, though a few cases could not be categorized. In his final sample there were 21 industries in which both management and labor interests could be clearly defined.

Before presenting his results, we note that the 1974 Trade Reform Act was controversial in a number of respects. Most significantly, it provided for the first time Presidential authority to negotiate liberalization of non-tariff barriers, such as import quotas and discriminatory government procurement laws, with other countries under the so-called "fast-track authority".[1] This authority, under which a U.S. President could present a negotiated trade agreement to Congress, which then had to approve or disapprove it within a certain time period and without modification, had already existed for tariff cuts. By extending it to non-tariff barriers the prospect of significantly greater multilateral trade liberalization became apparent. The Act also provided the foundation for American participation in the ensuing Tokyo Round (1974-79) of trade negotiations at the General Agreement on Tariffs and Trade, the precursor to the World Trade Organization. Thus, it could reasonably be viewed as heralding a sharp move toward freer trade in the United States.

In Table 9.1 we show Magee's industry list and the classifications of labor and management interests revealed through testimony. Again, if the SS theorem actually described policy preferences, we would expect all capital to support one position and all labor to support the other, independent of industry affiliation. Thus, the diagonal cells should be empty, while one of the off-diagonal cells should contain all entries and the other one should be empty. It is obvious from the table that this prediction was decisively rejected by Magee's categorization. Indeed, 14 of the 21 sectors saw both management and labor opposing the 1974 Trade Reform Act, five sectors experienced both management and labor favoring it, while just two registered a difference of opinion.

Table 9.1

From these results it seemed that rather than considering themselves to be mobile factors in a context in which free trade would redistribute income between them, capital and labor view themselves as specific to their particular industries. In this environment both factors preferred policies that they envisioned sustaining higher prices of the goods they produce. In short, fixed-factor models seemed

better at explaining lobbying for or against tariffs and quotas. This finding has been highly influential among trade economists. It suggests that in lobbying for trade policy, factor owners take a short-run view and that industries are natural opponents of each other rather than capital and labor in the larger sense. In short, the specific-factors model was heavily supported over the SS model.

While provocative, Magee's study raised more questions than it answered. An obvious question is why should these factors take a short-run approach to trade policy. Either factor owners really do care only about near-term impacts of policy changes as the specific-factors model would suggest, or they are irrationally myopic about the longer-term effects. In fact, this test could not readily sort out that question because in that era U.S. trade legislation came up for revision approximately every four years. Thus, rather than representing their interests on legislation that would permanently alter policy, management and labor may have truly seen the 1974 Trade Reform Act as temporary, biasing the results against SS and toward the specific-factors approach.

There are also methodological problems with Magee's study. Among many objections, two stand out. First, there are many other elements that affect citizens' views of trade policy that were not accounted for in this analysis. A central example is that people really are not just laborers or capital owners. Rather, workers themselves own assets in terms of housing and financial instruments. Thus, it is difficult to make a sharp distinction between types of factor ownership in the real world. Second, by using the testimony of industry associations and labor unions regarding a large piece of discrete legislation, Magee only indirectly measured preferences for protection or free trade.[2] It would be better to consider the opinions registered by individuals in surveys and correlated those views with basic economic characteristics.

Among the many studies that have extended Magee's approach to account for these problems we discuss two. Edward Balistreri (1997) adopted the HOV model of multiple goods and factors (described in the prior chapter of this text) to see if there was a correlation between individual Canadian citizens attitudes toward the Canada-U.S. Free Trade Agreement (CAFTA), which was under negotiation in the late 1980s, and their occupational employment.[3] He argued that occupations could be considered "factors" in that they are types of labor differentiated by skills that define an individual's working patterns. That is, an accountant is an accountant regardless of industry of employment. Further, he assumed reasonably that CAFTA was a permanent change in trade policy, because, once enacted, neither country could change its terms or give up membership except at very high cost. Balistreri claimed informally that abundant factors were likely to see their incomes rise, and scarce factors were likely to suffer declines in their incomes, if CAFTA were to be enacted. Thus, Balistreri's study was focused on whether individuals favored NAFTA based on their anticipated long-run income changes.

It is straightforward to adapt the HOV equations to a bilateral basis and compute the "revealed" abundances or scarcities of Canadian occupations relative to the same occupations in the United States, Canada's major trading partner. Thus, the author computed the sign patterns of the Canadian excess factor supplies for 21 broad occupations. To measure individual attitudes toward passage of CAFTA, he used an opinion survey of 2,470 people performed by York University in 1988. This survey asked Canadian respondents whether they favored or opposed the agreement, while also asking about a variety of individual characteristics. One characteristic was occupation of employment, with occupations then assigned to the 21 categories in the basic sample.

Armed with these data, Balistreri estimated a logit model to explain preferences toward CAFTA. In this model the dependent variable takes a value of one if the respondent favored CAFTA (and therefore expected a gain in real income) and a value of zero if she opposed the agreement. This binary variable was regressed on a dummy variable with a value of one if the person's occupation was revealed to be abundant and zero if the occupation was scarce. Control variables included age, education, income range, province of residency, union membership, and a dummy variable for whether the respondent felt financially better off than a year earlier.

Rather than presenting the precise quantitative results we simply discuss them here in qualitative terms. First, as individual incomes rose people gained a much stronger positive attitude toward CAFTA.

Similarly, as financial well-being increased citizens were significantly more inclined to support the agreement. Surprisingly, levels of education had no significant effects, though this is likely due to their strong multicollinearity with income and wealth. As may be expected, union members were highly significantly opposed to CAFTA. Finally, the variable measuring factor abundance had a positive and highly significant impact on attitudes toward CAFTA. From this evidence it seemed that Canadian citizens in relatively abundant occupations expected to be made better off from the passage of the trade agreement.

This finding is largely in accord with an extended view of the Stolper-Samuelson theorem, though there is no precise theory proving that HOV-revealed factor abundance yields those factors whose real incomes will actually rise in general equilibrium. As a matter of correlation, however, it makes sense. Thus, Balistreri unearthed convincing evidence that individuals make decisions about trade policies that are consistent with the relative supplies of their occupations. The primary difficulty with the analysis, however, is that occupations may not really be an appropriate measure of long-run factors, since people may change their type of employment over their careers. Further, to the extent that some occupations really are fixed in their use in particular sectors, such as fishing, farming, and mining, this test may support the specific-factors model as much as it does the SS theorem. Put differently, Balistreri did not test his version of Stolper-Samuelson against an explicit specific-factors alternative.

A second study that helped sort out the last problem was by Scheve and Slaughter (2001). In this analysis the authors consider U.S. workers either to be skilled or unskilled. There is ample evidence that the United States is relatively skill abundant. Thus, the long-run impact of trade liberalization should be to raise the real incomes of skilled workers and reduce those of unskilled workers. Thus, higher-skilled individuals should favor free trade and lower-skilled workers should favor protection. However, if individuals take a short-term view they would consider their incomes to be tied closely to price changes in the products made in their industry of employment. Scheve and Slaughter add a further complication, the extent of home ownership. In regions of the country in which production is concentrated in sectors with comparative disadvantage, more open trade should place downward pressure on housing prices, reducing the asset values and wealth of local citizens. Regions in which production is concentrated in export-oriented industries should see upward pressure on home prices from open trade and local citizens may be expected to favor that policy. Put in simplest terms, a skilled worker in those states, such as Michigan, Ohio and Indiana, that have experienced declining output and employment due to import competition would have a complex choice. She might favor open trade based on its long-run boost to skilled incomes. However, she might oppose it based on her industry of employment (such as automobiles or steel) and the potential loss in home value.

To study these possibilities the authors compile data from the 1992 National Election Studies (NES) survey, which asked many questions about political opinions of a large sample of individuals in the United States. One question pertained to their support for new trade barriers. An indicator variable with the value one for those favoring protection and the value of zero for those opposing it became the dependent variable to study. The NES respondents also listed their occupation, which was combined with average occupational wage data to measure their skill levels. A second skill measure was educational attainment. The Heckscher-Ohlin logic would suggest that higher-skilled individuals would oppose trade barriers.

Respondents also listed their industry of employment and the authors computed measures of trade orientation for those industries. The first measure was 1992 net exports (exports minus imports) as a share of output in the industry. Assuming positive (negative) shares correspond to sectors of comparative advantage (disadvantage), the specific-factors model would imply that workers in the former (latter) would oppose (favor) trade restrictions. A second measure was the industry average tariff rate, which were likely higher in sectors of comparative disadvantage. These variables were then appropriate for testing the Jones model.

Finally, respondents indicated their county of residence and whether they were homeowners or not. A dichotomous variable (one for ownership, zero for non-ownership) was interacted with a measure of industry production mix in each location to identify ownership in counties with higher-than-average tariff rates or higher-than-average import penetration rates. These were considered “county exposure”

indicators and those locations more exposed to import competition should favor increasing trade barriers.

The authors estimated a logit model in which the dichotomous variable indicating preferences for new trade restrictions was regressed on these independent variables. We present selected results in Table 9.2. These findings are interesting in several dimensions. First, it is clear that employees in occupations with higher wages oppose additional trade barriers. Thus, higher-skilled workers in the United States tend to favor more open trade. This result also came through when skills were measured by educational attainment: those with higher levels of education also significantly oppose further trade restrictions. Thus, the Heckscher-Ohlin/Stolper-Samuelson models do seem to matter for individual citizens. Second, the trade exposure of the industry of employment, measured here either by the tariff rate of net-import competition, did not significantly affect individual trade preferences. Indeed, in no specification did these variables have a statistical impact on attitudes. This result casts doubt on whether individuals think in terms of their short-run income effects in the specific-factors context. Third, although the import exposure of county of residence has no significant direct effect, the exposure combined with home ownership powerfully influences individual attitudes. That is, a homeowner in a county facing higher-than-average import competition is much more likely to favor additional barriers to trade than is a homeowner in a less exposed county. The potential impact of trade on wealth or asset values, therefore, is a strong determinant of how people view the personal benefits and losses of trade competition.

Table 9.2

Of the studies reviewed here the most sound on methodological grounds is Scheve and Slaughter (2001). Their results suggest strongly that individual preferences for trade policy arise from a factor basis, complemented by asset ownership, rather than a sectoral employment basis. In that context, people seem to take a longer-term and more rational view of their interests than was implied by Magee's (1980) findings. However, Magee may not have been that far wrong. It is possible that industry associations and labor unions have shorter time horizons and attempt to maximize some kind of benefits other than the long-run economic well-being of their factor owners. We will have more to say about this in Chapter 22 on political economy models of trade policy.

9.4 Summary

There are many possible formulations for factor-proportions models in which differences in relative endowments across countries link with differences in factor intensities across industries. The Heckscher-Ohlin model of the previous chapter is one and the Jones' specific-factor model is another. The Jones' formulation is popular in a number of applications including political economy models of lobbying for example. The fact that factors are allied within industries rather than across industries seems to have some intuitive appeal. The Jones' formuation also escapes from factor-price equalization which many find an improbable result from Heckscher-Ohlin.

While simpler than Heckscher-Ohlin in many ways, the specific-factors model is unfortunately more complicated in others. This is particularly true with regard to the role of the mobile factor and its role in creating comparative advantage. It is not the case that the role of the mobile factor can be discerned from factor intensities (as measured by factor shares) alone, we must also know the elasticities of substitution across industries to predict this factor's impact on trade.

Analogs to the Stolper-Samuelson and Rybczynski theorem are developed. Specific factors are the big gainers or losers from commodity price changes, which may be cause by trade liberalization or its opposite, increased protectionsim, or by change in the rest of the world's economy. The model suggests that specific factor owners such as physical capital or human capital that is industry specific (has little use outside their industry) or land and resources, will be particularly fierce lobbyists on trade policy.

The chapter concludes with short sections developing the CES and CET functions and showing a couple of applications they have within the context of the specific-factors model. While it may have seemed logical to include these in Chapter 2, we made the decision to postpone their introduction in order

to move things alone. The CES function will return to play an important role in the analysis of monopolistic competition in Chapter 12, and also in discussions about economic geography.

REFERENCES

Balistreri, Edward J. (1997). “The Performance of the Heckscher-Ohlin-Vanek Model in Predicting Endogenous Policy Forces at the Individual Level,” *Canadian Journal of Economics,* 30: 1-17.

Irwin, Douglas A. (1994). “The Political Economy of Free Trade: Voting in the British General Election of 1906,” *Journal of Law and Economics*, 37: 75-108.

Jones, R. W. (1971). "A Three-Factor Model in Theory, Trade, and History." in J. Bhagwati et al (eds.), *Trade, Balance of Payments, and Growth*, Chapter 1. Amsterdam: North-Holland.

Kaempfer, William H. and Stephen Marks (1993). “The Expected Effects of Trade Liberalization: Evidence from U.S. Congressional Action on Fast-Track Authority,” *The World Economy,* 16: 725-740.

Magee, Stephen P. (1980). “Three Simple Tests of the Stolper-Samuelson Theorem,” in Peter Oppenheimer, ed., *Issues in International Economics: Essays in Honor of Harry G. Johnson* (Stocksfield, UK: Oriel Press).

Mayer, W. (1974). "Short Run and Long Run Equilibrium for a Small Open Economy." *Journal of Political Economy* 82: 955-967.

Mussa, M. (1974). "Tariffs and the Distribution of Income: The Importance of Factor Specificity and Substitutability and Intensity in the Short and Long Run." *Journal of Political Economy* 82: 1191-1204.

Neary, J. P. (1978). "Short-run Capital Specificity and the Pure Theory of International Trade." *Economic Journal* 88: 477-510.

Rogowski, Ronald. (1987). “Political Cleavages and Changing Exposure to Trade,” *American Political Science Review*, 81: 1121-1137.

Scheve, Kenneth F. and Matthew J. Slaughter (2001). “What Determines Individual Trade-Policy Preferences?” *Journal of International Economics*, 54: 267-292.

Vanek, J. (1968), “The Factor Proportions Theory: the n-Factor Case”, *Kyklos* 21, 749-756.

ENDNOTES

1. Non-tariff barriers are analyzed in Chapter 19 of this text.

2. Many subsequent studies have analyzed the voting patterns of politicians in the United States, the United Kingdom, and elsewhere in order to understand how the political system generates trade policy. Prominent examples include Irwin (1994), Kaempfer and Marks (1993) and Rogowski (1987). However, this approach also fails to capture individual attitudes.

3. CAFTA later became the North American Free Trade Agreement (NAFTA) with the addition of Mexico.

Table 9.1 Positions on Protection and Free Trade of Capital and Labor

	Labor: Protection	*Labor: Free Trade*
Capital: Protection	Distilling Textiles Apparel Chemicals Plastics Rubber Leather Shoes Stone products Iron and steel Cutlery Hardware Bearings Watches	Tobacco products
Capital: Free Trade	Petroleum products	Paper Machinery Tractors Trucks Aviation

Source: Magee (1980).

Table 9.2 Estimates of the Determinants of Individual Trade-Policy Preferences					
Variable	*Model 1*	*Model 2*	*Model 3*	*Model 4*	
Constant	**1.567**	**1.552**	**1.596**	**1.578**	
Occupation wage	**-1.766**	**-1.759**	**-1.746**	**-1.738**	
Sector tariff		1.828		2.275	
Sector net ex share	-0.624		-0.653		
County exposure 1	-0.334	-0.301			
County 1*House	**2.195**	**2.182**			
Country exposure 2			-0.146	-0.136	
County 2*House			**0.780**	**0.779**	
No. of observations	1736	1736	1736	1736	
Note: coefficients in boldface type are significantly positive or negative at the 5 percent level.					
Source: Scheve and Slaughter (2001)					

Figure 9.1

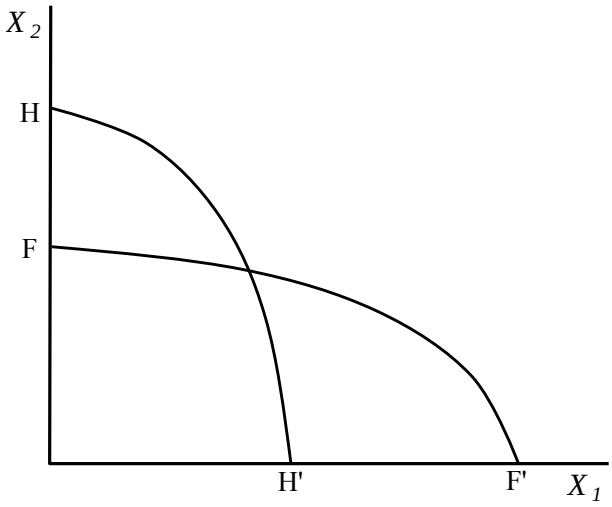


Figure 9.2

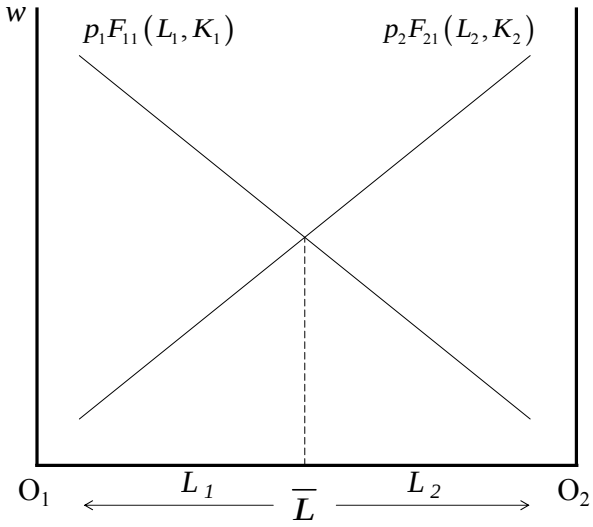


Figure 9.3

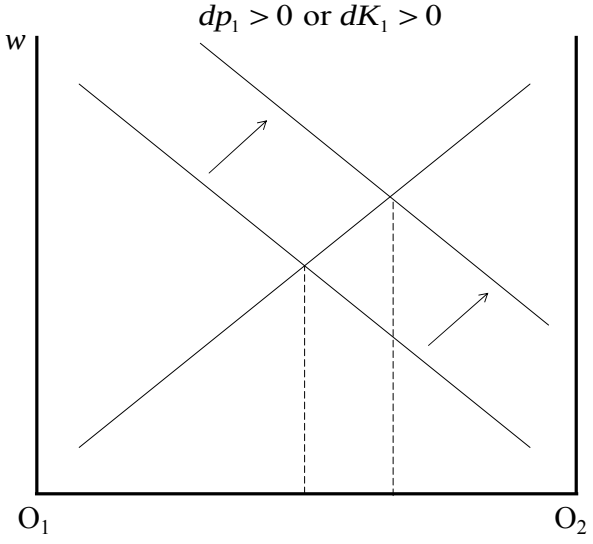
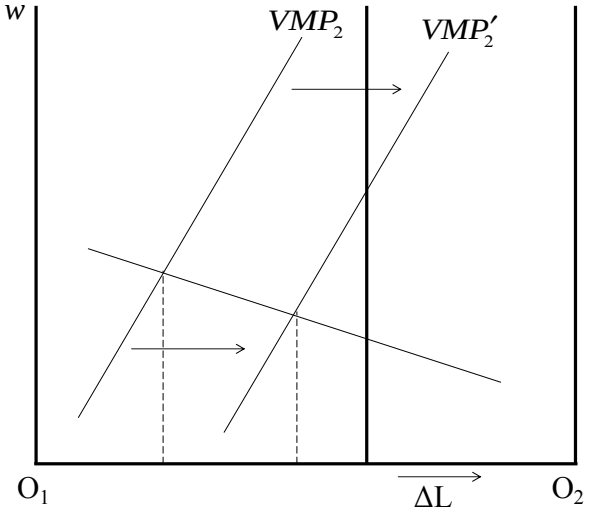


Figure 9.4



Chapter 10

DISTORTIONS AND EXTERNALITIES AS DETERMINANTS OF TRADE

10.1 Departures from our stylized world

Chapters 7, 8, and 9 discussed in some detail two of the five determinants of trade described in Chapter 6, namely, production function (technology) differences, and endowment differences. In this chapter we will discuss domestic distortions as a determinant of trade. The term distortion refers to departures from our highly stylized world which is characterized by (1) all agents are perfectly competitive and have no market power, which in turn generally implies that firms have constant returns to scale as we will discuss in the next chapter; (2) prices and quantities are freely determined by the forces of supply and demand; and (3) all interactions are though markets, meaning that there are no externalities. An externality in turn is an effect one agent (firm or consumer) has on another which is not priced through a market. Pollution is a classic example of a negative externalities, also called a spillover, in which the actions of one agent harm another and the latter is not compensated for this externality effect. A knowledge spillover in which the innovation of one firm is copied by other firms without the former being compensated is a classic example of a positive externality. Externalities, spillovers and other distortions are also referred to as "market failures".

The principal focus of the chapter is to understand how distortions and externalities by themselves can be causes of trade and whether or not distortion-induced trade leads to gains from trade. The chapter retains the focus on the positive theory of trade and does not focus on normative issues such as optimal government policy in the presence of market failures; some of that will be considered later in the book. Words like distortion and market failure inevitably sound like bad things and they indeed are relative to a best possible outcome. Keep in mind that the terms refer both to failures in which too much of something is produced such as pollution and in which too little of something is produced, such as knowledge.

There are two distinct parts of the chapter. The first part focuses on taxes and subsidies. The examination of taxes and subsidies is intended as an example of the effects that a wide range of government policies can have on trade. For example, environmental policies and regulations impact on firms costs, and hence have effects on outputs and trade similar to those produced by taxes. We are not asserting that commodity and factor taxes rank with factor endowments as a cause of trade, but we do believe that collectively, government policies have a much more profound impact on trade than is suggested in most international trade textbooks. One theme of the chapter is that government policies can generate trade, but it not necessarily beneficial trade.

As before, the approach will be to neutralize other factors so that a clear understanding of the specific effects of each can be obtained. Thus throughout the tax analysis, we will assume that either (A) we have a single country facing fixed world prices, or (B) there are two countries that are identical in all respects. They have identical technologies, identical factor endowments, identical homogeneous utility functions, and constant returns and perfect competition in production. In the absence of the distortions that we introduce, the two countries would have no incentive to trade, or alternatively, the free trade equilibrium would be identical to autarky.

10.2 Distinguishing among consumer, producer and world prices

When we introduce taxes and subsidies into the analysis, it becomes important to distinguish prices paid by consumers from prices received by producers. Once we introduce trade, consumer and producer prices must be distinguished from world prices: the prices at which the country can trade. Throughout this chapter, we will use the notation q to represent consumer prices, p to represent producer prices, and p^* to represent world prices. This notation will also be used in later chapters, such as the

chapter on tariffs.

In order to focus on trade issues, we will also make the assumption throughout the chapter that there is no government sector per se; the government returns all tax collections to consumers in lump-sum fashion and/or raises all subsidies by lump-sum taxation. We implicitly assume a very large number of consumers, with each consumer getting a check or a bill which gives to (or takes from) the consumer his or her share of taxes (subsidies). Each consumer regards their bill or check as being unaffected by their own purchases. For example, if a consumer pays $1 in sales tax, the consumer gets a refund of only $1/N of that amount, where N is the number of consumers (consider the refund if there are 100 million consumers). Thus each consumer does indeed regard the tax as raising prices, even though the tax is return to all consumers collectively. Similar comments apply to subsidies.

We will specify taxes and subsidies in an ad valorem (percentage of value) form throughout the chapter rather than in specific form. t will denote a tax and s a subsidy. *Ad valorem* taxes are quoted as rates. A sales tax of 5%, for example, would mean a tax rate of $t = .05$ in this context. (*Specific* taxes, on the other hand, are quoted in monetary units per unit of the good: the US gasoline tax is quoted in cents per gallon.) Thus the relationship between consumer and producer prices with a tax or subsidy is given as follows.

$$\begin{aligned} q &= p(1+t) > p \qquad \textit{tax} \\ q &= p(1-s) < p \qquad \textit{subsidy} \end{aligned} \tag{10.1}$$

A tax raises the consumer price above the producer price while a subsidy lowers the consumer price below the producer price. A tax rate $t = .05$ raises the consumer price 5% above the producer price: $q = p(1.05)$. When there are only two goods, the effects of a tax on one good are equivalent to a subsidy on the other good. In order to see this, consider the commodity price ratios resulting from a tax on X_1 versus a subsidy to X_2.

$$\begin{aligned} \frac{q_1}{q_2} &= \frac{p_1(1+t)}{p_2} > \frac{p_1}{p_2} \qquad \textit{tax on } X_1 \\ \frac{q_1}{q_2} &= \frac{p_1}{p_2(1-s)} > \frac{p_1}{p_2} \qquad \textit{subsidy on } X_2 \end{aligned} \tag{10.2}$$

We see from (10.2) that a subsidy to X_2 and a tax on X_1 induce the same "wedge" between the consumer and producer price ratios.

Figure 10.1 gives autarky equilibrium at point E when there is *either* a tax on X_1 or a subsidy on X_2. The assumption discussed above that the tax revenue is redistributed lump sum and the subsidy is raised by a lump sum tax is reflected in the fact the consumption and production bundles are the same even though, in the case of a tax, for example, the consumption bundle costs more than the value of those goods at producer prices. The consumers pay more than the producers receive because of the tax, but then they receive an income in excess of the value of production due to the fact that they receive the tax refund. Let D denote consumption (demand) and X denote production quantities. For a tax on X_1,

$$q_1D_1 + q_2D_2 = p_1(1+t)X_1 + p_2X_2 = \left[p_1X_1 + p_2X_2\right] + \left[p_1tX_1\right] \tag{10.3}$$

The left-hand side of (10.3) is consumer expenditure at consumer prices. The first bracketed term on the

right-hand side is income received from production (payments to factors of production), while the second term on the right-hand side is redistributed tax revenue. Thus consumer expenditure equals consumer income. A similar analysis of a subsidy simply requires us to change the sign of *t*.

Figure 10.1

Figure 10.1 illustrates the distortionary effect of the tax on X_1 or subsidy on X_2. Welfare is lower at E that at the undistorted competitive equilibrium at A. The producer price ratio p is tangent to the production frontier TT' while the consumer price ratio q is tangent to the indifference curve through E. The tax causes the consumers to perceive X_1 as more costly than it actual is to produce, or a subsidy to X_2 causes consumers to perceive X_2 as relatively cheaper than what it actually costs to produce.

The previous paragraph should not be taken to suggest that all taxes or subsidies are bad. First, governments usually raise revenues in order to provide public goods that are not or cannot be provided by markets. No account is taken of public goods in this analysis. Second, not all taxes are distortionary or as distortionary as the commodity tax shown here. For example, in the present model, an equal ad valorem tax on both goods would leave the relative consumer and producer prices equal. Such a *set* of taxes is non-distortionary.

Finally, some government policies are imposed to correct an *existing* distortion in the economy, such as an environmental externality. In such a situation, Figure 10.1 might accurately depict the effects of a pollution tax (on X_1) on production and trade, but the indifference curves no longer indicate welfare change, only that part of welfare derived from consumption. Welfare may be improving due to lower pollution (i.e., there is actually a third good, environmental quality, not shown in the diagram). More will be said about taxes in the presence of existing distortions later in the book.

10.3 Taxes and subsidies as determinants of trade: a small open economy

Suppose that the home country faces fixed world prices. Assume also that these prices just happen to be equal to home's autarky price ratio such that home does not choose to trade at these prices. This is completely unlikely, but we are simply following the strategy outlined in Chapter 6: "neutralize" all causes of trade except the one which we wish to examine. The situation is shown in Figure 10.2 where the autarky equilibrium A is also the free-trade equilibrium at price ratio p^*.

Once we introduce trade, we not only have to keep track of consumer and producer prices, but world prices as well. This in turn means that we have to specify whether a tax or subsidy is assessed on consumption or production. In the closed economy, it does not matter since production and consumption of each good are equal. But with trade, consumption and production are in general not equal, so it matters which one we are taxing. In this section, we will limit ourselves to looking only at production taxes and subsidies. These are not necessarily more common than consumption taxes, but space constraints limit the range of distortions we can deal with here. It is also true that focusing on production taxes helps build some intuition for the analysis of imperfect competition which begins in the next Chapter.

Consider a tax on the *production* of X_1 or a subsidy on the production of X_2. In this case, consumers face world prices, not producers. Consumer prices and world prices will be equal to one another, but not to world prices. The relationships among the three price ratios are given by

$$\frac{p_1(1+t)}{p_2} = \frac{q_1}{q_2} = \frac{p_1^*}{p_2^*} > \frac{p_1}{p_2} \tag{10.4}$$

The relationships in (10.4) are shown in Figure 10.2. The producer price ratio is now greater than the consumer and world price ratios, so production is shown as taking place at point X^t in Figure 10.2.

Consumption must take place along the world price ratio through X^t, and consumers now face world prices. Thus the consumption point is given by the tangency between an indifference curve and the price line p^* through point X^t. We show the consumption point as D^t in Figure 10.3. The production tax discourages production of X_1 and leads to a substitution in production toward good X_2.

Figure 10.2

Several important results are shown in Figure 10.2. First, it clearly demonstrates that government policies such as taxes and subsidies can generate trade. However, it shows equally clearly that *trade induced by the introduction of distortions is not beneficial trade*. In Figure 10.2, the country receives a welfare loss as a consequence of distortion-induced trade. Point X^* in Figure 10.2 would be the undistorted equilibrium yielding a utility level of U^*. The distorted equilibrium yields a utility level of U^t.

This is a very important results insofar as governments sometimes decide that it would be a good thing if the country produced and exported more of a certain good (e.g., "high tech" goods). We could think of Figure 10.2 as a situation generated by a government deciding that it must be good to produce and export X_2. By putting on a subsidy to the production of X_2 we do indeed get exports of X_2 and the government congratulates itself on the success of its project. However, exports generated by distortions are welfare reducing (put differently, the initial level of exports, zero, is optimal).

Welfare is reduced by this distortion, because producers make *privately* efficient choices given the prices they face, but they do not make efficient *social* choices when they do not face the true costs of producing the commodities. Again, we should separate this welfare result from the results concerning consumption and trade, since not all commodity taxes need be welfare reducing. Most countries have gasoline (petrol) taxes, for example, which have the beneficial effects of reducing pollution and traffic congestion.

10.4 Taxes and subsidies as determinants of trade: two identical countries.

Now we introduce a second country and explicitly return to our concept of the two- country, no-trade model. Suppose that we have two identical countries, both with production frontiers as in Figure 10.1, such that the point A in Figure 10.1 represents both the free-trade and autarky equilibria for both of the countries. Now country h imposes a production tax on X_1 or a production subsidy X_2. At the initial free-trade price ratio p^*, country f will wish to continue to produce and consume at A in Figure 10.3 (the same as A in Figure 10.1)) while country h will wish to shift production away from X_1 toward X_2.

This cannot be an equilibrium because there will be excess demand for X_1 and excess supply of X_2 at the initial prices. The price ratio p^* must rise, giving us a new equilibrium as shown in Figure 10.3. The rise in p^* induces country f to export X_1 and import X_2, producing at X^f and consuming at D^f in Figure 10.3. We see that a production tax on X_1 or subsidy on X_2 can indeed generate exports of X_2, but this is not welfare improving trade. Country h creates trade by its tax or subsidy, but it is not beneficial trade.

The interesting additional result that we get from Figure 10.3 is that *country f is made better off* by h's tax or subsidy. Recall from Chapter 5 that the ability of a country to trade at any prices other than its autarky prices can make it better off (and *will* make it better off if it has no distortions). The institution of the tax or subsidy in country h now allows country f an opportunity to trade at prices different than its autarky prices. This might also help us understand why country h has to be worse off. With the countries absolutely identical, there are no opportunities for mutual gains from trade. If the distortion makes country f better off, it must make country h worse off.

Figure 10.3

The implication here is that country f should be happy when it’s trading partner subsidies its exports to h. Intuitively, the subsidizing country h is selling for less than the cost of production to the

benefit of the passive country f. Happiness in f is rarely the reaction in practice, however, to a trading partner's subsidy. In some cases, a government may simply misunderstand this gift. But our result here is not general. In a Heckscher-Ohlin world of multiple factors of production, someone in country f is surely worse off and will understandably make a political fuss. Second, suppose that there are three countries, two of which export X_2 to the third. Then the first two, call them h and f are competitors, and a subsidy to X_2 by country h is going to drive down the price of X_2 for both countries, making country f worse off as well.

Again, we will not provide a detailed normative discussion here. Our purpose is simply to show that a distortion can serve as a basis for trade but also that trade generated by a distortion is not necessarily good trade.

10.5 Production externalities

As noted above, externalities are another source of "market failure" yet at the same time can imply an additional source of gains from trade. Such externalities come up in quite a number of contexts, ranging from pollution to intellectual property. In this section, we will look at positive production externalities among firms in an industry. As suggested earlier, this could result from knowledge spillovers in which the innovations of one firm are quickly copied by other firms without compensating the innovating firm. Many other cases are discussed in the literature, including the increases in the range of intermediate goods as an industry or country grows, an idea going back to Adam Smith's: "the division of labor is limited by the extent of the market". One interpretation of division of labor is increases in the number of specialized intermediate goods, a topic we will return to in Chapter 12.

Suppose that there is just a single factor of production labor, L, which is in fixed supply. Good X_2 is produced with constant returns to scale by a competitive industry. Good X_1 is produced by competitive firms who perceive themselves as having constant to scale, but the productivity of their labor inputs is positively related to the overall output of the industry. These competitive firms view total industry output as constant, much as we assume that competitive firms view the industry price as constant and unaffected by their own decisions. Let X_{i1} denote the output of an individual firm in industry 1 and let X_1 denote total industry output, the sum over all i firms. The production side of our economy is given as follows:

$$X_{i1} = (X_1^{\alpha})L_{i1} \qquad X_2 = L_2 \qquad \bar{L} = \sum_i L_{i1} + L_2 \tag{10.5}$$

where $0 \leq \alpha < 1$ is an externality parameter: $\alpha = 0$ is the special case of no externality, in which case the model reduces to the Ricardian model of Chapter 7. As just noted, each individual firm i in industry X_1 views total industry output as constant. In competitive equilibrium, each firm equate the value of the marginal product of labor to the wage rate, denoted w, as in the Ricardian model. Competitive equilibrium is then described by

$$p_1 X_1^{\alpha} = w \qquad p_2 = w \qquad \frac{p_1}{p_2} = \frac{1}{X_1^{\alpha}} \tag{10.6}$$

Total industry output in X_1 is given by summing the first equation in (10.5) over all i firms. We do this and then rearrange the equation to given total industry output X_1 as follows.

$$\sum_i X_{i1} = X_1 = X_1^{\alpha} \sum_i L_{i1} = X_1^{\alpha} L_1 \qquad X_1^{1-\alpha} = L_1 \qquad X_1 = L_1^{\frac{1}{1-\alpha}} \tag{10.7}$$

Since $\alpha < 1$, the exponent on the right-hand equation of (10.7) is greater than one: total industry output exhibits increasing returns to scale in its total labor input. Differentiate the middle equation in (10.7)

along with the equation for X_2 output, making use of the total labor supply constraint.

$$(1-\alpha)X_1^{-\alpha}dX_1 = dL_1 \qquad dX_2 = dL_2 = -dL_1 \tag{10.8}$$

Divide the first equation of (10.8) by the second and rearrange.

$$-\frac{dX_2}{dX_1} = (1-\alpha)\frac{1}{X_1^{\alpha}} \tag{10.9}$$

which is the slope of the production frontier, the marginal rate of transformation. As we noted back in Chapter 2, the production frontier is a convex function (the production set is non-convex) reflecting the increasing returns to scale in X_1: the denominator of (10.9) gets smaller as X_1 gets larger. The production frontier for our economy is shown as $\bar{X}_2\bar{X}_1$ in Figure 10.4.

Figure 10.4

Now combine (10.9) with the competitive pricing condition in (10.6). This gives us a relationship between the marginal rate of transformation and the equilibrium price ratio.

$$-\frac{dX_2}{dX_1} = (1-\alpha)\frac{p_1}{p_2} < \frac{p_1}{p_2} \tag{10.10}$$

Now we discover a second issue, in addition to the convexity issue, connected with a positive production externality. The competitive-equilibrium price ratio is not tangent to the production frontier, but rather cuts it as shown in Figure 10.4. The intuition behind this result is the fact that when an individual firm expands output a little, it confers a positive productivity effect on all other firms taken together. Thus the true or "social" marginal product of an additional worker hired is greater than the "private" marginal product of an individual firm. Or to put it the other way around, the private cost of an additional worker hired is more than the true social cost. The slope of the production frontier depends on the true social cost and so it is flatter than the price ratio, equal to private marginal cost. Competitive equilibrium is at a point like A in Figure 10.4, giving a welfare level of U^a.

While we don't want to get into a detailed normative analysis here, we should note that this is a case where an offsetting distortion could increase welfare. The first-best outcome in Figure 10.4 is at point S yielding welfare U^s. This could be achieved by a subsidy to X_1, in effect compensating for or "internalizing" the externality. If the externality is due to imperfect protection of intellectual property (the innovating firms is not compensated for benefits conferred on other firms), then added intellectual property protection will act to offset the distortion and move the country toward point S in Figure 10.4. This is an application of what is known as the theorem of the second best:

> Theorem of the second best: in the presence of one distortion, the imposition of an additional and offsetting distortion can improve welfare.

We will return to this idea later in the book.

10.6 Trade and gains from trade in the presence of production externalities

Let us abstract from the issue of the price line cutting the production frontier by assuming that there are positive externalities in both X_1 and X_2 of the same degree α. Then there will be a term $(1 - \alpha)$ in

both the numerator and denominator of (10.10) and these will cancel out, leaving the marginal rate of transformation equal to the price ratio. In this special case, we will have an autarky equilibrium at point $X^a = D^a$ in Figure 10.5, giving an autarky welfare level of U^a.

Figure 10.5

Now assume that there are two absolutely identical economies and let them trade. It is beyond the scope of this chapter to given a detailed analysis of the adjustment process but in short, the outcome shown in Figure 10.5 at X^a continues to be an equilibrium but it is *unstable*. A small perturbation can send the two countries off to corners, each specializing in only one of the two goods. One country could specialize in good X_2 and the other in X_1 and they could each trade half of their output for half of the other country's good, leading both counties to share a common consumption point at D^* in Figure 10.5. Here is our first instance of how there can exist gains from trade even between identical countries arising from increasing returns to scale.

The outcome shown in Figure 10.5 in which the equilibrium price ratio is exactly the chord connecting the two endpoints of the production frontier is a special case requiring strong symmetry assumptions. Suppose at this price ratio, the countries do not demand X_1 and X_2 in the proportions produced, but each country wants to consume a lot of X_1 and not so much X_2. Then at the price ratio shown in Figure 10.5, there will be excess demand for X_1 and excess supply of X_2. The outcome is going to have to be as shown in Figure 10.6: the relative price of X_1 will rise to clear the market. The country specializing in X_2, call that country h, consumes at D^h and country f specializing in X_1 consumes at D^f in Figure 10.6.

Figure 10.6

Note that the outcome shown in Figure 10.6 is not the only possibility. Reversing which country specializes in which good is also an equilibrium. Thus situations in which there are production externalities can be characterized by multiple equilibria. In such a situation, we want to note that this phenomenon has many implications, for development economics in particular. In the presence of multiple equilibria, a country doesn't want to end up in the "wrong" outcome: it wants to be country f in Figure 10.6, not country h. A more detailed analysis is beyond the scope of this chapter. In closing, refer back to Figure 10.4. Suppose one country has a free market and the other one "internalizes" the externality via intellectual property protection for example. The former country is at A while the latter one is at S in Figure 10.4. If trade opens up, the country at S has a lower price for X_1 and hence will tend to end up specializing in X_1: this country will be country f shown in Figure 10.6. Internalizing positive production externalities can give a country an advantage in a situation of multiple equilibria.

10.7 Summary

This chapter has turned our attention away from underlying production differences between countries, principally differences in technologies and in factor endowments. Instead, we neutralize these differences by assuming that technologies, factor endowments and demand are identical across two countries. This continues the methodology outlined in Chapter 6. Here, we look at distortions and externalities (which are themselves a class of distortions) to see how asymmetries in distortions across countries can generate trade and may or may not generate gains from trade for each of two countries.

One class of distortions is represented here by simple production taxes and subsidies. In such a situation, it is important to keep track of different sets of prices, in particular consumer, producer, and world prices. The principal result of this analysis is that production taxes or subsidies can generate trade between otherwise identical countries, but it is a welfare-reducing trade for the taxing/subsidizing country. This carries a very important lesson for policy makers: exports should never be confused with welfare. A production subsidy to a favored sector may indeed generate exports from that sector, but these exports are being sold abroad for less than the cost of production and hence are welfare worsening.

The second class of distortions addressed here are positive production externalities or spillovers, arising due to some failure on the part of firms to be able to capture returns for benefits they confer on rival firms. Examples include the lack of protection for intellectual property and innovations, and increases in productivity arising from the finer "division of labor" in a larger market. These issues will arise again in this book.

These positive production externalities imply aggregate increasing returns to scale, even though individual firms have constant returns technologies. As discussed back in Chapter 2, this can imply that the production frontier is convex (the production set non-convex) or "bowed in". This in turn can imply positive gains from trade for each of two identical countries: with each country specializing in only one sector, productivity rises and mutual gains from trade can exist. The situation is not simple however, and we touch briefly on issues like the existence of multiple equilibria, with the alternative equilibria having very different welfare implications for the trading partners.

The next two chapters continue to look at similar, even identical economies, and analyze how imperfect competition and increasing returns to scale offer added sources of gains beyond those arising from comparative advantage linked to differences between countries.

REFERENCES

Bhagwati, J. and T.N. Srinivasan (1983), *Lectures on International Trade*, Cambridge: MIT Press, Chapters 20-23.

Grossman, Gene M. and Estaban Rossi-Hansberg (2010), "External Economies and International Trade Redux", *Quarterly Journal of Economics* 125, 829-858.

Ethier, Wilfred J. (1979), "Internationally decreasing costs and world trade", *Journal of International Economics* 9, 1-24.

Kemp, Murray C. (1969), *The pure theory of international trade and investment*, New Jersey: Prentice Hall.

Markusen, James R. (1990), "Micro-Foundations of External Economies", *Canadian Journal of Economics* 23, 495-508.

Markusen, James R. and James R. Melvin (1981), "Trade, Factor Prices, and the Gains from Trade with Increasing Returns to Scale," *Canadian Journal of Economics* 14, 450-469.

Melvin, J. R. (1970). "Commodity Taxation as a Determinant of Trade." *Canadian Journal of Economics* 3, 62-78.

Melvin, J. R. (1969). "Increasing Returns to Scale as a Determinant of Trade", *Canadian Journal of Economics* 2, 389-402.

Figure 10.1

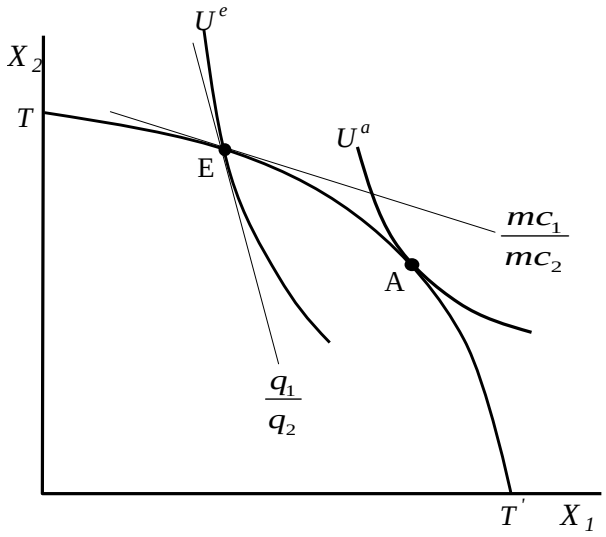


Figure 10.2

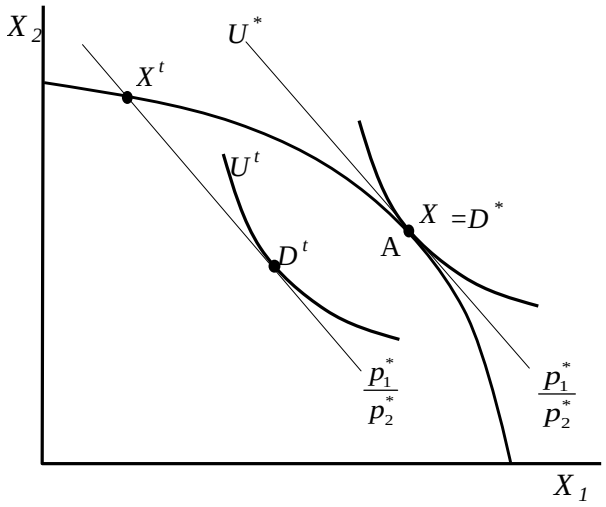


Figure 10.3

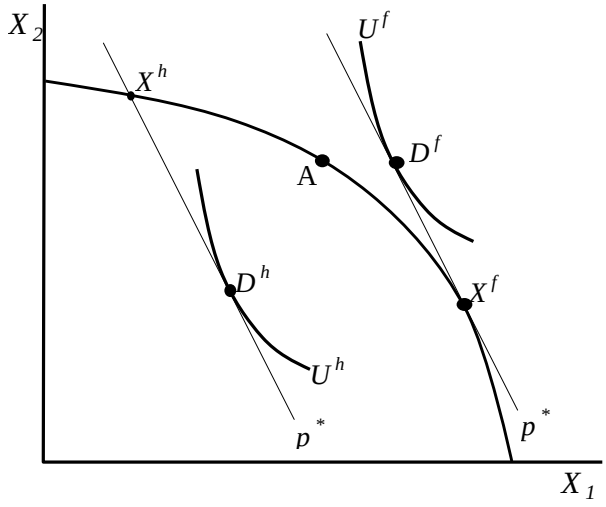


Figure 10.4

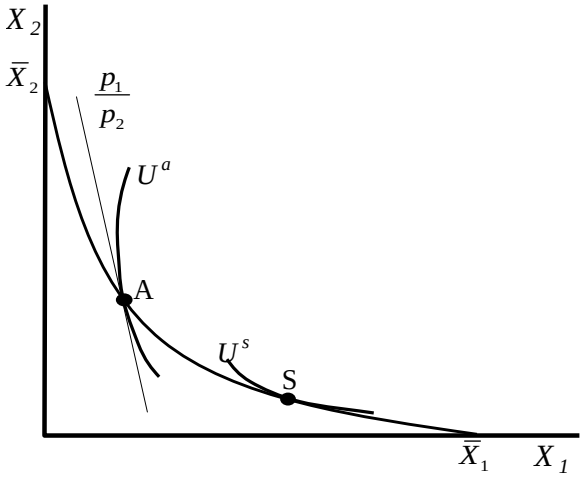


Figure 10.5

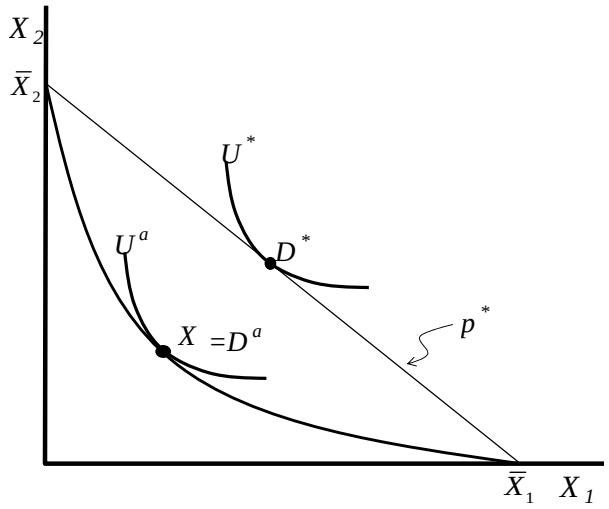
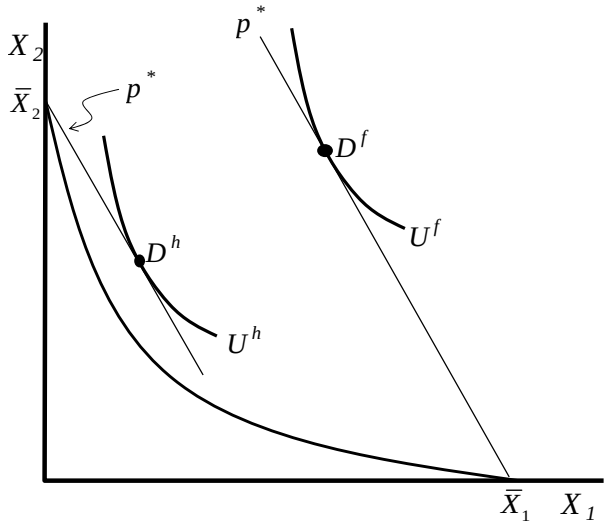


Figure 10.6



Chapter 11

IMPERFECT COMPETITION AND INCREASING RETURNS I: OLIGOPOLY

11.1 General discussion of increasing returns and non-comparative-advantage gains from trade.

In the previous chapter and also in Chapter 2 (e.g., Figures 2.6 and 2.7), we noted that increasing returns to scale, or scale economies for short, offer the opportunity for gains from trade even for identical economies. Trade and gains from trade can occur without any comparative advantage basis for trade. Trade economists believe that scale economies offer an important explanation for the observation of large volumes of trade between similar economies that we observe in trade data.. In this chapter and in the next, we will analyze these idea in a more thorough fashion.

In this chapter, we will focus on a case where there are two industries, one producing a good under conditions of perfect competition and constant returns to scale, and a second good produced under conditions of increasing returns to scale. This IRS good is assumed to be homogeneous, meaning the goods produced by different firms are identical from the consumers' point of view. In the next chapter, we assume that each firm in the IRS sector produces a good that is differentiated from those of its rival firms. In order to avoid some possible confusion, we partly depart from our usual notation here and denote the competitive good as Y and the IRS homogeneous good as X. This will prove useful in the next section when we have different varieties of X.

We will assume that the X firms have technologies in which there is an initial fixed cost of entering production and then a constant marginal cost of added output. We assume a single factor of production L (call it labor) which must be divided between the Y and X sectors and among firms in the X sector. Marginal cost is denoted by mc and total cost (tc) and average cost (ac) for an X firm are as follows:

$$tc = mcX + fc \qquad ac = \frac{tc}{X} = mc + \frac{fc}{X} \tag{11.1}$$

where mc and fc are *constants* (parameters), measured in units of labor. Figure 11.1 shows the cost curves of a single firm. It is important to note that average cost is always falling with increased output and it converges to marginal cost, but never quite equal to marginal cost. In mathematical terminology, average cost is a rectangular hyperbola, going to infinity as X goes to zero and approaching mc as X goes to infinity.

Figure 11.1

The consequences of this type of technology for competitive conditions, sometimes referred to as market structure are rather profound. Specifically, this type of increasing-returns technology cannot support perfect competition as a market outcome. This is typically explained by a proof by contradiction. Suppose that there are many small firms such that each firm regards market price as exogenous. If price equaled marginal cost as assumed in competitive theory, then each firm would be making losses since average cost is greater than marginal cost, so this cannot be an equilibrium. Suppose then that the competitive price exceeds marginal cost but again is viewed as a constant by each firm. Then each or any firm knows once it produces enough output, price will exceed average cost, so profits become positive. But the firm would not stop there, it maximizes profits at the constant price by producing an infinite output and so expands without bounds. But this cannot be an equilibrium and contradicts the assumption that firms are small. Thus the technology shown in (11.1) and Figure 11.1 is inconsistent with perfect competition.

The equilibrium outcome (assuming an equilibrium exists) must be one in which only a small number of firms are able to survive in equilibrium. Competitive will be imperfect, with each firm having influence over market price. The consequences of the technology in Figure 11.1 are shown for a single monopoly *X* firm in Figure 11.2. Also assume a very simple technology for *Y*, such that one unit of *L* produces exactly one unit of *Y: Y* = L_y. *Y* and *L* must have the same price and we will use *L* and therefore *Y* as numeraire, giving them a price of one. $\bar{L}$ is the economy's total endowment of *L*.

The production frontier for the economy, discussed briefly in Chapter 2, runs from $\bar{Y}$ to Y^1 to X^1 in Figure 11.2 where $\bar{Y}$ to Y^1 is the fixed cost *fc* measured in units of *Y* which is of course also units of *L* in this special case of $Y = L_y$. Suppose that the monopoly equilibrium is at point A in Figure 11.2, where the output of *Y* and *X* are Y^0 and X^0. Then the average cost of producing *X* at point A is given by the total amount of labor needed for *X* divided by the output of *X*.

$$ac = \frac{\bar{L} - L^0}{X^0} = \frac{\bar{Y} - Y^0}{X^0} \qquad (11.2)$$

Note that this is the slope of the dashed line in Figure 11.2, the line running from point A to the end of the production frontier $\bar{Y}$. The consequence of this is that, if the monopoly *X* firm is to at least break even, then the price line at equilibrium A must be at least as steep as the dashed line giving average cost. Such a price is shown at p^a in Figure 11.2, where *p* gives the price of *X* in terms of *Y* (and therefore in terms of *L*).

Figure 11.2

A crucial point is that the market equilibrium involves a market failure: the market outcome is not optimal. The optimum would be a tangency between an indifference curve and the production frontier, implying price equal to marginal cost, but we have shown that this is impossible because it involves the firm making losses. With increasing returns to scale that are "internal" to individual firms as we are assuming here, returns to scale are going to be inherently bound up with imperfect competition. Too little *X* is produced at too high a price.

That is a rather long preliminary that now allows us to get to the heart of the issue: the opportunity to trade is going to offer economies the possibility of pro-competitive gains from trade. Countries will gain from the benefits of increased competition that makes firms produce more at lower average costs (higher productivity) and lower prices. We will focus on putting together two identical economies. Within this setting we consider two cases, one in which the number of firms is fixed before and after trade opens up and one in which there is free entry or exit of firms in response to trading opportunities.

11.2 Pro-competitive gains: the basics

Any discussion of imperfect competition must start with some assumption about how firms react to one another, since each firm large relative to the market. In this chapter, we are going to assume Cournot-Nash (or Cournot for short) competition in which firms pick a quantity as a best response to their rivals' quantities. Cournot equilibrium occurs when each firm is making a best (optimal) response to its rivals' outputs. Algebraically, this will be modeled as each firm picking an output quantity holding rivals' outputs constant.

Revenue for a Cournot firm i and selling in country j is given by the price in j times quantity of the firm's sales. Price is a function of all firms' sales.

$$R_{ij} = p_j(X_j)X_{ij}. \qquad \text{where } X_j \text{ is total sales in market j by all firms} \qquad X_j = \sum_i X_{ij} \qquad (11.3)$$

Cournot conjectures imply that $\partial X_j/\partial X_{ij} = 1$; that is, a one-unit increase in the firm's own supply is a one-unit increase in market supply. Marginal revenue is then given by the derivative of revenue in (11.3) with respect to firm i's output (sales) in j.

$$\frac{\partial R_{ij}}{\partial X_{ij}} = p_j + X_{ij}\frac{\partial p_j}{\partial X_j}\frac{\partial X_j}{\partial X_{ij}} = p_j + X_{ij}\frac{\partial p_j}{\partial X_j} \qquad \text{since } \frac{\partial X_j}{\partial X_{ij}} = 1 \text{ (Cournot)} \qquad (11.4)$$

Now multiple and divide the right-hand equation by total market supply and also by the price.

$$\frac{\partial R_{ij}}{\partial X_{ij}} = p_j + X_{ij}\frac{\partial p_j}{\partial X_j} = p_j + p_j\frac{X_{ij}}{X_j}\left[\frac{X_j}{p_j}\frac{\partial p_j}{\partial X_j}\right] \qquad (11.5)$$

The term in square brackets in (11.5) is just the inverse of the price elasticity of demand, defined as the proportional change in market demand in response to a given proportional change in price. This is negative, but to help make the markup formula clearer we will denote minus the elasticity of demand, now a positive number, by the Greek letter $\eta > 0$. We can then write (11.5) as

$$\frac{\partial R_{ij}}{\partial X_{ij}} = p_j\left[1 - \frac{X_{ij}}{X_j}\frac{1}{\eta_j}\right] \qquad \eta_j \equiv -\left[\frac{p_j}{X_j}\frac{\partial X_j}{\partial p_j}\right] \quad (\textit{elasticity of demand}) \qquad (11.6)$$

The term X_{ij}/X_j in (11.6) is just firm i's market share in market j, which we can denote by s_{ij}. Then marginal revenue = marginal cost is given by:

$$mr_{ij} = p\left[1 - \frac{s_{ij}}{\eta_j}\right] = mc_i \qquad (11.7)$$

Marginal revenue in Cournot competition turns out to have a fairly simple form as shown in (11.7). The term s_{ij}/η_j is referred to as the "markup". As you will note, it looks something like the tax formulas we had in the previous chapter and indeed it is possible to think of the monopolist as putting on sort of a tax that raises price above marginal cost. If you refer back to equation (11.4) the markup relates to how much the market price falls when the firm increases its sales. When the demand elasticity is high, there is only a small fall in price when the firm expands output, and so the markup is small.

The role of the firm's market share is more subtle, but crucial to understanding the whole idea of pro-competitive gains. Suppose that we put two absolutely identical countries together in trade, each having just a single *X* producer. The firms could just continue to do what they were doing in autarky, and prices etc. would be preserved. But each firm will now understand that, if it increases output by one unit, the increase is spread over twice as many consumers and thus the price will fall by only half as much as it would if the monopolist increased supply by one unit in autarky. We sometimes say that the firm now

perceives demand as more elastic or that the *perceived elasticity of demand* η_j/s_{ij} is higher. This is reflected in (11.7) by the fact that s_{ij} goes from one in autarky to one-half when trade opens. The markup up falls, perceived marginal revenue increases, and each firm has an incentive to increase output.

Figures 11.3 and 11.4 work through the consequences of this. The production frontier shown in Figure 11.3 is the production frontier for each of the two identical countries. Point A is the autarky monopoly equilibrium in each country. When the countries are put together in trade, each firm perceives demand as more elastic according to (11.7) and expands output. Equilibrium is restored when the price of X is forced down so that (11.7) is again an equality for both firms with market shares both equal to one half. The new equilibrium must be at a point like T in Figure 11.3.

Figure 11.3

Note especially that this must constitute a welfare gain for both countries. Output was too low and the price was too high initially in autarky, essentially a distortion like those of the previous chapter. However, this is quite different in that it is an *endogenous* distortion and the opening of trade can be thought of as reducing that distortion. There is a welfare gain from trade (U^a to U^*)for the identical economies through a lower X price (p^a to p^*), higher X output (A to T), and more efficient production: the average cost of a unit of X is lower, firms move down their average cost curves in Figure 11.1.

Figure 11.4 shown a somewhat different and equally important case. Suppose that there might be more than one firm in each country initially, and that firms can enter or exit until profit are zero.[1] Then, referring back to equation (11.2) and Figure 11.2, the equilibrium price must be equal to average cost, given by the line connecting the production point in Figure 11.4 with the endpoint of the production frontier $\bar{Y}$. Let point A in Figure 11.4 be the initial autarky equilibrium for each country and let p^a denote the autarky price: the price equals average cost reflecting the entry or exit of firms until profits are zero. The distance $\bar{Y}Y^0$ in Figure 11.4 is now the combined or sum of the fixed costs of all firm active in the market in equilibrium.

Figure 11.4

Now put the two identical economies together in free trade as before. Each firm individually has an incentive to expand output: equation (11.7) continues to apply to pricing decisions. However, when they all do this, profits will be driven negative. This will cause the exit of some firms in each country until zero profits are re-established. This is represented by the outward shift of the linear segment of the production frontier in Figure 11.4 from $\bar{Y}Y^0$ to $\bar{Y}Y^1$: resources which were being used (uselessly) in fixed costs are now freed up for actual production. The new equilibrium will be at a point such at T in Figure 11.4. Note that it is possible now for the consumption of both Y and X to increase due to the efficiency gains of a larger world output being produced by fewer firms than the total number in the two countries in autarky. Welfare in each country rises from U^a to U^*.

A typical intelligent question at this point is how is firm exit consistent with more competition? The answer is that there is some exit in each country individually, but more left in free trade in total between the two countries than were in each individual country in autarky. For example, suppose that each country has four firms in autarky. Suppose that trade forces the exit of one firm in each country, leaving three in each country. That means that there are now six firms in total competing for the business of each consumer instead of four in autarky. Exit in each country individually is quite compatible with increased competition.

The nice outcomes (for free trade advocates at least) shown in Figures 11.3 and 11.4 have not been rigorously established, and then are a number of problems. First, the elasticity of demand which we denoted by η is generally not a constant: it will change as prices and total sales change. While the basic message is clear and some readers and professors will wish to move on, we now proceed to offer two special cases that have been widely used in the literature to more rigorously solve for the effects of trade.

11.3 Special case I: quasi-linear preferences

The first special (or rather specific) case has been widely used in industrial organization as well as in international trade. In the latter, it has been widely used to analyze what is known as "strategic trade policy", a topic treated later in the book. It will also be used later in analyzing multinational firms.

We can keep the notation simple at first by assuming just a single monopoly firm in each country, so n = 1, and by normalizing the population L to the number one, so X gives both total output of X, the output per firm, and the consumption per capita. Assume that the preferences of a representative consumer in each country are given by:

$$U = \alpha X - (\beta/2)X^2 + Y \tag{11.8}$$

The crucial property of these preferences are constant marginal utility for good Y and the importance of this will become clear in a minute. This linearity in Y has led some authors to call these preferences "quasi-linear". As above, one unit of factor L is need to produce one unit of good Y, and p denotes the price of X in terms of Y or L. Let Π denote profits of the firm and L (equal to one initially) the number of workers/consumers. Profits are viewed as exogenous by individual consumers. Then the budget constraint for the representative consumer is given by income (also measured in terms of L) equals expenditure.

$$L + \Pi = pX + Y \tag{11.9}$$

Substituting from the budget constraint for good Y in (11.8), we have the consumer's choice problem, where profits (Π_i) are viewed as exogenous by an individual consumer.

$$Max(X)\ U_i = \alpha X - (\beta/2)X^2 + L + \Pi - pX \tag{11.10}$$

The (inverse) demand function is given by the first-order condition, the derivative of (11.10) with respect to X, and is linear in X.

$$\frac{dU}{dX} = \alpha - \beta X - p = 0 \quad \Rightarrow \quad p = \alpha - \beta X \tag{11.11}$$

The feature of quasi-linear preferences that makes them so attractive is that demand does not depend on income and hence it does not depend on profits. This makes the model much easier to solve. On the other hand, the unattractive feature of these preferences is that the demand for X does not depend on income: there is a zero income-elasticity of demand for X, surely a totally unrealistic assumption. All added income at a fixed price for X will be spent on Y.

Consider autarky first. Let Π denote the profits of the domestic firm. These profits are revenues minus marginal costs (*mc* will be denote by just c) and minus fixed costs, denoted simply as f. Substituting in the demand function in (11.11) for p, we get

$$\Pi = pX - cX - f = [\alpha - \beta X]X - cX - f \tag{11.12}$$

The first-order condition for profit maximization gives:

$$\frac{d\Pi}{dX} = \alpha - 2\beta X - c = 0 \quad \Rightarrow \quad X = \frac{\alpha - c}{2\beta} \tag{11.13}$$

Now assume that two identical economies trade freely. Let the two countries and their firms be denoted with subscripts i and j. There are now twice as many consumers in the integrated world economy and the demand price depends on *per capita* consumption, not on *total* world consumption. That is, world price will be unchanged if there is twice as much output in the integrated world as there was in each country individually in autarky. We now have $L = 2$ and so the world price of X will be

$$p = \alpha - \beta(X_i + X_j)/L = \alpha - \beta(X_i + X_j)/2 \tag{11.14}$$

Profits for firm i are now given by

$$\Pi_i = [\alpha - \beta(X_i + X_j)/2]X_i - cX_i - f \tag{11.15}$$

Assume Cournot competition, so each firm makes a best response to the other firm's output, maximizing profits holding the other firm's output fixed. The first-order condition for profit maximization is given by

$$\frac{d\Pi_i}{dX_i} = \alpha - \beta X_i - \beta X_j/2 - c = 0 \tag{11.16}$$

There is a corresponding equation for firm j, giving two equations in two unknowns. Since we have assumed that the countries are absolutely identical, we know that the solution will be symmetric with both firms producing the same amount in equilibrium. Exploiting this symmetry, we can solve (11.16) for the Cournot output of the firm i (equal to the country j firm's output) by setting $X_i = X_j$.

$$X_i^* = X_j^* = 2\frac{(\alpha - c)}{3\beta} > \frac{\alpha - c}{2\beta} = X^a \tag{11.17}$$

where the right-hand inequality is the autarky output of each firm given in (11.13). Output of each firm expands by one-third ((2/3)/(1/2) = 1/3) when trade opens. This is the effect shown in Figure 11.3 and it must imply mutual gains from trade for the two countries.

Now let's turn to the case of free entry and exit. Again, let's just think about one market and think of trade between two countries as doubling the size of the market. This allows us to simplify the notation.

Assume that there are L individuals (which can again be normalized to one in autarky as we did above). We will also now need to keep track of the difference between output per firm and aggregate X output. Summing over the number of firms n, per capita consumption is

$$\sum_i^n X_i/L \tag{11.18}$$

The number of firms, n, is now *endogenous*.

Demand and profits for the ith firm are then given by

$$p = \alpha - \beta\left[\sum_j X_j/L\right] \tag{11.19}$$

$$\Pi_i = p_i X_i - cX_i - f = \left[\alpha - \beta\left[\sum_j X_j/L\right]\right]X_i - cX_i - f \tag{11.20}$$

The firm's first-order condition is the derivative of (11.20) with respect to X_i, holding all of the other firm's outputs constant. Marginal revenue minus marginal cost for firm i is given by:

$$MR - MC = \alpha - 2(\beta/L)X_i - (\beta/L)\sum_{j\neq i} X_j - c = 0 \tag{11.21}$$

Now we again know that there will be symmetry in equilibrium: the output of any active firm will be the same as that of every other firm. *X* will denote the output of an *individual* (not aggregate) "representative firm", and *n* the number of firms. All firms that are active in equilibrium will produce the same amount.

$$MR - MC = \alpha - (\beta/L)(n + 1)X - c = 0 \tag{11.22}$$

The second equation we need for equilibrium is the free-entry condition that will be associated with the number of firms. Then the zero profit condition is that the profits of the representative firm are exactly zero.

$$\alpha X - (\beta/L)nX^2 - cX - f = 0 \tag{11.23}$$

Multiply (11.22) through by *X*.

$$\alpha X - (\beta/L)(n + 1)X^2 - cX = 0 \tag{11.24}$$

We then have two equations (11.23 and 11.24) in two unknowns, *n* and *X*. Solving these two equations in two unknowns we get

$$X = \left[\frac{Lf}{\beta}\right]^{1/2} \tag{11.25}$$

Output per firm increases with the size of the economy (*L*). But with price equal to average cost, this must also mean that the equilibrium price of *X* falls. Finally, putting (11.25) into (11.22), we can solve for n

$$n = (\alpha - c)\left[\frac{L}{\beta f}\right]^{1/2} - 1 \tag{11.26}$$

The number of firms increases with the square root of the size of the world economy. A restriction that the economies are sufficiently big such that $n > 1$ in autarky is sufficient to imply that when *L* doubles the number of firms less than doubles. This in turn means that each country individually must have exit relative to the number of firms in autarky. It is much the same as our numerical example

in section 11.2 above: each country has some exit, yet the integrated world economy has more firms in total than the individual countries did in autarky. The situation is exactly like that shown in Figure 11.4.

In summary then, the combination of increasing returns to scale with imperfect competition means that there are gains from trade even for identical economies under the assumptions used here, regardless of whether or not there is a fixed number of firms (same in each country) or there is free entry and exit in response to the opening of trade. Note finally from Figures 11.3 and 11.4 that these gains from trade for identical economies are not associated with any net trade between the economies. Both countries are at point T in free trade. Yet (given our assumption of costless trade) there could be a lot of two-way gross trade flows, with some consumers arbitrarily buying *X* from the producer in the other country. This is referred to as intra-industry trade and the there is a high volume of such trade among the high-income developed countries.

11.4 Special case II: Cobb-Douglas preferences (can be skipped without loss of continuity)

One of the big limitations of the quasi-linear case is that it imposes the assumption that there is a zero income elasticity of demand for *X*. But surely a lot of the manufacturing and service industries that are characterized by increasing returns to scale and imperfect competition are producing goods with high income elasticities of demand. Let *X* again denote the output of an *individual X firm* and assume that there is a fixed number *n* of such firms. *Y* denotes the total output of *Y* as before. Suppose that preferences are Cobb-Douglas and given by

$$U = (nX)^{\alpha} Y^{1-\alpha} \qquad \text{with income } (I) \text{ constraint} \qquad I = \bar{L} + \Pi = pnX + Y \tag{11.27}$$

We treated this exact case earlier in Chapter 2. Continue to let the price of *Y* be numeraire, equal to one. We showed in Chapter 2 that the demand functions are

$$nX = \frac{\alpha I}{p} \qquad Y = (1-\alpha) I \tag{11.28}$$

In this case, if we compute the elasticity of demand for *X*, we will find that $\eta = 1$. This is going to greatly simplify our analysis, though note that we will need to be sure that there is more than one firm producing in each country in equilibrium; that is, the market share of a firm must be less than one in autarky in order for marginal revenue given by (11.7) to be positive. Subject to this restriction, the markup is just the firm's market share, which is just $1/n$. Marginal revenue equals marginal cost is given quite simply by

$$p(1 - 1/n) = c \qquad p = \frac{n}{n-1} c \tag{11.29}$$

So far, so good. But things get messy because income now matters for the demand for *X* and profits are part of income. This is exactly the mess that quasi-linear preferences avoids! Since *X* refer to the output of a single firm and, assuming all firms producing in equilibrium are identical, then *nX* gives total *X* output. Substituting the budget constraint from (11.27) into the demand function (11.28) and writing out the expression for profits, the aggregate demand for *X* is given by:

$$nX = \alpha(\bar{L} + \Pi)/p = \alpha(\bar{L} + n(pX - cX - f))/p \tag{11.30}$$

Rearranging the equation and making use of the second equation in (11.29) to eliminate the endogenous variable *p* gives us

$$(1-\alpha)npX + \alpha n(c)X = (1-\alpha)\frac{n^2}{n-1}(c)X + \alpha n(c)X = \alpha(\bar{L} - nf) \tag{11.31}$$

Dividing through by n(c) and multiplying one term by $(n-1)/(n-1)$, we have

$$\left[(1-\alpha)\frac{n}{n-1} + \alpha\frac{n-1}{n-1}\right]X = \frac{\alpha(\bar{L} - nf)}{n(c)} \tag{11.32}$$

and then an explicit solution for X in the case when *n* is fixed.

$$X = \left[\frac{\alpha(\bar{L} - nf)}{n(c)}\right]\left(\frac{n-1}{n-\alpha}\right) \tag{11.33}$$

Suppose now that we put two identical economies together, each with a fixed number of firms *n*. $\bar{L}$ doubles in (11.33) and so does *n* relative to autarky. Then the first bracketed term on the right-hand side of (11.33) does not change: both the numerator and denominator double. The second bracketed term increases given $\alpha < 1$. For example, let $n = 2$ in autarky and let $\alpha = 0.5$. Then the introduction of trade ($n = 2$ to $n = 4$) increases the second term from 2/3 to 6/7, an increase in output per firm of 29 percent. Once again, this is exactly the situation shown in Figure 11.3. Note from (11.29) that the price of *X* also falls as shown in the Figure 11.3.

Now let us consider the free entry and exit version of the Cobb-Douglas case. The marginal revenue, marginal cost equation is unchanged.

$$p(1 - 1/n) = c \tag{11.34}$$

The free entry or zero-profit equation is given by

$$pX = cX + f \tag{11.35}$$

and, with no profit income, the demand for *X* is given by

$$nX = \alpha\bar{L}/p \tag{11.36}$$

Multiply (11.34) through by *X* and then divide (11.35) by (11.34)

$$\frac{n}{n-1} = 1 + \frac{f}{cX} \quad \Rightarrow \quad \frac{n}{n-1} - \frac{n-1}{n-1} = \frac{f}{cX} \tag{11.37}$$

which gives us output per firm.

$$X = (n-1)\frac{f}{c} \tag{11.38}$$

Multiple both sides of (11.34) by X, and substitute for pX from (11.36).

$$p(1-\frac{1}{n})X = p\left(\frac{n-1}{n}\right)X = \left(\frac{n-1}{n}\right)\frac{\alpha\bar{L}}{n} = cX \tag{11.39}$$

Now substitute the expression for X in (11.38) to give us a solution for n, the endogenous number of firms.

$$\left(\frac{n-1}{n}\right)\frac{\alpha\bar{L}}{n} = c(n-1)\frac{f}{c} \qquad n^2 = \frac{\alpha\bar{L}}{f} \tag{11.40}$$

Take the square root of the right-hand equation to get n and then substitute this into (11.39) to get X.

$$n = \sqrt{\frac{\alpha\bar{L}}{f}} \qquad X = \left[\sqrt{\frac{\alpha\bar{L}}{f}} - 1\right]\frac{f}{c} \tag{11.41}$$

With free entry and exit, both the number of firms and the output per firm increase with the square root of the size of the market. Once again, thinking of trade as a doubling of market size, this means that there is some exit in each country individually with the opening of trade. The effects of trade for two identical economies are exactly those shown in Figure 11.4.

11.5 Summary

To this point in the book, we have concentrated on determinants of trade that involve differences between countries. Gains from trade involve exploiting these differences, such as producing and exporting according to comparative advantage. Now we come to a situation in which there can exist gains from trade even between identical economies and indeed we concentrate here on precisely this case. We did touch on the role of increasing returns to scale in the previous chapter, but now we turn to a fuller analysis and assume that scale economies occur at the level of the individual firm, sometimes termed "internal" economies of scale to distinguish them from the external economies of the previous chapter.

The situation quickly becomes complicated because internal or firm-level economies of scale are not compatible with perfect competition and hence the simple tools needed to analyze competitive models need to be extended. Increasing returns to scale and imperfect competition are inherently related to one another. This require introducing new methods and new tools to take into consideration firms with market power and the strategic interaction between such firms. In doing so, we simplified our theoretical economies in other ways, in particular on the factor market side, assuming only a single factor as in the Ricardian model of trade.

In this chapter, we focus on an industry in which firms produce a homogeneous good or alternatively the goods of the different firms are perfect substitutes. Firms compete according to the Cournot-Nash model, choosing quantities that are best responses to the outputs of other firms. Autarky equilibrium for a country is distorted, and we showed that this distortion looks a lot like a production tax as modeled in the previous chapter. Price exceeds marginal cost in market equilibrium, and too little is produced at too high a price to maximize social welfare.

We then open up two identical economies to trade and show that this generates a pro-competitive effect. In technical terms, firms perceive demand as more elastic and hence expand outputs in response to the opening of trade. But when all firms do this, industry output expands, the price and markup falls, average cost falls (or productivity improves) and social welfare increases.

We considered two versions of the model: one in which the number of firms is fixed before and after the opening of trade and one in which there is free entry and exit of firms in response to trade. The effects just mentioned in the previous paragraph are present in both cases, but the free entry/exit case is particularly interesting in that it may be possible for the identical economies to end up consuming more of both goods, the competitive good as well as the increasing-returns good. The important lesson is that trade does not offer welfare gains just based on differences between countries, it also offers gains to very similar countries in terms of more efficient production, lower prices, and high consumption quantities.

REFERENCES

Brander, James A. (1981), "Intra-industry trade in identical commodities", *Journal of International Economics* 11, 1-14.

Helpman, Elhanan. and Paul A. Krugman (1985). *Market Structure and Foreign Trade*, Cambridge: MIT press.

Horstmann, Ignatius J. and James R. Markusen (1986), "Up the average cost curve: inefficient entry and the new protectionism", *Journal of International Economics* 20, 225-228.

Markusen, James R. (1981), "Trade and the gains from trade with imperfect competition", *Journal of International Economics* 11, 531-551.

Melitz, Mark J. and Gianmarco I.P. Ottaviano (2008), "Market size, trade, and productivity", *Review of Economic Studies* 75, 295-298.

Venables, Anthony J. (1985), "Trade and trade-policy with imperfect competition - the case of indentical products and free entry", *Journal of International Economics* 19, 1-19.

Endnotes

1. The assumption that firms enter or exit until profits are exactly zero means that we are allowing the number of firms to be a continuous variable and not restricted to integer values. This can be puzzling at first, but it is a common trick in economic theory. It allows the problem to be formulated as equations rather than as a difficult (or impossible) to solve integer programming problem.

Figure 11.1

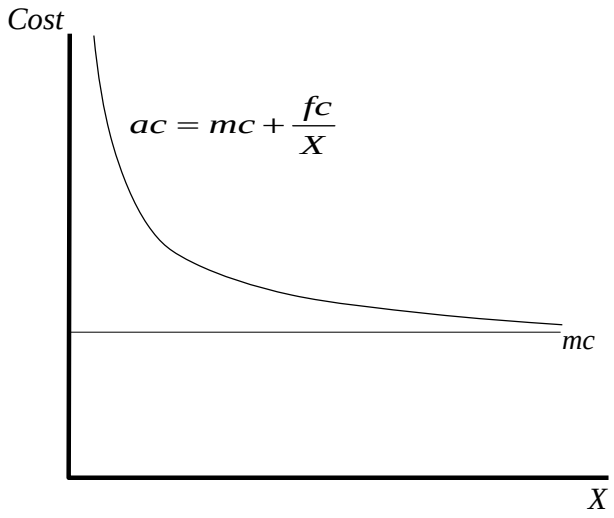


Figure 11.2

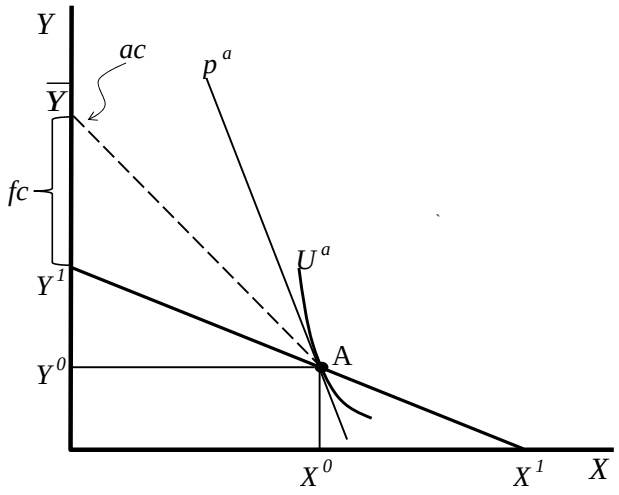


Figure 11.3

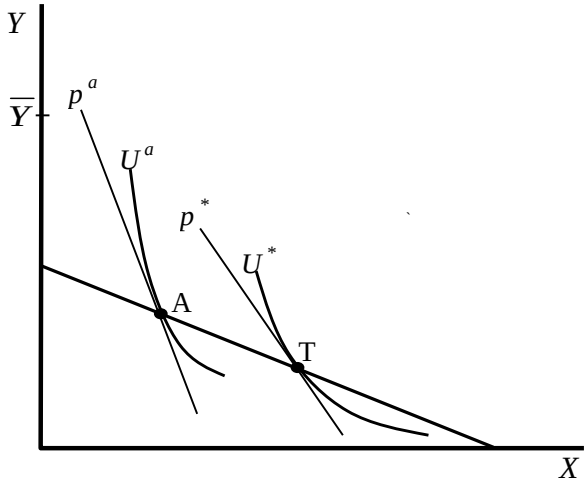
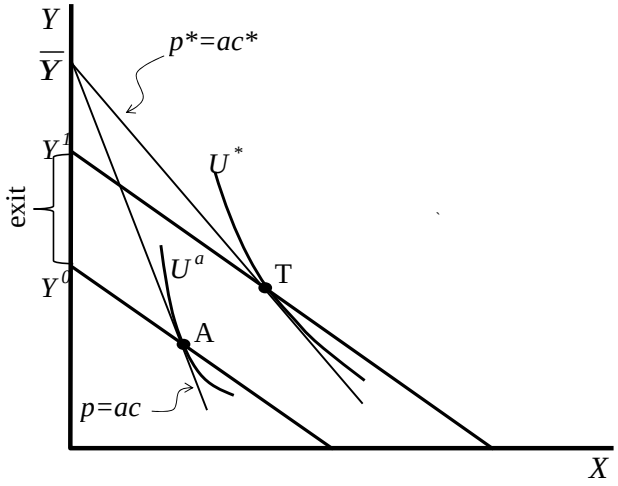


Figure 11.4



Chapter 12

INCREASING RETURNS AND IMPERFECT COMPETITION II: MONOPOLISTIC COMPETITION

12.1 Trade and gains from trade through increased product diversity

The previous chapter introduced economies of scale and imperfect competition as a determinant of trade and source of gains from trade. We concentrated on a homogeneous-good industry (firms produce identical products) and the idea that scale economies limit the number of firms in the market. This limitation leads to imperfect competition in equilibrium, with products marked up above marginal costs. Firms interact strategically with one another and consequently the effective enlargement of the market following the opening of trade leads to pro-competitive gains from trade. Firms move down their average cost curves and consumers benefit from lower prices for the same goods.

In this Chapter, we are again going to look at increasing return to scale and imperfect competition but from a different point of view. We are going to assume that firms produce differentiated products but that the market can support a relatively large number of firms such that there is minimal strategic interaction among the firms. Instead of trade resulting in pro-competitive gains, the same products at lower prices, trade will result in a greater variety of products at the same prices and this will raise consumers' welfare.

These assumptions, that products are different but there are many firms and minimal strategic interaction, are typically referred to as monopolistic competition. Monopoly refers to the fact that each firm produces a somewhat different product and hence will have influence over its market price even if there are literally hundreds of firms. Competition refers to the fact that there are sufficiently many firms such that there is little strategic interdependence among them.

The following preferences (utility function) are typically used to introduce monopolistic competition and are generally referred to as Dixit-Stiglitz (1977) preferences. This approach is also sometimes referred to as "love of variety" for a reason that should become clear. Let X_i refer to one good in a set of n goods. Ignoring any other sectors for the moment, utility is given by

$$U = \left[\sum_i^n X_i^{\alpha}\right]^{\frac{1}{\alpha}} \qquad 0 < \alpha < 1 \qquad \frac{1}{\alpha} > 1 \qquad \sigma = \frac{1}{1-\alpha} > 1 \tag{12.1}$$

Many of you will recognize this as a CES function that we introduced earlier in the book, where σ gives the elasticity of substitution among the varieties. It is in fact a special case in which all goods carry the same weighting in producing utility, and in which the elasticity of substitution is restricted to values greater than one. The latter assumption means that indifference curves intersect (all) the axes and so positive utility can be derived from a subset of goods. Indeed, the whole point of this approach is that only a subset of goods gets produced in equilibrium and that the number of goods available is endogenous.

The individual goods in (12.1) are said to be symmetric but imperfect substitutes. Symmetric means that they all have the same weight in producing utility as just noted and hence a consumer is indifferent between one apple and one orange. However, they are also imperfect substitutes meaning that variety is valuable: a consumer would rather have one apple and one orange than either two apples or two oranges. To see this last point, assume that each good that is produced is produced in the same amount, so the summation in (12.1) is just n, the number of goods, times X, which we will term the quantity of a "representative good".

$$U = \left[nX^{\alpha}\right]^{\frac{1}{\alpha}} = n^{\frac{1}{\alpha}}X \tag{12.2}$$

We see from (12.2) that utility has constant returns in the amount of each good consumer: double each X and utility doubles. But we also see that utility has increasing returns to scale in the range of product variety (henceforth just "variety"). Let n^0 and X^0 denote the number of goods and the representative quantity produced initially, and U^0 the initial level of utility. Then suppose that the number of goods doubles to $2n^0$ but the quantity consumed of each falls in half to $X^0/2$. Utility is then given by

$$U = (2n^0)^{\frac{1}{\alpha}}(X^0/2) = 2^{\frac{1}{\alpha}-1}(n^0)^{\frac{1}{\alpha}}X^0 = 2^{\frac{1-\alpha}{\alpha}}\left[(n^0)^{\frac{1}{\alpha}}X^0\right] = 2^{\frac{1-\alpha}{\alpha}}U^0 > U^0 \tag{12.3}$$

Equation (12.3) shown that welfare improves when the consumer has half as much of each of twice as many goods. Hence the term "love of variety".

Some intuition is provided in Figures 12.1 and 12.2, where we assume that each of two varieties are produced with increasing returns: a fixed cost and a constant marginal cost as in the previous chapter. Consider first Figure 12.1, where the production frontier is $\bar{X}_2X_2'X_1'\bar{X}_1$. It may be that in autarky, a country may wish have variety even though it is expensive in terms of having to pay the fixed costs for both goods, and hence prefers the autarky outcome shown as $X^a = D^a$ in Figure 12.1. Trade with the second identical country can then allow each country to specialize in one of the goods, trading half of its output for half of the output of the other country's good. Then both countries can consume at point D* in Figure 12.1.

Figure 12.1

On the other hand, the high fixed costs and the sacrifice of scale economies may mean that it is better to produce and consume just one good in autarky, which is the situation shown in Figure 12.2. Here the country achieves utility level U^a in autarky by producing either X_1 or X_2 but not both. This is a higher level of welfare than producing both goods: the added variety is not worth the sacrifice of quantity. Now let the two identical countries trade, with each specializing in one of the goods. Now they can trade to the common consumption point D^* in Figure 12.2, achieving a gain from trade.

Figure 12.2

Note the difference in the source of gains from trade between Figure 12.1 and 12.2. In the former case, the consumer gets more of the same goods, the source of gains in the previous Chapter. In the case of Figure 12.2, the consumer actually gets less quantity of a given good, but enjoys more variety. This is in fact exactly the outcome explored in equation (12.3). It is crucial to note in Figure 12.2 that trade does not result in increased output of any good that is produced initially. No firm moves down its average cost curve, and there is no increase in firm scale. Nevertheless, it is indeed scale economies that are responsible for the welfare gains since scale economies limit the number of goods produced initially.

To press this last point a little further, note that if there were no fixed costs and all goods were produced with constant returns to scale, the optimum under Dixit-Stiglitz preferences would be to have infinitely many goods produced in infinitely small quantities. This is just a logical extension of the argument behind equation (12.2). Increasing returns to scale make diversity costly and hence limit the range of goods in equilibrium.

Figures 12.1 and 12.2 illustrate what can happen, but they by no means prove that this is what will happen in a market environment characterized by imperfect competition.

12.2 A more formal approach to Dixit-Stiglitz and love of variety

We will assume that there are two sectors: sector *X* is composited of firms producing differentiated goods as above, and sector *Y* produces a homogeneous good with constant returns to scale. We will assume a very simple factor market structure much the same as in the previous chapter. There is only one factor of production which we will call labor, and we will use this as numeraire assigning a value of one to the wage rate, $w = 1$. We will assume that the consumer has Cobb-Douglas preferences between Y and X, and CES preferences among the *X* varieties. Much of what follows is based in Krugman (1979, 1981) and Helpman and Krugman (1985).

Total income is given by *L* when the wage is chosen as numeraire. We are also going to assume that all potential *X* varieties have the same cost function. This is a common assumption that, when combined with symmetry in demand, gives us the result that any good that is produced is produced in the same amount and sells for the same price. Henceforth *X* and p_x will denote the price of a representative good which are the same for all goods actually produced. The utility function and the budget constraint for the economy are given by:

$$U = \left[\sum_i X_i^{\alpha}\right]^{\frac{\beta}{\alpha}} Y^{1-\beta} \qquad \sigma = \frac{1}{1-\alpha} \qquad L = np_xX + p_yY \tag{12.4}$$

If you solve the optimization problem, the consumer's demands for *X* varieties and *Y* are

$$Y = (1-\beta)\frac{L}{p_y} \qquad X_i = p_{xi}^{-\sigma}\left[\sum_i p_{xi}^{1-\sigma}\right]^{-1}\beta L \qquad nX = \beta\frac{L}{p_x} \tag{12.5}$$

The demand response for a given variety in response to a change in its own price is a bit complex, since the variety's own price appears both as the first term on the right-hand side of the second equation of (12.5) but also appears in the summation term inside the square brackets. Thus the derivative of the demand for *X* with respect to its own prices must be found by using the differentiation of a product rule. However, it can be shown (though we will not do so here) that the effect of a change in a firm's price on the summation term in square brackets become extremely small as the number of varieties (firms) *n* becomes large. As a consequence, most work in this area assumes that an individual firm is too small to affect the summation term in (12.5), an assumption known as "large-group monopolistic competition. Assuming that the term in brackets in (12.5) is viewed as a constant by an individual firm, the price elasticity of demand for an individual good is given simply by σ, the elasticity of substitution among the *X* goods. Referring back to our derivation of marginal revenue in the previous chapter, the markup takes on the very simple formula 1/σ. The elasticity of demand and marginal revenue are then given as follows.

$$-\frac{p_x}{X}\frac{\partial X}{\partial p_x} = \sigma \qquad mr_x = p_x(1-1/\sigma) = mc_x \tag{12.6}$$

Turning to production, marginal cost for *Y*, marginal cost for *X*, and fixed costs of an *X* variety are denoted by mc_y, mc_x, and fc_x , respectively. The full general equilibrium model for a single economy is given by a set of inequalities with associated variables as described back in Chapter 4. First, there is are pricing equations for the *Y* industry and for each *X* variety. Second, there is a zero profit condition for each *X* variety, which is typically written as markup revenues equal fixed costs instead of the longer equation for revenues equal total costs. It is useful to think of fixed costs as a produced good, such as factor and equipment, and hence there is a pricing equation for factories (fixed costs). These three pricing inequalities are given as follows:

Inequality	Definition	Complementary Variable	
$p_y \le mc_y$	pricing for Y	Y	(12.7)
$p_x(1-1/\sigma) \le mc_x$	pricing for X	X	(12.8)
$(p_x/\sigma)X \le fc_x$	pricing for n (free entry)	n	(12.9)

Then there are three market-clearing conditions, which require that supply equal demand (strictly speaking supply is greater than or equal to demand). There is demand and supply for Y, for total X production, and for labor. These equations follow from (12.5) and are as follows.

$(1-\beta)L/p_y \le Y$	demand/supply Y	p_y	(12.10)
$\beta L/p_x \le nX$	demand/supply X varieties	p_x	(12.11)
$(mc_y)Y + n(mc_x)X + n(fc_x) = L$	demand/supply L	w	(12.12)

This model can be solved analytically due to the powerful advantages of the large-group monopolistic-competition assumption. Equations (12.8) and (12.9) can be solved for both X and p_x. Then these solution values can be used in (12.11) to get n. The solution values are:

$$X = (\sigma-1)\frac{fc_x}{mc_x} \qquad n = \frac{\beta L}{\sigma fc_x} \qquad nX = \frac{(\sigma-1)}{\sigma}\frac{\beta L}{mc_x} \tag{12.13}$$

Note from the first equation of (12.13) that the output of any good that is produced is a constant and that from the second equation that any expansion in the economy creates a proportional increases in variety n. Let X/L, the consumption of a representative variety per capita, be given by C. Then note from the last equation of (12.13) that nC is a constant:

$$nC = n\frac{X}{L} = \frac{(\sigma-1)}{\sigma}\frac{\beta}{mc_x} \tag{12.14}$$

Figure 12.3 plots n against C, and equation (12.14) is shown as a negatively-sloped curve in this Figure. Next, note from the second equation in (12.13) that n depends only on L and fixed parameters. Thus we show a second curve which is just a horizontal line in Figure 12.3 which gives the fixed value of n for a given value of L. The intersection of these two curves gives the number of varieties and consumption of a representative variety per capita. For a single economy, the outcome could be shown by variety level n^0 and variety consumption per capita by C^0.

Figure 12.3

Now repeat our usual experiment: put two identical countries together in free trade. This is represented by simply letting L double. The nC curve in Figure 12.3 does not shift as shown by (12.14), but the n curve shift up in proportion to L, doubling in value. If n doubles, then from (12.14) C must be cut in half. The new values in the open economy are C^1 and n^1 in Figure 12.3. Note that this is exactly what is analyzed in (12.3) and suggested in Figure 12.2 above. Equations (12.7) and (12.10) will show

that consumption per capita of *Y* remains unchanged after trade, and hence welfare increases in each country due to the variety effect.

12.3 Monopolistic competition in specialized intermediate inputs (basics idea, then algebra after (12.20) can be skipped without loss of continuity)

The basic idea behind love of variety has also been applied to intermediate inputs, starting with Ethier (1982). This has in turn been applied in a number of different contexts, including endogenous growth models, technology transfer through trade, and trade in producer services. Here we will analyze a much simplified Ethier model, in particular as extended to consider trade in final goods only versus trade in intermediates in Markusen (1989).

Suppose that there are two final consumption goods, *X* and *Y*, which are homogeneous and produced with constant returns to scale by competitive firms. Utility of the representative consumer is given by

$$U = U(X, Y) \tag{12.15}$$

There is a factor of production labor, *L*, which is in fixed supply. In addition, we assume a sector-specific factor *K* in the *Y* sector. The purpose of *K* is to generate a concave (bowed out) production frontier as we will discuss shortly. Good *X* is assumed to be costlessly assembled from differentiated or specialized intermediate goods S_i in a Dixit-Stiglitz fashion. The two production functions are given as follows, where σ is the elasticity of substitution as derived earlier in the book.

$$Y = G(L_y, \bar{K}) \qquad X = \left[\sum_i^n S_i^{\beta}\right]^{1/\beta} \qquad \sigma = \frac{1}{1-\beta} \tag{12.16}$$

Each S_i is produced with increasing returns to scale, consisting of the constant marginal cost and fixed-cost technology that we have now used many times. To reduce notation, one unit of *S* requires a single unit of labor. Labor requirements in *S* goods and the total labor supply constraint are given as follows, where *n* is the (endogenous) number of intermediates.

$$L_{xi} = wS_i + wF \qquad \bar{L} = L_y + nL_{xi} \tag{12.17}$$

Since each *S* enters (12.16) symmetrically and each has an identical technology, we can anticipate the result that any *S* that is produced is produced in the same amount and sells for the same price as any other *S*. Let superscript "a" denote a situation in which only the final *X* and *Y* goods can be traded and n^a the number of goods and S^a the amount of each *S* good in the "a" equilibrium. The *X* technology reduces to

$$X^a = \left[\sum_i^{n^a} (S_i^a)^{\beta}\right]^{1/\beta} = \left(n^a\right)^{1/\beta} S^a \tag{12.18}$$

Now again do our standard experiment where we put two identical economies together in trade. In order to illustrate the main idea, hold the amount of each *S* good produced constant and assume the number produced in each country constant. The number of intermediate goods used in each country in free trade is double the number in autarky *(n* = na)*, with the total output of each shared evenly between the two countries. Output of *X* in each country is now given by

$$X^* = \left[\sum_i^{n^*} (S_i^a/2)^\beta\right]^{1/\beta} = \left(2n^a\right)^{1/\beta}(S_i^a/2) \tag{12.19}$$

Simplifying the right-hand side, we can compare *X* output with intermediates trade to output under trade in final goods only given in (12.18).

$$X^* = 2^{\frac{1-\beta}{\beta}}(n^a)^{1/\beta}S_i^a = 2^{\frac{1-\beta}{\beta}}X^a > X^a \tag{12.20}$$

Allowing trade in intermediates increases productivity in *X* production as *X* producers now have access to a larger range of specialized intermediates, a greater division of labor.

Results are shown in Figure 12.4, where we assume that the diminishing returns to the fixed factor *K* in the *Y* sector outweigh the increasing returns to scale in the *X* sector, so the production frontier is concave. The frontier through A gives the frontier when only final goods can be traded. Both countries are identical by assumption, so there are no gains from trade: countries could trade but there is no benefit from doing so. Point A could represent the equilibrium production and consumption point for each country under goods trade only (the non-tangency of the price ratio with the frontier will be treated shortly).

Figure 12.4

If we do allow trade in intermediate goods (which could include specialized services), then the production frontier shifts to the one passing though F in Figure 12.4. Each country exports half of each of its inputs for half of each of the other country's inputs. With free trade in intermediatates, production and consumption for each country could be at a point like F in Figure 12.4.

As noted earlier, this formulation of the monopolistic competition model in a manner reminiscent of Adam Smith's division of labor has been used in a number of contexts including endogneous growth theory and the liberalization of trade in producer services. We now look at the issue of optimality of the market outcomes in this model; this is a somewhat more esoteric issue, and the remainder of the section may be skipped by some readers. Results of this are equally applicable to the more standard final-goods model treated in the previous two sections.

Suppose that the economy faces a fixed price of *X* relative to *Y*, denoted *p*. Then the optimal number of intermediates and the output level of each can be found by maximizing the value of final output of *X* minus input costs (representing the opportunity cost of labor in producing *Y*.

$$Max\ \pi^* = p\left[\sum S_i^\beta\right]^{1/B} - \sum (wS_i + wF) \quad \text{with respect to } n \text{ and } S_i, \text{ for all i} \tag{12.21}$$

The first-order condition with respect to S_i is given by applying the chain rule.

$$\frac{\partial \pi^*}{\partial S_i} = (p/\beta)\left[nS_i^\beta\right]^\alpha \beta S_i^{\beta-1} - w = pn^\alpha - w = 0 \qquad \alpha \equiv (1-\beta)/\beta \tag{12.22}$$

The first-order condition for the number of goods is given by the effect of adding one more *n*.

$$\frac{\partial \pi^*}{\partial n} = (p/\beta)\left[nS_i^\beta\right]^\alpha S_i^\beta - wS_i - wF = (p/\beta)n^\alpha S_i - wS_i - wF = 0 \tag{12.23}$$

If we solve these two equations in two unknowns, we get the optimal output of any intermediate that is produced. Note that w is just the marginal product of labor in the Y sector, G_l. From (12.22), we can also get the optimality condition for a tangency between the price ratio and the marginal rate of transformation along the production frontier.

$$S_i = \left(\frac{\beta}{1-\beta}\right)F \qquad p = \frac{w}{n^{\alpha}} = \frac{G_l}{n^{\alpha}} = MRT \tag{12.24}$$

Now turn to the outcome under a market solution. The price of an individual S is the value of its marginal product in producing X. This is in fact given in the first term on the right-hand side of the first equation in (12.22). Let r denote the price of an individual S.

$$r = (p/\beta)\left[nS_i^{\beta}\right]^{\alpha}\beta S_i^{\beta-1} = qS_i^{\beta-1} \qquad q \equiv p\left[nS_i^{\beta}\right]^{\alpha} \tag{12.25}$$

In the tradition of large-group monopolistic competition discussed above, assume that each individual S producer views q in (12.25) as fixed. Then each S producer maximizes the following expression with resepect to S_i viewing q as fixed.

$$Max\ \pi = (qS_i^{\beta-1})S_i - wS_i - wF \tag{12.26}$$

The first-order condition is given by

$$\frac{\partial \pi}{\partial S_i} = q\beta S_i^{\beta-1} - w = 0 \tag{12.27}$$

Secondly, the free entry condition to determine n is that each S producer's profits are zero.

$$qS_i^{\beta} - wS_i - wF = 0 \tag{12.28}$$

Solving (12.27) and (12.28), we get the market equilibrium amount of any S that is produced. Second, substituting the expression for q in (12.25) back into (12.27), we can get the relationship between the competitive price ratio and the marginal rate of transformation given in (12.24). These are given by

$$S_i = \left(\frac{\beta}{1-\beta}\right)F \qquad p\beta = \frac{w}{n^{\alpha}} = MRT < p \tag{12.29}$$

Comparing the optimum in (12.24) to the market outcome in (12.29) we see that any S that is produced is produced in the optimal amount. We also see that the market outcome involves a distortion between the price ratio and the marginal rate of transformation, much as in the case of a tax, or the external-economies case of Chapter 10, or the oligopoly outcomes in Chapter 12. This is the distortion between the price ratio and the slope of the frontier shown in Figure 12.5.

The conclusion of this exercise is that the market outcome is not first best: it produces the optimal output of any good that is produced, but too few goods are produced. The intuition behind this is essentially an externality argument. (12.28) gives the private profits of entry. But when one firm enters, it increases the productivity of every other firm holding prices constant. This is seen in (12.25): the value of the marginal product of an additional unit of S in producing X (r) is increasing in n, the division of

labor. This effect is not considered by an individual firm in its entry decision and hence there is a positive externality in the *X* sector. Note that when (12.28) holds with equality (the private profits from entry are zero), the "social" marginal product of an additional *S* given in (12.23) is strictly positive. There is thus a close analogy here between the present result and that in Chapter 10 in the section on production externalities.

12.4 The ideal variety approach to product diversity

Gains from trade through product differentiation can be looked at in a second way as well. While consumers may prefer diversity as just noted, consumers themselves may also have different tastes. Consumers may, for example, buy only one automobile each, but they have different views as to what is the "ideal" automobile for their tastes and income level. This approach to product diversity is thus labelled the "ideal variety" approach (Lancaster, 1980). Due to scale economies, no country can afford to produce a unique automobile for each consumer. Germany produces Volkswagens and Mercedes, and France produces Renaults and Peugeots, all of which have somewhat different characteristics from consumers' points of view. Trade in automobiles then occurs between France and Germanysince some Germans prefer Renaults or Peugeots and some Frenchmen prefer Volkswagens or Mercedes.

This situation is shown in Figure 12.5. Suppose that automobiles have only two characteristics: C_1 and C_2 (e.g., size and speed). There is a trade-off between these two characteristics such that if one wants a bigger car he or she must sacrifice some speed if the two are going to cost the same (e.g., Mercedes versus Porsche). Figure 12.5 shows three possible combinations of C_1 and C_2, denoted X_1, X_2, and X_3, corresponding to three different types of cars. Suppose that all three models could be produced at the same average cost for the same volume of production and that at this common cost, the amounts of C_1 and C_2 embodied in the cars are given by points (X_2^0, X_3^0, X_1^0) in Figure 12.5. The straight line through these points represents a sort of iso-cost line at a common scale of production.

Figure 12.5

Now suppose that we have two groups of consumers with distinct tastes (identical within each group). Consumer type-1 has a relative preference for characteristic C_1 and consumer type-2 has a relative preference for characteristic C_2. On the isocost line shown in Figure 12.5, an indifference curve for consumer type-2 is tangent to the isocost line at point X_2^0 and an indfference curve for consumer type-1 is tangent at point X_1^0. X_2 is then referred to as consumer type-2's *ideal variety* and X_1 as consumer type-1's *ideal variety*.

Variety X_3 could be referred to as a *compromise variety*. Note from Figure 12.5 that for X_3 to be equally attractive to our two consumer types, more "stuff" (e.g., stereo, air conditioning) would have to be added to it: to make the consumers indifferent between the compromise variety and their ideal varieties, we would have to offer X_3^1, which lies outside the isocost line. The proportional difference $(X_3^1 - X_3^0)/X_3^0$ in Figure 12.5 is sometimes called the *compensation ratio*: the proportional amount of extra "stuff" needed to make the compromise variety as attractive as the ideal varieties.

If there were constant returns to scale, the problem would be trivial: each consumer would get their ideal variety, no one would buy the compromise variety because it would cost more. However, the problem is far from trivial with increasing returns to scale. The production scale in producing the compromise variety is twice as large as in giving each consumer their ideal variety. Thus, adjusting for the benefits of larger production scale, variety , X_3^1, may in fact cost less than variety X_2^0 or X_1^0.

Suppose that there is only a single factor labor, *L*, with a wage one. We assume, as is typical in

this approach, that each consumer wants only one unit of the good but of course prefers their ideal variety. There are fixed and marginal costs for each variety. mc^0 is the labor needed for one unit of (X_2^0, X_3^0, X_1^0) and mc^1 is the labor need for one unit of X_3^1. Labor requirements are then

$$L_1^0 = L_2^0 = L_3^0 = mc^0 + fc \qquad L_3^1 = mc^1 + fc \qquad mc^0 < mc^1 \tag{12.30}$$

The relevant question is whether or not two ideal varieties (two times the first equation of (12.29)) costs more than two units of the compromise variety. The compromise variety is preferred if

$$2mc_3^1 + fc < 2mc_i^0 + 2fc \quad or \quad 2(mc_3^1 - mc_3^0) < fc \quad i=1,2,3 \tag{12.31}$$

The important point is that the compromise variety means incurring the fixed costs only once versus twice if each group gets its ideal variety.

Suppose that the inequality in (12.30) "marginally" holds, meaning that the left-hand side is less than the right-hand side by a very small amount. Then the compromise variety is preferred. Consider then our now-familiar experiment of putting two identical countries together. To give every consumer one unit, the total world labor requirements for ideal varieties are now twice the marginal costs shown in (12.30) and similarly for the compromise variety: no additional resources are needed for fixed costs. Total requirements (costs) and the condition for the compromise variety to be preferred are now

$$4mc_3^1 + fc < 4mc_i^0 + 2fc \quad or \quad 4(mc_3^1 - mc_3^0) < fc \quad i=1,2,3 \tag{12.32}$$

If (12.30) holds marginally, the inequality in (12.31) will *not* hold. It will be reversed and the ideal varieties will be preferred.

As in the love-of-variety approach, we see here that the ideal-variety approach means that there are gains from trade through increased product variety. The nature of those gains are different here. The idea is that trade allows each consumer type to get a product closer to their ideal variety embodying their ideal characteristic combination. In more general models with many consumer types and goods, the welfare statement in these models is typically that "on average", consumers get varieties closer to their ideal type.

In a world of imperfect competition that accompanies scale economies, we have to be cautious and note that just because something is preferred doesn't mean it is going to happen. These ideal-variety models get quite mathematically complicated, much more so than the simple love-of-variety model. For example, there may be a continuum of consumers who preferences are "located" at different points on a line or a circle. Such models are sometimes referred to as "location" or "address" models of product differentiation, but these embody the same basic ideas: a consumer's location or address on the circle is the consumer's ideal variety (see Helpman, 1981). These models have been perhaps more used in industrial-organization economics. The problem for international trade is that they quickly get very difficult in general-equilibrium settings.

12.5 Some useful algebra for Dixit-Stiglitz (may be skipped without loss of continuity but useful later for analyzing trade costs)

This section lays out the algebra required to solve for the demand function and price index (defined and discussed shortly) for the Dixit-Stiglitz utility function. In order to simplify it a bit, let's assume that there is just a single section producing differentiated *X* varieties. There is a straightforward extension to a two-sector model in which the other sector, *Y*, has constant returns to scale and perfect

competition, and there is Cobb-Douglas substitution between the X goods and good Y. In this situation, the consumer always devotes constant share of income to each sector, so we will start with an allocation of income M_x to the X sector.

Let X_c denote the utility derived from the X varieties; X_c is sometimes referred to as a composite commodity. X_c does have a price associated with it. This is a price index, the minimum expenditure necessary to buy one unit of composite good X_c. We will denote this price or rather price index as e_x. The value of X_c (the utility from X consumption) is defined as follows.

$$X_c = \left[\sum_i^n X_i^{\beta}\right]^{\frac{1}{\beta}} \qquad \sigma = \frac{1}{1-\beta}, \quad \beta = \frac{\sigma-1}{\sigma}$$

The consumer maximizes utility subject to a budget constraint. The maximization problem is written as a Lagrangean function and the first-order condition for an arbitrary good X_i are given as follows.

$$\max X_c = \left[\sum X_i^{\beta}\right]^{\frac{1}{\beta}} + \lambda(M_x - \sum p_i X_i) \Rightarrow \frac{1}{\beta}\left[\sum X_i^{\beta}\right]^{\frac{1}{\beta}-1} \beta X_i^{\beta-1} - \lambda p_i = 0 \qquad (12.33)$$

Divide the first-order condition in (12.33) for an arbitrary good i by the corresponding condition for good j.

$$\left[\frac{X_i}{X_j}\right]^{\beta-1} = \frac{p_i}{p_j} \qquad \frac{X_i}{X_j} = \left[\frac{p_i}{p_j}\right]^{\frac{1}{\beta-1}} = \left[\frac{p_i}{p_j}\right]^{-\sigma} \qquad \textit{since} \quad \sigma = \frac{1}{1-\beta} \qquad (12.34)$$

Now perform several steps. (1) write the second equation of (12.34) with X_j on the left. (2) multiply both sides of this equation by p_j. (3) sum this equation over all goods j. These three steps are

$$X_j = \left[\frac{p_i}{p_j}\right]^{\sigma} X_i \qquad p_j X_j = p_j p_j^{-\sigma} p_i^{\sigma} X_i \qquad \sum p_j X_j = M_x = \left[\sum p_j^{1-\sigma}\right] p_i^{\sigma} X_i \qquad (12.35)$$

Inverting this last equation, we have the demand for an individual variety i:

$$X_i = p_i^{-\sigma}\left[\sum p_j^{1-\sigma}\right]^{-1} M_x \qquad \sigma = \frac{1}{1-\beta}, \quad \beta = \frac{\sigma-1}{\sigma} \qquad (12.36)$$

Now we can use X_i to construct X_c and then solve for e_x, the price index, noting the relationship between α and σ. First, raise (12.36) to the power β.

$$X_i^{\beta} = X_i^{\frac{\sigma-1}{\sigma}} = p_i^{1-\sigma}\left[\sum p_j^{1-\sigma}\right]^{\frac{1-\sigma}{\sigma}} M_x^{\beta} \qquad (12.37)$$

Now sum over all of the i varieties of X.

$$\sum X_i^{\beta} = \left[\sum p_i^{1-\sigma}\right]\left[\sum p_j^{1-\sigma}\right]^{\frac{1-\sigma}{\sigma}} M_x^{\beta} = \left[\sum p_j^{1-\sigma}\right]^{\frac{1}{\sigma}} M_x^{\beta} \tag{12.38}$$

Now raise both sides of this equation to the power $1/\beta$ to get the composite commodity demand for X_c.

$$X_c = \left[\sum X_i^{\beta}\right]^{\frac{1}{\beta}} = \left[\sum X_i^{\beta}\right]^{\frac{\sigma}{\sigma-1}} = \left[\sum p_j^{1-\sigma}\right]^{\frac{1}{\sigma-1}} M_x = M_x/e_x \tag{12.39}$$

The demand for X_c must be the expenditure M_x on X_c, divided by the price index e_x. Examining the last equation of (12.39), this means that the price index is given by:

$$e_x = \left[\sum p_j^{1-\sigma}\right]^{\frac{1}{1-\sigma}} \tag{12.40}$$

Again, the price index e_x, also called the expenditure function, is the minimum cost or expenditure necessary to buy one unit of the composite commodity X_c. If all of the X varieties sold for the same price, the expenditure function (price index) in (12.40) simplifies as follows.

$$e_x = n^{\frac{1}{1-\sigma}} p \tag{12.41}$$

Note that the price index is homogeneous of degree one in the prices of the individual goods: if we double all prices we double to cost of buying one unit of X_c. However note that the price index is decreasing in the number of varieties available ($1 - \sigma < 0$). The is another way of thinking about the love-of-variety effect. The same utility derived from two apples or two oranges might be derived from (for example) 0.8 apples and 0.9 oranges. When more variety is available, the consumer can achieve the same utility (same value of X_c) by actually reducing total expenditure.

Finally, having derived e, we can then use equation (12.40) in (12.36) to get the demand for an individual variety.

$$X_i \equiv p_i^{-\sigma} e_x^{\sigma-1} M_x \qquad \textit{since} \qquad e_x^{\sigma-1} = \left[\sum p_j^{1-\sigma}\right]^{-1} \tag{12.42}$$

We now move onto a chapter introducing the existence of trade costs. The price index in (12.40) and the demand function in (12.42) will prove very useful in discussing trade costs in the context of monopolistic-competition models in the next chapter.

12.6 Summary

This is the second of two chapters on trade with increasing returns to scale and imperfect competition. The first (Chapter 11) focused on a pure case in which firms produced identical goods but

scale economies limited the number of firms such that individual firms took into account their strategic interactions with other firms in their output decisions. A principal result is that the larger economy created through trade offer welfare gains through pro-competitive and production-scale effects. This chapter focused also on a pure case in which firms produce somewhat different goods and there are a large number of firms such that each views the decision of rival firms as exogenous. A principal result is that the larger economy created through trade offers welfare gains through increased product or intermediate-good diversity: more products at the same prices rather than the same products at lower prices in Chapter 11. Of course, the two approaches can be combined, but the analysis becomes more difficult and is left to more advanced treatments (e.g., Melitz and Ottaviano (2008)) and indeed the ideal variety approach does involve variable markups and pro-competitive effects.

The two main approaches to product diversity are considered. The "love of variety" approach assume an endogenous set of symmetric but imperfectly substitutable products. Consumers are rewarded by a more diverse consumption bundle through trade. The "ideal variety" approach works rather differently. Consumers are assumed to differ in their views as to the ideal product, a product being a bundle of characteristics. Often in this approach, the consumer is assumed to buy just a single unit or nothing. As in the love-of-variety approach, diversity, in this case giving each consumer their ideal product, is costly in the presence of scale economies, so consumers accept compromise varieties in small autarky markets. The rewards to trade are that consumers, on average, get products closer to their ideal varieties.

While much of the literature has focused on final goods, an important variation of these models considers the differentiated products to be intermediate goods used in producing homogeneous final goods. Allowing trade in intermediate goods offers final producers higher productivity by increasing the division of labor. This ideal is a cornerstone of what is referred to as endogenous growth theory (e.g., Romer 1987) and has also been applied to trade in components and in producer services.

REFERENCES

Dixit, Avinash K. and Joseph Stiglitz (1977), “Monopolistic Competition and Optimum Product Diversity”, *American Economic Review* 67, 297-308.

Ethier, Wifred J., (1982), "National and International Returns to Scale in the Modern Theory of International Trade", *American Economic Review* 72, 389-406.

Helpman, Elhanan, (1981). "International Trade in the Presence of Product Differentiation, Economies of Scale and Monopolistic Competition. A Chamberlinian, Heckscher-Ohlin Approach." *Journal of International Economics* 11, 304-340.

Helpman, E. and P. Krugman (1985). *Market Structure and Foreign Trade*, Cambridge: MIT press.

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

Krugman, P. (1981), "Intraindustry Specialization and the Gains from Trade", *Journal of Political Economy* 89, 469-479.

Lancaster, K. (1980). "Intra-Industry Trade Under Perfect Monopolistic Competition." *Journal of International Economics* 10, 151-175.

Markusen, J. R. (1989). Trade in Producer Services and in Other Specialized Intermediate Inputs." *American Economic Review* 79, 85-95.

Melitz, Mark J. and Gianmarco I.P. Ottaviano (2008), “Market size, trade, and productivity”, *Review of Economic Studies* 75, 295-298.

Romer, Paul M. (1987), “Growth Based on Increasing Returns to Scale due to specialization”, *American Economic Review* 77, 56-62.

Figure 12.1

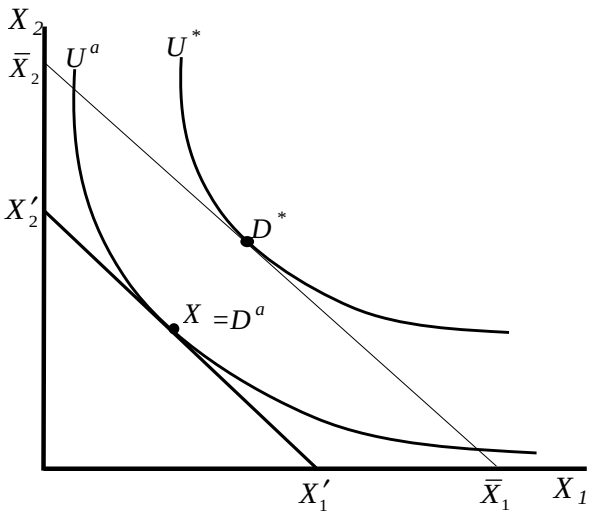


Figure 12.2

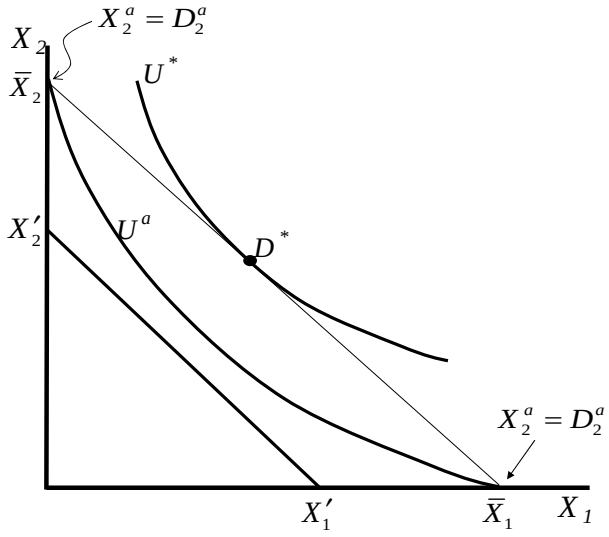


Figure 12.3

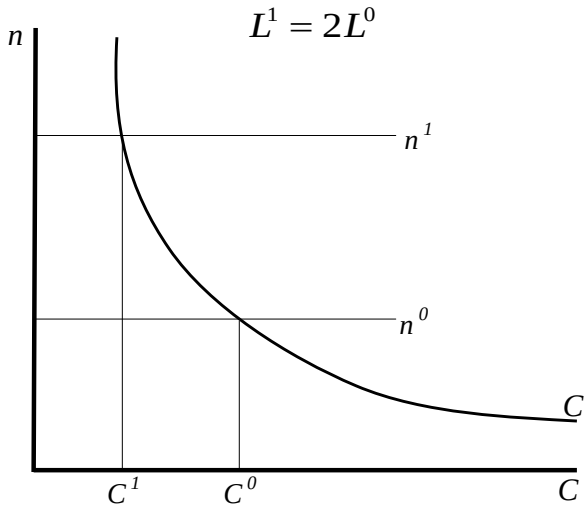


Figure 12.4

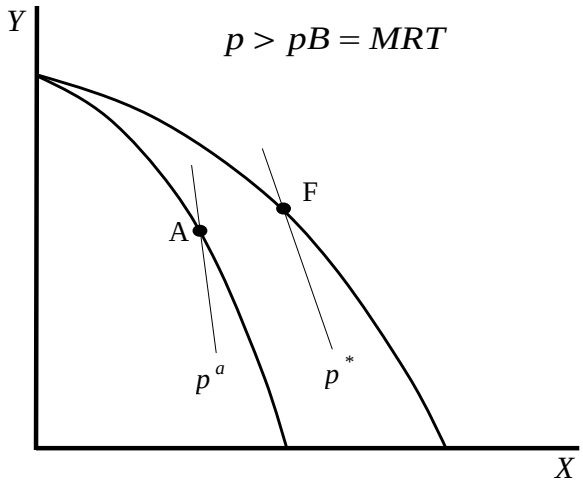
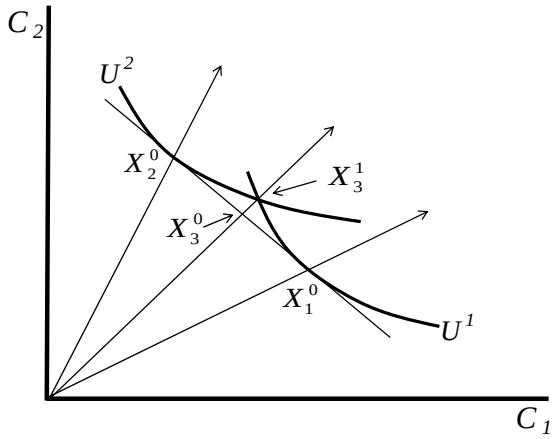


Figure 12.5



Chapter 13

TRADE COSTS, TRADE VOLUMES AND FIRM BEHAVIOR

13.1 Geography and trade costs

Much of what we have done so far presents a rather stark comparison between a country that does not trade at all and one that engages in completely free and costless trade. No one claims that this is a realistic comparison; rather, it is done for analytical simplicity which, in turn, permits a clear and simple presentation. The purpose of this chapter is to introduce costs of trade, which we can think of as shipping costs (it should also include time costs but this requires an explicitly dynamic treatment). Costs imposed by governments such as tariffs and quotas are quite different and will be treated in a later chapter.

Trade costs have traditionally been underplayed in international trade textbooks. We think that the reason is the supposition that trade costs are not terribly interesting: trade costs just put a country somewhere between free (costless) trade and autarky. Thus, there is not very much interesting to say. There is considerable truth in this and indeed we will illustrate this point in the next section. But even in a traditional comparative-advantage model such as the Heckscher-Ohlin model, we will note that trade costs must leave factor prices different between countries and hence there will be incentives for factors to migrate. In non-comparative-advantage models with oligopoly and monopolistic competition, trade costs can do some quite interesting things indeed. In particular, outcomes with positive but moderate trade costs are not "in between" free trade and autarky. Quite a large part of this chapter will be focused on these interesting cases.

Trade costs (unlike tariffs) require the use of real resources: labor, fuel and capital equipment such as ships and planes. In a sophisticated framework, we would want to model these as sectors which produce transportation services. In order to keep things manageable, we will avoid introducing an explicit transport sector in one of two ways. First (in section 3), we will use an essentially partial-equilibrium assumption and just say that there is some cost t to moving a good between countries. So if mc is marginal cost, the cost of supplying the foreign market is $(mc + t)$. In a one-factor model such as a Ricardian model, this would be in units of labor for example.

A second approach that is often taken is to assume what has become known as "iceberg" trade costs. This means that part of a good shipped abroad "melts" during transit: less arrives than is shipped. This is essentially the assumption that the transport technology uses the good itself and nothing else. We understand that the earliest example of this was due to an economist named von Tünen, who told the parable of a farmer taking grain to a market in a horse-drawn wagon. The horse has to be fed some of the grain, so less grain arrives in the market than leaves the farm: the transport cost is in terms of the good itself and there is no need to model a transportation service sector.

Suppose that the transportation cost rate or "melt rate" is τ, and so if X is shipped, $X/(1+\tau)$ arrives "unmelted." It is cumbersome to keep carrying this notation around, so generally trade economists simplify the transport cost by writing it on a gross basis, $t = (1+\tau)$. Then $t = 1$ is free trade, not $t = 0$, and $t = 1.5$ would be, for example, a melt rate of 50 percent. We will follow this convention.

If a firm ships a quantity X, then the amount that arrives in the foreign country is X/t. Suppose that the home price of the good is p and further assume that the firm cannot price discriminate between markets. Then export sales earn the firm the same price as domestic sales (we will explain price discrimination below). The earnings on export sales by the domestic firm must equal what the foreign importer pays. So what is the price in the foreign country? Let p^* be the foreign (importer's) price. Then the revenue balance condition must be:

$$pX = p^*(X/t) \quad => \quad p^* = pt \tag{13.1}$$

The left-hand side of the first equation is the revenues received by the exporter and the right-hand side is the amount paid by the importer. It follows that the price in the importing country is $pt \geq p$ with equality in free trade and a strict inequality when trade costs are strictly positive. In summary, if X units are shipped at an exporter's price of p, then (X/t) units arrive and sell for a price of (pt).

13.2 Trade costs and trade volumes in competitive, comparative-advantage models

In competitive, perfect-competition models where trade is based on comparative advantage, it is not inaccurate to say that trade costs between two countries leave each of them somewhere between autarky and free trade. This is shown more formally in Figure 13.1. Assume that a country has an excess demand curve for X_1 in free trade, given by the dashed line in the figure. p^a is the world price at which the country does not want to either import or export X_1, also equal to its autarky price ratio. Suppose that trade costs are incurred in both inward and outward directions. An example would be the port costs of loading and unloading ships. Then an importer of X_1 would have to pay p^*t,where p^* denotes the world price, and an exporter would only receive p^*/t in revenue

Figure 13.1

The country would be indifferent to importing X_1 if the world price is given by $p^* = (p^a/t)$ in which case the domestic price p is $p = p^*t = p^a$. The country would be indifferent to exporting X_1 if the world price is $p^* = p^a t$, in which case the domestic price p is given by $p = p^*/t = p^a$. This is shown in Figure 13.1, and we see that there is a range of world prices in which the country will not trade. At any world price p^*, the country will export less X_1 (or none at all) or import less X_1 (or none at all).

If we repeat this exercise for the other country, an outcome for two symmetric countries is shown in Figure 13.2. If p^* is the world price ratio p_1/p_2, then country h exporting X_1 faces (effectively) a price ratio p^*/t and country f importing X_1 faces a price ratio p^*t. Their production points are given by X_h and X_f in Figure 13.2, respectively, and their consumption points by D_h and D_f respectively. There is trade and there are positive gains from trade. If you refer back to Figure 8.2, you will see that the outcome is indeed between autarky and free (costless) trade.

Figure 13.2

Nevertheless, there are a couple of things to note. First, note from Figure 13.2 that each country is relatively "specialized" in consuming the same good as it is specialized in producing. This is true if preferences are identical and homogeneous across countries. Country h, for example, is relatively specialized in good X_1 and also relatively specialized in consuming it. This is just a reflection of the price differences inside the two countries: in each country, the export good is relatively cheap and the import

good is relatively expensive. Countries that produce relatively more food will consume relatively more food, even if preferences are identical.

This is sometimes referred to as a "home-market effect" or "home bias" as we will discuss below. However, there is a subtlety if you are interested. Consumption specialization refers here to quantities, but home-market effect is often used to refer to value or expenditure shares in consumption (price times quantity divided by income). In each country, the low quantity good is also the high-priced good so it does not follow that its expenditure share of consumption is lower. In fact, with Cobb-Douglas preferences, the share spent on each good is constant and so the two countries in Figure 13.2 would be observed to spend the same share of income on each good (see equation (3.4)). We won't comment more on this issue here.

The second interesting feature of Figure 13.2 relates to the Stolper-Samuelson theorem. In Chapter 8, we demonstrated that in autarky, each country has a relatively high price for its scarce factor. Under special circumstances, completely costless trade brings the price of each factor into equality across countries (the factor-price-equalization theorem). Consistent with the notion that costly trade is between autarky and costless trade in competitive, constant-returns models, the price of each factor will converge across countries but will not equalize in Figure 13.2. Thus each country will still have a high price for its scarce factor. This will have important implications for incentives for factors to migrate when possible, as we will see in a later chapter.

13.3 Trade costs, price discrimination, and trade volumes in oligopoly models

As hinted above, trade costs often create outcomes in models of imperfect competition and increasing returns to scale that are not "between" autarky and costless trade. We turn to some of these effects in this and the next few sections. In the current section, we use the simple oligopoly model of Chapter 11: linear demand, constant marginal costs, and firms produce identical, homogeneous products. We will assume that the two countries are absolutely identical in size, costs and preferences (to exploit symmetry in solving the model), that each country has a single firm in the *X* sector, and that there is no exit or entry of firms.

This model is often used to analyze price discrimination, defined here as the ability of firms to set different prices in different markets. A firm in one country can set one price for domestic sales and another price for export sales. When a firm sets a lower price for export sales than for domestic sales, this is one of many definitions of "dumping" which, in turn, is the subject of many trade disputes. (This is quite a weak definition of dumping: a stronger version is selling below costs, which firms will not do in this simple model.)

The demand for good *X* in market i is linear and depends on the supply of the domestic firm X_{ii} and the supply of the foreign firm X_{ji}.

$$p_i = \alpha - \beta(X_{ii} + X_{ji}) \tag{13.2}$$

Let π_{ij} denote the profits of firm i on it sales in market j. Profits for firm i on its domestic sales are given by

$$\pi_{ii} = p_i X_{ii} - cX_{ii} = [\alpha - \beta(X_{ii} + X_{ji})]X_{ii} - cX_{ii} \tag{13.3}$$

Let t denote a specific trade costs as discussed above (not iceberg costs: here $t = 0$ is costless trade). Profits of firm i on its export sales to j are given by.

$$\pi_{ij} = p_j X_{ij} - (c + t)X_{ij} = [\alpha - \beta(X_{ij} + X_{jj})]X_{ij} - (c + t)X_{ij} \tag{13.4}$$

The firm optimizes with respect to domestic and foreign sales independently given its ability to price discriminate. Firms behave in a Cournot fashion, choosing their optimal sales given the sales of the rival firm. The first-order conditions for profit maximization are given by:

$$\frac{d\pi_{ii}}{dX_{ii}} = \alpha - 2\beta X_{ii} - \beta X_{ji} - c = 0 \tag{13.5}$$

$$\frac{d\pi_{ij}}{dX_{ij}} = \alpha - 2\beta X_{ij} - \beta X_{jj} - c - t = 0 \tag{13.6}$$

Exploiting cost symmetry because the firms and countries are identical, we can solve (13.5) and (13.6) for the Cournot domestic and foreign sales of the firm i by setting $X_{ii} = X_{jj}$ and $X_{ij} = X_{ji}$.

$$X_{ii} = \frac{\alpha - c + t}{3\beta} \qquad X_{ij} = X_{ji} = \frac{\alpha - c - 2t}{3\beta} \tag{13.7}$$

The homogeneous good sells for the same price in the identical countries. Thus, the transport cost is fully absorbed by the exporter, and the export price (received by the exporter) is lower by t than the domestic producer price, hence the term "dumping"(Brander and Krugman 1983). Note from (13.7) that as long at $\alpha - c - 2t > 0$, each firm will serve both markets. Thus we observe the curious outcome of identical goods traveling in both directions between countries. This is often referred to as intra-industry trade, though the term is also widely used for closely related but differentiated goods as in the monopolistic-competition model.

If the results in (13.7) are substituted back into the demand function in (13.2), the domestic price (earned on local sales) and the export price (earned on foreign sales after covering trade costs are given by:

$$p = 2(\alpha - c)/3 + t/3 \qquad (p - t) = 2(\alpha - c)/3 - 2t/3 \tag{13.8}$$

Notice for future reference that trade costs essentially protect a firm in its home market in that the firm can raise its home price, but hurts the firm on its export sales by lowering the export price.

The two equations in (13.8) can be used to determine the net price that the firm in each country receives on its domestic and export sales. These are

$$(p - c) = (\alpha - c + t)/3 \qquad (p - c - t) = (\alpha - c - 2t)/3 \tag{13.9}$$

Using (13.7) plus (13.9), we can then solve for profits of the firm on its domestic and foreign sales.

$$\pi_{ii} = (p - c)X_{ii} = \beta X_{ii}^2 \qquad \pi_{ii} = (p - c - t)X_{ii} = \beta X_{ij}^2 \tag{13.10}$$

Now consider the same utility function discussed in Chapter 10, which gives rise to the linear demand function in the first place. Utility is given by

$$U(X) = \alpha(X_{ii} + X_{ji}) - (\beta/2)(X_{ii} + X_{ji})^2 + Y \tag{13.11}$$

where Y is a competitive good produced with constant returns to scale. Assume a single factor of production L, and assume that one unit of Y production requires one unit of labor, and that L or Y is numeraire with price one. The budget constraint for the economy requires that labor income plus profits equal expenditure on X and Y.

$$L + \pi_{ii} + \pi_{ij} = Y + p_{ii}X_{ii} + p_{ji}X_{ji} \tag{13.12}$$

Consumer surplus is generally defined as the utility derived from X consumption minus the amount that consumers pay for X. This is given by

$$CS = \alpha(X_{ii} + X_{ji}) - (\beta/2)(X_{ii} + X_{ji})^2 - pX_{ii} - pX_{ji} = (\beta/2)(X_{ii} + X_{ji})^2 \tag{13.13}$$

Substitute the budget constraint in (13.12) into (13.11). Utility is the sum of consumer surplus and profits, which is equal to

$$U_i = CS_i + \pi_{ii} + \pi_{ij} = (\beta/2)(X_{ii} + X_{ji})^2 + \beta X_{ii}^2 + \beta X_{ij}^2 \tag{13.14}$$

While (13.14) may look simple, it turns out to be complicated and not only non-linear in trade costs but also non-monotonic (e.g., rising over some range and falling over another range of trade costs). Figure 13.3 presents a simulation over trade costs on the horizontal axis (the results can be proved analytically and do not depend on the parameter values chosen). Total welfare (plotted on the right-hand vertical axis) is the sum of consumer surplus and profits (both plotted on the left-hand vertical axis. The left-hand end of the horizontal axis is free trade and the right-hand end is autarky: trade costs are prohibitive to trade at $t = 2$.

Figure 13.3

As trade costs fall from a prohibitive level (moving to the *left* along the horizontal axis), welfare actually falls over a range between approximately $t = 2.0$ and $t = 1.5$ and then begins climbing. Welfare in free trade is 18 percent higher than in autarky, normalized to a value of one in the Figure. The region of declining welfare is due to the fact that profits fall faster than consumer surplus rises. Interestingly, profits recover with further falls in trade costs when trade costs are low, near the right-hand boundary.

This result can seem puzzling at first, even to trained economists who would guess that welfare should rise with any fall in trade costs. The key is to remember that trade costs require real resources in transport. When trade costs are very high, a fall in these costs generates more trade and consumes more resources than when trade is prohibitively costly. Each firm has an incentive to invade the other firm's market, but when they both do so, resources are wasted in inefficiently cross-hauling the same good between countries. While consumer surplus rises, it is more than outweighed by the fall in profit income. When trade costs become small, further falls in costs generate a fall in total resources devoted to trade, and welfare unambiguously rises.

Finally, note that there is a distributional issue in Figure 13.3, as there so often is in our trade models. Over an intermediate range of trade costs, there is a conflict between profit income and consumer surplus from lower prices. The equity owners of the firm will not like falls in trade costs which expose them to more competition, while the consumers will benefit from the lower prices that this competition engenders.

13.4 Trade costs, inter and intra-industry trade in monopolistic-competition models

Trade costs in monopolistic-competition models also have interesting and complex effects, so much so that a whole literature, sometimes referred to as the "new economic geography", has arisen around the intersection of Dixit-Stiglitz models and trade costs. Let us return now to the iceberg costs discussed in the first section above. For a domestic firm, X_{ij}^{d} is the amount produced in country i and shipped to country j. Similarly, p_{ij} is the export price per unit in country i. Let t ($t \geq 1$) be the ratio of the amount of X exported to the amount that arrives "unmelted". Alternatively $1/t$ is the proportion of a good that "survives" transit (the proportion "unmelted"). If X_{ij} is shipped, the amount received in country j is X_{ij}/t.

Second, we again make the assumption that there is no price discrimination, so the home price of a good for local sales equals its export price. Thus we can use the notation p_i and p_j for the price of all goods *produced* in country i and country j respectively. The revenues received by the exporter are equal to the costs paid by the importer: $p_i X_{ij}$ is the revenue received by the exporter and X_{ij}/t are the number of units arriving in the importing country, so the price per unit in the importing country must be $p_i t$ ($p_i X_{ij} = (p_i t)X_{ij}/t$). Rather than introduce additional notation, we will therefore use X_{ij}/t and $p_i t$ as the quantity and price in country j of a country i variety exported to country j.

As in Chapter 12, we assume a two-level (nested) utility function in which there is Cobb-Douglas substitution between X varieties and Y, and a CES or Dixit-Stiglitz substitution between X varieties. For now, let labor be the only factor of production and one unit of labor produces one unit of Y. The utility function and the budget constraint are then given by

$$U = \left[\sum_i X_i^{\alpha}\right]^{\frac{\beta}{\alpha}} Y^{1-\beta} \qquad \sigma = \frac{1}{1-\alpha} > 1 \qquad L = np_x X + Y \tag{13.15}$$

If you solve the optimization problem, the consumer's demand for an individual home-produced X variety and the price index (cost of purchasing one unit of the composite X good) are given by

$$X_i = p_i^{-\sigma}\left[\sum_i p_i^{1-\sigma}\right]^{-1}\beta L = p_i^{-\sigma} e^{\sigma-1}\beta L \qquad e = \left[\sum_i p_i^{1-\sigma}\right]^{\frac{1}{1-\sigma}} \tag{13.16}$$

The price index for country i is given by:

$$e_i = \left[N_i p_i^{1-\sigma} + N_j (p_j t)^{1-\sigma}\right]^{\frac{1}{1-\sigma}} \tag{13.17}$$

Now separate varieties in country i into home produced and imported varieties, recalling that X_{ji} is the amount shipped from j to i and so the amount received and consumed in country i is Xji/t. Assuming that

X goods are produced in both countries, the demand functions for the various *X* varieties sold in country i are given by:

$$X_{ii} = p_i^{-\sigma} e_i^{\sigma-1} \beta L_i \qquad X_{ji}/t = X_{ij}/t = (p_j t)^{-\sigma} e_i^{\sigma-1} \beta L_i \tag{13.18}$$

where the second equation can also be written as:

$$X_{ji} = X_{ij} = p_j^{-\sigma} t^{1-\sigma} e_i^{\sigma-1} \beta L_i = p_j^{-\sigma} \phi e_i^{\sigma-1} \beta L_i \qquad \phi \equiv t^{1-\sigma} \tag{13.19}$$

The parameter ϕ ("phi") has been dubbed the "phi-ness" (mnemonic for "freeness") of trade: it takes a value of one when trade costs are zero ($t = 1$) and approaches a value of infinity as trade costs go off to infinity (remember $\sigma > 1$ so that the exponent of t in ϕ is negative: $(1 - \sigma) < 0$).

p_i will denote the home price of a representative good produced in country i (all goods produced will have equal home prices). Again assuming that both countries produce *X* goods (not trivial as we shall see), there are two equations: one for marginal revenue equals marginal cost and one for zero profits. The latter is given by markup revenues cover fixed costs as in Chapter 12. These are given by:

$$p_i(1 - 1/\sigma) = mc \qquad (p_i/\sigma)(X_{ii} + X_{ij}) = fc \tag{13.20}$$

There are two identical equations for a firm located in country j. Thus any good that is produced in either country must sell for the same price and be produced in the same quantity as a good in the other country, even if the countries are of different sizes, so $p_i = p_j$. Solve these two equations, eliminating p, to get the price and total output of any representative variety produced.

$$p = \frac{\sigma}{\sigma - 1} mc \qquad X = (X_{ii} + X_{ij}) = (\sigma - 1)\frac{fc}{mc} \tag{13.21}$$

Because the output quantity of a good is identical across countries, we can do the following.

$$X_{ii} + X_{ij} = X_{jj} + X_{ji} \qquad X_{ii} - X_{ji} = X_{jj} - X_{ij} \qquad \frac{X_{ii} - X_{ji}}{X_{jj} - X_{ij}} = 1 \tag{13.22}$$

Exploiting again the result that any good from country i has a home price equal to the home price of a good in country j, (13.18) and (13.19) allows us to write the numerator of the right-hand equation in (13.22) as

$$X_{ii} - X_{ji} = p^{-\sigma}(1 - \phi) e_i^{\sigma-1} \beta L_i \tag{13.23}$$

There is a similar expression for country j, and dividing (13.23) by the similar equation for country j allows us to write the right-hand equation of (13.22) as

$$\frac{X_{ii} - X_{ji}}{X_{jj} - X_{ij}} = 1 = \left(\frac{e_i}{e_j}\right)^{\sigma-1}\frac{L_i}{L_j} \qquad e_i^{\sigma-1} = \left[N_i p^{1-\sigma} + N_j (pt)^{1-\sigma}\right]^{-1} \tag{13.24}$$

Suppose that country i is bigger than country j. Then it must be true that country i is going to produce more varieties of *X* in equilibrium than country j: $N_i > N_j$. Each country spends the same fraction of its income on *X* varieties and since each variety is produced in the same quantity regardless of home country, then country i must produce more varieties (there is an exception to this in free trade, discussed shortly). The representative price of a good can be factored out of the price index expression in the right-hand equation of (13.24). This allows us to write the left-hand equation as:

$$\left(\frac{e_i}{e_j}\right)^{\sigma-1}\frac{L_i}{L_j} = \frac{N_j(1 + \phi N_i/N_j)}{N_i(1 + \phi N_j/N_i)}\frac{L_i}{L_j} = 1 > \frac{N_j/L_j}{N_i/L_i} \qquad \textit{if} \quad \phi < 1 \tag{13.25}$$

The inequality on the right follows from $N_i/N_j > 1$ and positive trade costs (the phi-ness of trade is less than one where $\phi = 1$ is costless trade). In other words, the result is that the large country not only produces more varieties in absolute terms, it produces more varieties relative to its size. The larger country is *relatively* more specialized in *X* goods.

13.5 The core-periphery model

In fact, it is not quite so simple. For small trade costs (t and ϕ near one but $t > 1$ and $\phi < 1$), the left-hand equations in (13.25) do not have a solution with positive values for both N_i and N_j. The true solution for small trade costs is that the large country i produces all *X* varieties and country j specializes in *Y* and exports *Y*, in exchange for *X*.

The result is shown in a numerical simulation in Figure 13.4. The trade cost t is shown on the horizontal axis and the share of all varieties of *X* produced in each country is shown on the vertical axis. Country i is three times the size of country j in the simulation. In free trade (t = 1), any distribution of the *X* industries is an equilibrium, though the small country will not be able to satisfy all world demand for *X* if the share of income spend on *X* is greater than 1/4. 1/4 of the varieties produced in j and 3/4 produced in i is one possible outcome. For small to moderate trade costs ($1 < t < 1.35$), country j producers cannot compete and all varieties are produced in the large country i.

Figure 13.4

The intuition for this result is that, at moderate trade costs, the demand for any variety produced in country j must be less than the demand for a variety produced in i: for a country i firm, most of the demand is in the large home market and only a small portion comes from export sales and vice versa for a country j firm. Thus, under free entry, the second equation of (13.20), the small-country firms make losses and do not enter. As trade costs continue to increase, the price index in country j, e_j, rises but the price index does not change in country i when it does not import any *X* goods. At some level of trade costs (about 1.35 in Figure 13.4), the demand for a local variety rises sufficiently that firms can begin to enter in country j (follows from 13.18 and 13.19). As trade costs continue to rise toward a prohibitive level (autarky), the share of firms in country i approaches 3/4, the same as its share of total income.

This result is rather extreme and depends, in particular, on the assumption of only one factor of

production. A more general result with a two-factor Heckscher-Ohlin model is found in Krugman and Venables (1991). The *X* sector is capital intensive and the *Y* sector is labor intensive, but assume that the two countries have the same relative endowments and differ only in size. The equivalent of Figure 13.4 for the two-factor model is shown in Figure 13.5. In free trade (t = 1), the share of firms in each country is no longer indeterminate. Each country will have a share of firms strictly in proportion to its size: if a country had a larger share of firms than relative size, this would drive up the price of capital and its firms would be uncompetitive.

Figure 13.5

As trade costs rise, the share of firms in the large country increases above its relative size, as shown in Figure 13.5, due to the larger demand for each large-country variety as discussed above. However, a rising price for capital means that the small country's firms are not driven out of business. As trade costs continue to rise, the price index *e* rises faster in the small country and this increases demand for the small country's goods. As in Figure 13.4, the shares of firms in the two countries approaches their relative size as trade costs become very large.

To summarize the result in Figure 13.5, the share of firms in each country is proportional to its share of world income both in free trade and in autarky in a two-factor model. In free trade, this is caused by factor market (cost) effects, while with prohibitive trade costs this is caused by product market (demand) effects. For intermediate trade costs, there is a divergence in shares and the outcome is clearly not "in between" the free trade and autarky outcomes. When countries do differ in size, trade costs lead to the simultaneous existence of intra and inter-industry trade: the large country is a *net* exporter of differentiated goods but also imports then while the small country is a *net* exporter of the homogeneous competitive good. Also note that each country is relatively specialized in the consumption of its own differentiated goods (in addition to obviously being specialized in producing them), another example of the "home-market effect".

Figures 13.4 and 13.5 will also have implications for per capita income and factor prices and, if factors can migrate, intermediate levels of trade costs can lead to further divergence between countries. This topic is postponed until Part III of the book.

13.6 Heterogeneous firms and firm-level export behavior

In all of the models with free entry and exit of firms that we examined to this point in the book, there are many potential entrants, all of whom are identical. A zero-profit condition determines or "cuts off" the number of firms that can be active in equilibrium, but who is active and who does not enter is entirely arbitrary. Firms have no individual identity just as the products in the Dixit-Stiglitz love-of-variety approach have no particular identity. In recent years, there has been great theoretical and empirical interest in so-called heterogeneous firm models, stimulated by Meltiz (2002). In this approach, there are many potential firms, not all of whom can successfully enter the market. Firms differ in their productivity or conversely in their marginal cost of production.

In the standard Melitz approach, the parable is a lottery in which firms pay an entry fee and then there is a draw in which each entrant into the lottery draws a marginal cost. These costs (or productivities) are then ordered from the firm with the lowest cost to that with the highest cost. In Figure 13.6, the curve labeled cumulative distribution function gives the total number of firms with a marginal

cost below a given number. At the left of the axis are the lowest-cost (most productive) firms and at the right is the highest cost firm, with a marginal cost denoted mc_0.

Again following the standard story, if a firm wants to enter domestic production after it gets its draw, it must pay an additional fixed cost which we will denote fc_d. If it also wants to export, there is an *added* fixed cost of setting up its foreign contacts and distribution network, denoted fc_x (so the *total* fixed costs for an exporting firm are $(fc_d + fc_x)$).

Assuming that there are a sufficient number of potential firms relative to the market size and that the distribution of costs is large enough, general equilibrium will establish some critical "cutoff" levels. These cutoff levels are marginal costs under which the firm with the cutoff cost breaks even. There are two critical cutoff costs in this model. First, the lowest-cost (most productive) firms will find it profitable to enter both the domestic and export markets. We have denoted the cutoff level for exporting firms as mc_x in Figure 13.6: the firm with this cost level breaks even on its export sales but earns positive profits on its domestic sales. Then there is a second cutoff cost at which a firm breaks even, serving only the domestic market. This is denoted as mc_d in Figure 13.6. Firms with lottery draws in the interval (mc_x, mc_d) serve only the domestic market and firms with costs in the interval (mc_d, mc_0) do not enter (do not produce) after getting their disappointing number.

Figure 13.6

This model has a number of empirically appealing features. Specifically, only a subset of firms engage in exports, and those firms are both larger and more productive than the strictly domestic firms. Both predictions are confirmed in the data. Figure 13.6 plots the distribution of exports across firms, with the strictly domestic firms exporting nothing.

Let firm d denote the marginal domestic firm with cutoff cost mc_d. Then this firm is characterized by a pricing equation and a zero-profit condition as follows.

$$p_d(1 - 1/\sigma) = mc_d \qquad (p_d/\sigma)X_d = fc_d \tag{13.26}$$

The marginal exporting firm earns zero profits on its export sales, which we denote X_x.

$$p_x(1 - 1/\sigma) = mc_x \qquad (p_x/\sigma)X_x = fc_x \tag{13.27}$$

In the right-hand equations of (13.26) and (13.27), substitute the demand equations given in (13.18) and (13.19), letting $\beta = 1/2$. Assume again that there are two identical countries, so in equilibrium the price index will be the same in each country. Thus we drop the subscript on the price index. The right-hand equations of (13.26) and (13.27) become

$$p_d^{1-\sigma} e^{\sigma-1}(L/2) = \sigma fc_d \qquad p_x^{1-\sigma} \phi e^{\sigma-1}(L/2) = \sigma fc_x \tag{13.28}$$

Now we can use the first equations of (13.26) and (13.27) to replace prices and with the cutoff marginal costs. The two equations in (13.28) become

$$\left[\frac{\sigma}{\sigma-1}mc_d\right]^{1-\sigma} e^{\sigma-1}(L/2) = \sigma fc_d \qquad \left[\frac{\sigma}{\sigma-1}mc_x\right]^{1-\sigma} \phi e^{\sigma-1}(L/2) = \sigma fc_x \tag{13.29}$$

Assume that both equations hold; that is, there are both exporting and domestic firms in equilibrium. Then divide the first equation by the second and rearranging. We get:

$$\left[\frac{mc_x}{mc_d}\right]^{\sigma-1} = \phi\frac{fc_d}{fc_x} < 1 \qquad \textit{iff} \quad \phi < \frac{fc_x}{fc_d} \tag{13.30}$$

The right-hand inequality is a boundary condition for the simultaneous existence of exporting and domestic firms: the added fixed cost of exporting must be sufficiently high and/or the phi-ness (freeness) of trade must be sufficiently low in order for $mc_x < mc_d$ and for strictly domestic firms to exist as in Figure 13.6.

Examining (13.30), it is clear that a *fall* in trade costs, an *increase* in ϕ, must increase mc_x relative to mc_d, closing the interval between the two cutoff costs.

$$\frac{d(mc_x/mc_d)}{d\phi} > 0 \tag{13.31}$$

This last result in (13.31) is actually quite weak. While the gap between the cutoff for exporting firms and the cutoff for domestic firms shrinks, this could be consistent with both going up and down and does not by itself prove that $dmc_x > 0$ and $dmc_d < 0$. But this is indeed the case. It can be demonstrated with a lot more algebra, so we must limit ourselves here to trying to show just the intuition behind the result.[1]

Arbitrarily consider country i, but remember both countries are identical. The logic of a falling t (rising ϕ) goes as follws. (1) country j firms that were initially exporting continue to export but their price in country i falls due to lower ϕ. (2) some firms in j that were initially domestic now find it profitable to enter exporting. For both reasons, the impact effect in country i is that the price index *e* must fall. (3) from the first equation of (13.29), we see that the fall in *e* means that the highest marginal cost domestic firms in country i must exit.

Baldwin (2005) shows that in the special case where $fc_x = fc_d$ (however implausible), it turns out that the number of varieties *consumed* in each country stays unchanged once all general-equilibrium adjustments have taken place. From the previous paragraph, this must mean that the number of domestic firms that exit in each country must be exactly balanced by the number of foreign domestic (initially non-exporting) firms which switch to exporting status. Let's just concentrate on this special case, though we want to emphasize that none of the key results about changes in cutoff productivities etc. rely on this assumption.

(4) given that the number of varieties consumed in a country i does not change with the rise in ϕ, this must mean that the price index *e* must fall. This is because the prices of initially imported varieties fall and the prices of the newly imported varieties are lower than the prices of the exiting domestic firms.

(5) from the first equation of (13.29), this in turn means that the cutoff marginal cost for domestic firms mc_d must fall.

The fact that e falls does not imply that the productivity cutoff mc_x for exporting firms rises in the second equation of (13.29). While e falls, ϕ rises. But indeed further analysis shows that mc_x must rise. Some domestic firms are becoming exporters: these switching firms have higher marginal costs than the highest marginal cost (cutoff) initially exporting firm. Thus ϕ rises in the second equation of (13.29) faster than e falls. Note that if this result is not true, it must mean that some initially-exporting firms stop exporting and switch to being purely firms after trade costs fall, a result that clearly seems to contradict our intuition.

These results are shown graphically in Figure 13.7. The effect of rise in ϕ (fall in t) shifts the cutoff mc_x to the right and the cutoff mc_d to the left. We also graph the initial sales per firm labeled "initial sales" in Figure 13.7 and "final sales" in the diagram. Firms up to the new exporting cutoff mc_x' produce more (the cutoff firm adds foreign sales to its existing domestic sales). Firms that remain purely domestic produce less and firms with costs above the new cutoff cost mc_d' exit.

Figure 13.7

The point about the results in Figure 13.7 is that trade liberalization is predicted to make firms in a sense more unequal. The more productive firms get larger, some middle firms shrink, and the initially least productive are forced to exit the market. We then have an addition source of gains from trade liberalization: the average productivity of firms in the market rises.

13.7 The gravity equation

An important empirical tool in international trade analysis is called the gravity equation. The term comes from physics and Newton's law of gravitational attraction. Let G be the gravitational attraction between two bodies, one of mass M_1 and the other of mass M_2. Let d denote the distance between them. α is some constant of proportionality. G is given by the formula

$$G = \alpha \frac{M_1 M_2}{d^2} \tag{13.32}$$

In the economics version of this equation, G is replaced by the volume of trade between countries 1 and 2 which we can denote by T_{ij}. M_1 and M_2 become the incomes (GDP) of the two countries. d continues to denote distance or some other measure of trade costs but there is no presumption that trade falls exactly with the square of distance. Then, economists usually take the logs of both sides of the equation so that it is linear in logs (useful for estimating regression equations). The gravity equation for trade between countries i and j is then

$$\ln T_{ij} = \alpha + \beta_1 \ln M_1 + \beta_2 \ln M_2 + \beta_3 \ln d \tag{13.33}$$

Empirical analysis consists of estimating the α and β coefficients via regression analysis. β_1 and β_2 should be positive and β_3 should be negative. This negative effect of distance (trade costs) should be apparent from Figures 13.1 and 13.2.

There are several theoretical drawbacks of the gravity equation, perhaps the most important of which is that it has no role for comparative advantage trade. Trade between two countries is independent of their factor endowment and technology differences, for example. But it works (fits) surprisingly well in practice, so well in fact that this long troubled trade economists. There have been many papers adding additional features to the gravity equation to reflect factors such as comparative advantage, institutions, common language and so forth and we cannot present an analysis of all these extensions here.

We do, however, want to comment on one feature of the gravity equation (also true in Newton's equation) which is that trade is predicted to depend not only on the total incomes of the two countries combined but also on the size difference. For a given total income ($M_1 + M_2$), gravity (13.32) and trade (13.33) are maximized when the countries are identical in size ($M_1 = M_2$) and minimized as one country goes to zero size. This is, in fact, a feature of the Dixit-Stiglitz love-of-variety monopolistic-competition model. The latter is, therefore, often advocated as a theoretical foundation for the gravity equation.

Suppose that we have two countries which are identical except for size and assume that trade is costless (let that be represented by $d = 1$). We don't need to assume a second sector *Y*, suppose all goods are *X* goods in the sense derived several times above. Let the total number of goods be normalized to one, the output of each equal to one, and let their prices equal one. If the countries are identical, each country will produce a number of goods equal to 1/2. Further, each country will consume half of each of its own goods and half of the output of each of the other countries good. A country's exports are then 1/2 of each of its goods, with the number of goods it produces equal to 1/2 (the world total is normalized at 1). The total two-way volume of trade will then be given by the sum of the two countries' exports:

Trade volume = (1/2)*(1/2) + (1/2)*(1/2) = 1/2

Now suppose that we hold the world size constant, but that one country has only 1/4 of the total endowment and the other country has 3/4. The small country will produce 1/4 of the total number of goods, with 1/4 of each being consumed at home and 3/4 being exported. The large country produces 3/4 of the goods, retaining 3/4 of each at home and exporting 1/4 to the small country. Total trade volume is now

Trade volume = (1/4)*(3/4) + (3/4)*(1/4) = 6/16 = 3/8 < 1/2

Total trade volume falls as the countries become more unequal in size holding total world size constant, exactly the prediction of the gravity equation in (13.33).

As noted above, there have been many extensions to the gravity equation via additional variables in the estimating equation (see Evenett and Keller (2002) for a review). In addition, it has been noted that different underlying economic models imply different predicted values and relative values for β_1 and β_2 in (13.33). These different predictions are then taken to data in order to help discriminate among alternative theories of trade (Feenstra, Markusen, and Rose (1999)). Unfortunately, we cannot go into more detail here but simply note that the basic gravity equation is often a starting point in empirical investigations.

13.8 Empirical Evidence

In the last three chapters we have introduced several concepts that depart markedly from the world of two industries in which homogeneous and perfectly competitive firms produce with constant

returns to scale. The world is more complex and we have developed models to explain various theoretical determinants and impacts of international trade. We may summarize the most important of these extended theories as follows.

1. Where there are large firms with market power that are protected from import competition, one important effect of trade is to reduce profit markups, generating a pro-competitive gain in welfare. This idea may be called the "imports as market discipline" hypothesis.

2. Many industries may be characterized as being relatively open to entry and populated by firms with some increasing returns to scale. At the same time, consumers have a taste for variety and firms can reduce costs by having access to a wide variety of intermediate inputs. Thus, international commerce may come in the form of *intra-industry trade*, in which countries both export and import similar but differentiated inputs and outputs.

3. Distance and other forms of trade costs seem to affect trade flows negatively. The gravity model of international trade is a convenient and successful econometric device for studying this factor.

4. Rather than firms being identical within an industry, they likely display a large degree of heterogeneity, which may have a considerable impact on international trade. If the effect of openness to trade is to expand output in the more productive firms while pushing the less production enterprises out of business there should be an overall increase in productivity.

In this section we present evidence on each of these four hypotheses, which continue to be heavily studied at the forefront of international trade analysis. The underlying literature is very large and we can only summarize some seminal articles here. Interested students may wish to read more widely.

Imports as Market Discipline

The idea is completely intuitive: if a country liberalizes its trade restrictions the domestic firms with market power that face more import competition should see their profitability decline. More precisely, the markup of price over marginal costs should go down. Intuition and evidence are different things, however, so economists have considered this problem in a number of countries that went through significant import liberalization episodes. Before considering the results, however, a number of initial issues need to be addressed.[1]

To begin, recall from Chapter 11 that for firm *i* selling in an oligopolistic home market its markup equation is

$$\frac{p_i}{mc_i} = \frac{1}{(1 - \frac{s_i}{\eta})} \tag{13.34}$$

Thus, a firm's markup increases as its market share gets bigger and decreases as demand elasticity falls. International trade may affect both of these parameters.

[1]For further perspective, see Tybout (2003)

For empirical purposes we would like data on transactions prices and marginal costs per firm. However, such figures at the firm level are rarely observed in available databases, so a proxy measure must be employed. The most frequently used is the price-cost margin *(PCM*), which is the proportional difference between a firm's sales and its expenditures on labor and intermediate (including raw materials) inputs.

$$PCM_{it} = \frac{p_{it}q_{it} - c_{it}q_{it}}{p_{it}q_{it}} = \frac{p_{it} - c_{it}}{p_{it}} \tag{13.35}$$

This expression shows the *PCM* for a firm i in a year t. If we assume that average cost (c_{it}) is constant in the neighborhood of existing sales, so that it captures movements in marginal cost, the *PCM* is a monotonic transformation of the standard markup in (13.34). We can therefore use the middle part of this equation, which requires data on sales and input costs, to proxy for the markup, keeping in mind this strong assumption.

Note that the difference between sales and costs of labor plus intermediate inputs includes both profits and returns to capital. Because we want to isolate the profit markup, regressions must control for the ratio of capital stocks to sales. Thus, the standard approach to analyzing the impact of import competition on markups is to estimate the following equation:

$$PCM_{it} = \beta_0 + \beta_1 KSALES_{it} + \beta_2 IMP_{it} + \sum_j \beta_j X_j + \varepsilon_{it} \tag{13.36}$$

Here, *KSALES* is the ratio of the capital stock to firm-level or industry-level sales, a measure of capital intensity. The X_j variables are other control variables, which may or may not vary by firm, industry, and year. Our greatest interest lies in the variable *IMP*, which indicates some measure of import competition faced by the firm or industry. The most frequently used measure is simply the import-penetration ratio, or imports in an industry divided by industry sales. It is generally not possible to define import competition faced by a specific firm, of course, since all domestic enterprises compete with imports. An important conceptual problem with import penetration is that it can rise or fall even if there is no change in trade policy. A better approach is to try to incorporate a variable that directly captures cuts in tariff rates or elimination of quotas, since the hypothesis is about changes in exposure to competition. However, it is often difficult to find such data.[2]

Among the control variables it is important to include some indication of how much market power exists, since it seems likely that industries and firms with more domestic market power are liable to experience sharper declines in markups. A standard measure of industry concentration is the Herfindahl Index, defined as $H = \sum_i s_i^2$, where s_i is the share of firm i in market sales. In cases where a few firms have very large shares this index will be larger than where many firms have small shares. In any case, the higher is *H* the more concentrated are sales within an industry, which is often taken to mean a high degree of market power. In cases where the analyst has firm-level data, the figures for sales shares for each firm

[2]In Chapters 18 and 19, we will describe available data sets covering tariffs and non-tariff barriers.

can be used instead in the regression. With respect to the regression, including an interaction term between *IMP* and either *H* or sales shares should capture whether the competition effect is greater for more concentrated industries or for firms with greater market power.

The initial econometric study of this kind was by James Levinsohn (1993), who studied data for Turkish firms in 10 manufacturing industries from 1983-86. In 1981 Turkey had a high degree of import protection, with an estimated average tariff of 49 percent and an extensive and complex system of import quotas, import licensing rules, export restrictions, and foreign-exchange regulations. Beginning in 1980, the export restrictions were lifted and in 1983-84 a strong dose of import liberalization took place. This policy ended licensing, reduced the number of import quotas, and cut average tariffs to 20 percent. Levinsohn's calculations suggested that there were three industries that were imperfectly competitive and had price markups significantly higher than 100 percent before the policy liberalization. All three experienced a decline in average firm markups after trade was made more open, with two of these reductions being statistically significant. Unfortunately, the author did not report the amounts by which the markups were reduced.

More systematic studies were published in a volume edited by Mark Roberts and James Tybout (1996). We report the basic results in Table 13.1. Four countries that underwent significant import liberalization are listed, including Chile, Colombia, Morocco and Mexico. The authors of these studies were asked to run regressions consistent with equation (13.36). This was done with both industry data and plant-level (firm-level for Morocco) data in periods following reforms of highly inward-looking trade policies. For example, Chile's average tariff rates exceeded 100 percent in 1973 and there were extensive deposit requirements and tax surcharges on imports. After 1973, a radical reform was imposed and the average tariff rate fell to 12 percent by 1979, with much of the regulatory system removed. Similar episodes emerged in the other nations in the table.

In the regressions in Table 13.1, the dependent variable is the price-cost margin. Looking first at the industry regressions, it seems that sectoral concentration (the Herfindahl index, *HERF*) increased PCMs significantly in Colombia and Mexico. For the first three countries, import competition is measured by *IMP*, the import-penetration ratio. This coefficient should be negative under the competition hypothesis, and it is for Chile and Colombia but not significantly. The coefficient surprisingly is positive and significant for Morocco. More interesting is the interaction term between *IMP* and *HERF*, which is significantly negative for Colombia. Among these three countries, then, there is evidence of the imports as discipline hypothesis in Colombia, but not the others, at the industry level. The Mexican regression uses *TAR*, the average tariff rate for an industry, as a policy measure of import competition. The higher is the tariff, the greater the protection. Perhaps surprisingly, industries with higher tariffs have lower PCMs in the data, while the interaction term between *TAR* and *HERF* is not significant. Overall, there seems tbe just marginal support for the hypothesis using industry data.

Table 13.1

There is more evidence for the hypothesis with the plant-level and firm-level data in the second set of regressions. There we find that a rise in a plant's or firm's market share raises its PCM, though this effect diminishes as the share grows larger, according to the negative coefficient on the squared share (*SHARE SQ)*. The *IMP* variable is not significant in the first three countries, but there is a strongly negative effect of the interaction term in both Chile and Colombia. Thus, it seems that plants with high market shares, and more market power find a larger reduction in their price-cost margins due to import competition than do smaller plants. In Mexico there is a similar large impact of *SHARE* on PCMs across

plants. Interestingly, the interaction term on *SHARE* and *TAR* is significantly positive, implying larger plants in industries with high tariff protection have higher PCMs. This is also consistent with the hypothesis that import competition forces domestic firms to reduce their markups.

An interesting variation on this theme is found in the study by Hoekman, et al (2004). These authors argue that one would expect the size of a country to matter in terms of how much competition affects markups. There are really two types of policy changes that expand competition for domestic firms with market power: reductions in domestic entry barriers through changes in regulation, such as anti-trust enforcement and liberalization of credit markets, and the easing of foreign entry barriers through lowering trade restrictions, such as tariffs.[3] Hoekman, et al hypothesize that the ability of tariff cuts to reduce domestic markups should be larger in smaller economies, simply because small economies are more dependent on trade and import a larger share of domestic consumption. On the other hand, relaxation of entry regulations should have a larger impact in bigger markets where competition among domestic firms would be more significant. They tested these ideas using data from a sample of 21 industries in 42 developed and developing countries over the period 1981 to 2000. They used an econometric method that permitted efficient estimation of markups in the first stage by correcting for endogeneity between markups and industry-level inputs (Olley and Pakes, 1996). Having purged the markups this way, they were regressed on a measure of the difficulty of domestic entry, which was the number of legal procedures required to register a new firm, and industry-level average tariffs in each country over time. Control variables included a national index of legal patent rights and financial market capitalization. A further control was country size, measured by geographical area in square kilometers. One can criticize this final variable since physical area does not imply a significantly large market if there is a small population density.

The authors calculated that the median price-cost margin in 21 developed countries was 1.60 (a 60-percent markup) and 1.91 in 21 developing countries, figures that were rather higher than those in prior literature. Developing countries generally had higher tariffs and more entry procedures than did developed countries. In their preferred second-stage estimation, the authors found that while the average tariff rate tended to raise domestic markups, the interaction term between size and tariffs significantly reduced them. That is, the effect of tariff cuts tends to fall as a country gets larger. Further, the interaction between country size and the number of entry restrictions had a significantly positive coefficient, so the effect of reducing entry restrictions these larger with market size. Both of these findings support the basic hypothesis. This study has not been replicated with firm-level data to our knowledge.

Overall, while the results of the studies reviewed here are not definitive, they do suggest that trade liberalization, especially in developing economies, often reduces the margins between prices and costs, particularly at the level of individual production plants. While more work could be done, this evidence does tend to support the underlying theory.

Intra-industry Trade and Product Variety

The phenomenon of intra-industry trade (IIT), in which a country both exports and imports similar goods within a sector (industry), is quite prevalent in global commerce. This may be easily seen in Table 13.2, which lists a measure of IIT for selected countries and broad industries in 2007. The

---

[3]For example, Konings, et al (2001) demonstrated that firm-level markups in the 1990s were higher in the Netherlands, where there was very little anti-trust activity, and Belgium, where policy was more strict.

measure is the so-called Grubel-Lloyd index, named for two trade economists who first extensively discussed the empirical concept of IIT in a famous book (Grubel and Lloyd, 1975). For any country, the formula for computing this index within an industry *j* is:

$$IIT_j = 100 * [1 - \frac{\left| EX_j - IM_j \right|}{EX_j + IM_j}] \quad (13.37)$$

For any industry or product this index can range from zero, where either exports or imports are zero and so all trade is *inter-industry* in nature, to unity, where exports and imports equal each other so all trade is *intra-industry*.[4] The figures in the table make it immediately clear that there is a high degree of two-way trade in nearly all industries and countries.[5]

There are some patterns worth noting that will guide our thinking about what factors cause IIT. First, note that the highest average indexes exist for Germany and the United Kingdom. These are two large economies with similar market sizes, manufacturing bases and factor endowments, while being close to each other within the European Union, an economic area that has free internal trade. Similarly, Canada has a high index in passenger vehicles, which is caused by extensive two-way trade with the United States in automobiles within NAFTA. In contrast, Japan and Australia have lower average indexes, reflecting the fact that they are quite distant from other developed markets. Second, excluding Japan and Australia there appears to be a relationship between per-capita income and IIT. The lowest average indexes exist for India, China, and Brazil. Mexico's IIT is somewhat larger, perhaps due to its close trade relationships with the United States and Canada. Indeed, Mexico's IIT is high in passenger vehicles as one effect of NAFTA has been to induce greater trade within the trade area in automobiles and automobile parts. Thus, income levels, size, similarity, and distance seem to national characteristics that matter in determining IIT.

Table 13.2

Third, comparing these indexes across sectors, the lowest extent of IIT is found in apparel and accessories. This is not surprising, since this industry is marked by constant returns to scale and labor-intensive production. Thus, international trade in this industry is largely driven by comparative advantage. Indeed, China and India have very small indexes in this industry because they have far greater exports than imports. Japan's small figure reflects its preponderance of imports over exports. Mexico's higher index of 63 percent IIT suggests an interesting phenomenon. Mexico tends to import large volumes of basic apparel products (unprocessed cotton, for example) from the United States, which it then assembles into finished apparel products for export back to the United States. In contrast, industrial inputs, such as organic chemicals and iron and steel products, have the highest IIT shares. This is because

[4]In 2007 several countries in the table had large trade imbalances, which can distort these figures. For example, the United States had a massive trade deficit, associated with macroeconomic factors, which would make the IIT ratios smaller than they would be in a situation closer to a full equilibrium. Thus, the calculations in Table 13.2 incorporate an adjustment for trade surpluses or deficits, as suggested by Acquino (1981).

[5]Interested students can download highly detailed international trade data from this source at http://comtrade.un.org/db/.

there are different types of intermediates produced in different locations, subject to increasing returns to scale, and then traded across borders, particularly among the developed economies. Industrial machinery, computers, passenger vehicles and scientific and professional instruments display smaller indexes, though they generally exceed 50 percent. Finally, we list alcoholic beverages, a sector in which product differentiation is significant and countries often trade varieties of wine, beer and spirits among themselves. These shares tend to be more moderate, though again the UK and Germany have high values. Assembling these observations, economies of scale, stage of processing, and scope for product differentiation seem to be important industry characteristics.

One obvious objection to these calculations is that the industries chosen are quite broad, so there may be high levels of IIT that exist only due to excessive product aggregation. Put differently, there are many sub-products within industry trade categories. These smaller groups might actually vary by factor intensities or other determinants of comparative advantage and not be subject to IIT. However, numerous studies have demonstrated that positive shares of IIT remain even at fine levels of product disaggregation. For example, consider the category "passenger vehicles" in Table 13.2. This category comes from the harmonized system (HS) trade classification of the United Nations and has the 3-digit number 871. As the number of digits expands, the product classification gets increasingly specific. Thus, IIT for the United States in 871 was 73 percent in 2007, as shown. However, in category 8711 (motorcycles, including mopeds) it was 86 percent. In category 871150 (motorcycles, including mopeds, with piston engines of cylinder capacity greater than 800 cubic centimeters) it was 77 percent. In category 871140 (motorcycles, including mopeds, with piston engines of cylinder capacity greater than 400 cubic centimeters but not more than 800 cubic centimeters) it was 15 percent. In category 871411 (saddles for motorcycles, including mopeds) it was 49 percent. For large economies, such as the United States, Germany and China, IIT shares remain positive at very high levels of disaggregation. In short, ITT indices remain high, even for extremely close substitutes.

As one might imagine, however, for smaller and poorer economies the extent of IIT diminishes more rapidly as products are disaggregated and often disappears. For example, in HS category 6401 (waterproof footwear with soles and uppers of plastic or rubber) Jamaica had an IIT index of 39 percent in 2007. But in category 640110 (waterproof footwear with soles and uppers of plastic or rubber incorporating a metal toe-cap) the index was just 0.3 percent. Beyond this level the index was zero because only either exports or imports existed or Jamaica did not trade the good. These "zeros" in trade associated with small size strongly suggest that increasing returns matter in determining actual trade flows.

Early studies of what explains the extent and structure of intra-industry trade relied on ad hoc regression specifications. For example, Balassa (1966) showed that trade in manufactures increased rapidly in Western Europe after the formation of the Common Market, suggesting that such forces as product differentiation, scale economies, and industrial restructuring were prominent results of economic integration among similar nations. In a later study, Balassa (1986) hypothesized that IIT between countries would rise with their levels of per-capita GNP and would fall as distance between trading partners increased and trade barriers became more restrictive. He also included indicators of shared borders and mutual membership in free-trade areas. Balassa analyzed disaggregated trade data for 1971 for 38 developed and developing countries, computing an IIT index for each nation. His results seemed to support his hypotheses, with coefficients that were statistically significant in the directions expected. In particular, the share of IIT seemed to be higher among countries with higher income levels, within the Common Market, and among nations that shared common borders.

A final empirical regarding IIT concerns product-variety gains. As shown in Chapter 12, trade liberalization is likely to induce increases in product variety via through imports. Table 13.3 shows the results of an interesting recent study of this question using U.S. tariffs and import data (Feenstra and Kee, 2007). As noted in the top panel, the United States reduced its tariffs on imports from Mexico, implementing NAFTA in the years after 1995. Average tariff rates across all merchandise sectors fell from 4.4 percent to 0.3 percent between 1990 and 2001, while there were large reductions in protection on agriculture, textiles and apparel, electronics and other sectors. At the same time, Mexico reduced its tariff rates against the United States considerably, which should have improved the its firms' access to varied intermediate inputs. As may be seen in the second row, over this same period the contribution of Mexican producers to imported varieties in the United States rose in each category. For example, in 1990 52 percent of agricultural varieties imported into the United States from any country also came from Mexico. That proportion rose to 67 percent by 2001. There was a particularly sharp rise in the proportion of imported electronics that came from Mexico.

Table 13.3

Data with respect to China are in the bottom panel. The United States reduced its tariffs on Chinese goods in most sectors over this period, partially in anticipation of that country's accession to the WTO in 2001. China also cut its tariff rates in most of these sectors. Again, there were marked increases in the proportion of U.S. imported varieties that came from China, including a rise from 28 percent to 63 percent in machinery and transport equipment.

Feenstra and Kee (2007) performed a simple econometric analysis of the contributions of U.S. tariff cuts to the growth in Mexican export varieties. They found that a one-percent reduction in the tariff rate induced a two-percent growth in varieties, a large impact in economic terms. Interestingly, their results also suggested a competition effect: increases in Chinese export varieties to the United States tended to reduce the growth of Mexican export varieties. Overall, however, tariff liberalization seems to have expanded the scope of Mexico's products shipped to the American market, with a consequent gain for American consumers.

The Gravity Model of Trade

We mentioned the famous gravity equation, and its wide empirical success in explaining international trade flows among countries, in Section 13.7. In fact, this equation can be derived in a straightforward way from the basic theory of monopolistic competition. Suppose that there are identical and homogeneous Dixit-Stiglitz utility functions as given in equation (12.1) in Chapter 12 and that all goods are differentiated substitutes. For the moment, assume there are no trade costs. Because preferences are identical and homogeneous, each country *i* will consume all goods produced everywhere and the share it consumes is equal to its proportion of world income. Thus, the total value of bilateral exports from country *j* to country *i* will be:

$$X_{ji} = s_i \sum_k p_k^j X_k^j = s_i GDP_j = \left(\frac{1}{GDP_W} \right) * \left(GDP_j * GDP_i \right) \qquad (13.38)$$

where X_k^j is production of variety *k* in country *i*. The immediate problem is that this expression does not

account for trade costs. The simple version of gravity assumes that distance reduces the amount of each good that arrives through imports:

$$X_{ji} = \left(\frac{1}{GDP_W} \right) * \left(\frac{GDP_j * GDP_i}{dist^{\lambda}} \right) \tag{13.39}$$

Taking natural logs of this equation leaves a regression specification like that in (13.33), with the constant term capturing the inverse of global GDP = GDP_w. Note, however, that in principle the exponents on the national GDP terms are both equal to one, so those coefficients should not be much different from unity.

Stated this way, the gravity equation is disarmingly simple. It simply states that bilateral trade between any two countries is determined by size and distance. The importer size matters because that determines its share of world consumption, while the exporter size matters because that determines its output volume, some of which can be exported. In terms of aggregate trade all the model requires is identical and homogeneous preferences and it can describe trade generated by any factor, including technologies, endowments, and product differentiation. It can be applied also at the sectoral or product level, in which case the assumption of differentiated products is important (recall that product-level trade is not determined in HO theory). In that case, the analyst would normally put importer GDP and exporter sectoral output on the right-hand side.

The gravity model can, in fact, be derived more rigorously as a general equilibrium, as in Bergstrand (1989) and Anderson and van Wincoop (2003). Here we provide a summary of the theory in the latter paper, which makes some useful point regarding distance and trade costs. Assume again that all countries share the same CES utility functions with substitution elasticity $\sigma > 1$. To make the exposition easier, suppose that the products made in each of *n* countries can be aggregated into a single good and that the utility function is described across country sources:

$$U_i = \left(\sum_j^n C_{ji}^{\alpha} \right)^{1/\alpha} ; \sigma = \frac{1}{1-\alpha} > 1 \tag{13.40}$$

Here, the C terms indicate consumption in country *i* the composite good from *j*. Maximizing this utility function subject to the income constraint $Y_i = \sum_j p_{ji} C_{ji} = \sum_j p_j t_{ji} C_{ji}$, where $t_{ji} \geq 1$ indicates trade costs, yields the following bilateral demand functions:

$$X_{ji} = \left[\frac{p_j t_{ji}}{P_i} \right]^{1-\sigma} Y_i \tag{13.41}$$

In words, the value of exports from *j* to *i* depends on price (and therefore costs) in the exporting country, bilateral trade costs, a price index in the importing country *i*, and GDP in the importing country *i*. This also applies for sales in *i* of its own good (X_{ii}), on which there are not trade costs. If there were just two countries in the world the price index would be identical to that in equation (13.17). However, with many

countries it depends on trade costs with all partners: $P_i = \left[\sum_j \left(p_j t_{ji} \right)^{1-\sigma} \right]^{1/1-\sigma}$

To make this a general equilibrium we can bring in a market-clearing condition for each exporting country: $Y_j = \sum_i X_{ji}$. That is, GDP equals the sum of domestic output plus export value. Using the trade demand functions in (13.41) these equations can be rewritten as

$$Y_j = p_j^{1-\sigma} \sum_i \left(\frac{t_{ji}}{P_i} \right)^{1-\sigma} Y_i , \quad \text{for all } j. \tag{13.42}$$

This system of equations can be solved for prices p_j in terms of GDPs, trade costs, and the price indexes. Note that the solutions will involve world GDP$_w$ ($Y_W = \sum_i Y_i$) and the share of each country in world GDP ($s_i = \frac{Y_i}{Y_W}$). It is also convenient to assume symmetric trade costs ($t_{ji} = t_{ij}$). Putting all of that together yields the following general-equilibrium gravity equation:[6]

$$X_{ji} = \left(\frac{Y_j Y_i}{Y_W} \right) * \left(\frac{t_{ji}}{P_j P_i} \right)^{1-\sigma} \tag{13.43}$$

This expression gives us more insight into how transport costs matter. The higher the costs between two nations, the lower bilateral exports. Moreover, *relative transport costs* matter as well, since they affect the price indexes in the denominator. For example, if country *i* faces an increase in its trade barriers with the rest of the world (raising P_i) it would increase its demand for imports from country *j*. We can also consider how distance might matter. Imagine that trade costs depend on distance: $t_{ji} = d_{ji}^{\lambda}$. The exponent is the contribution of distance to transport costs, while the term (1-σ) gives the ensuing impact on trade flows. Many other specifications of the linkage between distance and trade charges could be considered.

Taking logs of (13.43) would give us a fully specified econometric equation:

$$\ln X_{ji} = \beta_0 + \beta_1 \ln Y_j + \beta_2 \ln Y_i + (1-\sigma)\lambda \ln d_{ji} - (1-\sigma) \ln P_j - (1-\sigma) \ln P_i + \varepsilon_{ij} \tag{13.44}$$

---

[6]This model is general equilibrium in that it combines preferences and market clearing. But for simplicity it makes the odd assumption that goods exist as endowments and are not produced. Bergstrand's (1989) model incorporates supply behavior.

In this regression it would be ideal to identify both the elasticity of substitution (σ) and the distance parameter (λ). As a practical matter, however, it is generally quite difficult to measure the price indexes in the final two right-hand side variables, since they are complex functions of prices and transport costs with all other countries. As a result, many analysts simply control for these variables with fixed effects for exporters and importers.

Empirical Uses of Gravity

There are hundreds of published articles in the literature that estimate some form of the gravity equation as an explanation for bilateral international trade and investment flows, at both the aggregate and sectoral levels. We discuss just two of them here to offer a flavor of how they are used.

Gravity and Trade Blocs

One interesting study is by Frankel, et al (1998). Their interest was to see if countries that are mutually located within certain geographical areas have mutual trade flows that exceed what might be expected based on standard gravity variables. They were analyzing a significant policy question at the time: was the world diverging into large regional trading blocs that might be favoring intra-bloc trade over inter-bloc trade?[7] For this purpose they estimated the following equation:

$$X_{ji} = \alpha + \beta_1 \log(GNP_j * GNP_i) + \beta_2 \log(\frac{GNP_j}{POP_j} * \frac{GNP_i}{POP_i}) + \beta_3 \log(DIST_{ji}) + \beta_4 ADJ_{ji} + \beta_5 LANG_{ji} + \gamma_1 EC_{ji} + \gamma_2 WH_{ji} + \gamma_3 EA_{ji} + \varepsilon_{ji}$$

Here, the dependent variable is bilateral exports plus imports of all merchandise between country *j* and country *i*. Although this sum of two-way trade is not bilateral exports, as implied by the theory in (13.44), its usage is common in the applied literature and can be justified by simply adding the X_{ji} and X_{ij} observations for each pair. In their specification, the authors chose not to split apart the two countries' GDP values, thereby forcing the regression coefficient to be the same on both, which is an unusual practice. They also included the product of per-capita GNP levels in each country, arguing informally that as countries become richer they are more likely to demand more finely differentiated products and to specialize more in goods with increasing returns. Both factors should expand the value of bilateral trade.

Frankel, et. al. included the log of the distance between the largest cities in each of the two countries. They also added *ADJ*, which is a dummy variable that takes on the value of one where the countries share a common land border and zero otherwise. The idea is that countries that are adjacent may have developed commercial ties, perhaps through between-country movement of labor and capital, that expands trade beyond that generated by simple proximity. Similarly, a dummy variable *LANG* is included that takes on a value of one when the two countries share a common primary language or had earlier colonial linkages (for example, Peru and Venezuela were colonies of Spain). This is a simple means of testing whether similarity of language and culture expands trade or, alternatively, that a common language reduces transactions costs. This use of simple binary variables to capture impacts on trade other than those modeled by distance is common in the literature. Finally, *EC*, *WH*, and *EA* are dummy variables that are unity when the pair of countries both reside in the European Community,

[7]This is a primary issue in the analysis of preferential trade areas, which we address in Chapter 22.

Western Hemisphere, or East Asia, respectively. The model was estimated for 63 countries, pooling trade observations for the years 1970, 1980 and 1990, with fixed effects included for the final two years.

Table 13.4 lists the results for the main specification. All coefficients are statistically significant at the 99-percent level of confidence. Note that the elasticity of bilateral trade with respect to the product of country-pair GNP levels is 0.72, which clearly demonstrates that market sizes assert substantial impacts on trade. However, this coefficient is statistically less than 1.0, which is inconsistent with the theoretical gravity model. Distance is negative and highly significant, with an elasticity of 0.51. Thus, an increase of one percent in geographic distance between two countries that are not adjacent to one another reduces their mutual trade by 0.5 percent. This is a large economic impact and demonstrates that proximity is one of the primary determinants of trade. Sharing a common border substantially offsets this effect and, indeed, tends to double bilateral trade compared to other country pairs ($\exp(0.72) = 2.05$). In this context, it is no surprise that countries within the European Union trade much more among themselves than outside the area. Similarly, sharing a common language or colonial history tends to raise bilateral trade by 60 percent ($\exp(0.47) = 1.60$).

Table 13.4

The results also suggest that regional groupings have a significant impact on trade beyond simple distance. Thus, the EC dummy of 0.31 implies that two EC nations traded 36 percent more than two otherwise similar countries. The impact of location in the Western Hemisphere was similar in magnitude, though other specifications found that this effect started small and grew larger over time. Countries in the data sample located in East Asia experienced the largest expansion in bilateral trade relative to the basic gravity model. The authors speculate on these grounds that emerging regionalism in trade over this period was growing excessively in the sense that intra-regional trade may be displacing global trade. However, the simple gravity model adopted really is not suitable for making welfare or policy conclusions.

One can criticize the econometric approach on a number of grounds. The model is an ad hoc gravity specification; there are likely important omitted variables, including tariffs and other trade restrictions, and the use of aggregate trade data eliminates the possibility of considering whether the impacts of distance are larger on differentiated goods versus other kinds of goods. Still, these results strongly suggest that geography matters a great deal for international trade flows. Distance itself is an important inhibiting factor, but countries that are physically contiguous, located in the same region, and share a common language tend to trade at greater volumes than randomly selected countries under the gravity model.[8]

Gravity, Differentiated Goods and the Home-Market Effect

A second important study is by Feenstra, et al (2001), who pointed out that, when interpreted in terms of underlying theory, the results of estimating the gravity model for different types of goods can distinguish among theories of trade. To understand their thinking, go back to the model with differentiated products and monopolistic competition in Section 13.4. In particular, equation (13.25) demonstrated that a larger country must produce more differentiated varieties of good X than a smaller

[8]For a review, see Overman, et al (2003). The empirical literature finds that not only do trade volumes decline significantly as countries or cities become more remote from trading partners, so also do average factor productivities and real incomes.

country. In fact, it produces more differentiated products relative to its GDP than does the small country. From this follows the home-market effect: the responsiveness of exports of differentiated goods from country *j* to country *i* must be greater with respect to *j's* own GDP than to its partner's GDP. Putting this into an estimating equation the authors specified:

$$\ln X_{ji} = \beta_0 + \beta_1 \ln GDP_j + \beta_2 \ln GDP_i + \beta_3 \ln D_{ji} + \beta_4 ADJ_{ji}$$

$$+ \beta_5 LANG_{ji} + \beta_6 FTA_{ji} + \beta_7 REM_{ji} + \varepsilon_{ji}$$

Here, *X* refers to the value of bilateral exports from *j* to *i*, GDP is real GDP, D is the great-circle distance between capital cities in the two countries, ADJ is a dummy variable for a common land border, LANG is a dummy variable for common language, and FTA is a dummy variable equaling one if both countries are members of the same free trade agreement. REM is a variable capturing the physical and market-based remoteness of the importing country, conditional on the identity of the exporting nation. To some degree this variable captures the "relative transport costs" noted in the development of the gravity model above in (13.43). The variable measures the opposite of the GDP-weighted distance of the country pair from other nations and should have a positive sign in the regression.

This specification was estimated for three types of goods as defined by James Rauch (1999). There are *homogeneous* goods, such as raw materials and primary goods, that are traded in organized international exchanges. Next are *reference-priced* goods that have prices quoted by industry sources (such as trade magazines) but are not traded on exchanges. Finally there are *differentiated* goods, which are neither traded on exchanges nor have posted prices. The hypothesis set out by Feenstra, et al (2001) is that the coefficient on the exporter's GDP should be higher than that on the importer's GDP for differentiated goods. However, just the opposite would be anticipated for homogeneous goods to the extent that entry barriers in such industries induce firms to penetrate global markets through reciprocal dumping. The authors estimated the gravity equation for 1970, 1975, 1980, 1985 and 1990, using a sample of around 110 countries and products classified in the manner set out by Rauch.

As we show in Table 13.5, the regression estimates strongly support this hypothesis. To conserve space we show only the results for 1990 and only for differentiated and homogeneous goods. In the regression for differentiated goods the elasticity of exports with respect to own-country income is 1.12, while that for homogeneous goods is 0.54. Thus, it appears the gravity model is capable of identifying the role of product differentiation in international trade. In general, the supplementary variables such as language, FTA, and remoteness, matter somewhat more for differentiated goods than for homogeneous goods.

Table 13.5

There are many other uses of the gravity equation that could be covered here, but we will simply mention two current controversies in the literature. First, John McCallum (1995) estimated a standard gravity equation and discovered that the existence of national borders matter a great deal in trade. Specifically, trade between Canadian provinces was 22 times greater than between Canadian provinces and U.S. states, even if the provinces were close to the states in terms of distance. This "border effect" was reconsidered by Anderson and van Wincoop (2003) and their introduction of more comprehensive measures of multilateral trade restrictions reduced this effect considerably. Still, a surprisingly large

trade-diminishing border effect seems to exist (and has been found to exist within Europe also) and what explains this finding remains the subject of research.

Second, Andrew Rose (2004) used the gravity model to study whether membership of a country (or a pair of countries) in the World Trade Organization (or its predecessor, the General Agreement on Tariffs and Trade) actually caused its foreign trade volumes to rise. Qualitative descriptions of the WTO have long argued that membership in this multilateral organization enforces discipline to reduce tariffs, a major factor in globalization. However, Rose's gravity models could not find a consistent and significant impact of WTO membership on trade. This result was criticized by Subramanian and Wei (2007), who found in extended (and more careful) specifications of the gravity equation that WTO membership actually doubled world trade (raising it by some $8 trillion in 2000), though this effect varied widely among countries. This question also remains the subject of ongoing research.

Heterogeneous Firms and Export Behavior

A fourth subject of current empirical research is to study how individual firms differ in terms of size and productivity and to see if those characteristics are related to export activity. We begin our summary of the empirical literature by looking at some basic results on the nature of U.S. exporting firms.

Table 13.6, which reports some results from Bernard, et al (2007), offers a snapshot of how American manufacturing firms that export differ from manufacturing firms that do not. The top panel shows that only a small percentage of firms are exporters. Over all manufacturing industries, just 18 percent of firms exported at least one product to at least one market in 2002. Of those firms that did export, only 14 percent of total sales were shipped abroad. Interestingly, there is large variation across industries. At the low end, five percent of printing firms and eight percent of apparel firms were exporters, with foreign sales amounting to 14 percent of shipments in both cases. At the high end, 38 percent of both computer and electronic products firms and electrical equipment companies were exporters, with 21 percent and 13 percent export-to-sales ratios, respectively. Further, at 21 percent exports-to-sales, the ratio for computers and electronic firms was the largest among all industries. In short, theoretical models of trade in which firms have no identity, and therefore all are exporters, do not effectively capture reality. Moreover, even though the proportion is small, there are exporters in sectors of US comparative disadvantage, such as apparel. This finding is consistent with the idea of variety-oriented differentiation within each industry, though the percentage of exporters seems to rise with the extent of comparative advantage.

Table 13.6

The second panel demonstrates that U.S. exporting enterprises are considerably larger and more productive than non-exporters. These results are coefficients from simple regressions in which the dependent variable is the logarithm of one of the characteristics listed and the independent variable is a dummy for exporter status (taking the value one if a firm exported and zero if it did not). Each regression included industry fixed effects to control for differences in firm characteristics across industries. All coefficients are significant at the one-percent level. These are simply descriptive regressions that correlate export status with firm characteristics.

The data show that the average U.S. exporting firm employs 97 percent more workers than the average non-exporting firm in manufacturing. Despite this larger workforce, exporters are more

productive, as shown by the 11-percent higher value added per worker and three-percent higher total factor productivity (TFP). Exporting firms also pay a higher average wage and have larger capital stocks per worker and skills per worker (measured by non-production employees divided by total employment). Thus, it is clear that exporting firms tend to be larger and more productive. At the same time, they differ in capital and skill intensity, suggesting that basic comparative advantage determinants are important characteristics of trade. Similar results were found for Chilean firms by Alvarez and Lopez (2005).

The final two rows point out another important distinction. Compared with non-exporters, exporting firms on average sell 27 percent more product lines (defined at the 5-digit Standard Industrial Classification of disaggregated industries). This means that exporters have a wider range of products to sell. Economists refer to this outcome as a more pronounced *extensive margin* in that exporters compete in part on the basis of expanding the number of product lines. At the same time they have a larger *intensive margin* on average because their sales per product are larger by 73 percent. Because these concepts are important for thinking about how firms react to changes in policy, a precise definition is in order. The intensive margin of competition refers to changes in output per firm (or per plant or per product) in reaction to a shift in a policy environment, such as trade liberalization. The extensive margin refers to changes in the number of products sold in response to such policy changes. Another view is that the intensive margin refers to selling more of existing products in foreign markets that become more open to trade, while the extensive margin refers to selling more goods abroad or entering new markets. Note that exporters are more likely to compete on both margins than non-exporters.

There is one other important fact to mention, not shown in the table. International trade is extremely concentrated across firms. In 2000, for example, just one percent of U.S. trading firms by value (the sum of exports plus imports) were responsible for over 80 percent of the value of total trade, while the top 10 percent of enterprises accounted for over 95 percent of trade. This fact strongly suggests that firms are heterogeneous and that the largest also account for the most international trade.

One interesting question to pose at the outset is the direction of causality. Do more productive firms engage in exports or do firms become more productive because they export? Evidence from the United States strongly suggests that there is a *selection effect*: firms that are already more productive or larger tend to enter export markets. This suggests that there are sunk costs or fixed costs to enter foreign markets that only larger firms can absorb, as emphasized by Roberts and Tybout (1997) and Melitz (2003). However, there is little evidence to support the reverse idea that firms "learn by exporting" and see productivity gains from entering foreign markets. One exception is the study by Van Biesebroeck (2005), which finds evidence that exporting raises productivity among sub-Saharan African enterprises. This remains an active research area.

The more fundamental question raised by the theory of heterogeneous firms is whether more openness to trade, either by reductions in trade costs or through tariff cuts, tends to reallocate production among firms. Specifically, the primary prediction is that as countries become more integrated the competition should, in each market, force the least productive firms to exit and permit the most productive firms, the exporters, to expand. An alternative possibility is that firms will close their most inefficient plants and reallocate production to more productive locations.

These results are exactly what Nina Pavcnik (2002) found in a fascinating study of the reactions of Chilean manufacturing plants to trade liberalization. As noted earlier, Chile undertook a major reduction of trade barriers and liberalization of entry regulations in the late 1970s and early 1980s. Pavcnik's study is an excellent example of the importance of careful econometrics. Her goal was to

analyze whether productivity in individual manufacturing plants that faced greater import competition increased productivity by more than that in other plants. For this purpose, one needs a measure of efficiency, which she defined as TFP (total factor productivity) in a Cobb-Douglas econometric specification:

$$\log(y_{it}) = \beta_0 + \beta \log x_{it} + \beta_k \log(k_{it}) + \varepsilon_{it} \ , \ \varepsilon_{it} = \omega_{it} + \mu_{it} \qquad (13.45)$$

This equation states that gross output (y_{it}) of plant *i* at time *t* is produced by a vector of variable inputs (x_{it}), including intermediates, raw materials, and labor, plus the plant's capital input. In a naive model the residual term (ε_{it}) would be interpreted as TFP, or the output that cannot be explained by the combination of inputs. The problem is that this residual likely is composed of two parts: a plant-specific efficiency term (ω_{it}) and a random productivity shock (μ_{it}). The individual plant manager is aware of its efficiency term but the econometrician is not. Thus, incorporating it directly into the overall residual term will impart a simultaneity bias. This is because the manager will choose inputs based on her knowledge of plant efficiency, so the input variables are correlated with the error term. The result would be inconsistent ordinary least squares estimates of the coefficients and potentially large errors in calculating TFP. There is a further problem: plant and firm managers will decide whether to exit production based in part on this private information about efficiency. Thus, there is also a selection bias, in that plants that remain in operation (and therefore in the sample and estimation) would automatically be more efficient than exiting plants and this must be accounted for in the estimation.

To deal with simultaneity, Pavcnik incorporated a dynamic, non-parametric approach pioneered by Olley and Pakes (1996). In any period, each plant first makes a decision about whether to shut down operations, which is less likely for firms with larger capital stocks and also depends on expected market competition. In this theory, a plant will continue to produce and invest if its unobserved efficiency term (ω_{it}) exceeds a critical value depending on its existing capital stock. The efficiency term can then be expressed as a function of capital stock and current investment. Then this function can be substituted into (13.45) to explain gross output as a function of variable inputs, the capital stock, and a polynomial expansion in capital stock and investment terms. Incorporating this plant-specific and time-varying productivity relationship into (13.45) makes the resulting TFP estimation consistent with respect to simultaneity.

There remains the plant-selection bias. Pavcnik dealt with this bias by estimating the implicit threshold level of productivity below which a plant would exit by regressing the decisions of plants to continue producing on lagged investment values. Past investment should not affect current output but does influence future profitability, making it a good instrumental variable for exit decisions.

After making these adjustments for simultaneity and selection bias, Pavcnik estimated a second-stage TFP equation, in which gross output less variable inputs is a function of current and lagged capital stock and the observed probability of plant survival, computed across all plants in the data. This estimation used data on all Chilean manufacturing plants with at least ten workers over the period 1979 to 1986. Over 90 percent of these plants were actually single-plant firms. Based on trade patterns of the four-digit ISIC industry in which it produced, each plant was characterized as either export-oriented, non-traded, or import-competing. One initial result is that plant survival is a dynamic process in Chile. Of the plants in existence in 1979, 35 percent had ceased production by 1986. Of these, 13 percent were in export-oriented sectors and 40 percent were in the import-competing sector, suggesting that plant closures

were more common in import-competing sectors. In short, exit decisions are important in the wake of trade liberalization.

Turning to the econometric work, Pavcnik's estimated growth in productivity levels by plant could be compared by simple averages across all surviving plants and averages weighted by initial plant size. Her results found that the weighted- average productivity growth over seven years was positive in six of eight broad industries, ranging from 7.6 percent in machinery and equipment to 43 percent in chemicals. However, unweighted productivity grew only in food manufacturing and textiles. This result supports the view that firms adjust to greater import competition and export opportunities by shifting resources from less productive to more productive plants. Over time, the more productive plants produced a growing share of output in nearly all industries. For manufacturing as a whole, TFP rose by 19 percent, which was comprised of a 6.6 percent increase in within-plant efficiency and a 12.7 percent gain in productivity due to resource shifts among plants.

Her final task was to regress TFP across plants and time on year dummies, trade orientation, interaction terms between year dummies and trade orientation, and control variables. In her preferred specification the results were as follows. First, plants that exited in the period were on average 8.1 percent less productive than those that survived. Second, the coefficients on the interaction between import-competing orientation and the year fixed effects grew significantly over time. This implied that surviving import-competing plants became more productive, with efficiency gains associated with trade liberalization ranging from 3.0 percent to 10.4 percent. Finally, plants in export-oriented sectors did not experience a significant rise in productivity after the trade openness policies were adopted. As noted earlier, this likely reflects the fact that exporters were already more efficient and had little room for productivity growth.

To summarize, Chilean manufacturing firms reacted to greater import competition through a combination of plants shutting down, reallocation of resources to more productive plants, and higher productivity per surviving facility. All of these results are consistent with the notion of heterogeneous firms and responses to policy change.

Similar outcomes were found for U.S. manufacturing plants in a study by Bernard, et al (2006). Using detailed microeconomic data, they estimated the responses of plant productivity to reductions in industry-level import tariffs and measures of transportation costs from 1982 to 1997. They calculated that the average four-digit SIC industry saw a decline in shipping costs of 0.19 percent points per year in that period, and that average tariffs fell by more than 25 percent in the majority of sectors.

To summarize their result, the authors first regressed adjusted measures of TFP for each industry on trade costs (a combination of tariffs and transport charges) and found that a one-standard-deviation reduction in such barriers raised TFP growth by 0.19 percent per year, a significant outcome. Next, they found that the probability of a plant exiting production was positively and significantly affected by a fall in trade costs, suggesting that more openness to trade forces more plants out of existence. Specifically, a one-standard deviation decline in transportation costs increased the probability of exit by approximately five percent. On the other hand, a one-standard deviation fall in trade costs raised the probability that a non-exporting plant would enter foreign markets by 0.6 percent but the impact was significantly higher for higher-productivity facilities. In short, their main finding was that in response to falling transport costs, some plants, primarily lower-productivity non-exporters, were more likely to exit while higher-productivity non-exporters were more likely to start exporting. Each of these outcomes is consistent with a model of heterogeneous firm behavior.

13.9 Summary

Trade costs have long been somewhat neglected in both trade theory and empirical analysis. We conjecture that this may be due to an implicit assumption that trade costs are not inherently interesting: they simply put countries somewhere between autarky and free trade. We think this "between" statement is probably fair in models in which trade is motivated by comparative advantage. But even in this case there are important points to be made, such as unequal factor prices that give incentives for factors to migrate. More on this later in the book.

In models with increasing returns and imperfect competition however, there are quite a number of situations in which trade costs do not leave country in between autarky and free trade. We first consider a simple oligopoly model of the type analyzed in Chapter 11 and allowed firms to price discriminate between markets. We showed that trade costs in this environment can lead to two-way trade even in identical goods, a form of intra-industry trade, and to the pricing of exports below that of domestic sales (dubbed "reciprocal dumping" by Brander and Krugman (1983)).

Monopolistic-competition models generally assume that price discrimination is not possible (why, we don't know), yet several interesting features have been identified. First, countries that differ only in size will have different patterns of specialization in the presence of trade costs. The larger country will be relatively specialized in the differentiated goods sector, being a net exporter of differentiated goods to the small country and a net importer of the homogeneous competitive good. Thus intra and inter-industry trade will co-exist even though the countries are identical except for size. Each country will be relatively specialized in consuming the differentiated goods produced at home, sometimes referred to as the "home-market effect". This has further implications if some factors are mobile between countries and this will be discussed in Chapter 15.

Next, we look at a new class of models in which firms have heterogeneous productivities or costs. A full analysis of this is beyond the scope of this book . But firm heterogeneity is quite interesting in the presence of trade costs. In particular, a fall in trade costs leads to the exit of the least productive domestic firms, and the expansion of the most productive domestic firms into exporting as well. Falling trade costs now offer an additional source of gains from trade due to the rise in the average productivity of firms surviving in the market place. The heterogeneous firm models are popular in part because many of their predictions seem closely consistent with data, in particular observations that only a minority of firms export and firms that do export are larger and more productive than strictly domestic firms that do not export.

Finally, we present a brief analysis of the gravity equation, a widely-used empirical starting point for estimating the determinants of trade flows. Almost any theoretical approach provides a rationale for why trade flows decrease with the distance between trading partners, and the monopolistic competition model provides a nice rationale (theoretical foundation) for why trade flows should decrease with increased size *differences* between two countries, holding their total combined size constant.

REFERENCES

Alvaraez, Roberto and Ricardo A. Lopez, (2005), "Exporting and Performance: Evidence from Chilean Plants," *Canadian Journal of Economics* 38: 1384-1400.

Anderson, James E. (1979), "Theoretical foundation for the gravity equation", *American Economic Review* 1979, 106-116.

Anderson, James E. And Eric van Wincoop (2004), "Trade costs", *Journal of Economic Literature* 42, 691-751.

Aquino, Antonio (1981), "The Measurement of Intra-Industry Trade when Overall Trade is Imbalanced," *Review of World Economics*, 117, 763-766.

Balassa, Bela (1966), "Tariff Reductions and Trade in Manufactures among Industrial Countries," *American Economic Review*, 56: 466-473.

Balassa, Bela (1986), "Intra-Industry Specialization: A Cross-Country Analysis," *European Economic Review*, 30: 27-42.

Baldwin, Richard E. (2005), "Heterogeneous firms and trade: testable and untestable properties of the Melitz model", NBER working paper 11471.

Bergstrand, Jeffrey H. (1989), "The Generalized Gravity Equation, Monopolistic Competition, and the Factor-Proportions Theory in International Trade," *Review of Economics and Statistics*, 71: 143-153.

Bernard, Andrew B., J. Bradford Jensen, Stephen J. Redding and Peter K. Schott, (2007), "Firms in International Trade," *Journal of Economic Perspectives*, 21:105-130.

Bernard, Andrew B., J. Bradford Jensen, and Peter K. Schott, (2006), "Trade Costs, Firms and Productivity," *Journal of Monetary Economics*, 53: 917-937.

Brander, James A. and Paul R. Krugman (1983), "A reciprocal dumping model of international trade", *Journal of International Economics* 15, 313-321.

Evenett, Simon J. and Wolfgang Keller (2002), "On theories explaining the success of the gravity equation", *Journal of Political Economy* 110, 281-316.

Feenstra, Robert, James R. Markusen and Andrew Rose (2001), "Using the Gravity Equation to Differentiate among Alternative Theories of Trade", *Canadian Journal of Economics* 34, 430-447.

Feenstra Robert C. and Hiau Looi Kee, (2007), "Trade Liberalization and Export Variety: A Comparison of Mexico and China," *The World Economy*, 30: 5-21.

Frankel, Jeffrey, Ernesto Stein, and Shang-Jin Wei, (1998), "Continental Trading Blocs: Are They Natural or Super-Natural?" in Jeffrey Frankel, editor, *The Regionalization of the World Economy*

(Chicago: University of Chicago Press), 91-113.

Fujita, Masahisa, Paul R. Krugman and Anthony J. Venables (1999), *The spatial economy: cities, regions and international trade,* Cambridge: MIT Press.

Grether, Jean-Marie (1996), "Mexico, 1985-90: Trade Liberalization, Market Structure, and Manufacturing Performance," in Roberts and Tybout, 260-284.

Grubel, Herbert G. and Peter J. Lloyd, (1975), *Intra-Industry Trade: The Theory and Measurement of International Trade in Differentiated Products,* (New York: John Wiley).

Haddad, Mona, Jaime de Melo, and Brendan Horton, (1996), "Morocco, 1984-89: Trade Liberalization, Exports and Industrial Performance," in Roberts and Tybout, 285-313.

Hoekman, Bernard, Hiau Looi Kee, and Marcelo Olarreaga, (2004), "Tariffs, Entry Regulation and Markups: Country Size Matters," *Contributions to Macroeconomics,* 4, Article 8.

Konings, J. P. Van Cayseele, and F. Warzynski (2001), "The Dynamics of Industrial Markups in Two Small Open Economies: Does National Competition Policy Matter?" *International Journal of Industrial Organization,* 19, 841-859.

Krugman, Paul R. (1991), Increasing returns and economic geography, *Journal of Political Economy* 99, 483-499.

Krugman, Paul R. and Anthony J. Venables (1995), "Globalization and the inequality of Nations", *Quarterly Journal of Economics* 110, 857-880.

Levinsohn, James (1993), "Testing the Imports-as-Market-Discipline Hypothesis," *Journal of International Economics*, 35, 1-12.

Markusen, James R and Anthony J. Venables (2000), "The Theory of Endowment, Intra-Industry, and Multinational Trade", *Journal of International Economics* 52, 209-234.

Melitz, Mark J. (2003), "The impact of trade on intra-industry reallocations and aggregate productivity", Econometrica 71, 1695-1725.

Olley, S. and A. Pakes (1996), "The Dynamics of Productivity in the Telecommunications Equipment Industry," *Econometrica.* 64:1263-1297.

Overman, Henry G. Stephen Redding and Anthony J. Venables, (2003), "The Economic Geography of Trade, Production, and Income: A Survey of Empirics," in E. Kwan Choi and James Harrigan, editors, *Handbook of International Trade* (Malden MA: Blackwell Publishing, Ltd.)

Pavcnik, Nina (2002), "Trade Liberalization, Exit, and Productivity Improvement: Evidence from Chilean Plants," *Review of Economic Studies*, 69: 245-276.

Rauch, James R. (1999), "Networks versus Markets in International Trade," *Journal of International Economics*, 48: 7-37.

Roberts, Mark J. (1996), "Colombia: 1977-85: Producer Turnover, Margins, and Trade Exposure," in Roberts and Tybout, 227-259.

Roberts, Mark J. and James R. Tybout, (1996), *Industrial Evolution in Developing Countries: Micro Patterns of Turnover, Productivity and Market Structure* (Oxford: Oxford University Press).

Rose, Andrew K. (2004), "Do We Really Know That the WTO Increases Trade?" *American Economic Review* 94: 98-114.

Subramanian, Arvind and Shang-Jin Wei, (2007), "The WTO Promotes Trade, Strongly but Unevenly," *Journal of International Economics*, 72: 151-175.

Tybout, James R. (1996), "Chile, 1979-86:Trade Liberalization and Its Aftermath," in Roberts and Tybout, 200-226.

Tybout, James R. (2003), "Plant- and Firm-Level Evidence on "New" Trade Theories," in E. Kwan Choi and James Harrigan, editors, *Handbook of International Trade* (Malden MA: Blackwell Publishing, Ltd.)

Van Biesebroeck, Johannes, (2005), "Exporting Raises Productivity in Sub-Saharan African Manufacturing Firms," *Journal of International Economics* 67: 373-391.

ENDNOTES

1. Presenting the full general-equilibrium model is beyond the scope of this paper. There are a number of analytical complications and the solution involves taking integrals across firms over some probability distribution. The cost of entering the lottery is important in establishing an aggregate zero-profit condition and this in itself is complicated (recall the difficulty of having positive profits in general equilibrium that we discussed in Chapter 11). Firms enter the lottery up to the point where each has an expected payoff of zero. Firms with really bad draws don't subsequently enter production and so there ex post profits are negative and equal to the lottery entry cost, whereas the lucky firms make positive profits; but total profits ex post are zero. Incidentally, this assumes the existence of some sort of perfect underlying equities market, something rarely mentioned in this literature.

ENDNOTES

1. Presenting the full general-equilibrium model is beyond the scope of this paper. There are a number of analytical complications and the solution involves taking integrals across firms over some probability distribution. The cost of entering the lottery is important in establishing an aggregate zero-profit condition and this in itself is complicated (recall the difficulty of having positive profits in general equilibrium that we discussed in Chapter 11). Firms enter the lottery up to the point where each has an expected payoff of zero. Firms with really bad draws don't subsequently enter production and so there ex post profits are negative and equal to the lottery entry cost, whereas the lucky firms make positive profits; but total profits ex post are zero. Incidentally, this assumes the existence of some sort of perfect underlying equities market, something rarely mentioned in this literature.

Table 13.1 Regression Estimates of Impacts of Import Competition on Price-Cost Margins

			Independent Variables								
Country	*Period*	*Level*	*HERF*		*KSALES*	*IMP*	*HERF*IMP*	*TAR*	*HERF*TAR*	*Obs*	R^2
Chile	1979-85	Industry	0.184		-0.029	-0.032	0.531			196	0.85
Colombia	1977-85	Industry	1.069*		-0.001	-0.164	-0.740*			243	0.82
Morocco	1984-89	Industry	-0.314		-0.067	0.157*	0.312			108	0.98
Mexico	1985-90	Industry	0.162*		-0.066*			-0.088*	0.287	120	0.98
			SHARE	*SHARESQ*	*KSALES*	*IMP*	*SHARE*IMP*	*TAR*	*SHARE*TAR*	*Obs*	R^2
Chile	1979-85	Plant	2.469*	-3.225*	0.031*	0.009	-1.434*			22,174	0.45
Colombia	1977-85	Plant	1.507*	-2.327*	-0.530*	-0.063	-1.081*			51,340	0.01
Morocco	1984-89	Firm	0.798*	-0.64	-7.315*	0.157	-0.45			16,104	0.17
Mexico	1985-90	Plant	2.569*	6.376*	na			-0.02	1.969*	16,743	0.07

Notes: all regressions contain year and industry fixed effects. An asterisk indicates a coefficient that is significant at the 95% level of confidence.

Sources: for Chile, Tybout (1996); for Colombia, Roberts (1996); for Morocco, Haddad, et al (1996); for Mexico, Grether (1996).

Table 13.2 Intra-industry Trade by Commodity and Country, 2007

Country	Organic Chemicals	Iron & Steel Products	Industrial Machinery	Office Mach. & Computers	Passenger Vehicles	Prof. & Scient. Instruments	Apparel & Accessories	Alcoholic Beverages	Average
United States	84	81	59	85	73	61	16	44	63
Canada	79	76	71	46	85	66	32	39	62
Australia	8	72	43	24	38	65	11	45	38
Germany	87	93	57	77	57	78	60	75	73
UK	92	83	72	79	97	84	52	77	79
Japan	84	44	36	96	15	89	4	37	51
R. of Korea	81	91	94	64	14	58	60	62	65
Mexico	43	76	51	93	65	81	63	30	63
Brazil	68	48	89	13	94	23	51	29	52
China	51	75	61	52	77	57	4	49	53
India	85	81	52	27	23	41	2	99	51
Average	69	75	62	60	58	64	32	53	

Source: compiled by authors from United Nations, COMTRADE database

Table 13.3 Tariff Cuts and Growth in Import Varieties from Mexico and China to the United States, 1990-2001							
	U.S. Tariffs on Imports from Mexico						
	All Industries	Agriculture	Textiles & Apparel	Wood & Paper	Mining & Metals	Machinery & Transport	Electronics
1990	4.1	4.4	13	2.2	2.1	2.5	4.1
2001	0.3	0.8	0.4	0	0.7	0.1	0.1
	Percent of U.S. Imported Varieties from Mexico						
1990	52	42	71	47	47	66	40
2001	67	51	83	63	56	76	66
	U.S. Tariffs on Imports from China						
1990	5.8	1.4	13	6.6	6.4	5.1	5.1
2001	3.6	1.7	10.9	1.3	4.1	3	1.2
	Percent of U.S. Imported Varieties from China						
1990	42	30	79	52	31	28	35
2001	63	34	88	65	55	63	68
Source: Feenstra and Kee (2007), Table 1 and Table 2							

Table 13.4 Estimation of a Gravity Model with Trade Blocs	
Dependent Variable: log (bilateral exports plus imports)	
Independent Variable	Coefficient
Intercept	-9.70*
GNP* GNP	0.72*
PCGNP* PCGNP	0.23*
Distance	-0.51*
Adjacency	0.72*
Common language	0.47*
EC bloc	0.31*
WH bloc	0.31*
EA bloc	2.12*
1980 fixed effect	-1.01*
1990 fixed effect	-1.29*
Observations	4555
R^2	0.76
Note: * indicates coefficient is significant at one-percent level	
Source: Frankel, et al (1998), Table 4.2	

Table 13.5 Estimation of a Gravity Model and Differentiated Goods		
Dependent Variable: log (bilateral exports), 1990		
	Differentiated	*Homogeneous*
Exporter GDP	1.12*	0.54*
Importer GDP	0.72*	0.81*
Distance	-1.10*	-0.89*
Adjacency	-0.00	0.26
Language	30.69*	0.61*
FTA	1.73*	1.06
Remoteness	794*	384*
Observations	6367	5095
R^2	0.57	0.4
Note: * indicates coefficient is significant at one-percent level		
Source: Feenstra, et al (2001).		

Table 13.6 U.S. Export Firm Characteristics				
	% of Firms	*Mean Exports*		
	that Export	*as % of Sales*		
All Manufacturing	18	14		
Low (Printing)	5	14		
Low (Apparel)	8	14		
High (Computer & Electronic)	38	21		
High (Elect. Equipment)	38	13		
	Exporter			
	Premia			
Employment	97%			
Value Added per Worker	11%			
TFP	3%			
Wage	6%			
Capital per Worker	12%			
Skill per Worker	11%			
Number of Products	27%			
Shipments per Product	73%			
Notes: data are from 2002 U.S. Census of Manufactures except number of				
products and shipments per products, which are from 1997 Census.				
Source: Bernard, et al (2007)				

Figure 13.1

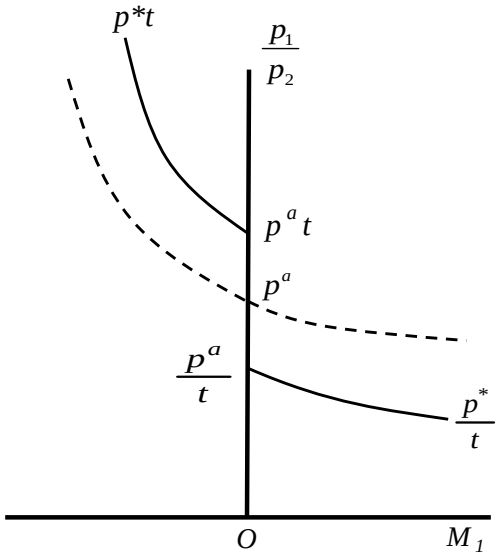


Figure 13.2

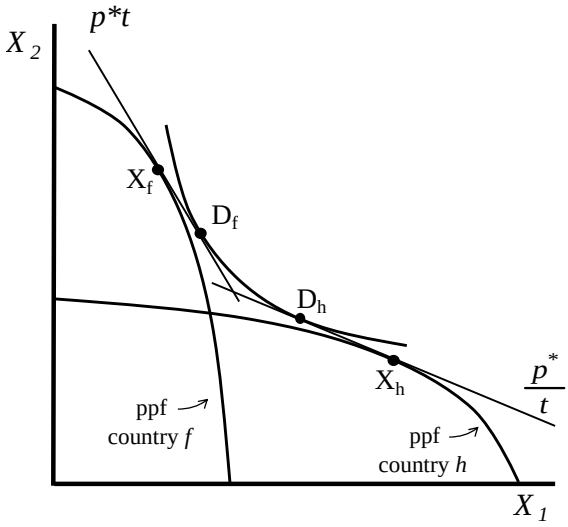


Figure 13.3: Trade costs in the "reciprocal dumping model

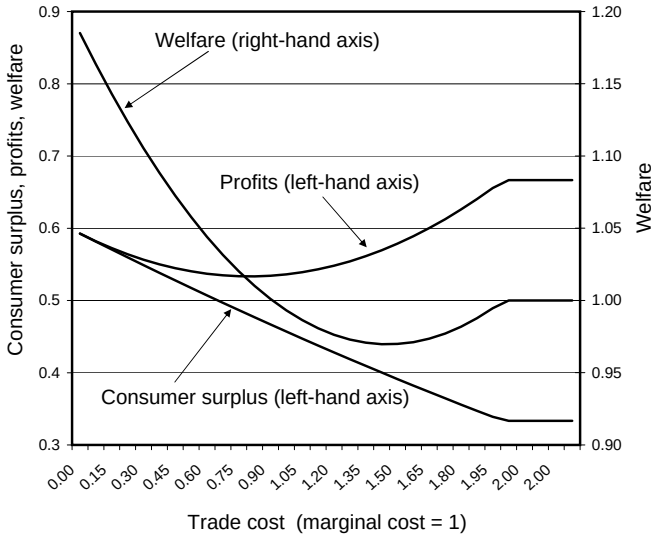


Figure 13.4 Trade costs and the home-market effect: one-factor model

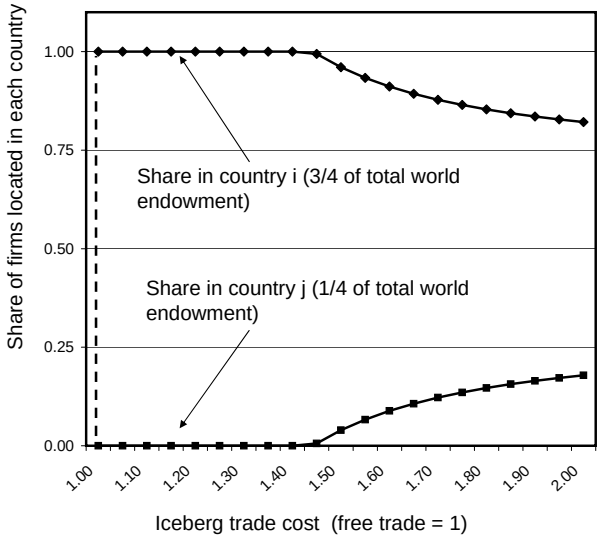


Figure 13.5 Trade costs and the home-market effect: two-factor model

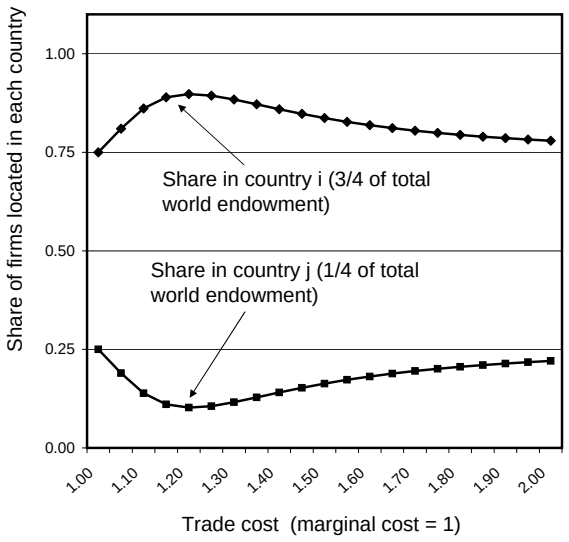
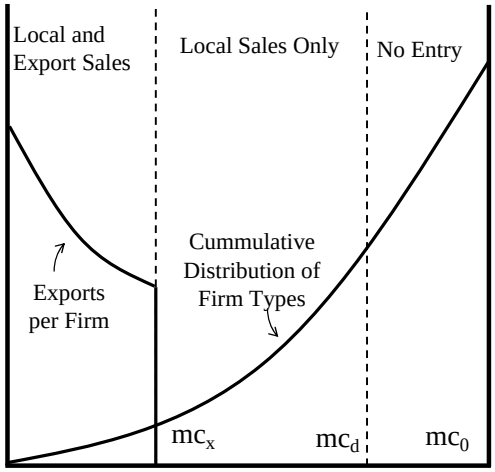
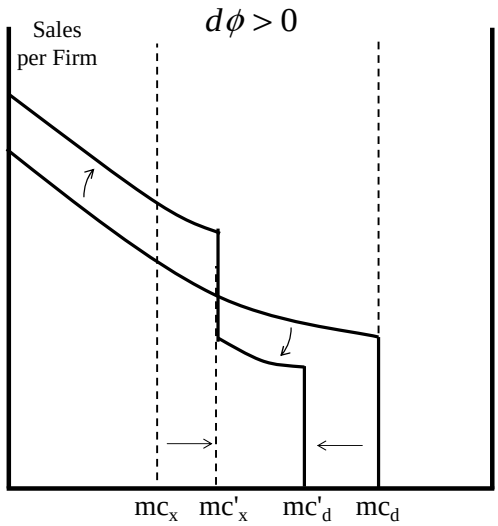


Figure 13.6



Marginal Cost

Figure 13.7



Chapter 14

PREFERENCES, PER-CAPITA INCOME AND PRODUCT QUALITY

14.1 Preferences and per-capita income as a determinant of trade

Our models to this point have focused on the production side of the general-equilibrium structure of economies to describe the causes of trade and the sources of gains from trade. This perspective is representative of trade theory in general, which has devoted the overwhelming share of its attention to production, almost neglecting consumption entirely. Many models assume that consumers have identical and homogeneous utility functions, regardless of location. Then if commodity prices were equalized by trade, consumers everywhere would demand goods in the same proportions. All trade would then be due to various differences in production among countries. As students have by now realized, this emphasis generates a markedly diverse and sophisticated set of predictions about the supply-side determinants of comparative advantage, such as differences in factor endowments, policy-based market distortions, and imperfect competition associated with increasing returns to scale.

There has been much less theoretical attention paid to demand-based determinants of trade. There has also been little empirical examination of the apparent presumption by trade economists that production differences are more important than consumption differences for explaining trade patterns. This phenomenon is curious in that the assumption of identical and homogeneous preferences across countries is easily shown to be questionable. For example, in 1991 59% of household consumption in Bangladesh was estimated to have been devoted to food. The corresponding shares of food in household budgets in other countries were: Indonesia 48%, Argentina 35%, Greece 30%, Japan 17%, and United States 10% (World Development Report, 1993). In contrast, budget shares allocated to medical care were: Bangladesh 2%, Indonesia 2%, Argentina 4%, Greece 6%, Japan 10%, and United States 14%. In part, these differences reflect variations in relative prices due to trade barriers and government policies. However, they clearly indicate that preferences are not homogeneous: as per-capita income rises, the relative share of income spend on food declines, and that on medical care rises. They might also suggest that tastes are not identical as well, though this cannot be rigorously inferred from share data alone. However, an obvious difference that would affect preferences is that the industrialized countries tend to have larger percentages of their populations at advanced ages, increasing the demand for health care. Accordingly, one purpose of this chapter is to bring together some strands of literature in order to examine the possible ways in which international differences in consumption patterns can influence trade flows.

Identical in production, different in tastes

It is surely true that countries differ in their aggregate preferences over different sets of goods and services. Figure 14.1 makes the point in the simplest possible way. Each of two countries have the same identical production frontier, but country h has a relative preference for good X_1 and country f has a relative preference for good X_2. Gains from trade are possible. Both countries could produce at the common point X, and trade to consumption points D_h and D_f for countries h and f respectively. Each country imports the good for which it has a relatively high preference in demand.

International trade economists have long been a bit uncomfortable with trade generated in this

way, since arbitrary preference differences may be a cheap way to explain a great deal of trade. Clearly, people in hot, tropical climates are going to eat different food, dress differently, and build their house differently from consumers in colder climates. Consumption patterns can also reflect differences in culture and religion. But economists tend to look for systematic factors that cause trade rather than somewhat arbitrary differences in preferences are unexplained by any systematic underlying factors.

Figure 14.1

Home bias in preferences

A somewhat less arbitrary and empirical-supported idea is referred to as "home bias", here used to mean something different than in Chapter 13. This is the notion that consumers have a systematic preference for home produced varieties of goods over foreign ones. This may be due to some subtle underlying product differentiation (packaging, size, color, spice and other characteristics of food, etc.) in which domestic producers target their products for domestic tastes. But even with perfectly homogeneous goods, consumers might prefer beef in a beef-producing country and rice in a rice-producing country.

This idea is shown in Figure 14.2 where country h has a comparative advantage in X_1 and country f has a comparative advantage in X_2, the two countries producing at points X_h and X_f respectively. Point A would be the common consumption point at world price p^* with homogeneous preferences and no home bias. Now suppose that consumers have a bias in favor of local goods. Then consumers in country h might want to shift from point A in figure 14.2 to point D_h, substituting away from imported good X_2 and toward locally produced X_1. Consumers in country f might want to shift from A to point D_f substituting its local X_2 for imported X_1.

Figure 14.2

Trade is reduced by home bias and there is a positive correlation between a country's pattern of specialization in consumption and its specialization in consumption; e.g., country h is relatively specialized in producing X_1 and also relatively specialized in consuming X_1.

One problem with this argument is that it is hard to discriminate empirically between this taste-based home bias and that arising from trade costs. Refer back to Figure 13.2. Here we see this same correlation between specialization in consumption and specialization in production, but it is occurring with countries having identical, homogeneous preferences.

One possible way to discriminate between taste-based and trade-cost-based causes of home bias is to use the value shares of goods in consumption instead of quantities. In Figure 14.2, in which the prices of goods are the same across countries, it is unambiguous that country h spends a larger share of its income on good X_1. But in Figure 13.2, the price of a good is high where its consumption quantity is low. If preferences are Cobb-Douglas, then we have seen several times that the shares spent on goods are constants: a ten percent higher price means ten percent lower quantity, so the share spent on each good is constant and independent of prices. In this special case, the shares of income spent on X_1 and X_2 by countries h and f in Figure 13.2 are the same, and the hypothesis of taste-based home bias can be rejected.

Non-homogeneous demand

A final contribution of preferences and demand to explaining trade flows lies in the role for per-capita income. Trade theory has long used the assumption of identical and homogeneous preferences between countries in order to focus on production as a determinant of trade, not because is a good assumption. In fact, probably every budget study ever done has shown very substantial deviations from homogeneity: some goods, such as food, have much larger budget shares at lower incomes while others such as medical care and recreation have much higher budget shares at higher incomes. Instead of countries differing in preferences, another pure case is that countries have identical preferences that a/re non-homogeneous. This is the assumption we will make in the next couple of sections.

A possible situation for a single consumer is shown in Figure 14.3 and we will use the underlying utility function in section 14.3 below. Suppose that there is some minimum consumption required for good X_2 in order to survive or before any X_1 will be purchased. This amount is denoted by the point γ in Figure 14.3. Once the consumer can afford this amount, consumption follows the path shown in Figure 14.3 with the slope of this income-consumption curve constant (linear Engel's curve) but depending on relative prices. With these preferences, assumed to be identical across two countries, the share spent on good X_1 will rise with per capita income and the share spent on X_2 falls. In alternative terminology, the income elasticity of demand for X_1 will exceed one and the income elasticity of demand for X_2 will be less than one: a doubling of income will less than double the consumption of X_2 at constant prices. X_2 is sometimes referred to as a "necessity" and X_1 is called a "luxury".

Figure 14.3

Table 14.1 presents some estimates of income elasticities of demand taken from Hunter and Markusen (1989). These are fairly broad categories, but even at this level of aggregation, it is clear that there are substantial deviations from homogeneity, which predicts all income elasticities should be one: a doubling of income doubles the demand for all goods and services under homogeneity.

Table 14.1

Identical but non-homogeneous preferences as in Figure 14.3 can serve as a basis for trade. Consider Figure 14.4 which shows two production frontiers on a per capita basis: these are alternative output bundles that one worker can produce, with workers more productive in country h. There is no pattern of comparative advantage but country h workers have an absolute advantage in both goods. A trading equilibrium at world price ratio p^* involves both countries producing the goods in the same proportions; their production points are denoted X_f and X_h respectively. But country h will have a relative preference for X_1, importing that good in exchange for X_2 and vice versa for country f.

Figure 14.4

The next section considers an important approach to trade that is based on non-homogeneous preferences. Following that, we incorporate these ideas into a model that combines this approach with elements of Heckscher-Ohlin and monopolistic competition in order to explain some important features of world trade flows.

14.2 The Linder Hypothesis

The analysis of the possible role of per-capita income in determining trade leads naturally to a discussion of the ideas of Swedish economist Staffan Linder (1961). Linder argued that the principles governing trade in manufacturing goods are not the same as those governing trade in primary products. He was quite prepared to support the idea that trade in primary products is determined by factor endowments. However, he argued against the notion that differences in factor endowments are the major determinants of trade in manufactured products. He chose, instead, to highlight the role of consumption demand, beginning his argument with an observation similar to the one we made in our chapter on increasing returns to scale: a large volume of trade exists between the developed countries. These countries have very similar factor endowments and thus, according to the Heckscher-Ohlin theory, we might not expect a large volume of trade between them. We must, therefore, look for a cause of trade other than factor endowments.

Linder contended that a manufactured good is created by an innovative entrepreneur in response to a perceived demand. A new manufactured good is, in other words, introduced only when an entrepreneur believes there is sufficient potential demand to warrant production. It is this perception of potential demand that triggers production rather than considerations of factor endowments. The second step in the argument is that entrepreneurs are most familiar with their home market. Barriers of distance, language, and culture indicate entrepreneurs are much less familiar with foreign markets and are unlikely to be able to perceive what sort of new products could be successfully introduced into foreign markets.

The third step follows from the first two. For a manufactured product to be produced in (and therefore potentially exported from) a country, there must exist significant home demand for the product. The range of manufactured goods produced in a country is, therefore, determined by domestic demand as much as by production considerations, such as factor endowments.

Suppose then that an entrepreneur perceives a home demand for a product and begins production. Where will he find export opportunities for additional sales? Linder argued that the best opportunities will be found in countries that have very similar demand patterns to the entrepreneur's home country. Thus, if an American invents a new communications device, such as a modem or telefax machine, and produces it for the home market, his best export opportunities will be found in Canada, Western Europe, and Japan. Similarly, entrepreneurs in Canada, Western Europe, and Japan will find that their best foreign markets will be in the United States and in each other's countries.

The final step in Linder's argument is that the countries with the most similar demand patterns for manufactured goods will tend to be those with similar per-capita incomes. People in countries with lower per-capita incomes may wish to buy relatively simple, inexpensive products. However, people in countries with much higher per-capita incomes may want more sophisticated devices, such as coffee makers with flashing lights, digital readout, and timers. Thus the volume of trade in manufactured goods will be highest among countries of similar per-capita incomes, such as the United States, Canada, Western Europe, and Japan. Predictions about which specific products each country would export are difficult to make, because such exports would depend on the history of entrepreneurial activity in each market. Overall, however, international patterns of income and demand determine the extent of trade in manufactured goods, rather than only factor endowments or technology. Similarity in demand is, in turn, related to similarity in per-capita income.

14.3 Integrating Linder, monopolistic competition and non-homogeneous preferences

An important part of Linder's ideas emerged as the monopolistic-competition approach to trade which we analyzed in Chapter 12. Individual entrepreneurs enter a market with unique varieties of differentiated goods, which are sold to consumers in all countries. Thus, there can be a substantial volume of trade even between identical countries. The most popular vehicle for these models is the Dixit-Stiglitz utility function and large-group monopolistic competition. However, this extensive literature left out an important part of Linder's theory, the role of per capita income. The monopolistic-competition approach to trade assumes homogeneous preferences over both differentiated goods and between these goods and another sector, usually a competitive, constant-returns sector.

This leaves the standard monopolistic-competition approach sufficient to explain the large volume of trade between similar "northern" (high per-capita income) countries, but it cannot contribute to explaining the small volume of trade between north and south (lower income developing countries) and between southern countries themselves, a point made in Markusen and Wigle (1990). Although much more empirical work is needed, a theoretical framework which combines Linder, monopolistic-competition and non-homogeneous preferences is found in Markusen (1986). Here, there is a standard two-good (*X* and *Y*) two-factor (*K* and *L*) Heckscher-Ohlin model. The *X* sector is capital intensive and produces differentiated Dixit-Stiglitz goods in the manner analyzed in Chapter 12. The big twist is that there is a "minimum consumption requirement" in the *Y* sector, giving us what is often referred to as the Stone-Geary utility function. Let lower case, *x* and *y* refer to *per-capita* quantities. Preferences of a single worker are given as follows.

$$u = \left[\sum x_i^{\alpha}\right]^{\beta/\alpha}(y-\gamma)^{1-\beta} \qquad 0 < \alpha, \beta < 1 \tag{14.1}$$

where γ is the minimum consumption requirement. Denoting x_c as a composite of the *X* varieties, the preferences in (14.1) are exactly like those shown in Figure 14.3, except X_1 in Figure 14.3 now becomes the X_c composite, and X_2 in Figure 14.3 becomes *Y*. Geometrically, we could think of γ as displacing the origin of the utility function to point γ in Figure 14.1. Preferences as Cobb-Douglas with respect to this displaced origin, *X* goods have an income elasticity of demand greater than one.

Following our work in Chapter 12, the *X* composite good and its price index (the unit expenditure function: the cost of buying on unit of x_c) are as follows.

$$x_c = \left[\sum x_i^{\alpha}\right]^{1/\alpha} \qquad p_c = \left[\sum p_i^{1-\sigma}\right]^{\frac{1}{1-\sigma}} \qquad \sigma = \frac{1}{1-\alpha} \tag{14.2}$$

Let 'i' denote the income of a single worker. Also, let good *Y* be numeraire so its price is equal to one. If we solve the optimization problem of maximizing (14.1) subject to income, an individual worker's demands for the *X* composite and good *Y* are given by

$$x_c = \beta(i-\gamma)/p_c \qquad y = \gamma + (1-\beta)(i-\gamma) \tag{14.3}$$

The expression (i - γ) can be thought of as the "discretionary" income left over after meeting the minimum consumption requirement of good *Y*. Consumers spend a constant fraction β of their discretionary income on *X* goods and a fraction (1 - β) on *Y* plus the minimum consumption requirement spending. Now multiply (14.3) through by L, the total number of workers and let I denote aggregate

income.

$$X_c = \beta(I - \gamma L)/p_c \qquad Y = \gamma L + (1 - \beta)(I - \gamma L) \qquad I = iL \tag{14.4}$$

Assume that the world consists of three countries: east (E), west (W), and south (S). E and W are identical and together they are called north (N). L_n workers live in N and L_s workers live in S. We can write the demands for the *X* composite and *Y* in N and S as follows:

$$X_{nc} = -\beta\gamma L_n/p_c + \beta I_n/p_c \qquad Y_n = \beta\gamma L_n + (1 - \beta)I_n \tag{14.5}$$

$$X_{sc} = -\beta\gamma L_s/p_c + \beta I_s/p_c \qquad Y_s = \beta\gamma L_s + (1 - \beta)I_s$$

First, assume that the north and south are completely identical (and of course E and W are identical halves of N). Suppose that we solve for the integrated world equilibrium and find that the wage-rental ratio *w/r* is that given in Figure 14.5. Both N and S have (identical) endowments at point *A* in that diagram. The corresponding output diagram is Figure 14.6, where *A* is the common production and consumption point of N and S, and the dotted line through *A* gives the Engels' curve through that point consistent with (14.5). There will be no inter-industry trade in *X* goods for *Y* but there will be intra-industry trade in the *X* varieties.

Figure 14.5 **Figure 14.6**

Now slide the endowments of N and S apart in Figure 14.5, transferring capital from S to N and transferring labor from N to S in the ratio *w/r* until the countries are both specialized: N is at point *B* and S is at point *C*. This movement preserves the integrated world equilibrium and keeps the factor prices constant. The production frontiers, originally identical at TT', now become SS' and NN' in Figure 14.6 for S and N, respectively. With identical and homogeneous preferences between countries, consumption would remain at point A in Figure 14.6. But that is not the case here, since per-capita income is rising in N and falling in S. Note from (14.5) that the intercept of the Engels' curve on the *Y* axis moves up in country S (more individuals to each satisfy their minimum consumption requirement) and the Engels' curve shifts down in N. These changes perfectly offset one another as seen from (14.5) so that the integrated equilibrium is preserved. Consumption will be at points D_s and D_n for S and N respectively in Figure 14.6.

North-south trade is purely inter-industry in Figure 14.6 with both regions specialized. The first thing we can conclude from Figure 14.6 is that the volume of north-south trade and hence the volume of inter-industry trade will be less than would be the case with homogeneous preferences. The second thing that must be true is that the North will be more specialized in consuming the differentiated *X* varieties than would be the case if preferences are homogeneous. However, combining this finding with the fall in north-south trade means the volume of trade between E and W is bigger than it would be with homogeneous preferences. More of each of east's production of a given variety goes to west rather than to south and similarly for west's varieties. As an example, with homogeneous preferences in Figure 14.6, east might send half of the output of each variety to south and one-quarter to west, retaining one-quarter at home and importing one-quarter of each of west's varieties. With non-homogeneous demand, east will send less than one-half of the output of each variety to south and more to west, and similarly for the exports of west. Thus non-homogeneity reduces the volume of north-south trade and increases the volume of east-west trade.

This result depends crucially on the assumption that the differentiated goods sector is both capital intensive and that these goods have a high-income elasticity of demand. These assumptions imply that each region has a relatively high demand for its own export good and, because the north's goods are differentiated and cross traded, they also imply a large north-north trade volume. The role of per-capita income, so important in Linder's story, is restored. There is not a great deal of empirical work on this issue, but Hunter (1991) find very good support for the idea that per capita income differences lead countries to be relatively more specialized in consuming their own export goods relative to a situation in which preferences are homogeneous. This in turn suggests that labor intensive goods tend to have lower income elasticities of demand than capital or skill-intensive goods. Stronger support for non-homogeneity and that its effects are consistent with this Markusen model is found in a recent paper by Cassing and Nishioka (2011).

14.4 Product quality and willingness to pay

All of our analysis of product diversity to this point in the book involves "horizontally" differentiated products with trade benefitting consumers by providing a larger range of products. But surely an important part of the gains from trade must be through quality upgrading. Trade can lead to better products rather than simply more products.

Product quality is the topic of this section. We will adopt the framework from Shaked and Sutton (1983, 1984). The Shaked and Sutton framework is complex and we cannot possibly do justice to the full theory here, so we will use a special case of the 1984 paper to talk about the benefits of trade. While we might have placed this section in the monopolistic-competition chapter, the theory relies very much on income differences across consumers and on the assumption that a consumer's willingness to pay for quality is increasing in income. So we have, rather arbitrarily, placed this discussion in this chapter due to the centrality of per-capita income to its story.

The theory proceeds as a three-stage game. First, firms decide whether or not to enter an industry. Second, the entering firms choose their respective qualities. In the third stage, firms compete by choosing prices given the prices of their rival firms. In keeping with the simple special case, assume that the costs of producing a higher quality product are entirely in added fixed costs (e.g., R&D), and that the marginal cost of production not only does not depend on quality. For simplicity we actually set marginal cost to zero.

The basic model has some flavor of the "ideal variety" story of Chapter 12. Consumers will only buy one unit of good *X*. However, now we assume that the consumers have identical preferences and that their preferred variety over a set of alternatives depends on their income level. Let a number of firms *n*, indexed by *k*, each sell a single product of some quality level, where quality is increasing in the index *k* (product X_k has better quality than product X_{k-1}). Let U_k denote the *quality* of product k. Let *I* denote the consumer's income, and let p_k denote the price of product *k*. Then a consumer of income I derives utility

$$U_k(I - p_k) \tag{14.6}$$

from consuming one unit of product quality U_k and purchasing $(I - p_k)$ of other goods.

Firms sell to a number of consumers who differ in income. The distribution of incomes is given by a uniform density function over a range of consumer "types", where type is synonymous with income

level. There is a density parameter *s* which we should think of as the *number* of consumers at each income level. I_r denotes the income of the richest consumer and I_p the income of the poorest consumer, so $(I_r - I_p)$ is the total range of consumer "types", with *s* of them at each income level in this interval.

Consider a consumer who is indifferent between consuming good *k-1* at price p_{k-1} or good *k* at price p_k. Then

$$U_k(I - p_k) = U_{k-1}(I - p_{k-1}) \quad => \quad p_k = \left[\frac{U_k - U_{k-1}}{U_k}\right] I + \frac{U_{k-1}}{U_k} p_{k-1} \tag{14.7}$$

The first equation can be interpreted as implicitly giving the level of income for which the consumer is indifferent between the two goods at their respective prices. Relative to the income of the consumer who is indifferent in (14.7), higher income consumers will strictly prefer the higher-quality good *k* and lower-income consumers will strictly prefer the lower-quality good *k-1*. The second equation rearranges the first, and can be interpreted as the "willingness to pay" for the high-quality good *k* when the low-quality good *k-1* is available at price p_{k-1}. Higher income consumers have a higher willingness to pay for *k* given the price of *k*-1.

Consider the final stage of the game first: firms have picked their quality levels and must decide on price, which is closely related to how many income types they serve (the density parameter *s* plays no role at the moment, but will reappear later). Suppose that there are just two products, X_k and X_{k-1}. The situation facing firm *k*, the high-quality producer is shown in Figure 14.7. Assume first that the product X_{k-1} sells for its marginal cost, which is zero: $p_{k-1} = 0$. Then the demand curve or willingness-to-pay curve facing firm *k* is given by the lower demand curve in Figure 14.7.

Figure 14.7

This demand curve is constructed by lining up the consumer types, with the *highest* income consumer on the *left* (and therefore highest willingness to pay), moving to lower income consumers as we move left to right. If the lowest income is $I_p = 0$, the demand curve continues all the way to the horizontal axis, and the length of the horizontal axis from the origin to the demand intercept is I_r, the total number of consumer types between I_r and $I_p = 0$. If the lowest income level is $I_p > 0$, then the demand curve cuts off at the income at the point $(I_r - I_p)$ since there are fewer number of consumer types. The lowest income type now has positive income and will be willing to pay a positive price for good *k* given $p_{k-1} = 0$.

Continue to assume that $p_{k-1} = 0$. If the firm only wants to sell to the richest consumer, (14.7) indicates that it can set a price $p_k = [(U_k - U_{k-1})/U_k]I_r$ given $p_{k-1} = 0$, and this is the intercept of the demand curve on the vertical axis. If firm k wants to sell to one additional consumer type with an income one unit lower than its current marginal or "cutoff" consumer type, then from (14.7) it must reduce its price by $(U_k - U_{k-1})/U_k$. If the firm wants to sell to a total of X_n consumer types, it must then offer a price

$$p_k = \left[\frac{U_k - U_{k-1}}{U_k}\right] I_r - \left[\frac{U_k - U_{k-1}}{U_k}\right] X_n = a - bX_n \tag{14.8}$$

$$a \equiv \left[\frac{U_k - U_{k-1}}{U_k}\right] I_r \qquad b \equiv \left[\frac{U_k - U_{k-1}}{U_k}\right]$$

which is the equation of the demand curve in Figure 14.7. Firm *k*'s profits are just its price times quantity, since marginal cost is zero, times *s*, the density (number of) consumers of each type *k*. Profits and the first-order condition for profit maximization (for a fixed quality) are given as follows:

$$\pi = s p_k X_k = s\left[(a - bX_k)X_k\right] \qquad \frac{d\pi}{dX_k} = s(a - 2bX_k) = 0 \qquad X_k = \frac{a/b}{2} \qquad (14.9)$$

From (14.8), note that (*a/b*) is just I_r, the point at which the demand curve in Figure 14.7 hits the horizontal axis (demand price is zero). Thus firm *k*'s optimal number of consumer types served is half the distance between the origin and I_r and its price is half the maximum willingness to pay. Note for future reference that, while *s* clearly increases total profits, it does not affect the optimal choice of price and range of consumer types served (again holding qualities fixed).

The important point to note is that, if the span of consumers' incomes is not too great, in particular if $I_p > I_r/2$ or $(I_r - I_p) < I_r/2$ in Figure 14.7 (the demand curve cuts off before reaching the half-way point between 0 and I_r), then firm *k* will set a price that covers the entire market. Firm *k-1* cannot sell anything even at a price $p_{k-1} = 0$. This is a natural monopoly case and contrasts sharply with the monopolistic-competition model of horizontal product differentiation.

Now assume that the range of incomes is larger so that $I_p < I_r/2$ or $(I_r - I_p) > I_r/2$ as shown Figure 14.7. Then firm *k-1* can cover some of the market at a positive price $p_{k-1} > 0$. From (14.7), this shifts up the demand curve for firm *k* as shown in Figure 14.7 since consumers' outside option (good *k*-1) is now more costly. Shaked and Sutton also show that if the range of incomes is in the interval $I_r/4 < I_p < I_r/2$, then exactly two firms can enter and no additional firm can enter and sell at a positive price. They term this a "natural oligopoly".

Having established conditions under which there are exactly two firms in the market, we can then complete the argument by analyzing what qualities are chosen in equilibrium. This is shown in Figure 14.8 where *F(U)* denotes the fixed cost of producing quality *U*. The slope of *F(U)* gives the marginal cost of an additional improvement in quality. Figure 14.8 then shows the Nash equilibrium choices for two firms, firm 2 being the high-quality firm and firm 1 being the low-quality firm. Curve Π_1 shows the profits before fixed costs for firm 1 given firm 2 chooses quality level U_2. Firm 1's choice satisfies the condition that the marginal improvement in profits from an additional quality improvement (slope of Π_1) equals the slope of the marginal cost of an additional quality improvement (slope of *F(U)*). Quality level U_2 similarly satisfies the optimality condition for higher-quality firm 2, given the choice of firm 1.

Figure 14.8

Finally, consider the opening of trade between two identical economies as we have done before. Trade doubles the density *s* of consumers at each point of the income distribution. Doubling the density of consumers increases the marginal return to a quality improvement, which means that a higher *s* must increase the slopes of Π_1 and Π_2 in Figure 14.8 in the region of positive slopes. Equating these slopes with the slope of the quality improvement curve *F(U)* in Figure 14.8 means that both firms will chose higher qualities than before trade. The benefits of trade are *quality upgrading* of a *fixed range* of

products, not more products of fixed quality as in the monopolistic-competition model of horizontal product differentiation.

14.5 Empirical evidence on preferences, quality and trade

Many of the questions raised in this Chapter are the subject of extensive current empirical study. Of particular interest are the roles of per-capita income differences and non-homogeneous preferences on international trade and the interactions between factor endowments, income levels, and quality in international trade.

Per-capita income and trade volumes

As described in Table 14.1, it is clear that the income elasticity of demand varies across goods. This means that the shares of a country's national expenditure spent on particular goods vary with levels of per-capita income. For example, Hunter (1991) listed figures showing that countries with annual per-capita income less than $750 in 1973 spent 50 percent of income on food, five percent on house furnishings, and four percent on medical care. In contrast, countries with per-capita incomes between $3,800 and $5,200 spent 17 percent on food, nine percent on house furnishings, and nine percent on medical care. Using a different set of commodity categories, Cassing and Nishioka (2011) demonstrated that a set of high-income developed economies in 2000 devoted about two percent of income to agricultural goods, seven percent to food products, and 23 percent to various business services. For a selection of low-income developing countries those shares were nine percent, 15 percent, and 14 percent respectively. On the basis of such figures it is fair to say that food products are a "necessity" good with low income elasticities and many services are a "luxury" good with high income elasticities.

This idea is quite different from the standard trade-theory assumption that preferences are identical and homogeneous across countries at all levels of income. How does this difference seem to affect the volume of trade? Hunter's (1991) early study is an interesting attempt to sort out this question. To understand her approach consider Figure 14.9. Suppose that country h has a low per-capita income and country f has a high per-capita income. Assume further that country h has a supply-based comparative advantage in good X_2, perhaps because that good is labor-intensive and h is labor-abundant, while f has a comparative advantage in good X_1. The line labeled C^* indicates what consumption points would be (C^{h*}, C^{f*}) if tastes were identical and homogeneous in these two countries and if they faced a common international price ratio P^*. Meanwhile, because of comparative advantage actual production points are at Q^h and Q^f. (Readers can imagine production possibility frontiers with tangencies to P^* at those points.) Finally, let tastes be non-homogeneous in the manner of Figure 14.3. Thus, in free trade actual consumption points are C^h and C^f in Figure 14.9, while the volumes of trade are given by the lines P^* through these production and consumption points. Thus, country h exports X_2 and country f exports X_1. However, it is immediately obvious that the impact of non-homogeneous preferences, making X_2 the necessity and X_1 the luxury, is to *diminish* the volumes of trade.

Figure 14.9

Hunter's study assumed that tastes across countries bear this simple non-homogeneous form and developed a means of neutralizing these income-based impacts on consumption. Her task was to estimate what the consumption levels (C^{h*}, C^{f*}) would be if preferences were homogeneous, noting that the additional trade generated would be a measure of the amount by which trade is reduced by non-

homogeneity. For this purpose, she put together a sample of 21 countries and 13 broad consumption goods using 1973 data. One obvious initial problem is that these data were not, in fact, generated under conditions of equal relative commodity prices, for at least two basic reasons. First, prices differ because, for reasons discussed in Chapter 1, wages and land prices vary significantly across countries, meaning that commodity prices measured at market exchange rates can severely misrepresent actual purchasing power. To reconcile this problem, Hunter used purchasing-power parity (PPP) exchange rates to compute implied relative prices of goods and services across nations. Second, trade costs and trade barriers generate differences in relative prices across nations even at PPP exchange rates. Here Hunter makes the strong assumption that once consumption patterns are neutralized for per-capita income differences the resulting demands arise from Cobb-Douglas preferences. In that case consumption shares depend only on income levels, not on relative prices, permitting her to ignore the existence of relative price differences in the final stage.

Thus, Hunter's approach was to estimate the values of actual consumption (C^h, C^f) using PPP-rate converted prices, by assuming that preferences are characterized by the utility function underlying equation (14.3) above but with multiple goods. Thus, per-capita consumption of each good was estimated in a linear econometric system across countries as a function of per-capita income and adjusted prices of all commodities. These fitted values, multiplied by population to achieve total consumption figures, were then "neutralized" by imposing a Cobb-Douglas specification of homogeneity. In this neutralization each country's consumption must be scaled up or down by its share of world income (in this case, the "world" consisted of the 21). The new consumption values (C^{h*}, C^{f*}), which were based on an expenditure categorization of goods, were mapped into standard production categories to make them consistent with corresponding output values. These figures were then used to compute "neutralized" trade flows as differences between the adjusted consumption data and outputs. These flows were compared to actual trade flows after the PPP adjustments. She calculated these trade flows on a net export basis (absolute value of exports minus imports) to sweep out the effects of intra-industry trade.

With this exercise, Hunter found that for the average country and commodity category the effect of consumption neutralization by imposing homogeneity was to raise trade flows by 29 percent. She interpreted that to mean that a simple alternative specification of demand – the linear quasi-homogeneous preferences along line C in Figure 14.9 rather than the strict homogeneity of line C* – was sufficient to diminish trade flows by a significant proportion. On this evidence it appears that non-homogeneity, in which income elasticities for different goods vary with per-capita income, is an important influence in world trade. Moreover, as noted in Figure 14.9, if low-income countries are initially net exporters of the necessity X_2 and high-income countries are net exporters of the luxury X_1, the non-homogeneity implies lower trade volumes than expected under homogeneous preferences. This provides an alternative hypothesis, strictly demand-based, for why the volume of trade between developed countries and developing countries may be relatively small, as predicted in the model in Section 14.3.

Astute readers will note that this same theory could be one explanation for the phenomenon of "missing trade" highlighted in Chapter 8. Indeed, Cassing and Nishioka (2011) recently respecified the Hecksher-Ohlin-Vanek equations to account for this possibility of quasi-homogeneous preferences that differ between developed and developing nations. Rather than go through their analysis in detail we note simply their main results. First, using a neutralization exercise similar to Hunter's the authors find that developing countries do consume relatively more labor-intensive goods (largely agriculture and food in their data) than would be expected under preference homogeneity. Second, they find that preference biases between rich and poor countries explain a larger proportion of missing factor trade than do differences in technology.

Both of these papers can be criticized in terms of data quality and the need to resort to extreme assumptions. However, they unearth intriguing findings and strongly suggest that consumption preferences vary with income levels in a way that diminishes the volume of trade in goods and, therefore, implicitly in factor services. More empirical work with alternative demand assumptions should help sort out the true contributions of preference differences and preference non-homogeneity to trade flows.

Product quality and trade

One of the more interesting areas of current empirical research concerns the relationships between international trade and apparent product quality. This question has both a supply side and a demand side. Do firms in more capital-abundant and skill-abundant nations produce higher quality goods on average? Do consumers in higher-income economies exhibit a higher willingness to pay for quality? If both of these statements are true we should see that trade among higher-income economies is disproportionately in goods with higher quality levels. Note that this is not a statement that such countries should have a high degree of intra-industry trade in horizontally differentiated goods, or products with similar quality levels, though that is the basic message of the monopolistic competition theory. Rather, it claims that countries should trade extensively in vertically differentiated goods, or products with different quality levels, and that trade among richer economies should reflect higher degrees of quality. The model in Section 14.4 provides one motivation for this idea.

There is ample evidence that the average quality of products increases on the supply side with the capital and skill endowments of the countries producing them. This point is made clearly by Schott (2004). Before considering his results, an obvious question should be asked: how can economists measure product quality in trade? The standard procedure, followed by Schott, is simply to look at highly detailed product categories within which goods may reasonably be assumed to meet specific needs of consumers or input buyers, then argue that higher-priced versions should reflect higher quality. The United States International Trade Commission, for example, makes available data on U.S. imports from, and exports to, all countries on product categories at the 10-digit level of disaggregation under various classifications.[1] The figures include both import and export values and export quantities. It is possible, therefore, to divide the dollar value of imports (measured as "free on board" or before any shipping charges or U.S. tariffs are added) by the corresponding quantities to get an average import price, or *unit value*, and similarly for exports. These prices may be compared across countries. Thus, if the unit value of a product from Germany is higher than that from Mexico, Schott considered the German version to embody higher quality.

To illustrate, consider a particular classification family within the harmonized tariff schedule (HTS), which is the system under widest use across countries. Four-digit category 6103 is titled "men's or boy's suits, ensembles, suit-type jackets, blazers, jackets, overalls, breeches and shorts, knitted or crocheted". While these are similar goods they are not specific products. In contrast, 10-digit subcategory 6103.1010.00 is "suits of wool or fine animal hair" and subcategory 6103.1060.10 is "jackets of cottons imported as parts of suits". These latter categories are so detailed that they plausibly refer to products that meet the same consumer need. By computing the unit values for each country that exported to the United States in a particular period, it is possible to correlate those prices with characteristics of the exporting nations. It should be noted, however, that there are a number of significant difficulties with this kind of data. First, only larger exporting countries are likely to export positive quantities in such detailed goods so this analysis tends to exclude small developing countries. Second, there is considerable noise in the variability of unit values and prices easily can vary by factors of four or more across import sources, which seems more than can be reasonably explained by quality differences. These problems likely arise

from measurement error and mis-classification of underlying transactions. Thus, analysts must eliminate from the sample unit values that are likely not meaningful. Finally, and most importantly, prices surely reflect more than just quality, including taxes in export countries and strategic pricing markups by imperfectly competitive firms.

Despite these problems, it is interesting to report Schott's primary findings. The author analyzed 10-digit product-line unit values over the period 1972-1994 from over 130 U.S. trading partners. These countries were defined as low-wage (L), middle-wage (M), and high-wage (H) based on real per-capita GDP. Specific 10-digit products were categorized as coming from L, M, or H nations solely or from combinations of such countries where, for example, a product arrived both from low-wage China and high-wage Canada. With this breakdown, one striking conclusion is that there seems to be relatively little specialization across products. In 1972, around 30 percent of U.S. imported products originated in all three types of countries (LMH goods). This share rose steadily to over 60 percent by 1994, strongly suggesting that countries of all income levels tend to produce the majority of goods even within detailed product groupings. This result seems sharply at odd with traditional trade theories in which countries tend to specialize across industries. Instead, it seems that there is significant within-industry and even within-product specialization.

Within-industry competition is consistent with models of monopolistic competition in horizontally differentiated goods. However, within-product competition that involves exporters at all levels of income suggests an additional important fact: there is considerable vertical differentiation in product quality. The next question is what explains this quality differentiation. Schott focused on supply-side determinants. Table 14.2 lists the basic results, which demonstrate convincingly that unit values rise as exporters have higher per-capita GDP, higher capital-labor endowments, and higher skilled labor endowments.[2] All of these coefficients are positive and highly significant in regressions that also included product-year fixed effects. The coefficient magnitudes suggest that a ten-percent increase in the specific characteristic raises unit values by 1.3 percent (GDP per capita), 4.4 percent (capital-labor endowment) and 5.0 percent (skilled labor share). From these results, it seems that average product quality rises as the exporting countries become richer, more capital-abundant, and more skill-abundant.

Table 14.2

Schott went further and related unit values to underlying industry production techniques. Using cross-country data from the United Nations he constructed capital-labor ratios for 28 3-digit manufacturing industries and assigned the traded goods in the import categories to these industries.[3] Using data for 1990, he regressed the log of unit values (by product and exporting country) on the log of capital-labor intensities (by country) for each industry, including product fixed effects. In these 28 regressions all the coefficients on labor-capital ratio were positive and only two were not highly significant. The magnitudes of the significant coefficients ranged from 0.18 (petroleum products) to 0.85 (electrical machinery). It is important to note that these industry categories cover many intermediate inputs as well as final goods. Thus, the differentiation of product quality varies sharply with the capital-labor ratios for both intermediate and final goods.

In summary, Schott's analysis of detailed price data in U.S. imports suggests that there is not much cross-product specialization but there is considerable within-product specialization. Moreover, these prices rise with incomes, capital, and skill endowments of exporters and also with the capital intensity of production. These results support the view that developed and developing countries may well lie within different cones of specialization, as analyzed in Chapter 8, but that the cones are defined by

products rather than broad industries.

The analysis unearthed an additional result of considerable importance. Specifically, the fact that average unit values rise with the per-capita GDP of exporters, and the within-industry capital-labor ratios of countries, strongly implies that export prices rise with levels of productivity. This outcome raises doubts about the notion that as firms become more productive they reduce price, as would be the case with models of monopolistic competition and heterogeneous firms. Rather, it seems that more productive firms specialize in higher-value and higher-quality goods. Thus, Schott's important paper raises a number of questions for further empirical work.

What about demand for quality? This is an important question for a number of reasons. For example, perhaps Schott's finding that higher-income economies produce higher-quality goods really is a demand-side phenomenon. If consumers prefer greater quality as their incomes grow the range of exportable products described in the Linder Hypothesis would become higher quality as domestic producers respond to that demand.

A recent paper by Hallak (2006) considered this question carefully. He offered a formalization of Linder's demand side by assuming that consumers' utility increases both with quantity and quality consumed of each type of good.[4] Specifically, he allowed Dixit-Stiglitz preferences for horizontally differentiated goods but augmented the preferences with a factor that shifted demand for quality upward as income increases. The theory is complex but can be summarized as follows. Suppose that Germany has a quality level θ_{Gz} for product z and Turkey has a quality level θ_{Tz}, while importer United States has a preference parameter for quality denoted γ_z^U and importer Argentina has quality parameter γ_z^A. Under quality-augmented Dixit-Stiglitz preferences, the United States and Argentina will gain utility from importing good *z* from both places but higher utility from the German good. Using this specification to determine *relative* bilateral import demands for each good z, we get (suppressing the *z* subscript):

$$\frac{USimportsG}{USimportsT} = \frac{N_G}{N_T}\left[\frac{p_G \tau_G^U / \theta_G^{\gamma^U}}{p_T \tau_T^U / \theta_T^{\gamma^U}}\right]^{1-\sigma}$$

Here, parameter $\sigma > 1$ is the elasticity of substitution among varieties of good z. This equation says that, as with monopolistic competition, the United States imports more from Germany as the number of firms increases in Germany, as German prices and bilateral trade costs (τ_G^U) fall, and as Germany quality rises, all relative to Turkey's similar factors. A similar ratio can be expressed for Argentina. Suppose we then compute a "ratio of ratios" by dividing the U.S. expression by its Argentine counterpart (which would eliminate the variables for firm numbers and prices, leaving just trade costs and quality levels). Doing so, we would easily find that relative bilateral demands are affected by both exporter quality levels and importer demands for quality. Specifically, if German quality in z is higher and Americans have higher demands for quality we would observe that the United States would import relatively more from Germany and Argentina relatively more from Turkey.

Hallak implemented these ideas econometrically by estimating bilateral import demand equations between 60 countries in 1995. Bilateral trade values were compiled for 114 four-digit sectors of the Standard International Trade Classification. For each sector, he regressed bilateral imports on exporter

fixed effects (to control for numbers of firms and prices), importer fixed effects (to control for importer expenditure levels), and trade costs, which he proxied by distance and a series of dummy variables for colonial ties, proximity, and the like. A final regressor was a measure of product quality, which he defined as import unit values in each 10-digit product category into the United States. The latter unit values were aggregated into a geometric-weighted index of product quality for each exporter in each sector. Again, there are complex problems in such an aggregation and in interpreting the results as a true quality measure. However, these indices are interesting. For example, the average price for differentiated goods made in Switzerland was 64 percent higher than in Canada and more than twice the level of Vietnamese differentiated goods. His data strongly suggest that average unit values are higher for richer exporters than poor exporters. Indeed, the correlation between GDP per capita and average prices of differentiated goods was significantly positive, at 0.45.

We simply describe the regression results. Distance and other measures of bilateral linkages had significant coefficients of the expected signs. Most interesting was the coefficient on exporter unit value, which should capture how much import demand goes up with an increase in quality. Across the 114 sectors there were 79 cases in which the coefficient on average unit value was positive (44 significantly so) and just 35 cases in which it was negative (15 significant). The median coefficient was 0.18 and, in a pooled regression across all trade sectors, the coefficient was 0.11 and highly significant. In brief, Hallak's results support the view that there is an independent and positive impact of product quality on demand for imports.

14.6 Summary

We suspect that it is accurate to suggest that both international trade theory and empirical analysis have concentrated far more (almost exclusively is not a big exaggeration) on production determinants of the direction and volume of trade. Yet there has never been convincing empirical evidence for this focus; it has been simply an article of faith that production rather than demand matters. Indeed, analysis of the importance of demand is so under developed that we cannot provide good evidence either way regarding this believe. An exception is Hunter (1991), who estimated that as much as 25 percent of trade can be associated with non-homogeneous preferences.

We began by describing ways that demand can matter: tastes can simply be different across countries in some sort of random way, or perhaps consumers have systematic bias toward domestically produced goods, a type of home bias. Much less arbitrary is the idea that preferences are non-homogeneous, and the shares of income spent on different categories of goods and services can vary substantially at different income levels. Poor households spend far more on food and far less on travel, recreation, and medical care than rich households. This has surely been convincingly demonstrated in every household budget survey ever undertaken. Differences in per-capita income then become a determinant of trade: countries can be specialized in consumption as much as in production.

An early advocate of this was Linder (1961) who argued strongly that the pattern of trade in manufacturers is a combination of product differentiation and non-homogeneous preferences. Oddly enough, the large literature on monopolistic competition addressed in Chapter 12 enthusiastically adopted the role of product differentiation, but abandoned Linder's crucial identification of the role of per-capita income. Later work (Markusen 1986) put this part of Linder back into the monopolistic-competition model and showed that if differentiated goods are both capital (or skilled-labor) intensive and have high income elasticities of demand, then there will be a higher level of north-north trade and a lower level of north-south trade than is predicted by standard models with homogeneous preferences. Later empirical

evidence gives good support to this idea.

In the final section, we turned to the issue of product quality. This relates to demand issues insofar as the standard assumption is that consumers with higher incomes having a higher willingness to pay for quality. We reviewed, in particular, the work of Shaked and Sutton (1983, 1984), which shows that trade can have an important benefit in leading to quality upgrading. Instead of having a larger range of products as in the monopolistic-competition approach, a model with endogenous quality choice can lead, under a number of special assumptions, to pure quality upgrading of existing products with no change in the range of products. The hypothesis of quality upgrading as a consequence of trade liberalization and integration seems to be receiving good empirical support in a range of recent articles.

REFERENCES

Cassing, James and Shuichiro Nishioka (2011), “Per-Capita Income and the Mystery of the Missing Trade”, working paper.

Choi, Yo Chul, David Hummels and Chong Xiang (2009), “Explaining Import Quality: The Role of Income Distribution”, *Journal of International Economics* 77, 265-275.

Dalgin, M., Vitor Trindade and Devashish Mitra (2008), “Inequality, Non-homothetic Preferences and Trade: A Gravity Approach”, *Southern Economic Journal* 74, 747-774.

Francois, Joseph F. and S. Kaplan (1996), “Aggregate Demand Shifts, Income Distribution, and the Linder Hypothesis”, *Review of Economics and Statistics* 78, 244-250.

Hallak, Juan-Carlos (2006), “Product Quality and the Direction of Trade”, *Journal of International Economics* 68, 238-265.

Hunter, Linda (1991), “The Contribution of Non-homothetic Preferences to Trade”, *Journal of International Economics* 30, 345-358.

Hunter, Linda and James R. Markusen (1988), "Per-Capita Income as a Determinant of Trade," in Robert Feenstra (editor), *Empirical Methods for International Economics*, Cambridge: MIT Press 89-109.

Linder, Staffan B. (1961). *An Essay on Trade and Transformation.* Stockholm: Almqvist & Wiksell.

Manova, Kalina and Zhiwei Zhang (2009), “Export Prices and Heterogeneous Firm Models”, Stanford University working paper.

Markusen, J. R. (1986). "Explaining the Volume of Trade: An Eclectic Approach." *American Economic Review* 76, 1002-1011.

Markusen, James R. and Randall Wigle (1990), "Explaining the Volume of North-South Trade," *Economic Journal* 100, 1206-1215.

Matsuyama, Kiminori (2000), “A Ricardian Model with a Continuum of Goods under Non-homothetic Preferences: Demand Complementarities, Income Distribution, and North-South Trade”, *Journal of Political Economy* 108, 1093-2000.

Mitra, Devashish and Vitor Trindade (2005), “ Inequality and Trade”, *Canadian Journal of Economics* 38, 1253-1271.

Schott, Peter K. (2004), “Across-Product versus Within-Product Specialization in International Trade,” *Quarterly Journal of Economics* 119, 647-678.

Shaked, A. and John Sutton (1983), “Natural Oligopolies”, *Econometrica* 51, 1469-1483.

Shaked, A. and John Sutton (1984), “Natural Oligopolies and International Trade”, in Henryk Kierzkowski (editor), *Monopolistic-Competition and International Trade*, Oxford: Oxford

University Press.

Summers, Robert and Alan Heston (1995), "The Penn World Table Mark 5.6," Center for International Comparisons at the University of Pennsylvania.

ENDNOTES

1. Interested students with some facility to work with interactive websites can find these data at http://dataweb.usitc.gov/. Earlier data (spanning 1972-1994) are available on CD-ROM disks for both imports and exports from The Center for International Data at the University of California-Davis; see http://cid.econ.ucdavis.edu/.

2.Capital and labor endowments were taken from the Penn World Tables Mark 5.6 (Summers and Heston, 1995), while skilled labor referred to the share of the population attaining secondary or higher education (Barro and Lee, 2000).

3.Interested students can find such data as the UNIDOT IND-STAT3 database at http://www.unido.org/index.php?id=o3699. They are quite useful for undertaking cross-country industry analysis.

4.This theory builds on Bergstrand (1989).

Table 14.1: Income elasticities of demand for various consumption goods and services

Food	0.45
Household furniture	0.76
Fuel and power	0.81
Education	0.87
Clothing and footwear	1.00
Beverages and tobacco	1.23
Other	1.25
Recreation	1.42
Transportation and communication	1.72
Gross rent	1.74
Medical	1.91

Source: Hunter and Markusen (1989)

Table 14.2 Relationship between Unit Values and Exporter Characteristics, 1972-94					
Independent variable	Log(unit value)	Log(unit value)	Log(unit value)		
Log PC GDP	0.134***				
Log Capital per Worker		0.435***			
Log Skill per Worker			0.501***		
Observations	214,852	214,852	214,852		
Unique products	12,024	12,024	12,024		
Unique countries	48	48	48		
R^2	0.77	0.78	0.77		
Notes: all regressions include product-year fixed effects. Coefficients are significant at the one-percent level					
Source: Schott (2004).					

Figure 14.1

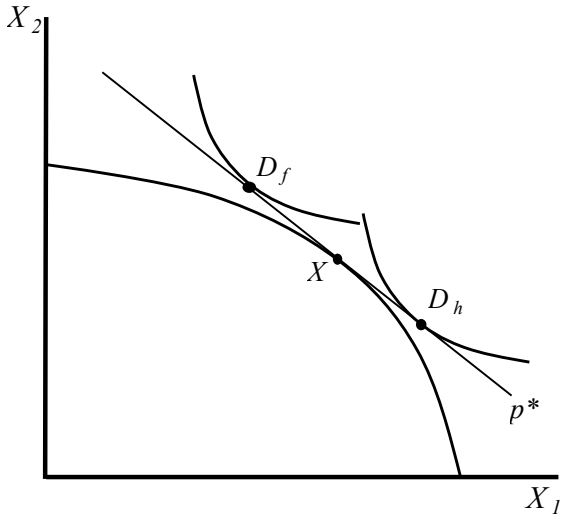


Figure 14.2

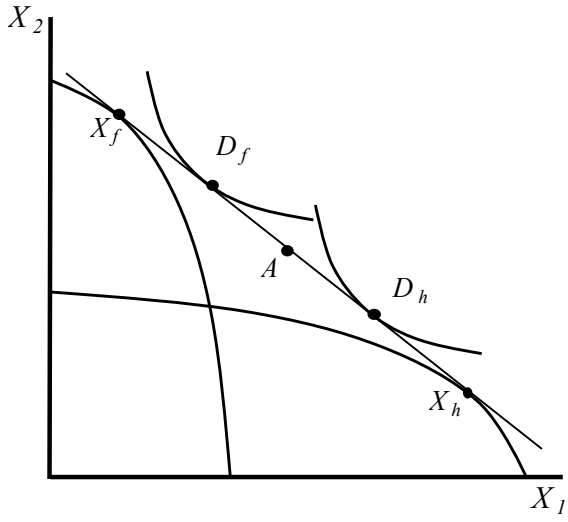


Figure 14.3

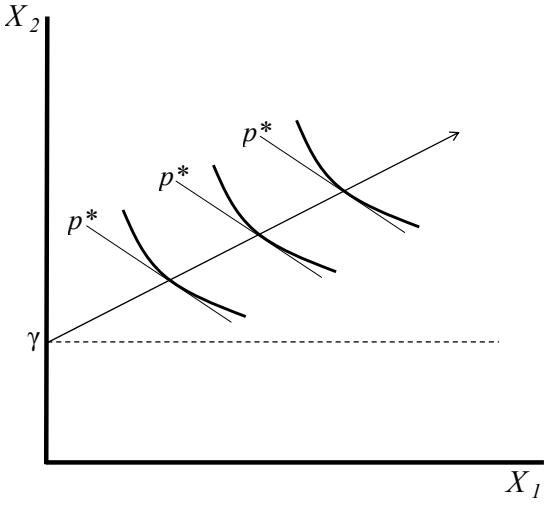


Figure 14.4

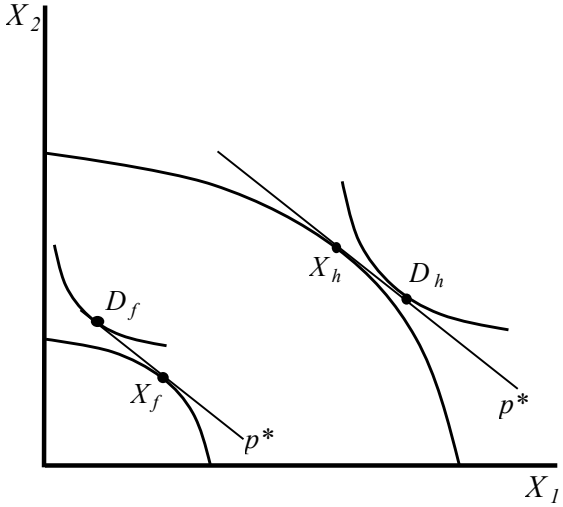


Figure 14.5

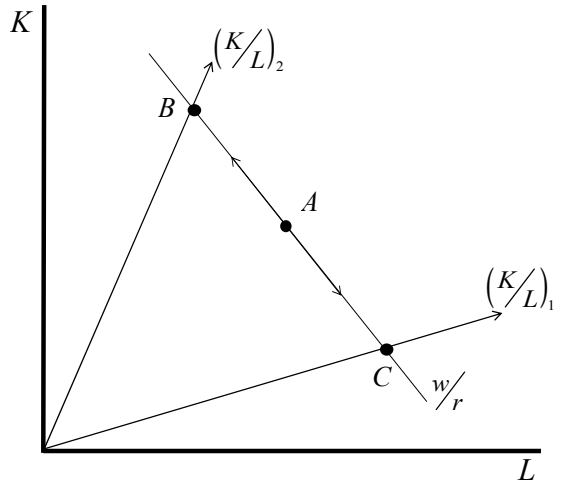


Figure 14.6

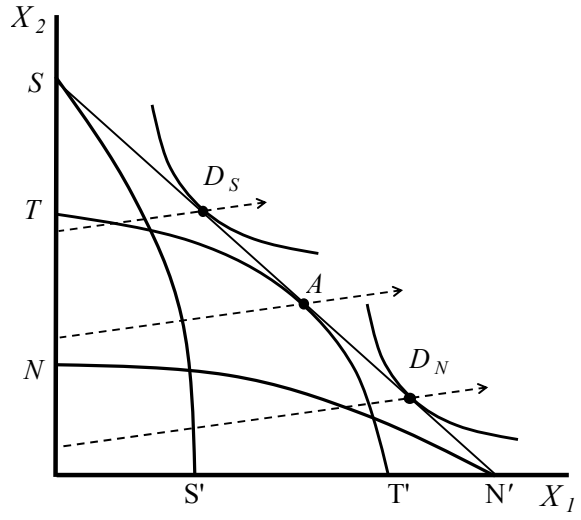


Figure 14.7

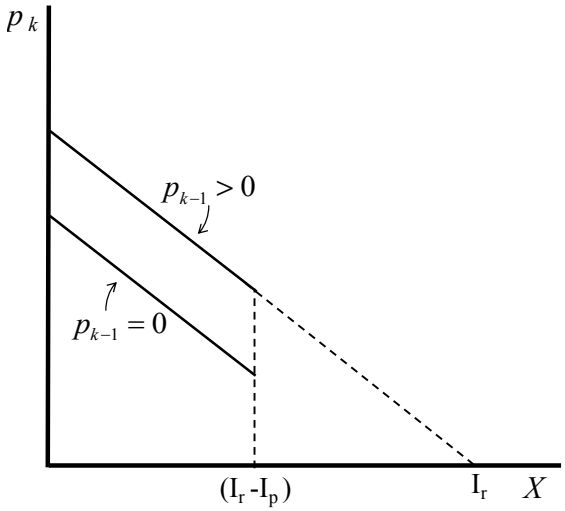


Figure 14.8

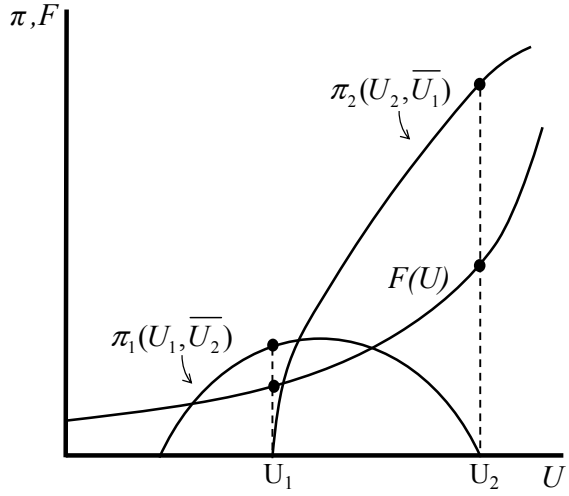
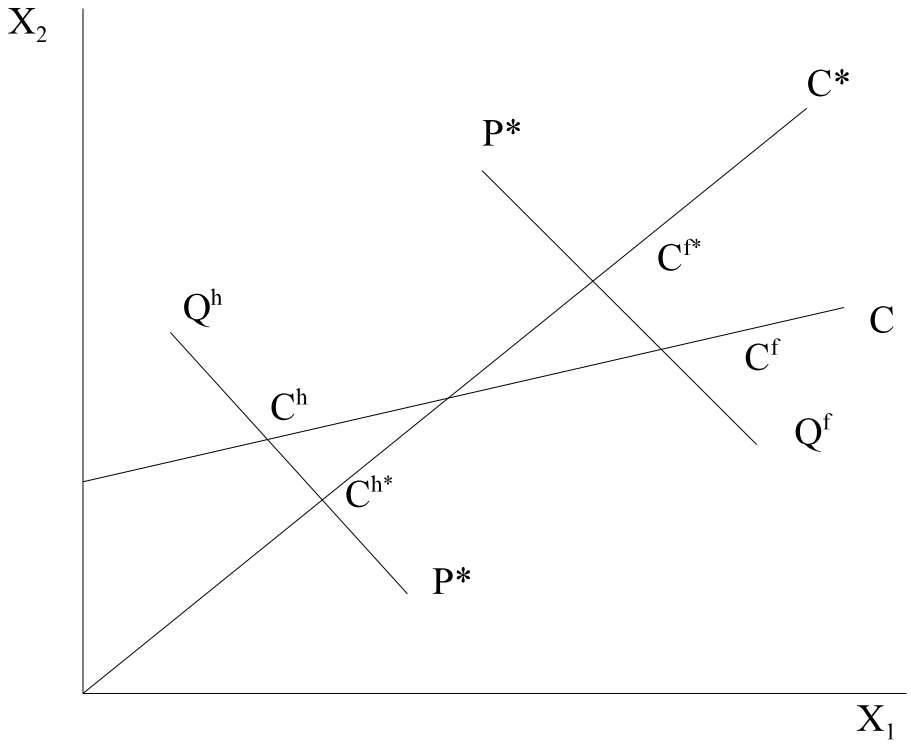


Figure 14.9



PART THREE

FACTOR TRADE, DIRECT FOREIGN INVESTMENT, OFFSHORING

Chapter 15

TRADE IN FACTORS OF PRODUCTION

15.1 Adding factor trade to goods trade

To this point, we have assumed that only goods are traded. This is obviously an unrealistic assumption in a world in which capital and, to a lesser extent, labor are mobile between countries. Anyone who has read history is aware of many examples of human migrations, such as those from Europe to North America in the nineteenth century. Those familiar with more contemporary business events are aware of the high degree of capital movements and foreign investment that have characterized the world economy since World War II.

It is thus apparent that some factors of production are mobile. What are the effects of such factor movements on world trade and welfare? This question is of particular interest in countries which are the source of foreign investment, as well as countries which are the recipients of investment. Some arguments seem to suggest that both source and host countries are worse off.

International trade economists generally distinguish between several types of factor movements: (1) foreign direct investment, (2) portfolio investment, and (3) labor migration. A direct foreign investment is defined as an investment in which the investor acquires a substantial controlling interest in a foreign firm or establishes a subsidiary in the foreign country. Foreign direct investment (FDI) thus involves ownership and/or control of a business enterprise abroad. Companies that engage in foreign direct investment are known as multinational enterprises or transnational corporations. FDI will be the subject of Chapter 16.

This chapter will be confined to an analysis of trade in factors, and is most relevant to portfolio investment and as well as labor migration. Portfolio investment occurs when an individual or company buys foreign bonds or purchases foreign stocks in quantities too small to gain control of a foreign firm. We will assume throughout the chapter that portfolio investment and labor migration occur in response to differences in wage and rental rates between two countries. (The motives behind direct foreign investment are more complex; thus an analysis of this type of investment is postponed until the next chapter.) Of course, portfolio investment and labor migration occur for reasons other than differences in factor prices. These reasons might include escape from war or repression in the case of labor migration, or uncertainty over future economic conditions in the case of capital mobility. With this caveat in mind, we restrict our discussion to factor movements motivated by factor price differences.

Portfolio capital movements and labor migration have very similar effects on production and trade. Thus, we analyze the two together, generally referring arbitrarily to capital as the mobile factor. The two may differ with respect to the repatriation of earnings. Capital owners generally tend to remain in their home country, repatriating their foreign earnings and consuming at home. Migrants generally spend most of their earnings in their new country. Yet migrants often repatriate substantial sums (e.g., guest workers in Europe, Mexican and West Indian farm workers in the United States) and capital owners sometimes move with their funds (e.g., Canadians retiring to Florida, employees sent abroad to work on a company investment project). Thus, we treat capital and labor movements as being essentially similar. First, we will prove a simple gains-from-trade theorem, assuming that the earnings of “immigrant” capital

or labor are repatriated or at least not counted as part of national income; i.e., national income is the income and consumption of the original (pre-factor trade) set of factor owners.

The gains-from-trade theorem of Chapter 5 showed that free-trade consumption is revealed preferred to autarky consumption under goods trade but no factor trade. The gains-from-trade theorem under factor as well as goods trade is a fairly straightforward extension of the proof in Chapter 5.

Let X_i denote production of good i and D_i denote consumption of good i. V_j will denote the *use* of factor j in production and $\bar{V}_j$ will denote the *endowment* of factor j. V_j and $\bar{V}_j$ can now differ due to imports or exports of factor i. $V_j > \bar{V}_j$, for example, means that the country is an importer of factor j. p_i denote goods prices and w_j is the price of factor j. Superscript * denotes free trade while superscript '*a*' denotes autarky.

Profit maximization in each industry i implies that the profits earned from the free-trade outputs and inputs must be at least as large as the profits from the autarky inputs and outputs at free trade prices. Summing over all industries, this gives us.

$$\sum_i p_i^* X_i^* - \sum_i \sum_j w_j^* V_{ij}^* \geq \sum_i p_i^* X_i^a - \sum_i \sum_j w_j^* V_{ij}^a \tag{15.1}$$

The sum of factor use on the left-hand side gives the total value of factor payments for factors used in the free-trade equilibrium.

$$\sum_i \sum_j w_j^* V_{ij}^* = \sum_j w_j^* V_j^* \tag{15.2}$$

The sum of factor use on the right-hand side is the value of the country's factor endowment (since no factors are traded) at free-trade prices.

$$\sum_i \sum_j w_j^* V_{ij}^a = \sum_j w_j^f \bar{V}_j \tag{15.3}$$

Substituting (15.2) and (15.3) into (15.1), the latter can be written as

$$\sum_i p_i^* X_i^* - \sum_j w_j^* V_j^* \geq \sum_i p_i^* X_i^a - \sum_j w_j^* \bar{V}_j \tag{15.4}$$

This can be re-arranged to yield

$$\sum_i p_i^* X_i^* + \sum_j \left[w_j^* \bar{V}_j - w_j^* V_j^* \right] \geq \sum_i p_i^* X_i^a \tag{15.5}$$

The trade balance condition in free trade can be written as the sum of the value of exports over all goods i plus the sum of the value of factor exports (the difference between each factor's endowment and use) over all factors j must equal zero. This is where we are making the assumption, briefly discussed above, that the income of imported factors is not considered part of domestic income while the income of exported factors is, as if both are fully repatriated to their country of origin.

$$\sum_i \left[p_i^* X_i^* - p_i^* D_i^* \right] + \sum_j \left[w_j^* \bar{V}_j - w_j^* V_j^* \right] = 0 \tag{15.6}$$

This can be rearranged to yield

$$\sum_i p_i^* X_i^* + \sum_j \left[w_j^* \bar{V}_j - w_j^* V_j^* \right] = \sum_i p_i^* D_i^* \tag{15.7}$$

Autarky market clearing condition is that the supply and demand of each good are equal.

$$X_i^a = D_i^a \tag{15.8}$$

Substitute (15.7) for the left-hand side of (15.5) and substitute (15.8) for the right-hand side of (15.5). The latter then becomes

$$\sum_i p_i^* D_i^* \geq \sum_i p_i^* D_i^a \tag{15.9}$$

Free trade consumption is revealed preferred to autarky consumption, which was to be proved. Free trade in goods and factors in a competitive, undistorted economy must be better than autarky.

15.2 Factor trade and goods trade as substitutes

A old and traditional idea in international trade theory is that trade in goods and factors are substitutes. This is credited to Mundell (1957), and indeed, the idea was so persuasive that little research followed on trade in factors for many years: there was just not very much interesting to say. The idea is that if trade in goods is based on differences across countries in factor endowments, then allowing factors to move will substitute for trade in goods.

There are, in fact, two different senses in which trade in goods and factors can be substitutes, both perhaps implicit in the previous sentence but not always aligned in practice. First, they can be substitutes in a welfare sense: the gains from trade can be captured either through trading factors or trading goods. Second, they can be substitutes in a volume-of-trade sense: allowing for trade in factors should reduce the volume of goods trade. This Chapter investigates both of these concepts.

It later turned out the Mundell focused on a very special case, though anchored in the very popular Heckscher-Ohlin model. In that model, goods trade only arises from difference in factor endowments, so it is pretty obvious that moving factors to make relative endowments equal is going to eliminate or substitute for goods trade. This does not diminish the importance of his contribution however, and we explore that in this section.

A very simple and intuitive approach to the problem is found in Jones et. al. (1986). Suppose that there is only a single produced good, X, and two factors of production, L and K. Suppose that countries h and f can only trade goods. It is clear in this very simple model that there is no way to capture any gains from trade: the countries must have at least two things to trade.

The situation is shown in the Edgeworth box of Figure 15.1. The dimensions of the box represent the total "world" endowments of capital and labor (*K* and *L*), the combined endowments of the two countries. The lower left corner is the origin for country *h* and the upper right-hand corner is the origin for country *f*. Isoquants are drawn for the production of *X* in each country. If the two countries have identical technologies, then the contract curve is the diagonal of the box, but identical technologies are not necessary for the argument.

Figure 15.1

Suppose that the endowment point for the two countries is at E in Figure 15.1. Here the home country is relatively well endowed with capital and the foreign country is relatively well endowed with labor. At *E*, the *X* isoquants of the two countries are not tangent (not drawn). The *X* isoquant for home is steeper than that for foreign, indicating that the autarky factor-price ratio *w/r* is higher in home than in foreign at the endowment point. This makes sense, since labor is the relatively scare factor in *h* and capital is the relatively abundant factor in *h*. The fact that factor prices are not the same in the two countries indicates that there are unexploited gains from trade, just as the inequality of commodity prices indicated unexploited gains from trade earlier in the book (e.g., Chapter 5).

Free trade in factors of production starting at point *E* in Figure 15.1 leads to an equilibrium at point A in Figure 15.1, where the factor-price ratios of the two countries are equalized at $(w/r)^*$. Country *h* exports capital, its abundant factor, and imports labor to reach point *A*. Country *f* exports labor, its abundant factor, and imports capital. The isoquants X_h and X_f are the free-trade *production* levels of *X* in the two countries, but are also the free-trade *consumption* levels since *X* is not traded. Both countries mutually gain from trade.

It is interesting and important to point out that the exchange of capital for labor is not the only possible way of moving from *E* to *A* in Figure 15.1. Suppose, for example, that capital and *X* can be traded, but labor is immobile. Then home can export capital in the amount *EB* in Figure 15.1 to foreign. Point *B* then becomes the *production* point for the two countries. But foreign must pay home for the capital, and hence foreign exports *X* in the amount *BA* in Figure 15.1. Point *A* continues to be the *consumption* point of the two countries, while *B* is the production point.[1]

This last result is interesting in light of controversies over the export of capital in countries like the United States. In the case just mentioned in the previous paragraph, the US loses production of *X* (jobs in manufacturing?), and imports *X* instead. This is in no way "bad". We end up at the same consumption point *A* in either case. Of course, factor price changes redistribute income, but these price changes are the same regardless of how we move from *E* to *A*.

To make the point more strongly, note that home could arrive at *A* by importing labor from foreign to arrive at *C*, which now becomes the *production* point. Home is now producing a lot of *X* which will make the critics happy. However, home must pay for the imported labor, exporting *X* to foreign (or giving the additional production to the foreign labor) in the amount *CA* to again arrive at the *consumption* point *A*. The level of production of *X* (jobs in manufacturing) is not of direct welfare consequences. As we have noted several times in the book, production must not be confused with consumption as a measure of the gains from trade.

This exercise emphasizes the idea that, when trade is based on differences in factor endowments in the first place, there are several alternative ways to capture the gains from trade. These alternatives

differ in what is traded, but they are substitutes in the sense that welfare is the same in the end. Trade in goods can be displaced by direct trade in factors.

Now consider the basic Heckscher-Ohlin model, with two goods (X_1 and X_2) produced by two countries (h and f) using identical technologies and factors *K* and *L*. Suppose initially that there is free trade (and no transport costs) in commodities but that factors cannot be traded. Assume that the factor endowments of the two countries are sufficiently similar such that both countries produce both goods in free trade. In this situation, free trade with equalize factor prices across countries, as we noted in Chapter 8 (Figures 8.5 and 8.6). There is no incentive to trade factors, all gains from trade are captured through commodity trade. However, there are two big assumptions underlying the factor-price-equalization theorem that we discussed in Chapter 8. First, endowments must be sufficiently similar such that both countries produce both goods. Second, commodity prices must be equalized by trade, ruling out any trade costs or tariffs.

Let us consider the first of these. Suppose that commodity prices are equalized by trade so that the unit-value isoquants for X_1 and X_2 are in the same position for both countries as shown in Figure 15.2. But the endowments of the two countries are sufficiently different (E_h and E_f in Figure 15.2) such that both countries are specialized in the free-trade equilibrium. Then each country's factor-price ratio (*w/r*) is the slope of the relevant isoquant at its endowment point. Country *h* has factor price ratio $(w/r)_h$ and country *f* has a factor-price ratio $(w/r)_f$. Each country has a high price for its scarce factor and a relatively low price for its abundant factor. If one or both factors are allowed to move, a country will export its abundant factor and/or import its scarce factor.

Figure 15.2

The general result is shown in Figure 15.3, where the factor-price-equalization set is as defined and discussed in Chapter 8 (Figure 8.6). If a country's endowments are inside this set, the price of each factor is equalized across countries and there is no incentive for factors to move and no additional gains from trade. If, for example, the endowment is at point E in Figure 15.3 however, then allowing factors to move will lead them to migrate in the direction of the arrows as shown: *K* will move from *h* to *f* and/or labor *L* will move from *f* to *h*. In general, these movements will displace commodity trade and hence trade in goods and factors are substitutes. However, trade in factors will allow for additional welfare gains from trade and hence adding factor trade is a welfare complement to goods trade. On the other hand, if we started with unrestricted factor trade and no goods trade, factors would be traded to the diagonal as in Figure 15.1 and there would be no additional gains from allowing goods trade. Thus, the situation is a bit more complicated than the simple statement that trade in goods and factors are substitutes, but this general notion is roughly correct at least with respect to the volume-of-trade definition of substitutes.

Figure 15.3

The second reason that factor-price equalization can fail in the Heckscher-Ohlin world is due to trade costs. Suppose that country *h* exports X_2 (i.e., *h* is relatively capital abundant, as in Figures 15.2 and 15.3) and that there are trade costs for both goods. These costs raise the price of a good in the importing country above the price in the exporting country. The price of X_1 in *h* must be higher relative to its price in f: $(p_1/p_2)_f > (p_1/p_2)_h$. This shifts the unit-value isoquant for X_1 in *h* inward: with the price of X_1 higher, fewer physical units of *K* and *L* are needed to produce one dollar's worth of X_1. Correspondingly, the X_2 unit-value isoquant is closer to the origin for country f, which imports X_2. This is shown in Figure 15.4.

Country *h* has an equilibrium factor-price ratio $(w/r)_h$ while Country *f* has an equilibrium factor-price ratio $(w/r)_f$. The important point is that each country has a high price for its scarce factor and a low price for its abundant factor. Indeed, we commented on this outcome in Chapter 13 on trade costs in connection with Figures 13.1 and 13.2.

Figure 15.4

Now suppose that we allow trade in factors. Country *h* will export capital and/or import labor. This causes the factor-ratios used in the countries to converge. This reduces commodity since the scarce factor in each country becomes less scarce and the abundant factor becomes less abundant. Factor prices in the two countries cannot begin to converge as long as there is commodity trade, because the trade costs maintain the commodity-price differences and that difference, in turn, determines the factor-price differences. Factor trade will continue until all commodity trade is eliminated. In Figure 15.5 we show something we call the "no trade set". This actually follows from Figure 13.1, where we note that there is a range of prices differences in the two countries such that the price differences are sufficiently small such that the countries will not want to trade goods. Beginning at endowment point E in Figure 15.5, factors will move toward the no-trade set and even continue to move toward the diagonal. Of course, this assumes that trade costs for factors are less than trade costs for goods.

Figure 15.5

This last point not withstanding, trade costs also suggest that trade in goods and trade in factors are substitutes in the volume-of-trade sense. However, they are not substitutes in the welfare sense: in the case shown, adding factor trade to goods trade leads to added welfare gains.

15.3 Factor trade and goods trade as complements

As suggested above, for many other causes of trade other than differences in factor endowments, trade in goods and factors can be complements in both the volume of trade and welfare senses (Markusen 1983). Consider a very different situation in which trade is caused by differences in production technologies.[2] Assume that we have two economies with the following characteristics: (1) Countries *h* and *f* have identical factor endowments; (2) Countries *h* and *f* have identical technologies for producing X_2, but country *h* has superior technology for producing X_1. It is assumed that country h's X_1 isoquants have the same shape as the *X* isoquants for country f, but that the former are renumbered so that more output is produced from the same inputs (this is called "Hicks-neutral" technical superiority, named after John Hicks).

The situation is shown in Figures 15.6 and 15.7. Assumptions (1) and (2) imply that both countries have an identical Edgeworth boxes and identical contract curves in Figure 15.7. However, their production frontiers differ, as shown in Figure 15.6. $X_{20}X_{f1}$ gives the production frontier for Country *f*. Country *h* can produce the same maximum amount of X_2 but more X_1, so country h's frontier is given by $X_{20}X_{h1}$.

Suppose that A in Figures 15.6 and 15.7 gives country f's production point in free-trade equilibrium. If country *h* allocated factors in the same way in Figure 15.6 (point B), country *h* would be at point A in Figure 15.7, producing the same amount of X_2 but more X_1. This cannot be an equilibrium for country *h* because the marginal cost of producing X_1 will be less in Country *h* relative to Country *f*.

Beginning at A in Figure 15.6, fewer factors are needed for an additional unit of X_1 in country *h* relative to country f, due to h's superior technology. Thus, if $p^* = MC_{f1}$ in Figures 15.6 and 15.7 at point A, then we must have $p^* > MC_{h1}$ at *B* in Figure 15.6 (which is also point A in Figure 15.7). The equilibrium for country *h* must be at a point like *C* in Figures 21.5 and 21.6.

Figure 15.6

Figure 15.7

If the countries are producing at *A* and *C* in Figures 15.6 and 15.7, we can conclude two things. First, country *h* must be exporting X_1 and importing X_2 (Figure 15.6). Second, the wage-rental ratio must be higher in country *h* (Figure 15.7) since the capital-labor ratios are higher in *h.* This is different from the case considered in the previous section; here each country will have a relatively *high* price for the factor used intensively in its export industry. If we permit factors to migrate, labor will flow into country *h* and/or capital will be exported, if X_1 is labor-intensive as shown in Figure 15.7. Similar comments apply, of course, to country *f.*

The result of this factor mobility is that each country becomes relatively better endowed with the factor used intensively in its export industry. This adds a Heckscher-Ohlin or factor-proportions basis for trade, which tends to reinforce the basis for trade caused by the difference in technology. Factor mobility can then lead to an increase in the volume of commodity trade. Country *h* will now export X_1 not only because it has superior technology but also because it is now relatively well endowed with labor.

In this simple model of trade based on differences in production technology, it thus turns out that factor movements and commodity trade are complements. This is true in both the trade volume and welfare senses: allowing factors to move starting with equal endowments increases the volume of goods trade as well as welfare. Although this may seem to be a very special case, it is true that the complementary relationship holds for a wide variety of models in which the basis for trade is something other than differences in factor endowments.

Consider as a second example of complementarity the simple, symmetric model of external economies of scale introduced in Chapter 10. Assume that scale economies are sufficiently strong that they outweigh factor-intensity effects, such that the production set is non-convex. Figure 15.8 draws the identical production frontiers for two identical economies. One country specializes in X_2 at point X_2^g (superscript g for goods trade) and one specializes in X_1 at point X_1^g, each trading half of its output for half of the other country's output so that they both reach the consumption point *D* in Figure 15.8. Figure 15.9 shows the factor market, with each country having the identical factor endowment *E*. While it is entirely arbitrary, assume country *h* produces output X_2 and country *f* produces X_1. Note from Figure 15.9 that each country will have a high price for the factor used intensively in its specialty good (given by the slopes of the isoquants at E).

Figure 15.8

Figure 15.9

Now suppose that the two countries can trade factors. Country *h* should import *K* and export *L* and vice versa for country *f*. Factor prices are equalized once country *h* reaches point A_2, producing X_2^f

(superscript *f* for factor trade) and country *f* reaches point A_1, producing X_1^f in Figure 15.9. The output of both goods has increased, and these production levels correspond to points in Figure 15.10: country *h* produces at X_2^f and country *f* produces at X_1^f. The two countries will now trade to both reach consumption point D^f in Figure 15.10. Factor trade has led to an increase in the volume of commodity trade, and the endowments of the countries have become more dissimilar due to factor trade. There are also welfare gains as is clear from Figure 15.10.

Figure 15.10

There are a fair number of other similar examples of complementarity, many explored in Markusen (1983). The point is that when the underlying cause of trade is something other than differences in factor endowments, equal relative endowments across countries does not exploit all gains from trade and allowing factors to move increases the volume of trade in addition to increasing welfare.

15.4 Agglomeration: combining monopolistic competition, trade costs, and mobile factors

During the 1990s and 2000s, there has been a great deal of interest in the nexus among monopolistic-competition, models, trade costs, and factor mobility. This is occasionally referred to as "new economic geography", a term the authors here don't much care for (what will we call the next wave of new stuff?). A central issue in this literature is the extent to which certain activities will naturally tend to agglomerate (firms clustering together in one location) versus spreading across countries.

In Section 15.3 regarding factor trade and goods trade as complements, we focused on the idea that equal factor endowments across countries might not be a stable situation if factors are allowed to move. If factor trade is introduced, factors may move to make countries more different in endowments than they were initially. Agglomeration can arise naturally, something quite different from a Heckscher-Ohlin world in which nothing would happen if countries are identical initially.

The mathematics of some of the newer models get complicated quickly, and much is beyond the scope of the book. What we will try to do is to present the intuition behind some of the basic concepts of this literature, beginning with a paper by Krugman (1991), followed later by the book of Fujita, Krugman and Venables (1999).

The basic Krugman model has a rather unusual factor structure, perhaps to give it analytical tractability. There are two sectors, a competitive, homogeneous agricultural sector *Y* which uses a type of labor that cannot be used in the other sector and cannot be internationally mobile. Thus there is a fixed, sector-specific factor in *Y*. The other sector *X* is the now familiar Dixit-Stiglitz monopolistic-competition sector. It also used a single factor of production and this type of labor may be internationally mobile. So there is no intra-country mobility but the *X*-sector labor may move abroad. In addition, there is an assumption that *Y* incurs no international trade costs and its price is equalized across countries, whereas *X* does incur a trade cost.

Let w_i denote the (nominal) wage of labor in country i in terms of the agricultural good *Y*, the latter's price being one and equal across countries. Recall that *X* varieties in each country have two equations associated with them: a pricing equation that determines output of a representative variety and a

free entry condition (markup revenues equal fixed costs) that determines the number of varieties. Let a unit of X require one unit of (mobile) labor, so marginal cost is just w_i. Then the two equations and resulting output per variety (X_{ij} is output produced in i sold in j) are given by

$$p_i(1 - 1/\sigma) = w_i \qquad (p_i/\sigma)(X_{ii} + X_{ij}) = wfc \qquad X_{ii} + X_{ij} = (\sigma - 1)fc \tag{15.10}$$

Any variety of X that is produced in either country is produced in the same amount. However, in the Krugman production structure, prices can differ across countries and thus so will wages. The first equation of (15.10) establishes a tight link between a locally produced X variety and the wage.

Now consider the price index in country i. Up to this point, we have used the term price index to refer to the price of buying one unit of the X "composite" good. Now, we will need to the price index for buying on unit of utility, which considers the prices of both X varieties and good Y. Let e_x denote the price index for the composite X good (previously we just used e) and let e denote the price index for utility. Using the same Cobb-Douglas function for utility between Y and X and the price index (expenditure function) for X goods, we showed way back in Chapter 2 (3.9) the over price index for utility:

$$U = \left[\sum_i X_i^{\alpha}\right]^{\frac{\beta}{\alpha}} Y^{1-\beta} \qquad e = e_x^{\beta} p_y^{1-\beta} \qquad e_x = \left[N_i p_i^{1-\sigma} + N_j (p_j t)^{1-\sigma}\right]^{\frac{1}{1-\sigma}} \tag{15.11}$$

Since we are using the price of Y as numeraire, $p_y = 1$, the overall utility price index for country i given in the second equation of (15.11) reduces to

$$e_i = \left[N_i p_i^{1-\sigma} + N_j (p_j t)^{1-\sigma}\right]^{\frac{\beta}{1-\sigma}} \tag{15.12}$$

The real wage of X sector workers, is perfectly measured by the nominal wage divided by the utility price index, also referred to as the ideal consumer price index. Making use of the first equation of (15.10), the real wage in country i is given by

$$\frac{w_i}{e_i} = \frac{p_i(1 - 1/\sigma)}{\left[N_i p_i^{1-\sigma} + N_j (p_j t)^{1-\sigma}\right]^{\frac{\beta}{1-\sigma}}} \tag{15.13}$$

This is where things get complicated, as well as interesting, so we are going to talk through the intuition and refer to some figures. Suppose first that the two countries are absolutely identical, including the assumption that the X-sector labor force is evenly divided between the two countries. Then (15.13) takes on the same value in both countries.

Now move a few workers from country f to country h. We assert that this will have the effect of driving down w_h and p_h and driving up w_f and p_f (the demand coming from the fixed number of agricultural workers in each country is crucial to these results). As a first step in understanding the consequences, hold the number of varieties in each country fixed. For country h (i = h) in (15.13), the nominal wage in the numerator falls. p_h in the denominator falls by the same amount as the numerator, but p_f rises. Thus even if the denominator falls when the two effects are combined, it cannot fall as much as the numerator. Holding the number and distribution of varieties constant, (15.13) must fall for country h the importer of labor and rise for country f the exporter of labor. In this case, the initial situation with

the mobile labor force distributed 50-50 and the countries identical is a *stable* equilibrium: any labor movement raises the real wage in the emigrant country and lowers it in the immigrant country.

However, that is not the end of the story. The movement of *X*-sector workers from country *f* to country raises the equilibrium number of varieties produced in country *h* and lowers the number produced in country *f*. From (15.12) and (15.13) we see that this has the effect of lowering e_h and raising e_f. For country *h*, a number of available varieties switch from being high-cost imports to low-cost local goods, increasing purchasing power in country *h* from a given nominal wage. By simple inspection of (15.13) therefore, we cannot tell whether or not the movement of workers from *f* to *h* raises or lowers the real wage in *h* relative to *f*. If the movement raises the real wage in *h* relative to f, we say that the initial symmetric equilibrium is *unstable*: the movement of a few workers from *f* to *h* will create an incentive for further movement.

The term "indeterminate" is some times used to mean we simply don't know (ignorance) and sometime it is used to mean different outcomes can occur depending on parameter values, a more positive type of statement. In the present case, it is the latter. The symmetric initial equilibrium will be stable or unstable depending on the level of trade costs and other parameters. Figure 15.11 shows some simulations of the basic Krugman model. The share of the mobile *X* sector workers is plotted on the horizontal axis and the real wage in country *h* divided by the real wage in country *f* is plotted on the vertical axis. Countries are identical when the worker share is 0.5 in the middle of the horizontal axis.

Figure 15.11

Figure 15.11 plots results for three levels of trade costs (these are gross trade costs: t = 1 is costless trade). A negatively sloped curve or portion thereof through the mid-point denotes a *stable* equilibrium: adding workers to *h* (moving to the right) lowers the real wage in *h* relative to *f*. A positive slope indicates adding some workers to *h* will lead to a divergence, higher wages in h, and hence to further movements away from the symmetric equilibrium. When trade costs are high (t = 3.0), the symmetric equilibrium is globally stable: migration will return the countries to the symmetric equilibrium if for some reason they start even very far away from it with workers concentrated in one country.

For an intermediate trade cost (t = 2.0), the symmetric equilibrium is *locally stable*, but there are two other crossings of the curve at a relative real wage of one. These two outer equilibria are *unstable*: any little movement toward the center will raise the wage in the labor-scarce country and movement will continue to the central symmetric equilibria. Any little movement from one of the asymmetric equilibria toward the boundary (i.e., movement away from the center) lowers the wage in the labor-scare country and a cumulative movement will continue until all mobile workers have deserted that country. At lower trade costs (t = 1.5) in Figure 15.11, the central symmetric equilibrium is globally unstable. Any little movement in either direction lowers the wage in the labor-scarce country relative to the labor-abundant country and a cumulative movement continues all the way to the boundary.

The structure of the model used here and popularized starting with Krugman (1991) is unusual in its factor-market structure, and this has led many trade economists to question its generality. To partly dissuade us from the view that this model's result rests solely on its extremely special nature, the present authors reformulated the model allowing for *Y* to again have a fixed (sector-specific) factor that is internationally, as well as intersectorally, immobile as before, but the mobile labor factor is used in both *Y* and in *X*. This means that more labor drawn away from the *Y* sector must bid up the nominal wage rate in terms of *Y*. Results of a simulation for this model are shown in Figure 15.12. The central symmetric

equilibrium for the case shown (t = 1.75) is unstable as it can be in the Krugman model for low trade costs. However, in this model there are always two other equilibria which are stable. These equilibria occur when the labor-scarce country begins to specialize in *Y* only. At that point, further losses of must raise the real wage in that country, so mobile labor will never desert the country entirely. These asymmetric equilibria are *stable* in the model of Figure 15.12 rather than *unstable* in Figure 15.11 (t = 2.0). Clearly, the essence of the Krugman result about the *possible* instability for identical countries does not hinge on the model's very special nature.

Figure 15.12

Markusen and Venables (2000) also show the same result shown in Figure 15.12 in the traditional two-good two-factor Heckscher-Ohlin structure. The symmetric equilibrium with countries identical is unstable to movements of the factor used intensively in the X sector, but a stable outcome is achieved before all of that factor deserts one country. As in Figure 15.12, the asymmetric outcomes are stable.

15.5 Summary

For many decades, the study of international trade focused on trade in goods, while analyses of factor mobility and trade in factors of production was generally left to inter-regional analysis and urban economics. This may have been due to the fact that the Heckscher-Ohlin model was the center piece of international economics. In that approach, trade in goods is induced by differences in factor endowments across countries, so trading factors directly tends to be a substitute for trading goods. The term substitutes is often used in both a volume-of-trade sense (allowing factor trade reduces the volume of goods trade) and in a welfare sense (gains from goods trade is reduced if factors are traded). There didn't seem to be much else interesting to say.

In the early 1980s, however, it was pointed out that the Heckscher-Ohlin model is essentially the only case in which this substitutability result holds. For almost any other cause of trade, trade in factors and goods can be complements in both the volume-of-trade and in the welfare sense. Examples include Ricardian differences in technology, production distortions such as taxes, external economies of scale and other factors. If countries have initially identical relative factor endowments, allowing factors to move both increases the volume of trade in goods and world welfare.

More recently, trade in factors and its relation to production specialization and trade in goods has received a lot of attention in the economic geography literature, which brings some of the tools of trade theory (especially monopolistic-competition) to traditional regional models with mobile factors. Of special interest in this literature are the possibilities of multi-equilibria, the creation of industrial agglomerations, and the possible instability of "spreading" (non-agglomerated or symmetric) equilibria. The interactions between trade costs, factor mobility, and production are complex. The movement of a factor used in a monopolistic-competition sector to one country may lower that factor's nominal wage, but it also lowers the price index for buying goods such that the real wage of the factor may rise. When that occurs, symmetric outcomes with identical regions can be unstable: if some factors arbitrarily move, they set off a cumulative movements that leaves industry agglomerated in one region and leaving the other region producing a competitive good (such as agriculture). Spatial differentiation and specialization, which is exogenously determined in Heckscher-Ohlin, becomes the endogenous outcome in this newer literature.

Empirical work on these issues is progressing. It is tricky insofar as the mere observation of industrial agglomeration does not prove that this agglomeration is due to the phenomena highlighted in the "new" economic geography. Many observed agglomerations are simply due to firms being drawn to the same immobile site-specific resources (e.g., ski resorts in Colorado, Hokkaido, or the Alps).

REFERENCES

Brackman, Steven, Harry Garretsen and Charles van Marrewijk (2009), *The New Introduction to Geographical Economics*, Cambridge: Cambridge University Press.

Combes, Pierre-Philippe, Thierry Mayer and Jacques-François Thisse (2008), *Economic Geography: The Integration of Regions and Nations*, Princeton: Princeton University Press.

Fujita, Masahisa, Paul R. Krugman and Anthony J. Venables (1999), *The spatial economy: cities, regions and international trade*, Cambridge: MIT Press.

Jones, R. W., I. Coelho and S. Easton (1986), "The Theory of International Factor Flows: the Basic Model, *Journal of International Economics* 20, 313-327.

Krugman, Paul R. (1991), Increasing returns and economic geography, *Journal of Political Economy* 99, 483-499.

Markusen, J. R. (1983). "Factor Movements and Commodity Trade as Complements." *Journal of International Economics* 14, 341-356.

Markusen, James R and Anthony J. Venables (2000), "The Theory of Endowment, Intra-Industry, and Multinational Trade", *Journal of International Economics* 52, 209-234.

Neary, J. Peter (1995), "Factor Mobility and International Trade", *Canadian Journal of Economics* 28, S4-S23.

Mundell, R. A. (1957). "International Trade and Factor Mobility." *American Economic Review* 47, 321-335.

Purvis, D. D. (1972). "Technology, Trade, and Factor Mobility," *Economic Journal* 82, 991-999.

Wong, Kar-Yiu (1995), *International Trade in Goods and Factor Mobility*, Cambridge: MIT Press.

ENDNOTES

1. A country's GNP (gross national product) refers to the income of its citizens, while GDP (gross domestic product) is the value of production in the country. In the presence of factor trade, the two are not the same. In the case just mentioned, point B in Figure 15.1 gives the two countries' GDP (production). Point A gives their GNP (consumption). For country *h*, GNP equals GDP plus the earnings of its capital now located in Country *f*. Conversely, GDP exceeds GNP for country *f*, since some of the value of production accrues to factors owned by country *h*.

2. This analysis is due to Purvis (1972) and Markusen (1983).

Figure 15.1 Equivalence of alternative types of trade

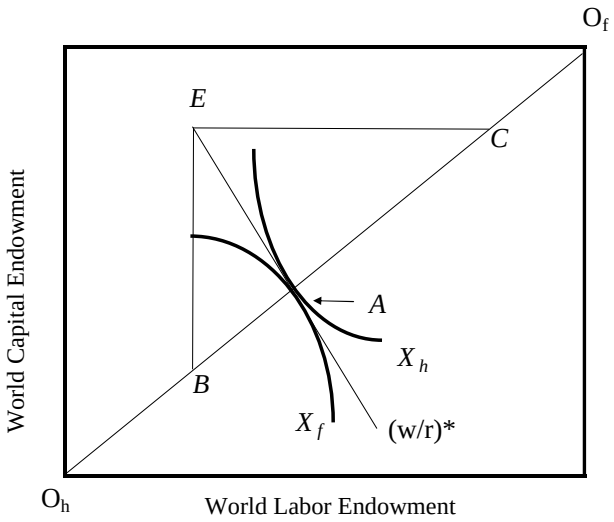


Figure 15.2: Specialization and Relative Factor Prices

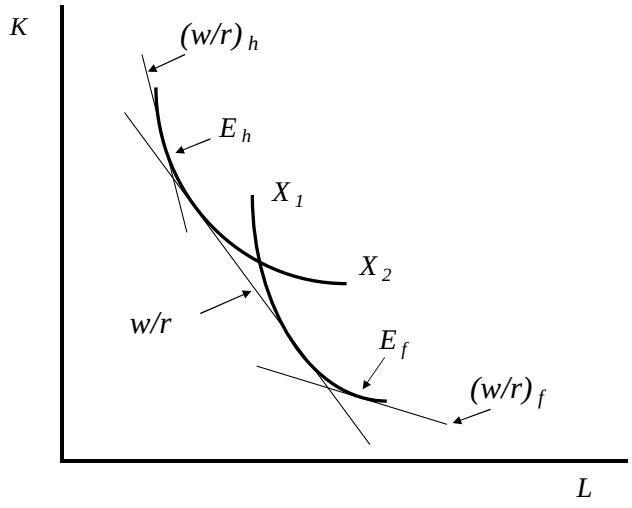


Figure 15.3: Factor trade outside the FPE set

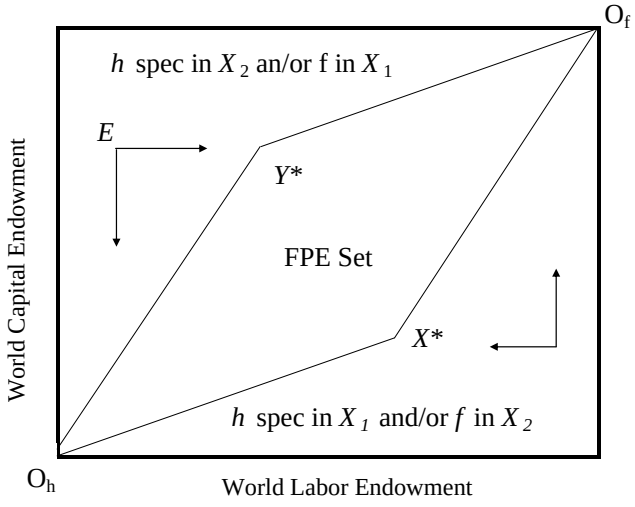


Figure 15.4: Trade Costs and Factor Prices

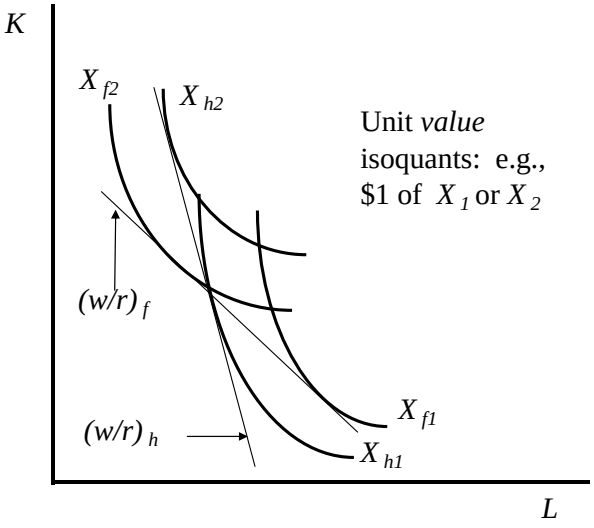


Figure 15.5: Factor trade with trade costs

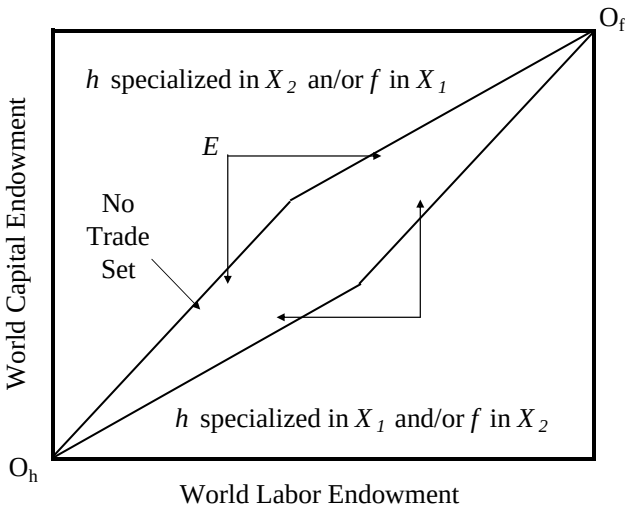


Figure 15.6: Country h has technical advantage in X_1

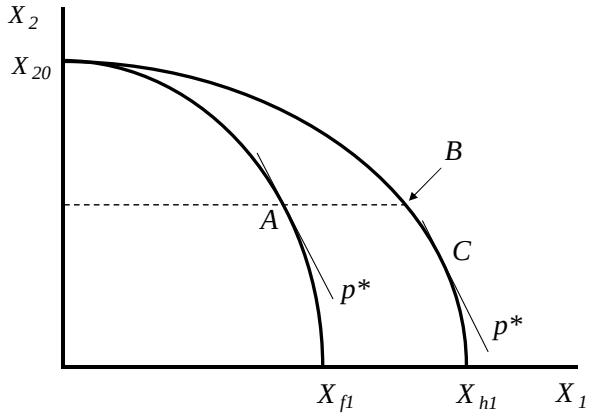


Figure 15.7: Factor prices with technology differences

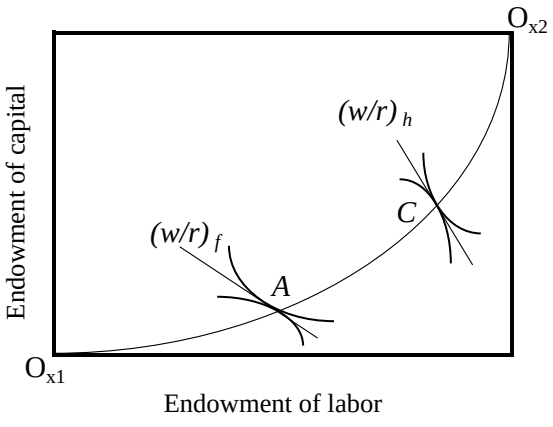


Figure 15.8: Specialization with identical countries

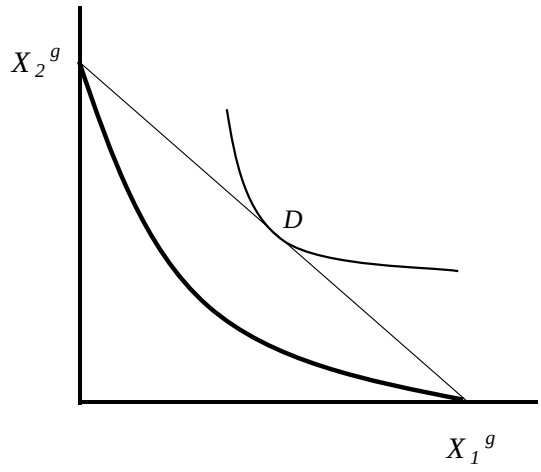


Figure 15.9: Adding factor trade

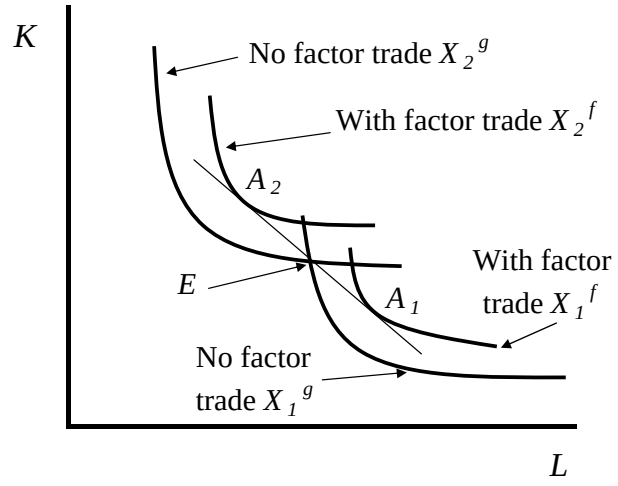


Figure 15.10: Factor trade and goods trade

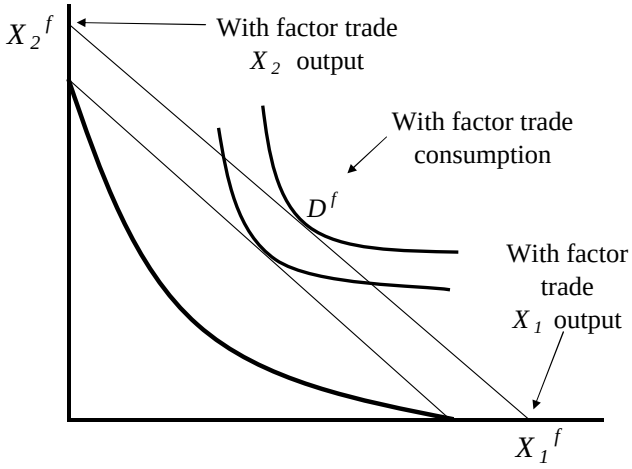


Figure 15.11: Agglomeration versus spreading in relation to trade costs

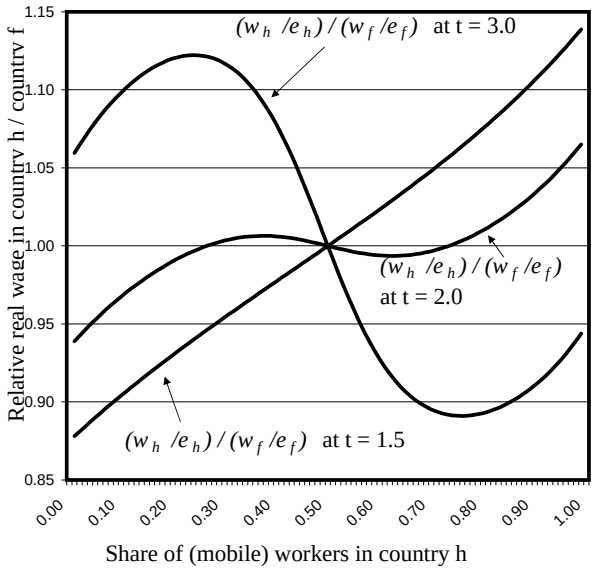
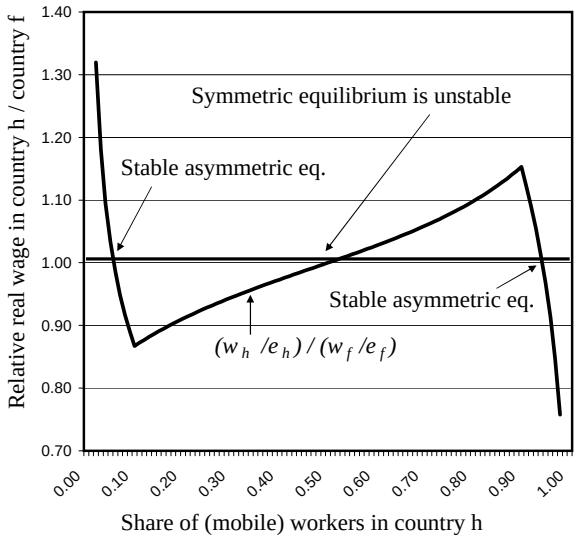


Figure 15.12: Alternative formulation: mobile factor also used in Y



Chapter 16

MULTINATIONAL FIRMS AND FOREIGN DIRECT INVESTMENT

16.1 Stylized facts, basic concepts

Multinational firms (MNEs) have become a crucial element in the modern world economy, and a few statistics are presented in Table 16.1 to document this. The first number documents that fact that sales of foreign affiliates of multinational corporations are nearly double the total value of world trade. Other number we have seen range up to a factor of five. The second number notes that value added in multinational affiliates is about 11 percent of world GDP (itself a value-added measure). The third number notes that one-third of all world exports originate in the foreign affiliates of multinational firms. The final number actually surprises some folks as being rather small: foreign affiliate exports are only 18 percent of their total sales. Most of the output of foreign affiliates is sold locally, a point that we must make sense of.

Table 16.1

Several decades ago, foreign direct investment (FDI) was viewed as simply a capital movement. Consistent with the dominant theory of the day, Heckscher-Ohlin theory, capital should flow from capital-rich, high-income countries to capital-scarce, low-income countries. During the 1980s, it became very obvious that the old view was at best inadequate and at worst simply wrong. Our first basic fact.

(A) FDI flows primarily from high-income developed countries to other high-income countries, not from capital-rich to capital-poor countries.

Table 16.2 gives some evidence on this. The top panel gives flows of new FDI for 2007, expressed as the *developed* countries' share of total flows, in both inward and outward directions. These figures are followed by the corresponding shares for the stock of FDI in 1990 and 2007. Clearly, the developed countries are the major source of FDI, but what is less appreciated is that they are also the major recipients of FDI. Our second basic fact.

Table 16.2

(B) Affiliate production is primarily for local sale, not for export back to the parent country.

Another myth that persists in the popular press and politics, though we hope not in the economics literature is that multinationals move production to poor countries to pay low wages and export the output back home. Table 16.3 breaks down sales of foreign manufacturing affiliates of US, Japanese, and Swedish parent firms into local sale and export sales, and imports are also listed. Note that two-thirds or more of foreign output of affiliates is sold locally. Table 16.4 has data from a more restricted sample for the US and Sweden. In 2003, local sales account for 60 percent of all sales, but that still leaves a large portion for export. Columns 2 and 3 reveal, however, that most of the exports do not go back to the parent country, they go to third countries. The point is that foreign affiliates are not primarily in the business of producing cheaply abroad for shipment back to the parent: they are primary in the business of producing abroad for local and regional markets.

Table 16.3 **Table 16.4**

Several other stylized facts will be important.

(C) FDI is attracted to large and high-income markets

This again fits well with the idea that production is largely for local or regional sale and, if there are fixed costs of setting up or acquiring new plants, then such investments are more likely to be observed in large and high-income markets. If multinationals were simply in search of low wages to produce for export back to the home country, then we should not observe firms to choose high-income markets and there should be little relationship to host-country size.

(D) There are high levels of intra-industry cross-investment, particularly among the high-income countries

Much has been written about intra-industry trade or "cross-hauling" over the last several decades. It seems to be less well known that exactly the same phenomenon is observed for affiliate sales. Combined with earlier statistics, these numbers emphasize that a satisfactory theory of multinational firms must be consistent with large volumes of cross-investment (intra-industry affiliate production) among similar large, high-income countries. It must also, of course, be able to explain the fact that firms from the high-income countries are net investors in developing countries.

The weight of these statistics also suggests that much if not most FDI is "horizontal" or "market seeking"; that is, foreign affiliates of multinational firms are doing much the same things in foreign countries as they are at home. The same products and services are produced, generally for local and regional sale. "Vertical" or "resource seeking" (e.g., low cost labor) investments involve the fragmentation of the production process into stages, with stages located where the factors of production they use intensively are relatively cheap. While not the dominant motivation for FDI, vertical investments may nevertheless be quite important in developing countries.

16.2 A basic organizing framework

Modern theory often begins with the premise that firms incur significant costs of doing business abroad, relative to domestic firms. Therefore, for a firm to become a multinational, it must have offsetting advantages. A limited but very useful organizing framework for inquiring into the nature of these advantages was proposed by John Dunning (1977). Dunning proposed three conditions needed for firms to have a strong incentive to undertake FDI.

Ownership Advantage: the firm must have a product or a production process, such that the firm enjoys some market power advantage in foreign markets.

Location Advantage: the firm must have a reason to want to locate production abroad rather than concentrate it in the home country, especially if there are scale economies at the plant level.

Internalization Advantage: the firm must have a reason to want to exploit its ownership advantage internally, rather than license or sell its product/process to a foreign firm.

Internalization is now often referred to as vertical integration, and the choice between an owned subsidiary versus licensing is often now referred to as the outsourcing decision, which is the same thing put the other way around.

The basic idea is shown in Figure 16.1. A firm wanting to produce in a foreign country for local sale has a location decision between producing in the foreign country or exporting to that country. We can refer to this as the "offshoring" decision. If the multinational chooses foreign production, it has the further choice between an owned subsidiary, perhaps some sort of joint venture, or an arm's length licensing contract. We can refer to this as the internalization or vertical integration versus outsourcing decision, whether or not to produce outside the ownership boundaries of the firm.

Figure 16.1

First consider ownership advantages. Evidence indicates that multinationals are firms that have high levels of R&D, large employment shares of marketing, scientific and technical workers, high levels of product newness and complexity, and product differented goods (Caves 2007, Markusen 2002). This suggests that multinationals are firms which are intensive in the use of knowledge capital. Knowledge capital is a broad term which includes the human capital of the employees, patents, blueprints, procedures, and other proprietary knowledge, as well as finally marketing assets such as trademarks, reputations, and brand names.

There are several reasons why the association of multinationals with knowledge-based assets rather than physical capital is appealing. First, the services of these assets may be easily used in distant plants, such as managers and engineers visiting those plants. Second, and more subtly, knowledge capital often has a joint-input or non-rivaled property within the firm. Blueprints, chemical formulae, or even reputation capital may be very costly to produce, but once they are created, they can be supplied at relatively low cost to foreign production facilities without reducing the value or productivity of those assets in existing facilities.

The sources of location advantages are varied, primarily because they can differ between horizontal and vertical firms. Consider horizontal firms that produce the same goods and services in each of several locations. Given the existence of plant-level scale economies, there are two principal sources of location advantages in a particular market. The first is the existence of trade costs between that market and the MNEs home country, in the form of transport costs (both distance and time), tariffs and quotas, and more intangible proximity advantages. The second source of location advantage, again following from the existence of plant-level scale economies, is a large market in the potential host country. If that market is very small, it will not pay for a firm to establish a local production facility and the firm will instead service that market by exports.

The sources of location advantage for vertical multinationals are somewhat different. This type of investment is likely to be encouraged by low trade costs rather than by high trade costs and by factor-price differences across countries. Low trade costs facilitate the intra-firm trade of intermediate and final goods and thus facilitate the geographic dis-integration of the production chain.

Internalization advantages (or outsourcing disadvantages) are the most abstract of the three. The topic quickly gets into fundamental issues such as what is a firm, and why and how agency problems might be better solved within a firm rather than through an arm's-length arrangement with a licensee or contractor. Basically, it is our view that internalization advantages arise from another property of

knowledge: in addition to being non-rivaled, it may be non-excludable. Local agents can learn the technology and management skills and defect to start rival firms. Further discussion is postponed until Section 16.6.

16.3 A simple monopoly model of location choice

In this section, we explore a basic model very similar to that used in Chapters 11 and 13 in order to demonstrate the crucial intuition as to the optimal international organization of a firm. Questions of outsourcing versus vertical integration (internalization) are left to a later section. There are two countries, home and foreign and one monopoly firm in country *h.* There is a linear inverse demand for the product where the intercept is α and slope is (*1/L*), *L* = market size. The price (p_i), quantity (X_i) and market size (L_i) in market i = *h, f* are related as follows, where the second equation is firm revenues (R_i) in market i.

$$p_i = \alpha - X_i/L_i \qquad R_i = p_i X_i = (\alpha - X_i/L_i)X_i \tag{16.1}$$

There is a constant marginal cost c_i in market i and a specific trade cost *t* between markets. Profits before fixed costs for a plant producing in market i and selling in i and a plant producing in country i and selling in country j are given by

$$\pi_{ii} = (\alpha - X_{ii}/L_i)X_{ii} - c_i X_{ii} \qquad \pi_{ij} = (\alpha - X_{ij}/L_j)X_{ij} - (c_i + t)X_{ij} \tag{16.2}$$

Taking the first-order conditions for profit maximization, given the optimal levels of domestic and export supply.

$$X_{ii} = \left(\frac{\alpha - c_i}{2}\right)L_i \qquad X_{ij} = \left(\frac{\alpha - c_i - t}{2}\right)L_j \tag{16.3}$$

The firm is headquartered in country *h* but may choose between three (discrete) alternatives. It can choose a single plant at home in country *h,* exporting to country *f* and we refer to this firm as a national or domestic firm (d). Second, it can have a single plant in country *f,* exporting back to *h,* a vertical (v) structure. Third, it can be a horizontal multinational with plants in both countries (m). There is a firm-specific fixed cost *F* and plant-specific fixed cost *G* where the latter must be incurred for each plant. This model thus displays firm-level scale economies: the total fixed costs of a two-plant firm are *(F + 2G)* which is less than the total fixed costs of two independent single-plant firms, *(2F + 2G)*. *F* thus represents the cost of creating a product or process, and this knowledge is a joint or non-rivaled input across plants.

If you substitute (16.3) into the two equations of (16.2) and add in fixed costs, you will find that the profits for each of the multinational firms three choice are given as follows (we did a derivation of this in Chapter 13).

Profits of a national firm: one plant at home (*h*) exporting to *f*

$$\Pi^d = \Pi_{hh} + \Pi_{hf} = \left[\frac{\alpha - c_h}{2}\right]^2 L_h + \left[\frac{\alpha - c_h - t}{2}\right]^2 L_f - G - F \tag{16.4}$$

Profits of a vertical firm: one plant in *f* exporting back to *h*

$$\Pi^v = \Pi_{fh} + \Pi_{ff} = \left[\frac{\alpha - c_f - t}{2}\right]^2 L_h + \left[\frac{\alpha - c_f}{2}\right]^2 L_f - G - F \qquad (16.5)$$

Profits of a horizontal firm: plants in both countries.

$$\Pi^m = \Pi_{hh} + \Pi_{ff} = \left[\frac{\alpha - c_h}{2}\right]^2 L_h + \left[\frac{\alpha - c_f}{2}\right]^2 L_f - 2G - F \qquad (16.6)$$

This is a surprisingly rich little model. Figure 16.2 shows profits from the three possible choices, holding total world demand constant, but varying the size of the two markets along the horizontal axis, with the parent country *h* small on the left and big on the right. If you stare at these three equations long enough and consider how each curve shifts (or does not shift) in response to some parameter change, you will conclude that a two-plant *horizontal* structure is more likely as:

Both markets are large	characteristic of markets
Markets of similar size	characteristic of markets
Marginal costs are similar	characteristic of markets
Firm fixed costs > plant fixed costs	characteristic of industry
Transport/tariff costs are large	geography/policy

Figure 16.2

Large markets mean that the added fixed costs of a second plant outweigh the higher variable costs of exporting. The intuition behind the second and third results is that if one market is much larger and/or production costs much smaller, then it pays to put a single plant there and export to the smaller/costlier market. For the third result, note that if we raise *F* and lower *G* in the same amounts, this increases the profits of a type-m firm, while leaving the profits of a type-d or type-v firm unchanged (the horizontal curve shifts up). Trade costs reduce the profits of a type-d or type-v firm but leave the profits of a type-m firm unchanged (the national and vertical curves shift down in Figure 16.2).

Despite its simplicity, this model fits the data well: horizontal firms will be important between similar, large (rich) markets in industries where knowledge capital is important (*F* is large relative to *G*).

A vertical structure is preferred to a national structure as:
Foreign market is larger
Foreign marginal cost is low
Low trade costs: vertical structure if $c_f < c_h$ even if country *f* is small

As noted above, if you are going to have a single plant, put it in the large and/or low-cost market. Note, however, that the importance of the local market size disappears as trade costs go to zero. As *t* converges to zero, the relationship between (16.4) and (16.5) is determined entirely by the low-cost location, and note that the horizontal structure will never be chosen at $t = 0$ ($G > 0$). As *t* goes to zero, the vertical structure is chosen if and only if country *f* is the low-cost location.

16.4 Monopolistic competition and the choice of exporting versus horizontal production

A crucial feature of the modern theory of the multinational, as noted earlier, is the notion of joint or non-rivaled knowledge-based assets of firms (these are sometimes also called firm-level or multi-plant economies of scale). Once a firm develops a product, process, or even brand identification and a reputation for quality, it can add additional plants abroad for significantly less additional fixed costs than is incurred for setting up the firm and first plant at home. In this section, we will consider a firm wanting to sell abroad, as well as at home, and limit its choices to exporting to the foreign market or becoming a two-plant horizontal multinational. We will use the monopolistic-competition model developed in Chapters 12 and 13.

To keep things simple, we will consider *identical countries*. Assuming, in addition, that marginal costs of production is the same for both domestic and multinational firms, the pricing equation in the model says that all varieties will have the same (domestic) prices in equilibrium, $p_i = p_j$, regardless of whether they are produced by a locally-owned firm or by a branch plant of a foreign multinational. Similarly, imported goods will have the same prices in each country. Let superscript d denote a domestic (one plant) firm and superscript m denote a two-plant horizontal multinational. As in Chapter 13, $\phi = t^{1-\sigma}$ is the "phi-ness" of trade where, t is the (gross) iceberg trade cost: $\phi = 1$ is free trade and $\phi < 1$ means positive trade costs. Let the first subscript on X denote country of firm ownership and the second the country of sale. The demand functions for the various X varieties (domestic production and imports) are given by:

$$X_{ii}^d = X_{jj}^d = X_{ij}^m = X_{ji}^m = p_i^{-\sigma} e_i^{\sigma-1} L/2 \qquad X_{ji}^d = X_{ij}^d = p_i^{-\sigma} \phi e_i^{\sigma-1} L/2 \tag{16.7}$$

Zero-profit conditions for d and m firms located in country i are markup revenues equal fixed costs. Let fc^d denote the fixed costs of a first plant and βfc^d denote the fixed costs of a two-plant multinational. The non-rivaled or multi-plant economies idea implies that $1 < \beta < 2$: a two-plant firm has less than double the fixed costs of a one-plant firm. The zero-profit conditions for one and two-plant firms are then

$$(p_i/\sigma)X_{ii}^d + (p_i/\sigma)X_{ij}^d \le fc^d \tag{16.8}$$

$$(p_i/\sigma)X_{ii}^m + (p_i/\sigma)X_{ij}^m \le fc_x^m = \beta fc^d \tag{16.9}$$

Using the demand functions for X_{ii} and X_{ij} above, these are:

$$p_i^{1-\sigma} e_i^{\sigma-1} L/2 + p_i^{1-\sigma} \phi e_j^{\sigma-1} L/2 \le \sigma fc^d \tag{16.10}$$

$$p_i^{1-\sigma} e_i^{\sigma-1} L/2 + p_i^{1-\sigma} e_j^{\sigma-1} L/2 \le \beta fc^d \tag{16.11}$$

Suppose that we pick values of parameters such that national and multinational firms can both just break even in the two identical countries: (16.10) and (16.11) both hold with equality. Then dividing the first equation by the second gives us the critical relationship between trade costs and fixed costs for indifference between being a national and a multinational firm.

$$\frac{(1+\phi)}{2} = \frac{1}{\beta} \qquad 2 > (1+\phi) = \frac{2}{\beta} > 1 \tag{16.12}$$

Lower trade costs (*higher* ϕ) must be balanced against higher firm-level scale economies (*lower* β) for firms to be indifferent between the national and multinational options. To put it differently, higher trade costs encourage firms to be multinationals and higher firm-level scale economies do the same. In line with our earlier discussions, we can think of high firm-level scale economies as being associated with knowledge and R&D intensive industries. Our results clearly have empirical predictions about what sort of firms should be multinationals and which industries should be dominated by multinationals. Tests of these ideas have confirmed the basic theory and we will discuss these later in the Chapter.

16.5 The knowledge-capital model

A recent development that integrates both horizontal and vertical motives for multinationals is Markusen's knowledge-capital model. The full definition of the model is found in Markusen (2002). The knowledge-capital model is a general-equilibrium approach that incorporates both horizontal and vertical motives for multinationals. The configuration of firms that arises in equilibrium depends on country characteristics (size, relative size, and relative endowments), industry characteristics (firm versus plant-level fixed costs or scale economies) and trade costs.

There are two goods, *X* and *Y* , and two factors of production, skilled and unskilled labor, *S* and *L*. There are two countries i and j. *Y* is produced with constant returns by a competitive industry and unskilled-labor intensive. *X* is produced with increasing returns by imperfectly competitive firms. There are both firm-level and plant-level fixed costs and trade costs. Firm-level fixed costs result in the creation of "knowledge-based assets".

There are three defining assumptions for the knowledge-capital model.

(A) *Fragmentation*: the location of knowledge-based assets may be fragmented from production. Any incremental cost of supplying services of the asset to a single foreign plant versus the cost to a single domestic plant is small.

(B) *Skilled-labor intensity*: knowledge-based assets are skilled-labor intensive relative to final production.

(C) *Jointness*: the services of knowledge-based assets are (at least partially) joint (non-rivaled) inputs into multiple production facilities. The added cost of a second plant is small compared to the cost of establishing a firm with a single plant.

There are three possible firm "types" that can exist in equilibrium in either country (so six firm types in all), and there is free entry an exit into and out of firm types.

Type m - horizontal multinationals which maintain plants in both countries, headquarters is located in country i or j.

Type d - national firms that maintain a single plant and headquarters in country i or j. Type d_i firms may or may not export to the other country.

Type v - vertical multinationals that maintain a single plant in one country, and headquarters in the other country. Type v_i firms may or may not export back to their headquarters country.

Various assumptions can be made about factor intensities and they make some quantitative difference to the results. The assumptions used to generate the diagrams attached below assume that the skilled-labor intensity of activities are

[headquarters only] > [integrated *X*] > [plant only] > [*Y*]

When countries are similar in size and in relative factor endowments, the model predicts that horizontal multinationals will be important. Firms will build plants in the foreign country to serve the local market instead of incurring trade costs on exports. We discussed this in the previous section for identical countries but now need to comment on this issue of size and endowment similarity. The intuition about the role of these characteristics is understood by considering what happens when countries are *dissimilar* in size or in relative endowments.

To a good degree, the intuition is seen in Figure 16.2 above. When one country is quite large and the other quite small, a national firm located in the large country will have an advantage since it spends little in trade costs, whereas a horizontal firm still has to bear the fixed cost of a second plant in the small country. Similarly, if the countries are similar in size but one country is skilled-labor abundant, a vertical firm located in the skill-abundant country has an advantage over the other two types: it can locate the headquarters in the skilled-labor-abundant country and the single plant in the unskilled-labor-abundant country. Note that this advantage for the vertical firm is reinforced if the skilled-labor-abundant country is also small. The vertical firm has the added advantage that locating its single plant in the large market abroad reduces on trade costs.

Figure 16.3 plots the results of a simulation solving for the number, types, and production of firms over a world Edgeworth box. The world skilled-labor endowment is on the axis running toward the northwest and the unskilled-labor endowment on the axis running to the northeast. We could also think of this as a “composite”, including physical capital, land and resources as unskilled labor. The total equilibrium volume of affiliate production is shown on the vertical axis, where affiliate production is defined as the output of foreign plants of horizontal and vertical multinationals (i.e., the output of a horizontal firm’s home plant is not included in affiliate production). In the center of the box where the two countries are identical, exactly half of all world X output is affiliate output: all firms are horizontal multinationals and their home and foreign plants are identical, thus half of each firm’s output is affiliate output.

Figure 16.3

However, affiliate sales can be even higher, and this occurs in Figure 16.3 when one country is both small and skilled-labor abundant. In this case, most or even all firms are vertical multinationals with their headquarters in the small, skilled-labor-abundant country. This, in turn, means that all (or virtually all) plants and all production are in the other (large) country, and so all of world output is classified as affiliate output and sales.

Figures 16.4 and 16.5 help sharpen our intuition by showing two restricted versions of the model. Figure 16.4 eliminates the possibility of vertical firms (the first plant must be located with the

headquarters) so only national firms and horizontal multinationals can exist. Consequently, we should observe multinationals when the countries are similar in size and in relative endowments. Figure 16.5 does the opposite, ruling out horizontal multinationals by eliminating firm-level scale economies: the fixed costs of a two-plant firm are double the fixed costs of a single-plant firm. In this case, there is no multinational activity between identical countries, and activity is maximized when one country is both small and skilled-labor abundant.

Figure 16.4 Figure 16.5

A nice thing about Figures 16.3-16.5 is that they provide clearly testable hypotheses, and further hypotheses follow from the size of trade costs relative to firm and plant-level scale economies. We will discuss this later in the chapter, but the consistent empirical finding is that the data give far more support to the horizontal model. The purely vertical case in Figure 16.5 is overwhelming rejected and the world looks much more like Figure 16.4 than 16.5: multinational activity is concentrated among similar, high-income countries. Multinational activity and investment from high-income to low-income countries is of much less importance, especially relative to what seems to be the popular impression.

16.6 Outsourcing versus internalization (vertical integration)

The final node in the decision tree shown in Figure 16.1 concerns the choice to maintain ownership of a foreign production facility or to outsource/license a foreign firm to produce for the multinational. While this is an old question in the international business literature, it has only recently attracted interests from international trade economists. There are two principal approaches and we will present simple versions here.

Some of the first formal models of the internalization decision were published in the late 1980s and 1990s, and draw their empirical motivation from the strong association of multinationality with knowledge-based assets, such as those described in the previously (see Caves 2007 and Markusen 2002). On the one hand, these assets (or the services thereof) are easily transferred overseas, such as providing a blueprint, chemical formula, or procedure to a foreign plant. On the other hand, the same characteristics that make it easy to transfer these assets make them easily learned by foreign managers, agents, or licensees. Once the agent sees the blueprint or formula, he or she could defect to produce the product in a new firm. Knowledge is non-excludable, at least after some period of time.

About the same time as this last set of papers appeared, an important advancement to the internalization question was being developed by several authors, and we will review and explain a key paper by Antrás (2003). These authors substitute the term outsourcing for the converse term internalization. Their approach is sometimes termed the “property-rights” approach to the firm. The new literature combines a number of separate elements that together produce a coherent model that offers clear empirical predictions. The first element is the assumption that production requires “relation-specific investments”, meaning that a multinational and a foreign individual or firm must incur sunk investments prior to production that have no outside value if the relationship breaks down. The second element is the assumption of incomplete contracting: certain things, such as agent effort, are simply not contractible; alternatively, any contract on these items is not enforceable. The assumptions of sunk investments and non-contractibility lead to a third problem, which is ex-post “hold up”. What happens after production occurs cannot be contracted ex ante, so each party has some ability to negotiate ex post and to prevent the other party from fully utilizing the output.

The first set of papers, focusing on the non-excludability of knowledge which is learned by a local agent or manager, can be explained by a simple version of Markusen (2002). To efficiently explain the key idea, we will focus only on the issue of foreign production versus exports in order to focus clearly on the role of non-excludability of knowledge.

(1) The MNE introduces (or attempts to introduce) a new product every second time periods. Two periods are referred to as a "product cycle". A product is economically obsolete at the end of the second period (end of the product cycle).

(2) The probability of the MNE successfully developing a new product in the next cycle is $1/(1+r)$ if there is a product in the current cycle, zero otherwise (i.e., once the firm fails to develop a new product, it is out of the game). The probability of having a product in the third cycle is $1/(1+r)^2$, etc. Ignore discounting over time, but note the r is very much like a discount rate of interest. $r/(1+r)$ is the probability of failing to develop a new product: the higher r, the higher the risk the firm fails.

(3) The MNE can serve a foreign market by exporting, or by creating an affiliate to produce in the foreign market. Because of the costs of exporting, producing in the foreign country generates the most potential rents (profits).

(5) Any local manager learns the technology in the first period of a cycle and can quit (defect) to start a rival firm in the second period. Similarly, the MNE can defect, dismissing the manager and hiring a new one in the second period. The (defecting) manager can only imitate, but cannot innovate, and thus cannot compete in the next product cycle.

(6) No binding contracts can be written to prevent either partner from undertaking such a defection. We will assume that the MNE either offers a self-enforcing contract or exports.

R- Total per period licensing rents from the foreign country.

E- Total per period exporting rents ($E < R$).

F- Fixed cost of transferring the technology to a foreign partner. These include physical capital costs, training of the local manager, etc.

T- Training costs of a new manager that the MNE incurs if it dismisses the first one (i.e., if the MNE defects).

G- Fixed cost that the manager must incur if he/she defects. This could include costs of physical capital, etc.

L_i- Licensing or royalty fee charged to the subsidiary in period i (i = 1,2).

V Rents earned by the manager in one product cycle: $V = (R-L_1) + (R-L_2)$.

V/r- Present value of rents to the manager of maintaining the relationship.

The manager ("a" for agent) has an individual rationality constraint (IR): the manager must earn non-negative rents. The manager also has an incentive-compatibility constraint (IC): the manager must

not want to defect in the second period: second-period earnings plus the present value of earning from future products (if any) must exceed the single-period one-shot return from defecting.

$$(R - L_1) + (R - L_2) \geq 0 \qquad \text{IR}_a \qquad (16.13)$$

$$(R - L_2) + V/r \geq (R - G) \qquad \text{IC}_a \qquad (16.14)$$

where $V = (R - L_1) + (R - L_2)$ is the present value to the manager of the future rents, if there are any. $(R - G)$ is the payoff to unilaterally defecting.

The MNE similarly has an individual rationality constraint (IR): the MNE must earn profits at least equal to the profits from exporting. The MNE also has an incentive-compatibility constraint: the MNE must not want to defect (fire the manager) in the second period.

$$L_1 + L_2 - F \geq 2E \qquad \text{IR}_m \qquad (16.15)$$

$$L_2 \geq R - T \qquad \text{IC}_m \qquad (16.16)$$

Combine the IC constraints.

$$R - T \leq L_2 \leq G + V/r \qquad (16.17)$$

The firm's objective is to minimize V subject to this incentive compatibility. Making V as small as possible subject to (16.17), gives us:

$$2R - L_1 - L_2 = V = r(R - T - G) \geq 0 \qquad \text{(rent share to the manager)} \qquad (16.18)$$

Our first result is then that, if $R \leq G + T$, the MNE captures all rents in a product cycle, henceforth referred to as a rent-capture (RC) contract, and the agent's IR_a constraint holds with equality. This occurs when

(1) The market is relatively small.
(2) Defection costs for the MNE (T) are high.
(3) Defection costs for the manager (G) are high.

If $R > T + G$, there is no single-product fee schedule that will not cause one party to defect. In this case, the manager's IR_a constraint does not hold as a strict equality: that is, the MNE shares rents with the manager and the amount of rent sharing is given in (16.18). This is a credible commitment to a long-term relationship that we could think of as a subsidiary. However, it is costly for the multinational and if it gets too costly then the multinational will choose exporting instead: dissipating some rent is preferable to sharing a larger total. This is the inefficiency caused by the lack of contractibility of knowledge, and may lead (will lead for many parameter values) the firm to make a welfare-inefficient choice to export and dissipate total surplus rather than share a larger surplus with the local agent.

As noted above, the "property-right" approach works rather differently. Here, we present a much simplified version of Antrás (2003). The idea is that the firm and the local agent must make ex ante investments that are not contractible. The multinational invests capital K and the agent invests labor L. Ex post, they divide the surplus via the Nash bargaining solution, with the firm getting share s and the

agent the share (1-*s*). We cannot go through an analysis of Nash bargaining here, but many readers may know from industrial organization or game theory that the equilibrium share is a function of each party's "bargaining power" and its outside option. Antrás assumes the firm has a bargaining power parameter of at least 1/2. Ownership is defined as a property right to anything left (e.g., an intermediate input) in the event of bargaining breakdown.

Under outsourcing, denoted with the subscript "o", any intermediate output produced is worthless to both in the event of a breakdown. Thus, the outside option of both the firm and the agent is zero even though the agent owns what is left. Under FDI, denoted with the subscript "v" for vertical integration, the multinational has some use for what is left, giving the MNE an outside option. If $\phi > 1/2$ denotes the multinational's bargaining power and δ denotes the share of the total potential rent (under a successful contract), the Nash solution gives the multinational the following shares.

$$s_o = \phi \qquad s_v = \delta + \phi(1-\delta) \qquad s_v > s_o > 1/2 \tag{16.19}$$

The second equation is a common result in the bargaining literature: the firm gets its outside option, plus its bargaining share of the total minus the sum of the outside options (the agent's outside option is zero).

Let the prices of K and L equal one. Profits from the project are given by

$$\Pi = \left(\frac{K}{\beta}\right)^{\beta}\left(\frac{L}{\gamma}\right)^{\gamma} - K - L \qquad \beta + \gamma < 1 \tag{16.20}$$

Knowing that there will be holdup and ex-post bargaining with equilibrium share *s*, the firm and the agent respectively maximize the following, choosing the input they control.

$$\max_K \; s\left(\frac{K}{\beta}\right)^{\beta}\left(\frac{L}{\gamma}\right)^{\gamma} - K \qquad \text{Firm chooses } K \tag{16.21}$$

$$\max_L \; (1-s)\left(\frac{K}{\beta}\right)^{\beta}\left(\frac{L}{\gamma}\right)^{\gamma} - L \qquad \text{Agent chooses } L \tag{16.22}$$

The first-order conditions for capital (chosen by the firm) and labor (chosen by the agent) are given by

$$\left(\frac{K}{\beta}\right)^{\beta-1}\left(\frac{L}{\gamma}\right)^{\gamma} = \frac{1}{s} = (1+t_k) \qquad t_k \equiv \frac{(1-s)}{s} \tag{16.23}$$

$$\left(\frac{K}{\beta}\right)^{\beta}\left(\frac{L}{\gamma}\right)^{\gamma-1} = \frac{1}{1-s} = (1+t_l) \qquad t_l \equiv \frac{s}{(1-s)} \tag{16.24}$$

The last equality in each line is to emphasizes that the agent problem here is much like having a tax on capital of $t_k = (1\text{-}s)/s$ and a tax on labor of $t_l = s/(1\text{-}s)$. Indeed, the first-order conditions are exactly those of a single integrated firm maximizing profits subject to these input taxes. Ex-post holdup is like each party being able to tax the other. Given (16.19), vertical integration is effectively a lower tax on capital than outsourcing, and vertical integration is effectively a higher tax on labor. A first-best outcome (i.e., by an integrated single decision maker) would be to set both equations to one (the true prices of capital

and labor) rather than to something greater than one. Rearranging (16.23) and (16.24) we get

$$(K/\beta)^{1-\beta} = s(L/\gamma)^{\gamma} < (L/\gamma)^{\gamma} \qquad \text{firm's choice of } K \text{ for a given } L \qquad (16.25)$$

$$(L/\gamma)^{1-\gamma} = (1-s)(K/\beta)^{\beta} < (K/\beta)^{\beta} \qquad \text{agent's choice of } L \text{ for a given } K \qquad (16.26)$$

where the quantities on the far right of each expression are the first-best outcomes.

Antrás notes that (16.25) and (16.26) can be thought of as reaction or best-response functions for the firm and agent, respectively, and the situation is shown in Figure 16.6.[1] When s = 1 and (1-s) = 1 in the two equations respectively, we have the first-best outcome in which the two reaction functions F^* (firm) and S^* (agent) cross at point A. Both outsourcing and vertical integration shift in both reaction functions. However, since $s_v > s_o$, vertical integration shifts F^* less (to F_v) than outsourcing does (to F_o) in Figure 16.6. Conversely, outsourcing shifts S^* less (to S_o) than vertical integration does (to S_v) in Figure 16.6. The vertical integration equilibrium is at point B in Figure 16.6 while outsourcing leaves us at C. Both outcomes are inefficient, but vertical integration has a relatively less under-utilization of capital while outsourcing has a relatively less under-utilization of labor.

Figure 16.6

It may be intuitive that a firm in a capital-intensive industry will choose vertical integration while a firm in a labor-intensive industry will choose outsourcing, but this requires a fair bit more algebra to prove. The two first-order conditions can be solved to yield

$$\frac{K}{L} = \frac{s}{1-s}\frac{\beta}{\gamma} \qquad \frac{L}{\gamma} = \frac{1-s}{s}\frac{K}{\beta} \qquad (16.27)$$

These can be substituted back into the first order conditions to give the equilibrium inputs

$$K = \beta\left[s^{1-\gamma}(1-s)^{\gamma}\right]^{\frac{1}{1-\beta-\gamma}} \qquad (16.28)$$

$$L = \gamma\left[s^{1-\beta}(1-s)^{\beta}\right]^{\frac{1}{1-\beta-\gamma}} \qquad (16.29)$$

The profit level for the firm in (16.20) is then given by:

$$\Pi = (1-\beta)\left[s^{1-\gamma}(1-s)^{\gamma}\right]^{\frac{1}{1-\beta-\gamma}} \qquad (16.30)$$

The choice between vertical integration or outsourcing then reduces to evaluating the ratio

$$\frac{\Pi_v}{\Pi_o} = \left[\frac{s_v^{1-\gamma}(1-s_v)^{\gamma}}{s_o^{1-\gamma}(1-s_o)^{\gamma}}\right]^{\frac{1}{1-\beta-\gamma}} \qquad (16.31)$$

Given the assumption that $s_v > s_o \geq 1/2$, it is true that $s_v(1-s_v) < s_o(1-s_o)$: the function

$s(1-s)$ reaches a maximum value of 1/4 at $s = 1/2$ and then decreases as s grows larger (or shrinks for that matter). As the share on labor becomes small (the industry is very labor intensive), (16.35) reduces to

$$\gamma \Rightarrow 0 \qquad \frac{\Pi_v}{\Pi_o} \Rightarrow \frac{s_v}{s_o} > 1 \tag{16.32}$$

so vertical integration is chosen by a capital-intensive firm. At a value of $\gamma = 1/2$, implying a value of $\beta < 1/2$, we have

$$\gamma = \frac{1}{2} \qquad \frac{\Pi_v}{\Pi_o} = \left[\frac{s_v(1-s_v)}{s_o(1-s_o)}\right]^{\frac{0.5}{1-\beta-\gamma}} < 1 \tag{16.33}$$

Thus outsourcing is chosen by a labor-intensive firm. Antrás then presents empirical evidence, using intra-firm versus arm's-length trade, that gives good support to his model.

16.7 Summary

A number of empirical observations, particularly the fact that high-income countries are both the major sources but also the major recipients of multinational investment has led economists to re-think the old idea of foreign direct investment as simply capital flows, moving from where capital is abundant to where it is scarce (developing countries). A cornerstone of the new theory is the existence of firm-level (or multi-plant) economies of scale arising from the joint input or non-rivaled property of knowledge-based assets. Blueprints, processes, managerial techniques and so forth are costly to create but once created, can be applied to multiple plants at very low additional cost.

Combined with trade costs and differing factor intensities across activities (e.g., skilled-labor intensive headquarter services versus less skilled-labor-intensive production), knowledge-based firm scale economies generate rich models and predictions about what sort of firms will arise between what sorts of country pairs. Further, the predictions of the models are empirically testable and have generated good support in formal econometric estimation.

In parallel with this new learning is a literature on outsourcing versus vertical integration, previously termed internalization (the converse of outsourcing). In one strand of literature, the same properties of knowledge-based assets that lead firms to expand abroad in the first place create difficulties in contracting and in securing property rights over the knowledge. This can lead firm to choose vertical integration (subsidiaries) or exporting when they would, in principle, like to choose outsourcing (licensing) on the basis of cost. A more recent approach focuses on what is referred to as the property-rights theory of the firm and to the existence of hold-up problems arising from relation-specific investments. In the approach reviewed here (Antrás), capital-intensive firms choose vertical integration while labor-intensive firms choose outsourcing. Empirical work is in its early stages, but results are supportive to the theory.

REFERENCES

Antrás, Pol (2003), “Firms, Contracts, and Trade Structure”, *Quarterly Journal of Economics* 118(4), 1375-1418.

Braconier, Henrik, Pehr-Johan Norbäck, and Dieter Urban (2005), “Reconciling the Evidence on the Knowledge-Capital Model.” *Review of International Economics* 13(4): 770-86.

Brainard, S. Lael (1997), "An Empirical Assessment of the Proximity-Concentration Tradeoff between Multinational Sales and Trade", *American Economic Review* 87(4): 520-544.

Carr, David L., James R. Markusen, and Keith E. Maskus (2001), “Estimating the Knowledge-Capital Model of the Multinational Enterprise.” *American Economic Review* 91(3): 693-708.

Caves, Richard E. (2007). *Multinational Enterprise and Economic Analysis*. Cambridge: Cambridge University Press, third edition.

Davies, Ronald B (2008), “Hunting High and Low for Vertical FDI”. *Review of International Economics*, 16(2), 250–267, 2008

Davies Ronald B., Per-Johan Norbäck, A. Tekin-Koru (2009), “The Effect of Tax Treaties on Multinational Firms: New Evidence from Microdata”, *World Economy* 32 , 77-110.

Dunning, John H. (1973). “The Determinants of International Production”. *Oxford Economic Papers* 25: 289-336.

Ethier, Wilfred J. and James R. Markusen (1996), “Multinational Firms, Technology Diffusion and Trade,” *Journal of International Economics* 41, 1-28.

Helpman, Elhanan (1984), "A Simple Theory of Trade with Multinational Corporations", *Journal of Political Economy* 92(3): 451-471.

Markusen, James R (1984), “Multinationals, Multi-Plant Economies, and the Gains from Trade”, *Journal of International Economics* 16(3-4): 205-226.

Markusen, James R. (2002), *Multinational Firms and the Theory of International Trade*. Cambridge: MIT Press.

Markusen, James R. and Anthony J. Venables (2000), "The Theory of Endowment, Intra-Industry, and Multinational Trade", *Journal of International Economics* 52(2): 209-234.

ENDNOTES

1. The curvature of the two reaction functions follow from the assumption that $\beta + \gamma < 1$, which in turn implies that $\gamma/(1-\beta) < 1$ and $\beta/(1-\gamma) < 1$.

Table 16.1: World statistics, 2007

Affiliates sales as a share of world exports	1.82
Value added of affiliates as a share of world GDP	0.11
Affiliate exports as a share of world exports	0.33
Affiliate exports as a share of affiliate sales	0.18

Source: UNCTAD World Investment Report

Table 16.2: Developed countries as source *and* destination for FDI: developed countries' share of world totals

	FDI inflows	FDI outflows
2007	0.66	0.85
	FDI inward stock	FDI outward stock
1990	0.73	0.92
2007	0.69	0.84

Source: UNCTAD World Investment Report

Table 16.3: Local sales, export sales, and imports of foreign affiliates, 2007

	Affiliate local sales as as share of total sales	Affiliate exports as a share of total sales	Affiliate imports as a share of affiliate sales
United States	0.72	0.28	0.06
Japan	0.65	0.35	0.43
Sweden	0.78	0.22	0.16

UNCTAD World Investment Report 2008, Annex Tables B.12, B.15, B16

Table 16.4 Sales by US and Swedish manufacturing affiliates: shares in total, 2003 / 1998

	local sales	export sales back to parent country	export sales to third countries
USA			
2003	0.60	0.13	0.26
Sweden			
1998	0.65	0.08	0.27

Source: Markusen (2002), Davies, Norbaeck, Tekin-Koru (2009)

Figure 16.1: Decision tree for FDI

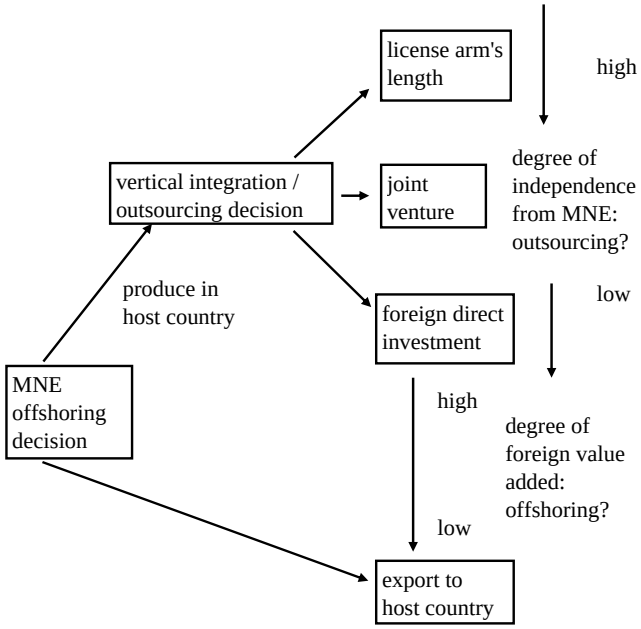
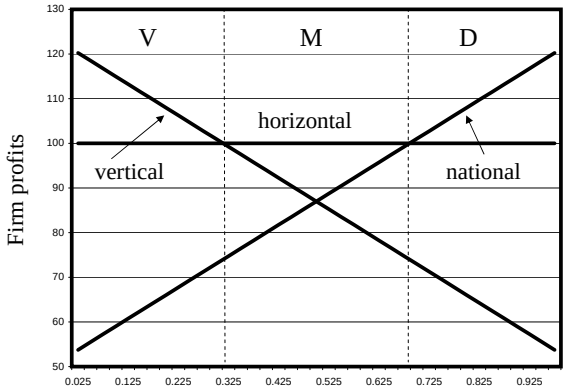


Figure 16.2: Relative size differences and choice of regime



Share of demand in country i

Figure 16.3: Affliate sales in the knowledge-capital model

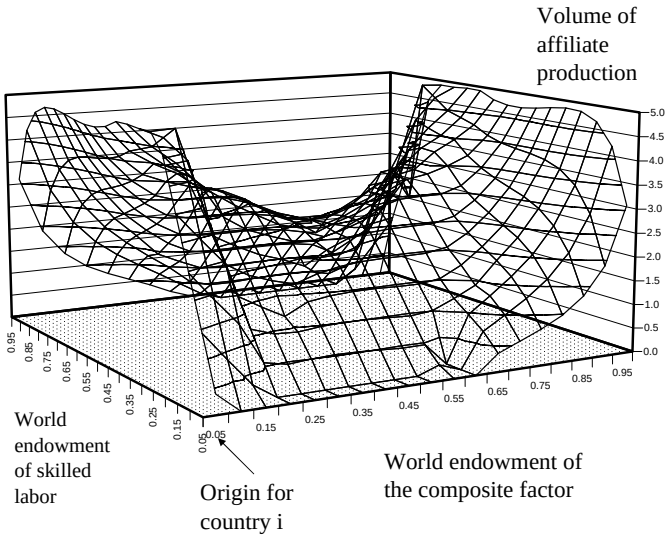


Figure 16.4: Affliate sales in the knowledge-capital model, restricted to horizontal firms

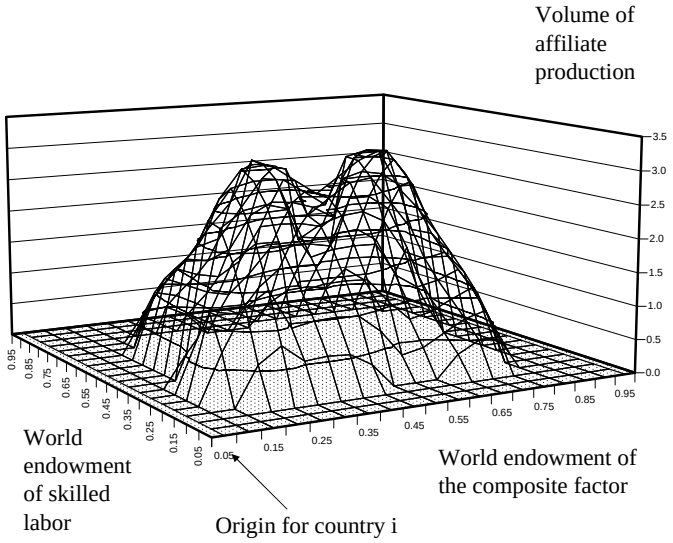


Figure 16.5: Affliate sales in the knowledge-capital model, restricted to vertical firms

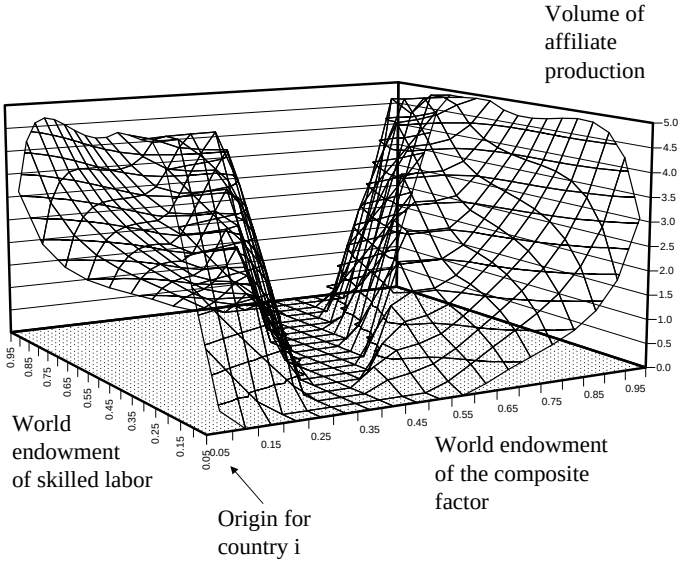
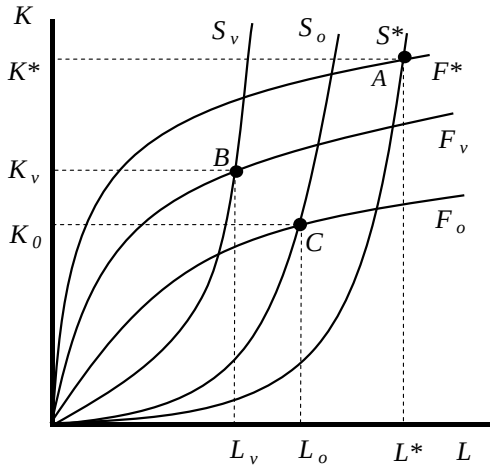


Figure 16.6: vertical integration versus outsourcing



A - first best, B - vertical integration, C - outsourcing

Chapter 17

FRAGMENTATION, OFFSHORING, AND TRADE IN SERVICES

17.1 Stylized facts, basic concepts

There are a few topics that have attracted a good deal of attention but which do not fit neatly into any simple model of trade. One of these is the fact that more things seem to be traded in the last decade or two, including intermediate goods and various types of services. Indeed, twenty years ago, many trade economists defined services as more or less non-traded.

A couple of terminology points first. There is a distinction between new goods and newly-traded goods. The former, of course, refer to things that simply did not exist before, while the latter refer to things that existed as non-traded goods or services which have now become traded. The latter is the topic of this chapter. The breaking up of the production process for a final good into different stages which are physically located in different countries has been referred to by a number of names, but we will use the term "fragmentation". Vertical specialization is an alternative name for fragmentation, though this has been used in a more narrow sense. The introduction of newly traded goods is sometimes referred to as the expansion of trade at the "extensive" margin whereas more trade in the same goods is often called expansion at the "intensive" margin. By whatever name, trade economists have argued persuasively that much of the great increase in trade over the last decades has come by expansion at the extensive margin (Hummels, Rapoport, and Yi 1998; Hummels, Ishii, and Yi 2001; Yi 2003).

Often intermediate goods are produced in one or several countries, while final assembly occurs in another country. Of course, final assembly may also occur in many countries which is the stuff of horizontal multinationals discussed in Chapter 16. In many cases such as in electronics, the "upstream" intermediates or components are more capital and skilled-labor intensive while "downstream" assembly is less-skilled-labor intensive but there are surely examples which run the other way. Finally, we again draw the distinction between "outsourcing", which we reserve for production outside the ownership boundaries of the firm and "offshoring" which we reserve for geographical location of production outside the home country.

There are obvious advantages that fragmentation offer to the world economy. Think back to Chapter 15 where we used the concept of the factor-price-equalization set. If the countries' endowments are outside this set, then trade in goods alone cannot exhaust all gains from trade. In that Chapter, we discussed how additional gains from trade can be captured by allowing trade in factors. The same is true in the presence of increasing returns to scale as we discussed in Chapter 15. Trading intermediate goods and services also allows for additional gains, and in Chapter 16 we advanced the view that multinational firms could be thought of primarily as vehicles for the intra-firm trade in services such as management, technology, marketing, and finance.

There is no general theory of fragmentation at this time and it is probably fair to say that the literature consists of a lot of fairly specific situations. There few, if any, general welfare results about fragmentation and the expansion of trade at the extensive margin. Specifically, it is impossible to predict that all countries will be made better off by adding more traded goods, never mind the further complication of factor-price changes within economies. The fundamental problem is that newly-traded

goods and services introduce general-equilibrium effects such that some individual countries could be made worse off through adverse price changes. An example is taken from Markusen and Venables (2007). Suppose that our country is really good at making shirts, a product produced in two stages: fabric (textiles) and cut-and-sew. Our country is not great at either of those individually, but good enough at each to be very competitive on the world market. Now fragmentation of the industry is made possible. Suppose that we find that some countries are really good at textiles and some others really good at cut-and-sew: neither group was good at the other activity initially which is why we were so competitive. With trade in intermediates allowed, the world price of shirts is going to fall and our country is going to experience an adverse terms-of-trade change. In spite of our best efforts to adapt, we could be worse off.

The next two sections of this Chapter will look at two quite specific examples of fragmentation from three recent papers. The first example is from Markusen and Venables (2007) and Markusen (2010) while the second is from Grossman and Rossi-Hansberg (2008). The two provide an interesting contrast. Section 4 provides a somewhat general gains-from-trade theorem with respect to the effects of newly-traded goods and services and we do indeed see the crucial role of changes in the prices of existing traded goods for establishing welfare results. Section 17.5 looks at the issue of trade and FDI in services and the so-called four modes of trade.

17.2 Fragmentation and newly-traded intermediate goods

The first special case we will consider might be termed "conventional", with no normative connotation good or bad, in that it follows a number of general formulations in the literature. Examples are Deardorff (2001, 2008), Jones and Kierzkowski, (2001), and a series of articles in Arndt and Kierzkowski (2001). Empirical analysis is found in Feenstra and Hanson (1996, 1997). Initially-traded final goods are assembled from intermediate goods and services or "components" and may also use primary factors directly. Later innovations in transportation or management then allow some of the components to be directly traded.

An example is shown in Figure 17.1, which sticks to the Heckscher-Ohlin tradition for familiarity: there are two final goods, X_1 and X_2 and two primary factors K and L (or these could be skilled and unskilled labor). There are three intermediate goods, A, B, and C which are produced from primary factors L and K. A is the most capital intensive, B is in the middle, and C is the most labor intensive. X_1 is costlessly assembled (no added primary factor requirements) from A and B, and X_2 is costlessly assembled from B and C. All production functions as assumed to be Cobb-Douglas, so the value shares of factors in intermediates and value shares of intermediates in final production are constants. Figure 17.1 gives these shares. Including intermediate use, X_1 is capital intensive overall with factor shares 65 for capital and 35 for labor in the simulation we present shortly. X_2 is the mirror image, with shares 35 for capital and 65 for labor.

Figure 17.1

We cannot present a great many cases here and will stick with just one specific example to illustrate some general ideas. We assume three countries: 1, 2 and 3. Country 1 is capital (skilled labor) abundant with an endowment ratio K/L = 85/15 and country 3 is labor abundant with an endowment ratio K/L = 15/85. Country 2 is in the middle, with an endowment ratio 50/50. The model is symmetric, both with respect to production and with respect to countries. Note that the country 1 and 3 endowment ratios differ more from one than the factor intensities of A and C which are 80/20 and 20/80 respectively. This

is important for some results, but we will not comment more on this in more detail.

Table 17.1 shows some simulations of this model. The model is benchmarked (parameters picked) so that the integrated world equilibrium we have discussed before gives each country a welfare level of one and all factor and commodity prices are one. The first column of results gives results when X_1 and X_2 are freely traded, but trade in intermediates is not allowed. Because countries 1 and 3 have extreme endowment ratios, they are outside the factor-price equalization set, specialized in X_1 and X_2 respectively. Each of these countries has a high real price for its scarce factor and a low real price for its abundant factor, and each country has a welfare level (0.885), which is lower than in the integrated equilibrium. Because of the symmetry in the model, relative prices of X_1 and X_2 are one, and so country 2 is indifferent to trade and enjoys the same welfare level as in the integrated world equilibrium.

Table 17.1

The second column of results in Table 17.1 allows free trade in *A* and *C*, what we could call a "symmetric" fragmentation. Countries 1 and 3 become the sole producers of *A* and *C* respectively which are better suited to their factor endowments, and country 2 becomes specialized in producing *B* only. Countries 1 and 3 have a welfare gain, but have the usual Stolper-Samuelson effect, which raises the real income of the abundant factor and lowers the real income of the scarce factor. Perhaps the most unanticipated result of Table 17.1 is the large welfare gain for country 2: it is not the case that the country with the average world endowment is unaffected by the symmetric fragmentation.

The intuition for the large welfare gain for country 2 can be explained in relatively simple supply and demand terms. In the base case, we noted that country 2 is indifferent to trading goods only and, when trade in *A* and *C* open, it remains indifferent as long as prices all remain at one. Countries 1 and 3 are not at all indifferent. They are inefficient producers of *B*, so at prices of unity they want to export *A* and *C* to country 2 and import finished X_1 and X_2 in exchange. This drives down the prices of *A* and *C* relative to *B*, which creates a welfare gain for country 2 in the middle: country 2 imports *A* and *C* cheaply from countries 1 and 2 respectively, assembles these with local intermediate *B*, and exports finished X_1 and X_2 to 1 and 3 respectively. In this simulation, country 2 imports *A* and *C* at the price 0.91 while the domestic price of *B*, its specialty, is 1.10. Both factors share fully in this gain for country 2.

The third column of results in Table 17.1 shows the results for an asymmetric fragmentation in which only intermediate good *C*, the most labor intensive, is traded. *A* and *B* are not traded. Here we see a couple of interesting results. First, note that the relative prices of X_1 and X_2 are going to change: the relative price of X_1 rises and X_2 falls. Country 3 specializes more in *C*, and this drives down the price of *C* and hence of X_2 in general equilibrium. In this particular example, this negative terms-of-trade effect is actually strong enough to reduce the welfare of country 3 relative to the benchmark. The Stolper-Samuelson theorem again applies here, and the scarce factor *K* in country 3 suffers a loss as in country 2 as the theorem predicts. The opening of trade in *C* is pure trade creation for country 2 (it did not trade initially), so the gains-from-trade theorem fully applies country 2's welfare rises.

Perhaps the most interesting result in the third column of Table 17.1 is that not only does country 1's welfare increase, but both factors gain. The reason for this is that country 1 does not produce or import *C* either before or after fragmentation. Country 1 produces X_1 from domestically produced A and *B*: its relative factor prices are pinned down by its domestic specialization in X_1. Thus, its welfare level and the real prices of both factors rise by the same proportion.

Table 17.1 is a very specific and special case, but it shows both the opportunities and the pitfalls that are involved in expanding trade at the extensive margin. All countries can gain, but it is possible that some countries might lose due to an adverse terms-of-trade effect. We will shown in section 17.4 that this is, in fact, a very general proposition: a country can only lose if it suffers a deterioration in the price(s) of its initially exported good(s), otherwise it is assured of gain. Table 17.1 also shows that it is possible, though not inevitable, that all primary factors in a country might gain: country 2 in the second column and country 1 in the third column.

17.3 Fragmentation and trade in "tasks"

A recent paper by Grossman and Rossi-Hansberg (2008) has proposed a rather different structure which they term "tasks". A simplified version of their formulation is shown in in Figure 17.2. There are again two final goods, X_1 and X_2 and two primary factors, L and K. A task uses a single factor, either capital or labor. There are many labor tasks and many capital tasks. In our simplified version here, we assume that there are just two labor tasks, L_1 and L_2, and two capital tasks, K_1 and K_2. Each final good requires all labor tasks in equal amounts, and it is further assumed that there is no substitution possible among these tasks; e.g., X_1 might require exactly one unit of both L_1 and L_2 with no substitution possible. The same is true for capital tasks.

Figure 17.2

Where the final goods differ is in the ratio of labor to capital tasks, and hence there remains a Heckscher-Ohlin type of factor intensity. A numerical example is given in Figure 17.2. X_1 is assumed to be capital intensive, requiring 65 units of each of tasks K_1 and K_2 and 35 units of each of tasks L_1 and L_2 to produce 400 units of X_1 if all prices are initially normalized at unity. X_2 is the labor-intensive mirror image: 65 units of each of tasks L_1 and L_2 and 35 units of each of K_1 and K_2 are required.

There are two important differences between this and the formulation in the previous section. First, a country cannot, by definition, have a factor-endowment ratio that is more extreme than some intermediate (task) that is potentially traded. Tasks use only a single, primary factor and hence there are tasks that are potentially tradeable that have a more extreme factor intensity than any country (or at worse a tie, if there is a country that has only one factor). Second, if only a subset of labor or capital tasks are tradeable, then free trade in the subset is quite different from simply allowing labor or capital migration, as in Chapter 15. A trabeable L task must, for example, be combined in fixed proportions with a non-tradeable domestic L tasks produced by a domestic worker.

This last property is crucial in certain results that Grossman and Rossi-Hansberg emphasize in their paper and so further explanation is warranted. Suppose we have a small country 1 facing fixed world prices, but it is capital abundant and specialized in X_1 , so has a wage rate higher than the rest of the world. If task L_1 becomes tradeable, our country will, of course, import L_1. But each unit of imported L_1 must be combined with a unit of L_2 produced by a domestic worker. Let's do the thought experiment of holding domestic factor prices constant in country h. Production of X_1 is now strictly profitable since the labor "composite" input (tasks L_1 and L_2 in fixed proportions) is now cheaper: a combination of the lower cost of importing L_1 and the existing wage for task L_2. Capital is in fixed supply, so firms will demand more labor until the imported L_2 is equal to the domestic labor force, all of whom now do task L_2.

This expansion in X_1 output will raise the relative and real price of the composite capital input (in

fixed supply) and lower the price of the composite labor input. However, although the final price of the *composite* labor input may be below its price before task trade, the former is a combination of the lower import price of L_1 and the higher domestic wage paid for task L_2. Thus it may well be that the final price of domestic labor is higher than before task trade, and this is indeed a finding that Grossman and Rossi-Hansberg highlight. When a limited number of tasks can be traded, the domestic factor performing those tasks (competing with the imports) may be made better off. Grossman and Rossi-Hansberg make an interesting analogy between this trade in a limited number of tasks and technical change. In our simple case just discussed, the composite *L* task becomes cheaper at a given domestic wage rate, akin to a technical improvement in the production of the composite *L* task.

In order to provide some concreteness with the minimum possible algebra, suppose that the *composite* capital task (K_1 and K_2 in fixed proportions, perfect *complements*) and the *composite* labor task (L_1 and L_2 in fixed proportions, perfect *complements*) are perfect *substitutes* in producing final good X_1. This is equivalent to saying that there are two alternative ways of producing X_1:

$$X_1 = \left[\min\left(K_1, K_2\right) + \min\left(L_1, L_2\right)\right] \quad \Rightarrow \quad X_1 = \min\left[K_1, K_2\right] \quad X_1 = \min\left[L_1, L_2\right] \qquad (17.1)$$

Let *r* and *w* be the domestic prices of capital and labor, and *t* the price of imported L_1. The *fixed* world price of X_1 is denoted p_1. Superscript *n* denotes no trade in tasks while superscript *t* denotes trade in L_1 permitted. The cost functions corresponding to (17.1) and the price-equals-marginal-cost conditions for country 1 are as follows:

$$p_1 = (r + r) \qquad p_1 = (w^n + w^n) \qquad p_1 = (w^t + t) \qquad t < w^n \qquad (17.2)$$

The second equation applies in the absence of trade in tasks and the third equation with imports of task L_1. In this very special case, it is clear that allowing trade in L_1 does not affect the return to domestic capital. However, with $t < w$, it must *raise* the price of *w*, the wage of the domestic labor that (superficially) competes with imported task L_1: $w^t > w^n$.

While Grossman and Rossi-Hansberg put a great emphasis on this interesting result, it is of course a special case. Table 17.2 presents some simulations of our simplified version of their model. In the top panel of the Table, the countries are the same size and have the same factor endowments as countries 1 and 3 of the previous section. X_1 and X_2 are assumed Cobb-Douglas as in the previous section while the two labor tasks are required in fixed proportion and the same for capital tasks. Factor shares etc. are chosen to be the same as in the previous section, so the benchmark when the countries trade in X_1 and X_2 only produces exactly the same outcome as in Table 17.1. The two models are completely equivalent when only X_1 and X_2 are traded.

Table 17.2

The first counterfactual simulation in Table 17.2 (second column of results) allows trade in tasks L_1 and K_2, a symmetric fragmentation similar to that in the previous section where we allowed *A* and *C* to both be traded. Both countries gain and the Stolper-Samuelson property again appears: there are real gains to each country's abundant factor and real losses for each country's scarce factor. The second counterfactual allows only trade in L_1, with the results reported in the third column of Table 17.2. Both countries gain, but country 1, the importer of task L_1 gains less. Lower down in the Table we see the reason: the added production capacity in country 1 and lower capacity in country *2* moves the terms of trade against country 1: the relative price of X_1, its export good, declines. The direction of factor-price

changes is the same as in the symmetric fragmentation though the magnitudes are different.

Panel B of Table 17.2 conducts the experiment we discussed above: country 1 is small (everything is the same except the factor endowment of country 2 is multiplied by 100). Country 1's welfare and the prices of both factors are higher in the benchmark now because the relative price of X_1, country 1's export good, is higher in the benchmark. The first counterfactual reported in the second column of Panel B is again the symmetric fragmentation of allowing trade in both L_1 and K_1. Country 1 gets a huge welfare increase, but again we see that the scarce factor labor suffers a real income loss. Capital exports from country 1 reduce the capital available for local production and so lower the marginal product of domestic labor in spite of being able to import L_1. The third column of Table 17.2 shows the result emphasized above in which allowing trade in L_1 only actually raises the prices of both factors of production in country 1. It also demonstrates that the result does not rely on the perfect substitutes assumption used to explain the result above (Table 17.2 assumes that L and K composites are Cobb-Douglas substitutes exactly the same as Table 17.1).

As a final point, both Tables 17.1 and 17.2 report the proportional change in the volume of trade in the counterfactuals. In all cases, these numbers are positive, which certainly concurs with intuition: more things traded means more trade. For exactly this reason, we might note that Markusen and Venables (2007) and Markusen (2010) emphasize that this is not a general result and that it is easy to construct cases in which some countries will trade less as a consequence of fragmentation. A simple example should suffice. Suppose that there are two goods, wheat and cars. Suppose that country 1 is generally good at car production but simply cannot produce tires. Then if only final goods are traded, country 1 will be forced to export wheat in exchange for cars. Now allow tires to be traded. Country 1 can now just import the tires and pay for this by exporting a few cars and/or a little wheat. The volume of trade falls. Markusen and Venables show that this is by no means pathological and it occurs for significant subsets of countries in multi-country simulations.

17.4 A gains-from-trade theorem

General gains-from-trade results in trade theory are hard to obtain. The basic textbook result only proves that free trade is better for a country than autarky or, somewhat more generally, any level of restricted (but not subsidized!) trade is better than autarky. However, there is no general result that says that more trade is better than less (but positive) trade. This is due to the possibility of adverse price changes; e.g., further liberalization by a large country can make it worse off. However, this is exactly the type of result we are seeking here: under what conditions will the introduction of an additional set of trade possibilities to an already-existing set of traded goods improve welfare? Our approach follows the classic "revealed preference" methodology that we have used several times. The theorem itself is a much simplified version of a more general proof in Markusen (2010).

X denotes a vector of final goods quantities, with the vector p denoting their prices. Similarly, Z denotes a vector of intermediate goods with prices q. In our explicit model above, X (now denoting X_1, X_2) and Z (now denoting A, B, and C) were disjoint sets. The important distinction two-fold: (a) only X goods enter into welfare-producing consumption and (b) we will *define* the absence of fragmentation to mean that goods cannot be traded for intermediate use. That is, the Z goods are initially non-traded.

Superscript f denotes prices with fragmentation (trade in intermediates allowed), while superscript n denotes prices in the initial no-fragmentation equilibrium. Thus vectors of *world* final

goods prices with and without fragmentation are p^f and p^n respectively. q^f denotes the world prices of the Z goods when they can be traded and q^n denotes their domestic prices when they are non-traded. In the simplified proof here, we assume that the X goods are all freely and costlessly traded (again, see Markusen (2010) for a more complete proof with trade costs). Thus the domestic prices of the X goods must equal world prices.

The notation that we have been using, X for production and D for consumption is now a little awkward, so we make a modification to avoid confusion. Just X and Z will be used: a subscript "o" on an X or Z quantity denotes production (output) of the good and a subscript "c" denotes consumption. For an intermediate good, "consumption" means its use as a domestic input in a final good. Thus for both final and intermediate goods, a positive value for production minus consumption of the good indicates it is an export good. We will also simplify the notation a bit: rather than writing lots of summations, we will use a shorthand such as

$$\sum p_i X_i \equiv pX$$

We assume a competitive, undistorted economy, so production efficiency applies: the value-added of the economy in regime j (j = f, n) must be maximized at that regime's prices. Specifically, the value added in regime f must be greater than or equal to regime-n value added when the latter is evaluated at regime-f prices. Total value added for the economy is given by the value of final goods production minus intermediate usage ("consumption" of intermediates) plus the value of intermediate goods production. The production efficiency condition is then given by

$$(p^f X_o^f - q^f Z_c^f) + q^f Z_o^f \geq (p^f X_o^n - q^f Z_c^n) + q^f Z_o^n \tag{17.3}$$

This rearranges to an inequality on final production plus net exports of intermediates.

$$p^f X_o^f + (q^f Z_o^f - q^f Z_c^f) \geq p^f X_o^n + (q^f Z_o^n - q^f Z_c^n) \tag{17.4}$$

The last term in parentheses on the right-hand side is zero: intermediates are not traded in the n-regime, so production and consumption are the same for each good.

$$p^f X_o^f + (q^f Z_o^f - q^f Z_c^f) \geq p^f X_o^n \tag{17.5}$$

Now we introduce the balance of trade constraints for each regime, which require that the value added in production equals the value of final consumption.

$$p^f X_o^f + (q^f Z_o^f - q^f Z_c^f) = p^f X_c^f \tag{17.6}$$

$$p^n X_o^n = p^n X_c^n \tag{17.7}$$

Substitute the right-hand side of (17.6) for the left-hand side of (17.5)

$$p^f X_c^f \geq p^f X_o^n \tag{17.8}$$

Now add the right-hand side of (17.7) and subtract the left-hand side, and also add and subtract the term $p^f X_c^n$. (17.8) then becomes

$$p^f X_c^f \geq p^f X_o^n + p^n X_c^n - p^n X_o^n + p^f X_c^n - p^f X_c^n \tag{17.9}$$

$$p^f X_c^f \geq p^f X_c^n + p^f (X_o^n - X_c^n) - p^n (X_o^n - X_c^n) \tag{17.10}$$

$$p^f X_c^f \geq p^f X_c^n + (p^f - p^n)(X_o^n - X_c^n) \tag{17.11}$$

The classic revealed-preference criterion for gains from fragmentation requires that the left-hand side of (17.11) is greater than or equal to the first term on the right-hand side: the value of regime-f consumption is revealed preferred to regime-n consumption valued at regime-f prices. A *sufficient* condition for this to be true is that the second additive term on the right-hand side of (17.11) is non-negative.

Inequality (17.11) tell us that a sufficient condition for the added trade in the intermediate *Z* goods to improve welfare $(p^f X_c^f \geq p^f X_c^n)$ is that world prices of the initially-traded *X* goods don't change $(p^f = p^n)$. Actually, all that is needed is that the world *relative* prices of the *X* goods don't change (a numeraire rule could be that the *X* prices sum to one, so the price vectors p^f and p^n will be the same before and after *Z* trade). The absence of terms-of-trade changes for the *X* goods could be referred to as a "neutral" or "symmetric" fragmentation, emphasizing that these are *definitions* (e.g., a symmetric fragmentation is one that leaves relative final-goods prices unchanged).

The absence of relative price changes for initially-traded goods rules out the sort of example we posed above: a country is good at shirts but the result of fragmentation is that efficient fabric and cut-and-sew producers are able to exploit their process advantage and cause a fall in the world relative price of shirts. This welfare loss cannot happen if the relative price of shirts to other initially-traded goods doesn't change and in fact the country will generally benefit by exploiting some differences between its (old) domestic prices for fabric and cut-and-sew and the new world prices.

The condition that world prices do not change is very restrictive and "overly" sufficient. Consider again the last term in (17.11): if this is non-negative, then it is sufficient for gains from fragmentation. Let $X_e^n \equiv (X_o^n - X_c^n)$ denote the vector of net exports in regime n (no trade in intermediates), also referred to as the "initial" net export vector. The last term in (17.11) can be written as either of the following

$$(p^f - p^n)X_e^n \geq 0 \quad \textit{or} \quad p^f X_e^n \geq 0 \quad \textit{since} \quad p^n X_e^n = 0 \text{ by trade balance} \tag{17.12}$$

The first expression is just a simple correlation between the changes in domestic final-goods prices and the initial net export vector. We will have gains from fragmentation if, "on average", the prices of initially-exported goods rise and the prices of initially-imported goods fall; that is, there is an "aggregate" terms-of-trade improvement. Noting that the value of the initial trade vector at initial prices is zero (trade balance: equation (17.7)), we have the second expression. If, after fragmentation, the country were to retain its initial net export vector, it would run a trade surplus if the inequality is strict. It

could then improve its welfare by cutting some exports and/or increasing some imports to restore trade balance.

Unfortunately, the inequality in (17.12) is unlikely to hold for all countries. If there are only two countries, then it cannot hold for both, because the elements of their net export vectors are equal and opposite in sign. One country must have an aggregate terms-of-trade deterioration. However, we must emphasize again that (17.12) is a *sufficient* condition for $(p^f X_c^f \geq p^f X_c^n)$, which is itself a *sufficient* condition for gains from trade. It is easy to produce numerical examples where a country benefits substantially from fragmentation and outsourcing in spite of a terms-of-trade loss and all countries gain.

17.5 Trade and foreign direct investment in business services

Services cover a very wide range of items and industries, and we have made a decision here to restrict our analysis to business services, which are generally intermediate goods (services supplied to other businesses), consistent with the overall focus of this chapter. It is indeed the case that trade and FDI in business service has increased greatly in the last decade or two. Statistics are presented in Table 17.3 for the United States and a very thorough analysis is found in Francois and Hoekman (2010). It is also clear that much of the increased activity is mediated by multinational firms, whether that is actual FDI (we measure sales of foreign affiliates) or intra-firm trade between parent and affiliate. Table 17.3 shows that both trade and FDI have increased faster relative to all trade and that they now account for a very significant portion of US international economics activity.

Table 17.3

One reason to study trade and FDI in business services, in addition to the rate of growth, is that they are often subject to very different restrictions than trade in goods. Barriers to trade and investment in services can be roughly broken down into what we could call "natural" economic costs and "policy-imposed" costs. For the former, we include communications and transport costs (workers flying between countries), language, customs, time zones, the need for face-to-face interaction, and so forth.

Policy-induced barriers, henceforth "barriers" for this section, to trade in services take diverse forms and therefore affect service suppliers' cost functions differently. Regulatory policies, in addition to explicit and implicit barriers to trade in services, generally fall into one of four basic categories. First, there can be quantity-based restrictions imposed on services suppliers that explicitly restrict the volume of services imported, similar to a quota. The use of a fixed number of licenses available or access to only certain firms or sectors also falls into this category. If a "quota" type policy is only applied to imported services, then we would expect to see more multinationals establishing affiliates in the market (all other things equal). This is similar to a "tariff jumping" activity discussed in traditional theories of the multinational firm.

Second, there are numerous barriers to establishment that restrict foreign supply of services due to the high costs of establishing a commercial presence. Policies regarding licensing procedures, requirements, and fees can be prohibitive. Bureaucratic red tape, requirements for local management, or lack of transparency all have detrimental effects on the fixed costs of establishing commercial presence for multinationals. These may create a substitution effect, with multinationals supplying the service from abroad rather than establishing an affiliate.

Third, barriers to trade and establishing commercial presence in services may take the form of restricting the use of inputs. This category can include restrictions on workers, required percentages of locally produced material inputs, as well as barriers or limits on the use of networks or media for promotion and/or marketing purposes. These policies can greatly increase the costs of operations for foreign suppliers and may be prohibitive to entering the market.

The fourth category of restrictions encompasses the various domestic regulatory barriers that take many forms and are often overlooked when discussing impediments to trade and investment. These include policies regulating professional qualification, residency and citizenship restrictions, obligatory membership in local professional association, juridical requirements, and limitations of inter-professional cooperation. While the policies and regulations may not explicitly target foreign firms, they often have this effect in practice. Regulations on professional qualifications are important domestic policies so as to guarantee a level of skill and professionalism to consumers. However, when these policies require residency, citizenship, or involve re-certification for professionals with comparable certifications from another country, they become costly to foreign multinationals. While they are not explicit barriers to trade, these policies severely increase the fixed costs (time and money) for a firm establishing commercial presence and may be prohibitive as well.

It is beyond the scope of this Chapter to present a detailed analysis of these issues, but we would like readers to at least be aware of the basic ideas involved. Figure 17.3 presents a schematic of the general principles. Final consumption is assumed to be composed of two industries: manufacturing and agriculture. Business services are an intermediate input into both of these industries which also use primary factors of production, skilled and unskilled labor. Following closely on our modeling of multinational firms in the previous Chapter, business services can be decomposed into a headquarters and affiliate offices. Banks, finance, and insurance companies, for example, may have a headquarters in New York or London, and have offices throughout the world. Similar to a horizontal multinational (what many services multinationals are), a firm has one headquarters but possibly many offices.

Figure 17.3

Figure 17.3 notes the differences in trade in services and foreign direct investment in services. In the upper part of the chart, we note that trade in services occurs when there is an international geographic separation between the service firm and the service purchaser in agriculture or manufacturing. In the lower part of the Figure, we note that FDI in services occurs when there is a geographic separation between headquarters and one or more offices. Either type of separation can occur without the other; for example, a horizontal firm tends to have geographic separation between headquarters and some offices, but not between the offices and service buyers.

As briefly discussed above, quite a range of natural and policy restrictions impact on whether or not it is easy and/or legal to fragment headquarters from offices and fragment the service provider from final user. Figure 17.4 characterizes these restrictions in terms of what are know as the four modes of service trade from the World Trade Organization's GATS: general agreement on trade in services. The schematic in Figure 17.4 has two countries, north and south, and the headquarters of a service firm headquartered in the north is represented by the central box. A possible northern office and southern office are represented by the two boxes to either side.

Figure 17.4

The Figure shows four possible ways in which services can be provided by the offices to north or south manufacturing and agricultural firms. The channel on the upper left has the headquarters and office both in the north and providing services to northern firms. We mark this as "always feasible" since it is a strictly domestic activity. Mode 1 of the GATS is defined as cross-border trade in services, which is the channel on the lower left of Figure 17.4: the northern office of the northern firm supplies services to southern firms. The northern firm may also have a southern office, which is by definition FDI in services. This is known as Mode 3 trade in services in the GATS terminology. This southern office can supply services to local firms, which is represented by the channel on the lower right of Figure 17.3. Finally, the southern office may be able to supply services to northern firms which would require both Modes 1 and 3 being feasible. An example of this last type of supply is call centers which are located in developing countries such as India. A customer in the north needing service from a northern firm finds their phone call re-routed to a call center in India.

For completeness, GATS mode 2 is when the buyer travels to the seller's country of origin. Traditionally, this category has been dominated by tourism. But the extensive margin of Mode 2 is expanding as in "medical tourism". Patients in high-income countries are finding that they can get high-quality surgery or dental work done in places as diverse as Hungary, India, and Mexico for a fraction of the cost of having it done at home. We have seen websites in Ireland, for example, that advertise dental holidays in Hungary, in which you get round-trip plane tickets, two days in a resort, and all your dental work done for less than just the latter would cost in Dublin. GATS Mode 4 is the international movement of person to work abroad, typically for limited periods of time, rather than permanent migration. There is obviously some overlap and fuzziness between this and Mode 3. If Gregorz the Polish plumber goes to work in Denmark, that is Mode 4; if Gregorz Plumbing Incorporated opens a one-man office in Copenhagen, that is Mode 3.

As suggested earlier, both natural and policy-imposed restrictions can make it costly or even prohibitive for a firm to engage in one or more of the modes shown in Figure 17.4. If face-to-face contact is required, for example, then Mode 1 can be prohibitively costly. If countries restrict the right of foreign firms to establish or acquire local subsidiaries (known as right of establishment), then Mode 3 is difficult. The principle of national treatment means that foreign firms are supposed to be treated exactly the same as domestic firms, but this is often violated in practice (de facto) by rules requiring managers to be local citizens and so forth.

Finally, arguments about classifying a barrier and its degree of restrictiveness abound between countries. Part of the difficulty is that many service industries are, by their very nature, highly regulated. Consider banking, finance, insurance, architecture and construction, telecommunications, legal services and medical services. These tend to be regulated in all countries and while foreign firms see such regulations as barriers, the local firms and governments always tend to claim that they are non-discriminatory against foreign firms.

17.6 Summary

It has been documented that a good deal of the expansion of trade over the last decade or two has been through the extensive margin, meaning trade in goods and services which were not previously traded, rather than through the intensive margin, trading larger quantities of the same things. Much of this new trade seems to be in intermediate goods and services, as firms fragment the production process to take advantage of favorable factor prices or other considerations across geographic locations.

The theoretical literature on fragmentation and outsourcing is now extensive, but there is no general model on which we can rely. We presented two special cases from recent literature that posit different mechanisms as to how we might think about fragmentation and newly-traded goods. These are interesting but even in these rather restrictive special cases, results still depend on choices of parameters such as country sizes and factor intensities.

A simplified gains-from-trade theorem pinpoints one difficulty in trying to obtain general results. This difficulty is that, when new goods and services become traded, the prices of existing goods change in general-equilibrium, which leads to terms-of-trade effects for the countries involved. A country could lose through fragmentation because the components and stages of production for something the country was originally very good at can now be individually done more efficiently by several other countries. It is easy to create specific examples in which everyone gains but unfortunately also not impossible to create examples in which someone loses.

We closed with a discussion of trade and FDI in intermediate business services, a fast-growing component of total world economic activity. The theoretical issues here are much the same as in Chapter 16 on multinational firms and the choice between exporting and FDI. What makes trade and FDI in business services of particular interest is that many of the natural and policy barriers to trade and investment differ substantially from those restricting trade in goods.

REFERENCES

Arndt, Sven W. and Henryk Kierzkowski (editors) (2001), *Fragmentation: New Production Patterns in the World Economy*. Oxford University Press, Oxford.

Brainard, S. Lael. and Susan. Collins (editors.) (2006), *Brookings Trade Forum 2005: Offshoring White-collarWork* (Washington, DC: Brookings Institution).

Deardorff, Alan V. (2001), "Fragmentation across Cones," in Sven W. Arndt and Henryk Kierzkowski, eds., *Fragmentation: New Production Patterns in the World Economy*, Oxford: Oxford University Press, 35-51.

Deardorff, Alan V. (2008), "Gains from Trade and Fragmentation", in Steven Brakman and Harry Garretsen, eds., *Foreign Direct Investment and the Multinational Enterprise*, CESifo Seminar Series, Cambridge: MIT Press, 155-169.

Feenstra, Robert C. and Gordon H. Hanson (1996), "Globalization, Outsourcing, and Wage Inequality", *American Economic Review*, 86, 240–45.

Feenstra, Robert C. and Gordon H. Hanson (1997), "Foreign Direct Investment and Relative Wages: Evidencefrom Mexico's Maquiladoras", *Journal of International Economics*, 42, 371–93.

Francois, Joseph and Bernard Hoekman (2010), "Service Trade and Policy", *Journal of Economic Literature* 48, 642-692.

Grossman, Gene. M., Estaban Rossi-Hansberg (2008), "Trading Tasks: A Simple Theory of Offshoring", *American Economic Review* 98, 1978-1997.

Hanson, Gordon. H., Raymond J. Mataloni, and Matthew J. Slaughter (2005), "Vertical production networks in multinational firms", *Review of Economics and Statistics* 87, 664-678.

Hummels, D., Rapoport, D., Yi, K-M. (1998), Vertical Specialization and the Changing Nature of World Trade. *Federal Reserve Bank of New York Economic Policy Review* 4, 79-99.

Hummels, David, Ishii, J. and K.-M. Yi (2001), "The Nature and Growth of Vertical Specialization in World Trade", *Journal of International Economics* 54, 75-96.

Jones, R. W., Kierzkowski, H. (2001), A Framework for Fragmentation, in: Arndt, S.W.. Kierzkowski, H., (Eds), *Fragmentation: New Production Patterns in the World Economy*, Oxford University Press, Oxford.

Leamer, Edward and James Levinsohn (1995), "International trade theory; the evidence", in: Gene Grossman and Kenneth Rogoff (editors), *Handbook of International Economics, Vol. 3*. North-Holland, Amsterdam, 1339–1394.

Markusen, James R. (2010), "Fragmentation and offshoring: a general gains-from-trade theorem and some specific cases", working paper.

Markusen, James R. and Anthony J. Venables (2007), "Interacting factor endowments and trade costs: a multi-country, multi-good approach to trade theory", *Journal of International Economics* 73, 333-354.

Markusen, James R. and Bridget Strand (2009), "Adapting the Knowledge-Capital Model of the Multinational Enterprise to Trade and Investment in Business Services", *World Economy* 32, 6-29.

Ng, F., Yeats, A. (1999), "Production sharing in East Asia; who does what for whom and why", World Bank Policy Research Working Paper 2197, Washington, DC.

Venables, Anthony J. (1999), "Fragmentation and multinational production", *European Economic Review*,43, 935-945.

Yeats, A. (1998), "Just how big is global production sharing?", World Bank Policy Research Working Paper 1871, Washington, DC.

Yi, K-M. (2003), "Can Vertical Specialization Explain the Growth of World Trade?", *Journal of Political Economy* 111, 52-102.

Table 17.1: Simulation results
Markusen (2010)

	Benchmark No trade in Intermediates	Trade in A (most K-int) C (most L-int)		Trade in C (most L-int) No trade in A, B	
	Level	Level	Prop change from bench	Level	Prop change from bench
Welfare country 1 (K abundant)	0.885	0.903	0.020	0.944	0.067
real price of labor	2.064	1.560	-0.244	2.202	0.067
real price of capital	0.676	0.789	0.167	0.722	0.067
Welfare country 2	1.000	1.103	0.103	1.035	0.035
real price of labor	1.000	1.103	0.103	0.894	-0.106
real price of capital	1.000	1.103	0.103	1.177	0.177
Welfare country 3 (L abundant)	0.885	0.903	0.020	0.849	-0.041
real price of labor	0.676	0.789	0.167	0.750	0.109
real price of capital	2.064	1.560	-0.244	1.420	-0.312
Price of X	1.000	1.000		1.068	
Price of Y	1.000	1.000		0.937	
Proportional change in trade volume			1.492		0.861

Table 17.2: Simulation results
Grossman / Rossi-Hansberg (2008)

Panel A: Countries symmetric	Benchmark No trade in Intermediates	Trade in L1 and K1		Trade in L1	
	Level	Level	Prop change from bench	Level	Prop change from bench
Welfare country 1 (K abundant)	0.885	0.994	0.123	0.923	0.043
real price of labor	2.065	1.395	-0.324	1.838	-0.110
real price of capital	0.677	0.923	0.363	0.762	0.126
Welfare country 2 (L abundant)	0.885	0.994	0.123	0.964	0.089
real price of labor	0.677	0.923	0.363	0.786	0.161
real price of capital	2.065	1.395	-0.324	1.974	-0.044
Price of X	1.000	1.000		0.922	
Price of Y	1.000	1.000		1.085	
Proportional change in trade volume			0.089		0.436

Panel B: Country 1 small (relative endowments = panel A:)	Level	Level	Prop change from bench	Level	Prop change from bench
Welfare country 1 (K abundant)	1.146	2.033	0.774	1.396	0.218
real price of labor	2.674	0.433	-0.838	2.983	0.116
real price of capital	0.877	2.315	1.640	1.116	0.273
Proportional change in trade volume			0.329		0.507

TABLE 17.3: US Foreign Affiliate Sales and Cross Border Trade

Outward US Foreign Affiliate Sales-all countries

	1999	2005	% change
Total Sales-All Industries	2316654.8	3276024.4	41.41%
Total Private Services	353200.0	528000.0	49.49%
Information	63236.5	97069.9	53.50%
Finance & Insurance	86337.1	140341.6	62.55%
Finance	32330.4	43847.0	35.62%
Insurance	54006.6	96495.7	78.67%
PST	65290.2	97490.9	49.32%

Sales in Total Private Services = 15.25% and 16.12% of All Industries Sales in 1999 and 2005 respectively
PST- Professional, Scientific, and Technical Services
All data are in millions of 2000 US dollars

Inward Foreign Affiliate Sales-all countries

	1999	2005	% change
Total Sales-All Industries	1831561.1	2213172.5	20.84%
Total Private Services	293500.0	389000.0	32.54%
Information	46440.3	48138.5	3.66%
Finance & Insurance	95840.7	104308.3	8.84%
Finance	15651.8	25458.9	62.66%
Insurance	80188.9	78849.4	-1.67%
PST	15757.0	49648.7	215.09%

Sales in Total Private Services = 16.02% and 17.57% of All Industries Sales in 1999 and 2005 respectively
PST- Professional, Scientific, and Technical Services
All data are in millions of 2000 US dollars

Cross-Border Trade- All countries-Exports

	1999	2005	% change
All Industries	1287247.6	1586285.5	23.23%
All Industries-Affiliated	194697.1	186463.5	-4.23%
All Industries-Unaffiliated	1092550.5	1399822.0	28.12%
Total Private Services	265100.0	368000.0	38.82%
Total Private Services-Affiliated	32952.4	44441.2	34.86%
Total Private Services-Unaffiliated	73245.5	101278.7	38.27%
Financial Total	17789.3	31626.3	77.78%
Financial-Affiliated	4087.2	4312.0	5.50%
Financial-Unaffiliated	13702.2	27314.3	99.34%
Insurance Total*	3119.2	6011.5	92.73%
BPT Total	54683.1	73911.2	35.16%
BPT-Affiliated	26379.5	37062.1	40.50%
BPT-Unaffiliated	28303.5	36849.1	30.19%

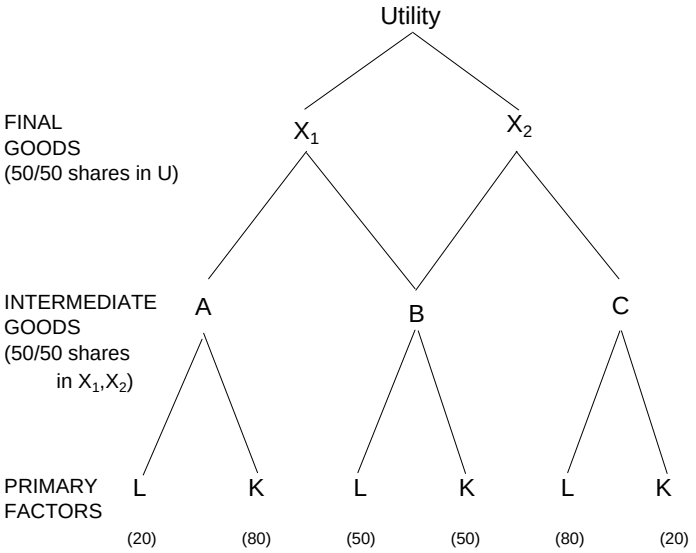
Exports of total private services = 20.59% and 23.19% of trade in all industries in 1999 and 2005 respectively
*Insurance transactions are considered unaffiliated by BEA
BPT-Business, Professional, and Technical Services

Cross-Border Trade- All countries-Imports

	1999	2005	% change
All Industries	1543920.8	2177244.7	41.02%
All Industries-Affiliated	186215.3	231946.0	24.56%
All Industries-Unaffiliated	1357705.5	1945298.8	43.28%
Total Private Services	183000.0	282000.0	54.10%
Total Private Services-Affiliated	25670.2	35340.6	37.67%
Total Private Services-Unaffiliated	31048.8	53285.4	71.62%
Financial Total	9623.2	11105.6	15.40%
Financial-Affiliated	6130.7	5192.0	-15.31%
Financial-Unaffiliated	3492.5	5913.6	69.32%
Insurance Total*	9593.9	25064.2	161.25%
BPT Total	28238.2	42913.2	51.97%
BPT-Affiliated	19462.0	29868.1	53.47%
BPT-Unaffiliated	8776.1	13045.1	48.64%

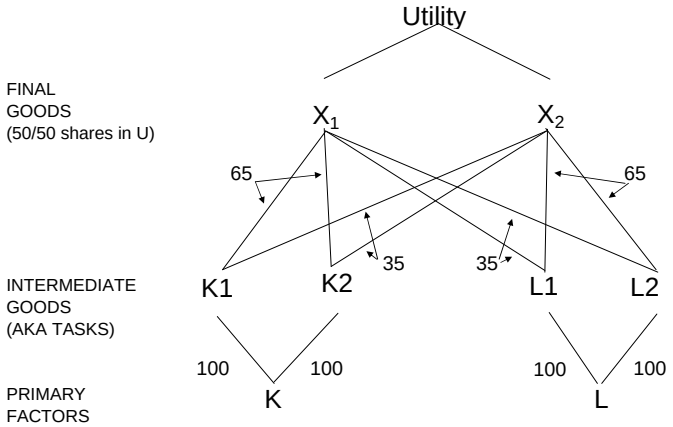
Imports of total private services = 11.85% and 12.95% of trade in all industries in 1999 and 2005 respectively
*Insurance transactions are considered unaffiliated by BEA
BPT-Business, Professional, and Technical Services

Figure 17.1: Structure of production
Markusen / Venables (2007)



FINAL GOODS	ALWAYS TRADABLE AT COUNTRY-SPECIFIC TRADE COST
INTERMEDIATE GOODS	NONE / SOME / ALL TRADABLE AT COUNTRY-SPECIFIC TRADE COST
PRIMARY FACTORS	NOT TRADABLE

Figure 17.2: Structure of production
Grossman / Rossi-Hansberg (2008)



FINAL GOODS	ALWAYS TRADABLE
INTERMEDIATE GOODS	NONE / SOME / ALL TRADABLE
PRIMARY FACTORS	NOT TRADABLE

Figure 17.3: Structure of production
Trade and FDI in intermediate business services

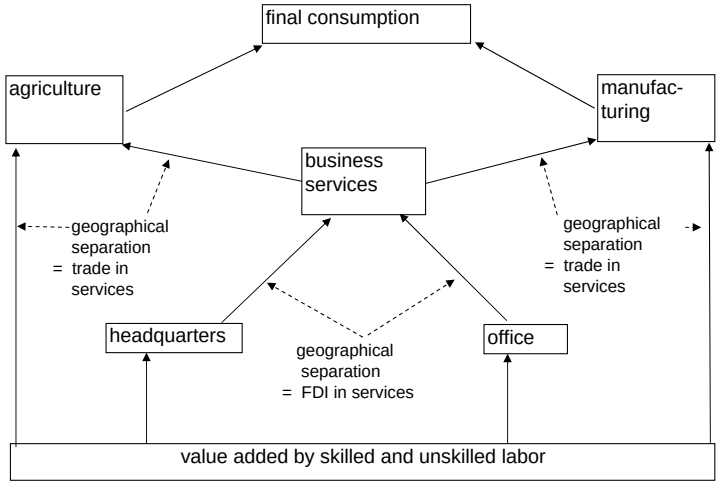
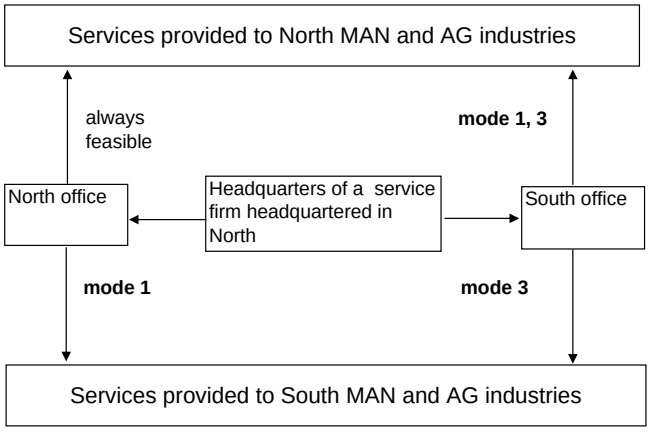


Figure 17.4: Modes of trade in services for a northern service firm



MAN: manufacturing, AG: agriculture

PART FOUR

TRADE POLICY

Chapter 18

TARIFFS AND TRADE SUBSIDIES IN COMPETITIVE TRADE MODELS

18.1 Tariffs, welfare and factor prices in a small economy

In Chapter 5, we contrasted a free-trade equilibrium with an autarky equilibrium, where a country does not trade at all. Both of these extremes are virtually unheard of in practice. Instead, a country will engage in trade but the government of that country will erect various barriers to restrict trade. The most common of these barriers are taxes levied on the importation of foreign goods. These taxes are commonly referred to as "tariffs," but they are simply a form of commodity taxation. Tariffs are sometimes levied on exports as well as on imports, and there are many cases where governments actively promote exports in ways that can be thought of as export subsidies. While there are other forms of trade restriction, this Chapter will concentrate on tariffs. Other barriers, such as import quotas, will be discussed in the following Chapter.

For now, there are essentially two reasons that governments may choose to levy taxes on trade. The more important objective is to protect the operations of domestic industries that compete with imports. Taking a Heckscher-Ohlin framework, for example, we would expect import restrictions to be more severe in sectors that intensively use an economy's scarce factors. In a specific-factors' framework, the tariff could protect both labor and capital owners in these import-competing industries. The second objective would be to raise revenues for the government. This practice is common in many developing countries, where it is easier to tax international trade at the border than to establish broad-based income taxes. Indeed, many exporters of primary products tax their foreign sales for revenue purposes. However, trade taxes are relatively unimportant as sources of revenue for developed economies.[1]

Consider first an import tariff and assume that good X_1 is the import good. Assume perfect competition and constant-returns to scale. Increasing returns and imperfect competition raise quite different issues and are the subject of Chapter 20. Let p denote domestic prices (producer and consumer prices), and let p^* denote world prices. An import tariff on X_1 leads to the following relationship between domestic and world prices.

$$\frac{p_1}{p_2} = \frac{p_1^*(1+t)}{p_2^*} \qquad \frac{p_1}{p_2} > \frac{p_1^*}{p_2^*} \tag{18.1}$$

Bear in mind, however, that trade must balance at world prices. The balance-of-trade constraint, first introduced in Chapter 4, is given by

$$p_1^*(D_1 - X_1) + p_2^*(D_2 - X_2) = 0 \quad \text{or} \quad p^* = \frac{(X_2 - D_2)}{(D_1 - X_1)} \qquad p^* = \frac{p_1^*}{p_2^*} \tag{18.2}$$

Graphically, the production and consumption points must be connected by the world price ratio.

What does equilibrium with an import tariff look like? (1) the MRT (slope of the production

frontier) and the MRS (slope of an indifference curve) must equal the *domestic price ratio*. (2) the production and consumption points must be connected by the *world price ratio* as just noted. Assume for the moment that the country is small, so that it faces fixed world prices. The result is shown in Figure 18.1, where X^t is the production point and D^t is the consumption point. For reference, we include the free-trade outcome, which has X* as production and *D** as consumption. For both outcomes, the production and consumption points are connected by the world price ratio, which is the trade-balance condition.

Figure 18.1

What are the effects of the tariff for a small, competitive economy? (1) the trade volume is reduced: both exports and imports decrease and resources are shifted out of the export sector X_2 and into the import-competing sector X_1. Similarly, on the consumption side, consumers reduce purchases of the now-more-expensive import good, and consume relatively more (not necessarily absolutely more) of the export good. (2) welfare is clearly reduced. Later in the chapter, we will prove that this is a general result and not just some artifact of this particular diagram. (3) there is going to be income redistribution inside the economy quite apart from the aggregate losses. If the underlying economy is a Heckscher-Ohlin economy, the scare factor, the factor used intensively in the import good X_1, will gain and the abundant factor will lose. If the underlying economy is a specific-factor's economy, the specific factor(s) in X_1 will gain and the specific factor(s) in X_2 will lose.

These results are far from dry, technical points. Governments make big mistakes in not understanding these points. Governments are often responding to intense pressure from import-competing industries and often fail to realize that giving into their request harms other industries and causes many more losses than the gains to the lobbying industries. Resource are transferred out of export industries, the industries in which the country has a comparative advantage, and the losses to factors of production outside the import industry exceed any gains to the latter. In a sense, failure to understand tariffs is largely a failure to understand general equilibrium.

18.2 Two key equivalences

A point that is to grasp is that an import tariff on X_1 is exactly equivalent to an export tax on X_2.[2] Formally, recall that an import tariff on X_1 raises the domestic price of X_1 above the world price ($p_1 > p_1^*$) while leaving the domestic price of X_2 equal to the world price ($p_2 = p_2^*$). In ratio terms, the effect of the tariff is to set $p > p^*$. For its part, an export tax establishes the following relationship between the domestic price of X_2 and the world price: $p_2 = p_2^*(1 - t)$. Thus, the tax drives a wedge between the domestic and world prices of X_2 ($p_2 < p_2^*$) while leaving the domestic price of X_1 equal to the world price ($p_1 = p_1^*$). In ratio terms, the effect of the export tax again is $p > p^*$.

$$\frac{p_1}{p_2} = \frac{p_1^*}{p_2^*(1 - t)} \qquad \frac{p_1}{p_2} > \frac{p_1^*}{p_2^*} \tag{18.3}$$

Thus, an import tariff and an export tax have the same effect on the domestic price ratio, and since it is only the ratio that matters, both have the same effect on production, consumption, and factor prices (income distribution). Again looking only at the information in Figure 18.1, we would find it impossible

to tell if the equilibrium at D^t was generated by an import tariff or by an export tax.

Equivalence 1: An import tariff on one good is equivalent to an export tax on the other good.

It generally seems hard to convince people of this result and we are not sure why. Often, the conjecture is that an import tariff and an export subsidy do the same thing, but this is completely wrong. An export subsidy, which we will return to in a minute will expand trade above the free-trade level, whereas an import tariff will clearly reduce trade as shown in Figure 18.1.

A second important equivalence is that an import tariff is equivalent to the *combined* policy of a consumption tax on the import good and a production subsidy for that good. A tariff is a double distortion, distorting both the consumer and the producer decisions. Consider again equation (18.1) and for emphasis let the consumer price of X_1 equal q_1 though keeping in mind that $q_1 = p_1$. Then, from (18.1), the relationship between the three prices is given by

$$p_1 = q_1 = p_1^*(1 + t) \qquad p_1 = q_1 > p_1^*$$

The fact that producers can sell above the world price of X_1 is exactly what would happen if the country subsidized X_1 production. They receive the world price p^* plus the subsidy payment p^*t. For consumers, the fact that they have to pay more than the world price is exactly the same as a consumption tax: they pay the world price p^* plus the tax payment p^*t.

Equivalence 2: an import tariff is exactly the same as a combined policy of a subsidy to producers of the import-competing good plus a tax on consumption of the import good.

As noted above, this emphasizes that a tariff is a double distortion. An important policy implication is that if the goal of a policy (normatively rightly or wrongly) is to protect production of the import-competing good, the optimal policy is just a production subsidy which avoids also distorting the consumption decision. This is shown in Figure 18.2. X^t and D^t are again the production and consumption points under a tariff. But if we only distort production with a production subsidy rather than a tariff, then consumers face the true world prices and can make an optimal decision. In this case, consumers will choose point D^* which is on the world price ratio through X^t so that trade balances, but their marginal rate of substitution is equal to the true world price ratio. Clearly, D^* is on a higher indifference curve than D^t. This establishes that, if a country only wants to distort (protect) production, a production subsidy is less costly (less welfare costly) than the double-distorting tariff.

Figure 18.2

There is one important caveat to this argument. The tariff raises revenue for the government whereas the subsidy costs the government expenditures. This can be a problem when the public (not to mention the government itself) does not appreciate general equilibrium. However, even for the trained economist, the argument for the superiority of the tariff rests on the assumption that some other non-distortionary tax can be used to raise the funds for the subsidy. In many if not most cases, there is no such alternative policy and hence the tariff, bad as it is, might be better. In fact, many developing countries use tariffs to raise revenue and not primarily for protecting domestic industries (many tariffs are on goods not produced domestically) because they are administratively easy to use and less distortionary than alternatives.

18.3 Export subsidies

While import tariffs cause a small, undistorted competitive economy to trade too little, an export subsidy causes it to trade too much. The fact that both policies are bad follows from an earlier proof in Chapter 2 showing the efficiency of the competitive economy: free market trade is just right. An export subsidy s on good X_2 can be written as follows:

$$\frac{p_1}{p_2} = \frac{p_1^*}{p_2^*(1+s)} \qquad \frac{p_1}{p_2} < \frac{p_1^*}{p_2^*} \tag{18.4}$$

Figure 18.3 shows the effects of an export subsidy on X_2 for a small economy facing fixed world prices. The free-trade outcome is shown for reference, with X^* and D^* denoting the free-trade consumption and production points respectively. The subsidy to X_2 exports raises the domestic producer price above the world price and producers respond by shifting production to X^s in Figure 18.3. Think of this response as determining income and then find the consumption point on the world price ratio through X^s (trade balance). As in the case of a tariff, the domestic price is distorted as well: consumers will economize on the now more expensive export good X_2 and buy more X_1. The point on the world price ratio at which the MRS is equal to the domestic price ratio is shown as point D^s in Figure 18.3.

Figure 18.3

Clearly, there is now too much trade. One way to think about the non-optimality of an export subsidy is that it involves selling to foreigners for less than the cost of production. This is not a good strategy. Yet there often seems to be confusion on this point on the part of politicians and others, thinking that exports are a good thing for their own sake and thus it seems good to stimulate and increase exports. Tracing the argument just made, however, that a subsidy is selling to foreigners below cost, is generally fairly convincing.

18.4 Gains-from-trade with many goods, trade taxes, and subsidies

Several questions arise following the graphical presentation above. First, is distorted trade always better than autarky or, rephrased, is some trade always better than none? Second, is free trade always better than distorted trade; that is, are distortions always and everywhere bad? Let's consider the first question, that of distorted trade versus autarky.

Let domestic prices be $p^d = p^*(1+t)$, where p refers to any good, an import or export or even a non-traded good. Whether or not is a tax or subsidy, depends on whether the good is an export or an import. (A) If X_i is imported, $t_i > 0$ is an import tariff, $t_i < 0$ is an import subsidy. This is the case we consider in the first section above. (B) If X_i is exported, $t_i < 0$ is an export tax, $t_i > 0$ is an export subsidy. Remember that an export tax lowers the domestic price below the world price, and thus $t_i < 0$ on an export good is a tax as in (18.3).

Stars * denote world prices and superscript '*a*' denotes autarky. Producers respond to domestic prices, not world prices, so the value of output with trade is maximized at domestic producer prices.

$$\sum_i p_i^*(1+t_i)X_i^* \;\geq\; \sum_i p_i^*(1+t_i)X_i^a \tag{18.5}$$

Autarky market clearing and trade balance conditions are the usual

$$X_i^a = D_i^a \qquad \sum_i p_i^* X_i^* = \sum_i p_i^* D_i^* \tag{18.6}$$

Now we do three things. First, use the first equation in (18.6) to substitute autarky consumption for production on the right-hand side of (18.5). Second, move the left-hand side of (18.5) over to the right, entering it with a minus sign. Third, add the following term to both sides of (18.5).

$$\sum_i p_i^*(1+t_i)D_i^*$$

Inequality (18.5) then becomes

$$\sum_i p_i^*(1+t_i)D_i^* \;\geq\; \sum_i p_i^*(1+t_i)D_i^a + \sum_i p_i^*(1+t_i)(D_i^* - X_i^*) \tag{18.7}$$

Apply the balance-of-trade constraint in (18.6) to the right-hand term in (18.7) which allows this to be simplified. Then use the definition of domestic prices to replace the $p^*(1+t)$ terms with p^d. (18.7) then simplifies to

$$\sum_i p_i^d D_i^f \;\geq\; \sum_i p_i^d D_i^a + \sum_i p_i^* t_i (D_i^* - X_i^*) \tag{18.8}$$

This is a convenient expression to understand, since the term on the far right of (18.8) is just total trade tax revenue: taxes minus subsidies. If a good is imported $(D - X) > 0$, then $t > 0$ is a tax. If $t < 0$, then it is an import subsidy. Similarly, if the good is exported $(D - X) < 0$, then $t < 0$ is a tax and so forth. Taxes on either imports or exports are positive terms on the far right and subsidies on either imports or exports are negative terms. Thus this right-hand term is total net trade tax revenue.

Welfare of the trade consumption bundle must be compared to the autarky bundle at domestic (distorted) prices p^d. Consumers don't face world prices. Thus gains from trade are ensured in spite of the distortions if the left-hand side of (18.8) is greater than the first term on the right-hand side: trade is revealed preferred. A sufficient condition for this to be true is the term on the far right, net trade tax revenue, is positive. Then the left-hand side must be greater than the first term on the right.

Gains from trade 1: Distorted trade is better than autarky if total net trade tax revenue is positive: taxing trade cannot make the country worse off than in autarky but subsidizing trade might.

What we can conclude from this is that (on net) taxing trade still beats autarky. However,

subsidizing trade could make a country worse off than in autarky. Think again of the intuition that trade subsidies are selling below cost. We can understand the idea by referring back to the production tax/subsidy discussion in Chapter 10. There, we presented a simple case where world prices just happen to equal undistorted domestic autarky prices. Using a production subsidy must make the country worse off. The same idea applies here to an export subsidy rather than a production subsidy.

A second case to consider is free trade versus distorted trade. Now things are a little more tricky because world prices might change due to the country's trade distortions and there are balance-of-trade conditions for both free and restricted trade. Let p^r denote the vector of *world* prices under trade restrictions. p^f now denotes world prices in free trade.

Production efficiency in free trade gives us

$$\sum_i p_i^f X_i^f \geq \sum_i p_i^f X_i^r \tag{18.9}$$

The balance of trade condition in free trade is given by

$$\sum_i p_i^f X_i^f = \sum_i p_i^f D_i^f \tag{18.10}$$

and the balance of trade constraint in restricted trade

$$\sum_i p_i^r X_i^r = \sum_i p_i^r D_i^r \quad or \quad -\sum_i p_i^r X_i^r + \sum_i p_i^r D_i^r = 0 \tag{18.11}$$

Do three things: first, use (18.10) to substitute for the left-hand side of (18.9). Second, add the right-hand equation (equal to zero) of (18.11) to the right-hand side of (18.9). Third, add and subtract the following term from the right-hand side of (18.9).

$$\sum_i p_i^f D_i^r$$

Inequality (18.9) then becomes

$$\sum_i p_i^f D_i^f \geq \sum_i p_i^f D_i^r + \sum_i (p_i^f X_i^r - p_i^f D_i^r) + \sum_i (-p_i^r X_i^r + p_i^r D_i^r) \tag{18.12}$$

Simplifying the last two summations on the right hand side of (18.12) reduces this inequality to

$$\sum_i p_i^f D_i^f \geq \sum_i p_i^f D_i^r + \sum_i (p_i^r - p_i^f)(D_i^r - X_i^r) \tag{18.13}$$

The last term on the right-hand side of (18.13) is called a *terms-of-trade effect*. The imposition of the distortions may change world prices. If this term is positive or zero, then free trade must be preferred to distorted trade for the reasons just discussed above. For a small economy, defined as usual as an economy that cannot affect world prices, this last term must be zero ($p^r = p^f$) and hence completely free trade is optimal. However, suppose that the country is large and that trade taxes push down world prices ($p^r < p^f$) of import goods ($D - X > 0$) by restricting domestic demand, and/or export taxes push up world prices ($p^r > p^f$) of export goods ($D - X < 0$) by restricting supply. Then the summation on the far right of (18.13) can be negative, and it is not guaranteed that free trade is better than distorted trade. This relates to the monopoly power in trade idea that we will consider below.

On the other hand, import subsidies could push up the prices of import goods [($p^r - p^f > 0$) and ($D - X > 0$)] and export subsidies could push down the world prices of the country's export goods [($p^r - p^f < 0$) and ($D - X < 0$)]. Then the term on the right-hand side of (18.13) is positive and free trade is revealed preferred to distorted trade.

> Gains from trade 2: (a) Free trade is superior to restricted trade for a "small" (price taking) economy. (b) Restricted trade may be superior to free trade for a large country if the trade restrictions sufficienctly improve the terms of trade. (c) Once again free trade is superior to distorted trade when the distortions are subsidies.

18.5 Monopoly power and the "optimal" tariff

The previous section hinted that for a large country with an ability to influence world prices, free trade may not be optimal. The idea is that a country's export and import-competing industries may be composed of many small firms that view themselves as price takers, yet the industries as a whole have some market power in trade that is not exploited in a free-market equilibrium. An import or export tariff then becomes an instrument that mimics a monopoly supplier restricting imports or exports in order to raise their prices. The tariff becomes similar to a markup that a monopolist imposes as we discussed in Chapter 11.

Figures 18.4 and 18.5 show the situation, where country h imports good X_1. M_{h1} and E_{f1} give the import demand and export supply of countries h and f respectively in Figure 18.4. Similar to our earlier chapter on trade costs, a tariff will shift down the import section of country h's excess demand function. We can see this from Figure 18.1: the tariff leads country h to want to import less and export less at any world price ratio than in free trade. Curve M_{h1}^t gives country h's tariff-shifted import demand curve in Figure 18.4. Note that this has the effect of pushing down the equilibrium world price of X_1, from p^{*f} (for free trade) to p^{*t} with the tariff.

Figure 18.4

This change in price is a terms-of-trade improvement for country h. The relative price of its import good has fallen (or relative price of its export good has risen). The volume of trade has fallen which is not a good thing, but it is possible that the improvement in its terms of trade may more than compensate and make country h better off relative to free trade (assuming no retaliation by the other country, a point we return to later). This possibility is illustrated in Figure 18.5. Line p^* gives free trade and X^* and D^* are the free-trade production and consumption points, respectively. p^{*t} gives a new price

ratio after the tariff, with X^t and D^t denoting the tariff-distorted production and consumption points respectively. The improvement in prices outweighs the fall in trade volume so country h is better off. It must be emphasized that possibility is limited for country h: a continued increase in its tariff rate will eventually drive the country back to autarky and we know that free trade is superior to autarky.

Figure 18.5

Let us now consider the algebra of tariffs. What follows is a common technique in analyzing taxation problems that you might encounter in a public economics course, for example. First, the utility function of the (aggregate) consumer is given as follows where *D* represents demand for a good.

$$U = U(D_2, D_1) \tag{18.14}$$

The production frontier of the economy is given by an implicit function as follows.

$$F(X_2, X_1) = 0 \tag{18.15}$$

Let $p = p_1/p_2$ be the relative price of good 1 in terms of good 2 as we have done before (or alternatively the price of good 2 is chosen as numeraire with price equal to one). Superscript * denotes the world price ratio. The balance-of-trade constraint for the economy is given by

$$M_2 + p^* M_1 = 0, \qquad M_i \equiv D_i - X_i \qquad M_1 > 0 \qquad p^* = p_1^*/p_2^* \tag{18.16}$$

where M_i is the excess demand for good i: $M_i > 0$ for an import and we assume that this country imports good 1. The domestic imports of good 1 are foreign exports, denoted E_1^*. We assume that E_1^* is an increasing function of the world price ratio *p**. Let *G(p*)* be referred to as the foreign excess supply function (foreign exports of X_1 are our imports).

$$M_1 = E_1^* = G(p^*) \qquad G' \geq 0 \tag{18.17}$$

Finally, the relationship between the domestic and world prices ratio is given by the tariff distortion.

$$p = p^*(1 + t) \tag{18.18}$$

Now take the total differential of the utility function (18.14)

$$dU = U_2 dD_2 + U_1 dD_1 \tag{18.19}$$

Divide through by U_2, effectively giving welfare changes in terms of good 2. Call this *dW* to distinguish it from *dU*.

$$dW \equiv \frac{dU_1}{U_2} = dD_2 + \frac{U_1}{U_2} dD_1 = dD_2 + p dD_1 \tag{18.20}$$

The ratio of marginal utilities U_1/U_2 is the marginal rate of substitution and hence equal to the domestic

price ratio in equilibrium.

Now do the same for the transformation function (18.15), noting that the ratio F_1/F_2 must be the marginal rate of transformation and hence the domestic price ratio in equilibrium.

$$F_2 dX_2 + F_1 dX_1 = 0 \qquad dX_2 + \frac{F_1}{F_2} dX_1 = dX_2 + p dX_1 = 0 \tag{18.21}$$

Do the same for the balance-of -trade constraint (18.16).

$$dM_2 + p^* dM_1 + M_1 dp^* = 0 \tag{18.22}$$

Finally, note the relationship between import changes and production and consumption changes, and also differentiate the foreign excess supply function in (18.17).

$$dM_i = dD_i - dX_i \qquad dM_1 = dE_1^* = G' dp^* \qquad (p - p^*) = p^* T \tag{18.23}$$

Substitute the first equation of (18.23) into the welfare differential (18.20).

$$dW = dD_2 + p dD_1 = dM_2 + dX_2 + p_1 dM_1 + p dX_1 \tag{18.24}$$

Substitute for dX_2 from (18.21) and for dM_2 from (18.22) and replace M_1 with E_1^*

$$dW = -p^* dE_1^* - E_1^* dp^* + p dE_1^* - p dX_1 + p dX_1 \tag{18.25}$$

Now collect terms (and the last two cancel) and use the relationship between domestic and world prices in (18.23)

$$dW = (p - p^*) dE_1^* - E_1^* dp^* = (p^* t) dE_1^* - E_1^* dp^* \tag{18.26}$$

The change in welfare from a tariff is composed of a volume-of-trade effect (first term in the right-hand expression) and a terms-of-trade effect (second term in the right-hand expression). The volume-of-trade effect is negative: $dE_1^* < 0$. The intuition for this is that, with the domestic price of X_1 higher than the world price, a one-unit reduction in imports causes a loss of value to the economy of p, the domestic price, but saves the economy only p^* in import costs. Thus, there is a surplus loss of $(p - p^*)$ on each unit of imports lost.

The terms-of-trade effect, the right-most term in (18.26) is positive: the tariff forces down the world price ratio p^* and hence, with the minus sign, this term contributes positively to welfare. Then we have one negative term and one positive term; what can we say? Consider beginning in free trade and raising the tariff t from zero. First, note that the volume-of-trade effect is zero near free trade at $t = 0$. Thus at free trade only the second term contributes to welfare and hence at first welfare improves as the tariff rises from zero. As we continue to raise the tariff, the first (negative) term gets larger while the

trade volume E_1^* falls: the first negative term gets larger and the second positive term gets smaller. At some point, welfare as a function of the tariff rate turns negative and eventually the economy hits autarky, clearly worse than free trade.

Figure 18.6 plots welfare as a function of the tariff rate. W^* denotes free trade and W^a denotes autarky. As per the previous paragraph, the welfare first rises with the tariff, reaches a maximum at W^o ('o' for optimum) and then declines toward W^a. At the optimum, the welfare curve is flat and hence a small change in t has no effect on welfare. Thus, the optimal tariff is found by setting (18.26) equal to zero.

Figure 18.6

Equation (18.26) can be solved for the optimal tariff, the one that achieves the maximum welfare at W^o in Figure 18.6, given by t^o in that graph. Replace dE_1^* with $G'dp^*$. (18.26) becomes

$$dW = (p^* t G' - E_1^*)dp^* \tag{18.27}$$

Optimizing with respect to the world price ratio by setting this to zero gives the optimal tariff.

$$t^o = \frac{E_1^*}{p^* G'} = \frac{1}{\eta_s^*} \qquad \eta_s^* \equiv \frac{p^*}{E_1^*}\frac{dE_1^*}{dp^*} \qquad since \qquad G' = \frac{dE_1^*}{dp^*} \tag{18.28}$$

The optimal tariff formula is given by one over the foreign elasticity of export supply, η_s^*. The less elastic is foreign supply, the higher the optimal tariff.

A few points to conclude this section. First, note that for a small country for which a tariff cannot change world prices, the optimal tariff is zero: there is only a negative volume-of-trade effect. This is an alternative proof that free trade is optimal for a small country. Second, note that the optimal tariff is just a *formula*, it is not a *value* (number). We have no idea what its value is, and this can only be informed by detailed simulations using applied general-equilibrium models. Finally, we put optimal in quotes in the section heading to emphasize that "optimal" is being used in a somewhat narrow sense here. It is unilaterally optimal, but it is a beggar-thy-neighbor policy: it must make trading partners worse off and it can be proved that losses to the trading partner exceed the gains to the tariff-imposing country. If the trading partner(s) retaliate, then even the tariff-imposing country is likely to be worse off. Clearly, this is not a policy to be encouraged in practice, and superior outcomes to unilateral protection should be achievable through negotiations and cooperation.

18.6 Tariffs and the theory of the second best

As we have often noted through the book, many results depend on the important assumption that there are no distortions in the economy. Distortions include domestic taxes and subsidies, external economies, imperfect competition and so forth: many of the things we studied in Chapters 10-12. For present purposes, the result that tariffs are harmful for a small open economy relies on the assumption that

there are no other distortions in the economy. If there are distortions, it may be the case that tariffs could be used to offset these distortions and thereby increase welfare. This possibility is an application of what is known in economics as the "theory of the second best." There are two equivalent versions of this general proposition; one is just the contra-positive of the other.

> Second best 1: in the present of one (or more) distortions, adding a further distortion that acts to offset the first one can improve welfare.
>
> Second best 2: in the presence of more than one distortion, removing one of the distortions can make the country worse off.

Figure 18.7 presents an example of this. Following our discussion in Chapter 10, suppose that the X_1 sector is characterized by external economies of scale. This means that the free market outcome is not optimal and there will be under-production of X_1 in equilibrium. In Figure 18.7, the world price ratio is p^*, and X^* and D^* are the free-trade production and consumption points respectively.

Figure 18.7

Now consider the effects of an import tariff on X_1, holding the world price ratio constant. This will raise the producer price of X_1 and shift the economy around the production frontier to X^t. This is a positive outcome: the tariff is, in part, a production subsidy to X_1 as discussed earlier and thus is an offsetting distortion to the external economies distortion. However, as we also noted above, the tariff distorts consumption as well, so the consumption point will be something like D^t in Figure 18.7, where p^t is the domestic price ratio.

As we have drawn Figure 18.7, welfare clearly improves. We could do better by simply subsidizing X_1 production instead of using the tariff, the latter having the unfortunate secondary effect of distorting consumption as well. That of course assumes that the subsidy costs can be raised by some sort of non-distortionary taxation, often not a realistic assumption. This problem notwithstanding, the lesson is that the most direct and limited offsetting distortion is the one that improves welfare. In fact, the result is that using a tariff (or trade subsidy) to correct a domestic distortion is a “third best” outcome because of its double-distortion property.

The second-best policy in Figure 18.7 is often advanced under the term *infant-industry* argument. The notion is that it is privately unprofitable to start an industry, but with a little government help, it can become profitable in the long run. Economists are generally very skeptical of this argument and immediately point to the fact that government funds have an opportunity cost to other projects. If capital markets work well and private and social rates of return are equal, then if the project is privately unprofitable it is socially unprofitable as well.

Yet it surely is the case that there are market failures, particularly in developing countries. Often under-developed capital markets means that the private cost of funds does exceed the social costs (e.g., rates of interest paid on government bonds). Another source of market failures, related in some ways to external economies, are termed coordination failures. For example, no one wants to build a factory because there is not a reliable electrical supply, and no one wants to build a quality power plant because there are too few customers. Putting this and the previous paragraph together, we conclude that there surely are situations where second-best intervention is justified, but all cases need to be evaluated with care and skepticism. Once again, trade remedies are rarely second best, they are typically “third best”.

18.7 Effective protection

One fundamental prediction from the theory thus far is that a tariff provides protection from imports, encouraging expanded domestic production of the protected commodity. Implicitly, this theory assumes that the tariff is the only tax that directly affectscosts and prices of the good in question. Such an assumption is sensible for goods that are produced solely by non-traded primary inputs, such as capital and labor, and for goods that require intermediate inputs that are freely traded internationally. However, most commodities are produced with the use of intermediate goods that are themselves subject to trade taxes. Thus, for example, a tariff on imported steel would raise costs and lower output in the automobile sector. In general, a manufacturer will be better off the higher are tariffs on imports that compete with her outputs and worse off the higher are tariffs on her imported inputs. The term "effective protection" refers to the fact that all such tariffs need to be taken into account in computing the net protective effect of the tariff structure as a whole.

Because our concern is to take into account intermediate inputs, the notion of effective protection actually refers to the positive or negative stimulus to *value added* in production of a commodity. Value added per unit of output, *v*, is the difference between the price of a final good and the cost of purchasing intermediate inputs. As such, it measures the portion of the value of output that is available for payments to primary inputs. For example, if the price of an automobile is $15,000 and the cost of acquiring the steel, leather, glass, rubber, and so on needed to produce the car is $10,000 there remain $5,000 that may be used to pay for wages, the costs of capital (e.g., profits and interest), and the costs of land (e.g., rents). Value added thus captures the costs of primary inputs. In this case, value added makes up 33% of the gross value of the car. If the tariff structure combines to expand value added relative to free trade, it effectively raises payments to these primary factors.

The “standard” tariffs we have been dealing with are now referred to as *nominal protection*: t^n is the protection offered to the output price.

$$p = p^*(1 + t^n) \qquad t^n = \frac{p - p^*}{p^*} \tag{18.29}$$

Effective protection, and specifically the effective tariff t^e, is defined as the protection offered to value added. Consider industry i and assume that value added is just payments to labor. Value added is then output price minus the value of purchased inputs of other goods X_j. Let a_{ij} denote the amount of good X_j needed to produce one unit of good i; a_{il} is the amount of labor needed to produce one unit of X_i w is the wage rate, and p^* gives world goods prices. The value added per unit of X_i production in free trade is then given by

$$v_i^* = a_{il}w = p_i^* - \sum_j a_{ij}p_j^* \tag{18.30}$$

Now consider all tariffs on all goods, and replace the world prices in (18.30) with tariff-distorted domestic prices.

$$v_i = a_{il}w = p_i^*(1 + t_i^n) - \sum_j a_{ij}p_j^*(1 + t_j^n) \tag{18.31}$$

The effective tariff is defined as:

$$t_i^e = \frac{v_i - v_i^*}{v_i^*} = \frac{p_i^*(1+t_i^n) - \sum_j a_{ij}p_j^*(1+t_j^n) - p_i^* + \sum_j a_{ij}p_j^*}{p_i^* - \sum_j a_{ij}p_j^*} \quad (18.32)$$

Divide through the numerator and denominator on the right by p_i*, and use the notation

$$\frac{a_{ij}p_j^*}{p_i^*} \equiv \sigma_{ij} = \text{share of input j in the value of i}$$

The effective tariff rate is then given by

$$t_i^e = \frac{v_i - v_i^*}{v_i^*} = \frac{(1+t_i^n) - \sum_j \sigma_{ij}(1+t_j^n) - 1 + \sum_j \sigma_{ij}}{1 - \sum_j \sigma_{ij}} \quad (18.33)$$

which simplifies to

$$t_i^e = \frac{v_i - v_i^*}{v_i^*} = \frac{t_i^n - \sum_j \sigma_{ij}t_j^n}{1 - \sum_j \sigma_{ij}} \quad (18.34)$$

Consider a couple of special cases. First, suppose that there is a tariff protecting industry i and all other tariffs are zero. Then the effective tariff exceeds the nominal tariff: $t_i^e > t_i^n$. This is a sort of leverage effect (the tariff is "levered up") to use a finance term. The protection to value added exceeds the nominal tariff. Second, assume that all tariffs on all goods are the same $t_i^n = t_j^n \ \forall j$. In this case, the common tariff rate factors out of (18.34), and the remaining terms in the numerator cancel with the denominator: the effective and nominal rates are equal: $t_i^e = t_i^n$. Finally, assume that there is no tariff protecting the X_i industry but that at least one input tariff is positive. In this case, the effective tariff for industry i is *negative*. Export industries, for example, are (correctly) classified as losing from the tariff system as a whole.

The concept of effective protection is very appealing because it captures the general-equilibrium idea that an industry can be affected by all tariffs in the economy, not just its own protective tariff. The notion that export industries are classified with negative protection is appealing. However, these calculations, at least the simple formula presented above, is fraught with general-equilibrium difficulties as it is trying to add general-equilibrium calculations to the traditional nominal tariff. We can see, for example, that the effective protection formula assumes that the σ_{ij} do not change in the movement between free trade and protection, nor do the world prices of any goods. Neither assumption is a very good one, though the second can be justified for a small economy.

Somewhat deeper is the question of how to interpret the welfare effect of protection to value added. In the Heckscher-Ohlin model, for example, labor and capital are mobile across industries and, in general-equilibrium, each factor earns the same return in all industries. Thus, protection to value added cannot be interpreted as protecting the incomes of individual factors and indeed the Stolper-Samuelson theorem says that one factor must loose in all industries. Perhaps the case where effective protection has a clearer interpretation is in the specific-factors model. Take an extreme case in which all value added is returns to specific factors that have no alternative use outside the industry and no factors from other industries can enter (perhaps a short-run interpretation of the effects of tariffs). Then there is a pretty clear link between the effective tariff rate and individual incomes. Similarly, the effective tariff might measure something about income distribution in imperfectly competitive sectors with no firm entry or exit: effective protection is, in large part, protection to profit income.

18.8 Tariffs versus transport and transactions costs, foreign ownership

One important point to bear in mind, infrequentlymentioned in other trade textbooks, is the fact the our analysis of tariffs assumes that they are collected costlessly, and the revenues are returned lump sum to the consumers. In this section, we emphasize the importance of the former assumption and note that, if a chunk of tariff revenue is burned up in red tape and the costs of the taxation system (collectors, accountants, lawyers), then tariffs can be very much worse indeed than our diagrams suggest. Recall that the balance of trade condition for the economy is given at world prices.

$$\sum_i p_i^* X_i = \sum_i p_i^* D_i \tag{18.35}$$

Adding and subtracting terms for tariff revenue, this can be written as

$$\sum_i p_i^* (1 + t_i) D_i = \sum_i p_i^* (1 + t_i) X_i + \sum_i p_i^* t_i (D_i - X_i) \tag{18.36}$$

The tariff-distorted prices $p^*(1+t)$ are domestic prices, which we can denote by p^t, and the last term on the right-hand side of (18.36) is tariff revenue: price times the tax rate times the net import volume.

$$\sum_i p_i^t D_i = \sum_i p_i^t X_i + \sum_i p_i^* t_i (D_i - X_i) \qquad \sum_i p_i^* t_i (D_i - X_i) = [\textit{tariff revenue}] \tag{18.37}$$

This has a graphical interpretation in terms of Figure 18.1. The left hand side of (18.37) is the value of consumption at D^t in Figure 18.1, valued at the domestic price ratio p^t and $p^t X^t$ is the value of production at the domestic price ratio. Equation (18.37) tells us that the difference between the two is interpreted as tariff revenue: the added spending power that household received when refunded the tariff revenue.

Suppose that costless tax collections and redistributions are not possible, and make the extreme opposite assumption. Suppose that all tax collections are paid out in salaries to tax collectors who have withdrawn from productive work to collect the taxes. Alternatively, tax collections are equal to delays and other red tape which raise the costs to the foreign supplier. Then there is effectively no revenue to redistribute: tax collects are just equal to the real resource costs (or lost production of goods) of imposing and collecting the tax. Figure 18.8 draws in the costless tariff-collection equilibrium shown in Figure

18.1. The simplest case to consider is that the tariff revenue is entirely consumed in the customs infrastructure provided by the government (think of these costs as being borne in goods so that X^t continues to represent gross output in Figure 18.8). Note from (18.37) that with no tariff revenue to redistribute, the value of consumption and production must balance at domestic prices. Hence while the production point X^t remains the same as in the case of the (costless) tariff, the consumption point must now lie on the domestic price ratio p^t at point D'. The case of pure transactions costs or "red tape" in which tariff collections equal real resources (labor) needed to make the collections has a much worse welfare effect than a tariff.

Figure 18.8

The welfare effect shown in Figure 18.8 is also the correct analysis of comparing a transport cost (which also uses real resource) to a tariff of the same rate t. A researcher using partial-equilibrium analysis might conclude that a tariff and a trade cost of a certain percent are the same thing. This is quite false: the tariff leads to point D^t while the trade cost leads to point D' in Figure 18.8.

We wish to briefly mention one final point before moving on, which is the existence of foreign-owned factors of production in the economy. Our analysis assumes that all changes in factor returns go to domestic consumers. If there are foreign owned factors of production capturing some part of the returns, then the situation may be better or worse than in the standard tariff analysis. Suppose that, for example, there is foreign capital in the country and that a tariff is put in place protecting import-competing capital-intensive goods. The Stolper-Samuelson theorem states that this will increase the real return to capital, but now a portion of the increased return is captured by foreign capital owners. In this case, the welfare implications of the tariff are worse than in the standard analysis.

18.9 Summary

Import tariffs and export taxes raise the costs of trade and, at least for small countries, reduce welfare below the free trade level. In many models such as the Heckscher-Ohlin model, they also redistribute income inside the country which is the most obvious explanation for the fact that we have so many trade barriers in our modern world. In the Heckscher-Ohlin case, an import tariff protects the country's scare factor and leads to real income gains for the scare factor. The abundant factor is unambiguously worse off. More will be said about this in Chapter 21.

Two important equivalences are noted. First, a tariff is equivalent to an export tax. Individuals who advocate taxing imports and subsidizing (or at least facilitating) exports don't understand that these policies, in fact, cancel each other out. Second, a tariff is equivalent to a production subsidy and a consumption tax on the imported good. It is important to understand that a tariff or export tax is a double distortion. In a situation where a country is determined to protect an industry, a production subsidy alone is the preferred policy. We also noted that an export subsidy is equally bad: it amounts to a country selling to foreigners for less than the cost of production. In the absence of any other distortions, the competitive market economy picks the most efficient amount of trade.

A general analysis in which there are many goods and an arbitrary number of trade taxes and subsidies is presented. Quite a neat result emerges, which is that the distorted outcome of trade must be superior to autkary if total net tax collections (taxes minus subsidies) is positive. If trade is on net

subsidized, however, the distorted trading equilibrium could be worse than autarky. This analysis then set the stage for looking at a large country with monopoly power in trade. In this case, the free market outcome involves many national firms and consumers behaving as competitors and hence the ability to act as a collective monopolist is lost. Tariffs or export taxes essentially cause the country to behave "as if" it were a monopolist. An import tariff lowers the demand for foreign goods thus forcing down their prices and restricts the supply of domestic exports, thereby forcing up their prices. Problems with this argument are mentioned.

In a qualification to much that proceeded it, we present the theorem of the second best, which in our context says that in the presence of an existing distortion, adding a second offsetting distortion may improve welfare. There are a number of caveats to this, including the fact that using a trade instrument to correct a domestic distortion is not the best choice, a sort of "third best" option. The concept of effective protection is discussed in the next section. It is an important idea but one which is beset by some general-equilibrium difficulties.

The Chapter concludes with a short section on the relationship between tariffs and transactions costs or trade costs. While in partial equilibrium these things are similar, they are very different in general equilibrium. Hyper-efficient tariffs with no administrative costs return the revenue to the consumer and use no real resources. Suppose, we suggested, that tariff revenue is completely exhausted in paying the salaries of tax collectors and that they must be drawn out of productive work or exhausted in the use of goods to maintain customs' infrastructure. The welfare consequences of this type of tax is much worse than an administratively-costless tariff.

REFERENCES

Bhagwati, Jagdish N. (1971), "The Generalized Theory of Distortions and Welfare." In J. Bhagwatti et al. (eds.), *Trade, Balance of Payments and Growth: Essays in Honor of Charles P. Kindleberger*, Amsterdam: North-Holland.

Bhagwati, J. N., and R. A. Brecher (1980), "National Welfare in the Open Economy in the Presence of Foreign-Owned Factors of Production." *Journal of International Economics* 10, 103-115.

Graaff, J. de V. (1949), "On Optimal Tariff Structures." *Review of Economic Studies* 16, 47-59.

Johnson, H. G. (1954), "Optimal Tariffs and Retaliation." *Review of Economic Studies* 21, 142-153.

Jones, Ronald W. (1967), “International Capital Movements and the Theory of Tariffs and Trade”, *Quarterly Journal of Economics* 81, 1-38.

Lerner, Aba. (1936), "The Symmetry Between Import and Export Taxes." *Economica* 11, 306-313.

Markusen, James R. and Randall Wigle (1989), "Nash Equilibrium Tariffs for the U.S. Canada: The Roles of Country Size, Scale Economies, and Capital Mobility," *Journal of Political Economy* 97, 368-386.

ENDNOTES

1. For example, tariff revenue amounts to only 0.01 percent of total government revenue in the United Kingdom, 0.02 percent in Germany and 1.56% in the United States. On the other hand, tariffs provide the government of Argentina with 13.31 percent of total revenue and 40.90 percent of all revenues to the government of Ghana. The source for these figures is International Monetary Fund (1886).

2. This point is credited to Lerner (1836), who assumed competitive markets in its proof.

Figure 18.1

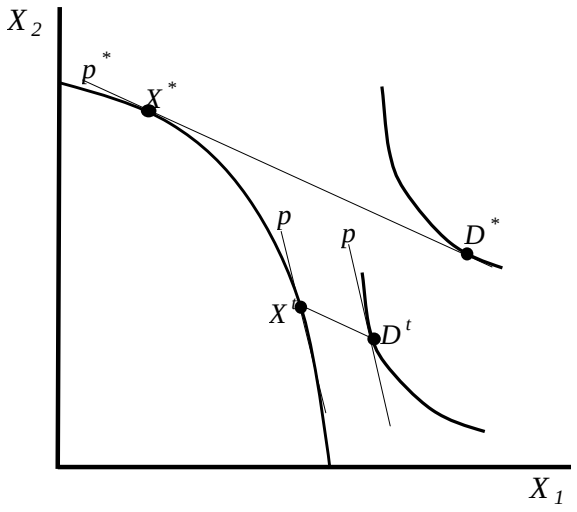


Figure 18.2

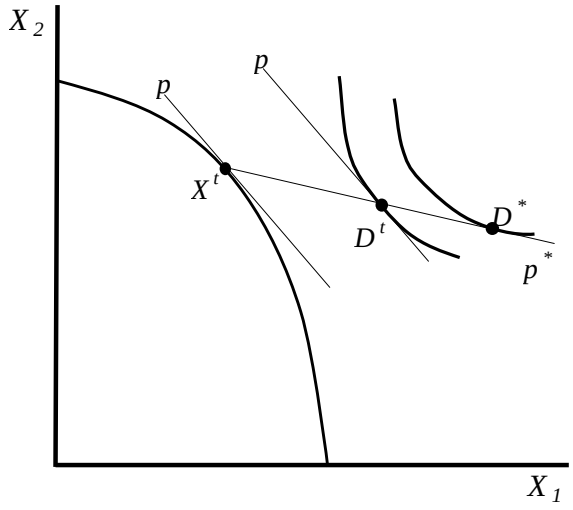


Figure 18.3

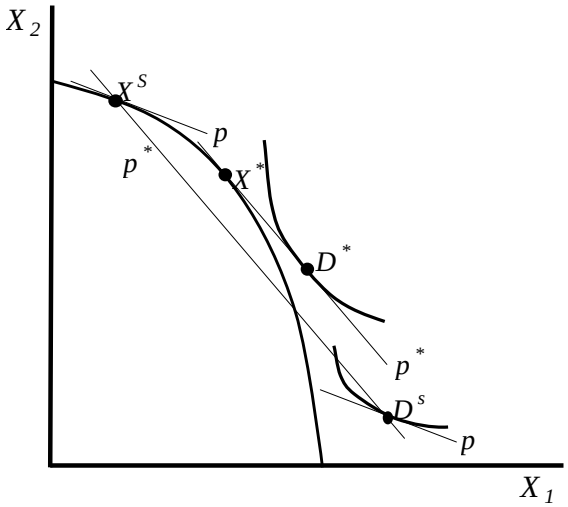


Figure 18.4

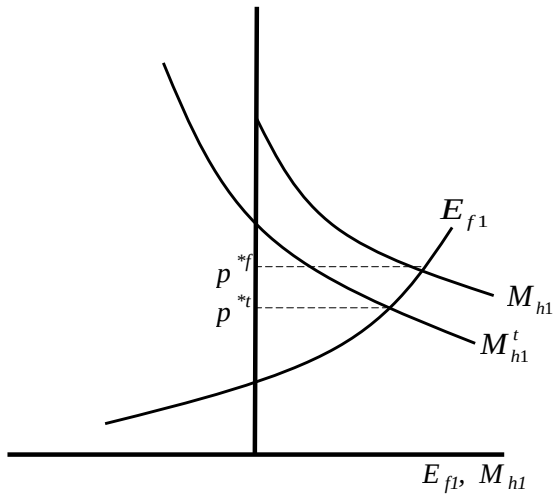


Figure 18.5

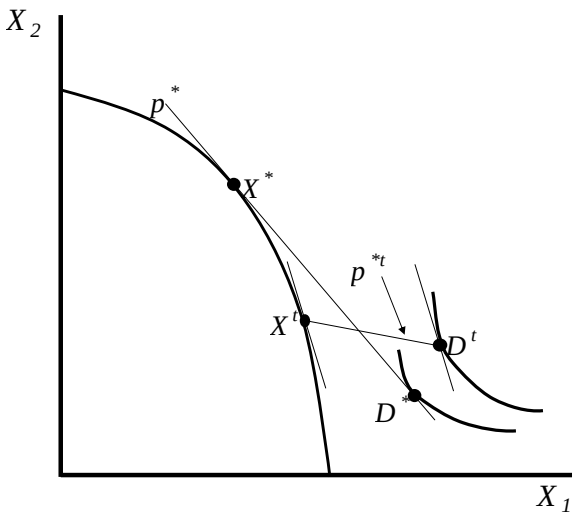


Figure 18.6

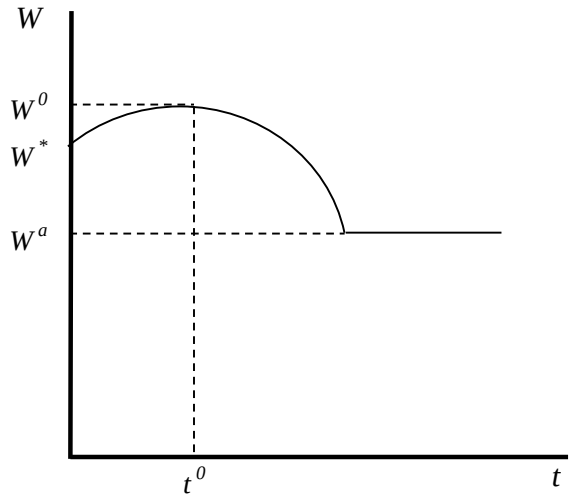


Figure 18.7

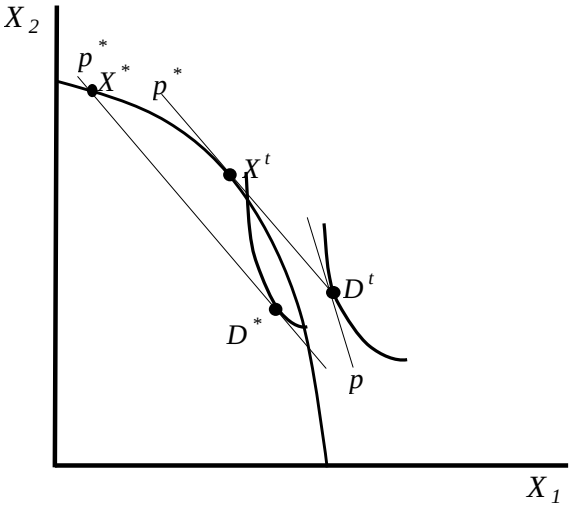
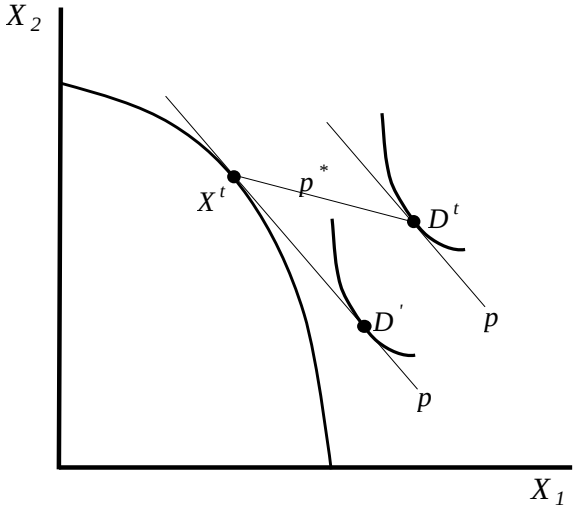


Figure 18.8



Chapter 19

QUOTAS AND RELATED BARRIERS

19.1 Quotas and equivalent tariffs in a small economy

Restrictions on markets can operate on either the price side or on the quantity side of the market. Tariffs and other taxes operate on the price side and can be termed price instruments: they drive a wedge between the price in the domestic market and in the world market. Quota is a general name for restrictions that operate on the quantity side of a market and could be termed quantity instruments. Usually, quotas set an upper bound on the amount that can be bought and sold in a market, but there are examples of minimum quotas as well. In the case of international trade, a quota typically restricts the quantity (number of units) of a good that can be imported.

Figure 19.1 shows the situation in a very simple partial-equilibrium diagram in which good M_1 is imported by a country, p_1 is its domestic price, and D_1 is the demand curve for imports. Suppose that the country is small and that the good is available at a fixed world price of p^*. Then the equilibrium quantity of imports is M_1^* at a price of p^*. Suppose now that the country imposes an import tariff of t, making the domestic price $p = p^*(1+t)$ in Figure 19.1. This will reduce the equilibrium level of imports to $\bar{M}_1$ in Figure 19.1 The hatched rectangle in the figure has a simple interpretation: it is tariff revenue. The vertical distance is the tariff collection per unit of imports (p^*t) and the horizontal distance of the rectangle is the quantity of imports $\bar{M}_1$. Tariff revenue is $p^* t \bar{M}_1$.

Figure 19.1

Suppose instead that the government announces an import quota, setting the maximum allowable imports at the level $\bar{M}_1$. The domestic market-clearing price must be at the intersection of this quantity constraint with the demand curve and hence the domestic price must be $p^*(1+t)$ as shown in Figure 19.1. The hatched rectangle in Figure 19.1 is still relevant. It represents the difference between the domestic willingness to pay for the constrained level of imports minus the foreign willingness to supply. This rectangle is referred to as *quota rents* or *Ricardian rents*, the same Ricardo as in Chapter 7. The word "rent" here refers to the price that something (in this case good M_1) commands in excess of its opportunity cost. p^* is the opportunity cost per unit: what the foreign export could get selling the good elsewhere. $p = p^*(1+t)$ is the price the good commands in the domestic market. Hence p^*t is the Ricardian rent per unit of imports and $p^* t \bar{M}_1$ is the total Ricardian rents created by the restriction.

Ricardian rents are similar to profits arising, for example, from the restrictions of output by a monopolist that we studied earlier. The difference is important in many contexts. Ricardian rents arise from scarcity value and the term is generally used to apply in situations where that scarcity is not due to market power. Take, for example, land in urban areas. Land close to the city center is in scarce supply relative to more distant land and hence the closer land commands a much higher price than in the hinterland. This price premium generally has nothing to do with market power. It is due to scarcity value. Similarly, land with zoning (planning or construction) permission near the city center is worth much more than land that cannot be developed. It is the scarcity value of the planning permission that

creates the Ricardian rents. Thus the point illustrated in Figure 19.1 is a very general idea that applies in many areas of economics.

In the tariff case, we just assume that the tariff revenue goes to the government, though corruption might leave to some diversion of those revenues to certain individuals. Who gets the rents associated with the scarcity value created by the quota? That depends on how the quota is introduced and enforced and that is the subject of the next section.

19.2 Distribution and dissipation of quota rents

When a government decides to constrain an import good by a quota, it must have some administrative plan as to how it will enforce this. A simple declaration (thou shalt not...) is probably not going to work. The typical way that a quota works is by the government issuing some sort of licenses. The importer of a good can pay the world price for the good, but must also surrender a license to import to the customs' inspector at the port. Let's attack this question by way of example.

> Quota auction: the government can print up a number of licenses (e.g., in units or hundreds or thousands of units of M_1) and then conduct an auction and sell them off. Quota rents accrue to the government treasury.

In very simple settings with lots of buyers and sellers and very good information about supply and demand, this method of instituting the quota is considered virtually the same as a tariff. Competitive bidders will be willing to bid up to the difference between the supply and the demand price for each unit, which is exactly what the tariff rate is under the tariff alterative. In Figure 19.1, the hatched rectangle is collected by the government as tariff revenue at tariff rate t or collected by the government in the revenues from the auction of licenses to import quantity $\bar{M}_1$. (1) the quantity of imports is the same, (2) the revenue raised is the same and (3) the revenue accrues to the government in both cases, so the tariff t and auction quota $\bar{M}_1$ are equivalent.

A second method of instituting the quota is to simply give the licenses to favored parties inside the country. These could be domestic importing firms who are allocated licenses on the basis of their historical market shares or, all-too-commonly, they are given out as an instrument of corruption to family or friends of those in power (or used to buy off political enemies).

> Quota give-away: the government gives the license to favored individuals and firms inside the country and those firms/individuals capture the quota rents.

One could argue that at least this method keeps the rents inside the country and that it only differs from the quota auction in the distribution of rents among firms and individuals in the country. The same thing could, of course, occur with a tariff if the tariff revenues are given to favored individuals or firms. In both cases, the tariff revenue or the quota rents could end up in Swiss bank accounts and constitute a major vehicle for corruption in many developing and transition economies.

A third method of instituting the quota and therefore of allocating the quota rents has become known as a "voluntary export restraint" (VER). In this case, the government asks the foreign government to limit exports to the home country and leaves it up to the foreign country to determine how to allocate

licenses in their own country; it could be via auction there or the licenses could be given to local (foreign) firms on the basis of their historic market shares. The rectangle of quota rents in Figure 19.1 then goes to the foreign government or firms and is lost to the domestic economy entirely.

> Voluntary export restraint: the home government asks the foreign government to limit its exports to the quota level and hence transfers all rents to the foreign government or firms.

The word "voluntary" is a rather odd word choice, since it really isn't voluntary at all. Generally, this method is used to buy-off opposition and retaliation by the foreign government by giving the foreigner a type of compensation in form of quota rents. It is essentially saying that we are going to hurt your firms but will compensate you by giving you rents associated with the policy.

Finally, it is possible that some, much or indeed all of the quota rents may be dissipated in wasteful activity. This is the worst outcome from the world's point of view. No one gets the rents. Suppose, for example, that the government says that it is going to give away the licenses on a first-come first-served basis at 17.00 (5pm) Wednesday. This is similar to announcing under-priced sports or concert tickets going on sale at a specific time, we are just taking it to the extreme and assuming that the announced price is zero. Some buyers will take a day off work and queue up for many hours, possibly days, in order to get a license. In equilibrium the quota rents, the difference between the valuation by the buyers and the price of the import good (concert or sports tickets) will be exactly offset by the value of lost wages and production in the economy. Alternatively, the licenses might be allocated to the individual or firm who spends the most time lobbying the government, so a cadre of high-priced lawyers spends their time in Washington lobbying instead of doing productive work. Jagdish Bhagwati has referred to this as "directly unproductive activity" (DUPS).

> Directly unproductive activity: the government gives away the licenses on a first-come, first-served basis or on the basis of who lobbies most. Quota rents are dissipated in the value of production lost in queuing and/or lobbying.

From the home country's point of view, this is essentially the same as the voluntary export restraint, but from the whole world's point of view it is worse as just noted. The VER transfers the rents to foreigners but in the case of DUPS no gets any rents. Now we turn to a more formal analysis of this question of the relationship between tariffs and quotas under alternative quota allocation rules.

19.3 An algebraic example

Let's consider a specific example of a quota and the distribution of rents. To keep it as simple as possible, suppose that the country produces only good X_2 in the amount $\bar{X}_2$. It trades this for X_1 at fixed world prices. Suppose that preferences are Cobb-Douglas over consumption quantities D_1 and D_2. Since D_1 is not produced domestically, we have $D_1 = M_1$,where M_1 is imports. If t denotes the import tariff or tariff equivalent of a quota on good X_1 and I denotes income, then preferences and the demand for imports (using earlier results) are given by:

$$U(D_1, D_2) = D_1^{\alpha} D_2^{1-\alpha} \qquad D_1 = M_1 \qquad M_1 = \frac{\alpha I}{p_1^*(1+t)} \tag{19.1}$$

Income I is given by the value of production of X_2 at world prices plus tariff revenue. Letting the world prices of X_1 and X_2 be equal to one, income is given by:

$$I = p_2^* \bar{X}_2 + p_1^* t M_1 \qquad let\ p_1^* = p_2^* = 1 \qquad I = \bar{X}_2 + t M_1 \tag{19.2}$$

Replace I in the right-hand equation of (19.1) with the right-hand equation of (19.2). Import demand is then:

$$M_1 = \frac{\alpha \bar{X}_2}{1+t} + \frac{\alpha t M_1}{1+t} \qquad M_1^c \equiv \frac{\alpha \bar{X}_2}{1+t} \qquad M_1^r \equiv \frac{\alpha t M_1}{1+t} \tag{19.3}$$

As indicated in (19.3), the first term can be thought of as imports by "consumers" paid for by factor income from producing X_2 and is denoted M_1^c. The second term is imports paid for by tariff revenue or quota rents by whomever gets that and is denoted M_1^r. This breakdown of imports is shown in Figure 19.2. p^* is the world prices ratio as usual (normalized to one here). Production is fixed at $\bar{X}_2$. D denotes total consumption: D is on the world price ratio through the production point but the marginal rate of substitution is the distorted domestic price ratio $p = p^*(1+t)$. D^c in Figure 19.2 gives consumption out of factor income with imports given by M_1^c (middle equation in (19.3)); the difference between total imports M_1 and M_1^c is imports paid for by tariff revenue or quota rents (right-hand equation in (19.3)): M_1^r

Figure 19.2

Rearrange the first equation of (19.3) as follows.

$$M_1 - \frac{\alpha t M_1}{1+t} = \frac{\alpha \bar{X}_2}{1+t} \qquad \frac{(1+t) - \alpha t}{1+t} M_1 = \frac{\alpha \bar{X}_2}{1+t} \tag{19.4}$$

We then have an expression for imports as a function only of parameter values.

$$M_1 = \frac{\alpha \bar{X}_2}{1 + t - \alpha t} \tag{19.5}$$

Now divide the expression for imports out of factor income in (19.3) by total imports in (19.5). This gives us imports from factor income as a share of total imports. Subtracting this from one gives the share of imports from tariff revenue or quota rents. These are

$$\frac{M_1^c}{M_1} = \frac{1 + t - \alpha t}{1+t} \qquad \frac{M_1^r}{M_1} = \frac{\alpha t}{1+t} \tag{19.6}$$

The importance of tariff revenue or quota rents increases with the importance of the import sector, measured by the share parameter α.

Now we wish to make a couple of points about the non-equivalence of tariffs and quotas once we move away from an initial situation where they are set to have the same effect. Let point M_1 in Figure 19.2 represent a quota in this amount of imports, denoted by $\bar{M}_1 = M_1$. We can then solve for the tariff equivalent using (19.5).

$$1 + (1-\alpha)t = \frac{\alpha\bar{X}_2}{\bar{M}_1} \qquad (1-\alpha)t = \frac{\alpha\bar{X}_2}{\bar{M}_1} - \frac{\bar{M}_1}{\bar{M}_1} = \frac{\alpha\bar{X}_2 - \bar{M}_1}{\bar{M}_1} \tag{19.7}$$

which gives us the tariff equivalent of $\bar{M}_1$ denoted by t^q.

$$t^q = \frac{\alpha\bar{X}_2 - \bar{M}_1}{(1-\alpha)\bar{M}_1} \tag{19.8}$$

One thing to note is that the tariff equivalent t^q rises or falls as the economy expands or contracts, measure by increases or decreases in the productive capacity of the economy $\bar{X}_2$. When the economy grows under a tariff, added demand for the import good can be accommodated: shifting the D_1 curve out in Figure 19.1 allows imports to expand along the line $p^*(1+t)$. However, under a quota in the amount $\bar{M}_1$ in Figure 19.1, all adjustment to the increased demand must come on the price margin: the new outcome will be vertically above the old one on the vertical quota constraint $\bar{M}_1$. This is equivalent to raising the tariff as indicated in (19.8) following an increase in $\bar{X}_2$. Conversely, a fall in production and therefore income in (19.8) reduces the tariff equivalent of the quota. Once we reach the point where $(\alpha\bar{X}_2 - \bar{M}_1) = 0$, the quota is no longer binding and the tariff equivalent is zero. In summary, tariffs allow adjustment in import *quantities* as the tariff changes (for a small economy all adjustment is in quantities) whereas a quota forces all adjustment onto domestic *prices*.

Similarly, if we go back to the last equation of (19.1) and let the world price vary, taking on values different than one, then the tariff equivalent of the quota (the value of t in (19.1) that keeps imports constant) varies. A world price increase reduces the tariff equivalent and, eventually, the latter falls to zero when the quota is no longer binding.

Now let's turn to the distribution of quota rents, and consider rents going to the domestic government versus being given to the foreign country as in a VER. Suppose first that we are using a tariff but that now we give the tariff revenue to the foreigner. Import demand would then be given by the second equation of (19.3): imports would equal M_1^c and the situation would be "as if" the foreign country raised its export price from one to $(1+t)$. Consumption would be at point D^c in Figure 19.2 and welfare is clearly lower than under the tariff, in which the revenue is returned to the government or to some firms or individuals inside the country.

However, D^c in Figure 19.2 is not the equilibrium if we started with the quota $\bar{M}_1$ in Figure 19.2 and now give the quota rents to the foreigner. At point D^c, imports have been reduced, so the quota is no longer binding. The domestic price ratio p must rise, which is the same as saying that the tariff equivalent

falls, until imports once again reach the quota constrained level $\bar{M}_1$. The tariff equivalent for the VER is found by going back through our algebra and eliminating tariff revenue from the demand equation. Solving for the tariff equivalent under the VER gives

$$t^{ver} = \frac{\alpha \bar{X}_2 - \bar{M}_1}{\bar{M}_1} < t^q \tag{19.9}$$

which as indicated is less than the quota-equivalent tariff t^q under the quota with revenue going to the domestic government. This difference is referred to as an "income effect" and can be thought of in terms of Figure 19.1. Giving the revenue to the foreigner reduces income and so shifts the demand curve D_1 down/left in Figure 19.1. The tariff equivalent needed to hold imports at $\bar{M}_1$ falls as the demand curve shifts down.

19.4 (Non) Equivalence of tariffs and quotas, other related policies

There are a number of situations in which tariffs and quotas are not equivalent. We have already touched on several of these which we can review quickly and then mention a few others. First, if the economy grows (or shrinks) then a tariff and quota, while initially equivalent under an auction, will no longer be equivalent. Shifts in the demand curve in Figure 19.1 will move the economy along the foreign supply curve (quantity adjustment) with a tariff whereas a quota means that adjustment will be entirely in price changes needed to clear the market at a fixed quantity. A sufficient fall in demand means that the quota becomes non-binding.

The effects of shifting world prices (movement up or down of p^* in Figure 19.1) has related but distinct consequences. Suppose again that the tariff and the quota are initially set to be equivalent. Under a tariff, the equilibrium will slide up or down the domestic demand curve leading to changes in both price and quantity. Under a quota, an increase in the world price will simply raise or lower the quota rent. The tariff equivalent value t^q will absorb the change in world price and the domestic consumer price will remain unchanged as long as the quota is binding. We also noted the important difference between a tariff and a VER quota in terms of rent distribution.

When the import good is supplied by a foreign monopoly or domestic import-competing supply is from a domestic monopolist, a tariff and quota are not equivalent. Suppose the market is supplied by a foreign monopolist. Then the imposition of a binding domestic quota will lead the foreign supplier to immediately raise price until the import price p^* is equal to the quota-constrained domestic price. The government can auction licenses if it likes, but it will find that the market-clearing price is zero: the foreign monopolist will capture all rents, even with an auction.

When the domestic import-competing good is supplied by a domestic monopolist against foreign competitive suppliers, the situation is a little more subtle. Interpret the horizontal axis as including both imports and domestic supply. In the situation of Figure 19.1 with a fixed world price, the domestic monopolist has in fact no market power under a tariff and the outcome will be the same as the competitive outcome. Any reduction in supply by the domestic monopolist will just be replaced by more foreign supply, so the domestic firm has no market power. With a quota, the domestic firm can now raise the domestic price by cutting supply, since consumers cannot buy any more from the foreign competitive

suppliers. Thus the quota confers market power on the domestic monopolist where there was none before.

Another thing that has been documented in the literature is the tendency or incentive for foreign suppliers to engage in quality upgrading under a quota as opposed to a tariff. An *ad valorem* tariff just increases the duty in proportion to the price of the import. But a quota gives clear incentives for a foreign supplier to substitute more expensive products or models, since the added price of the latter is not taxed. A quota gives suppliers an incentive to substitute high-margin goods for lower ones.

Finally, there is an issue about tariffs versus quotas when the foreign country engages in retaliation in the face of home-country protection. Once again, quotas come off very much worse in the comparison. Suppose that the home country puts on an import quota with licenses distributed by auction. The incentives for the foreign country are now much like the situation with a foreign monopoly supplier just discussed (the foreign government now takes on the role of the monopoly supplier). The foreign government should retaliate by putting on an export quota equal to the home import quota. This raises the export price to the market-clearing domestic price in the home country, and the auction price for the home-country quota will be zero. A more detailed analysis will show that a non-cooperative Nash equilibrium under quotas, in which each country has no incentive to change their quota, involves much less trade (in fact it can be autarky) and a much larger welfare loss than a non-cooperative Nash equilibrium in tariff rate.

Other forms of protection, particularly contingent protection in which tariffs or quotas are instituted in response to particular events, include policies such as anti-dumping, safeguards, and countervailing duties. These are discussed in Chapter 21.

19.5 Summary

As every student of economics understands, there is a price side and a quantity side to every market. Tariffs, export taxes and subsidies, and other taxes operate on the price side of the market, and we have shown that this will lead to significant quantity adjustments as, for example, the import price of a good rises with a tariff. Quotas and other regulatory policies operate on the quantity side of the market, generally fixing the maximum quantity of imports in the case of an import quota. This then implies that adjustments must occur on the price side on the market. Price instruments lead to quantity adjustment, whereas quantity instruments lead to price adjustment.

A tariff creates tariff revenue, the difference between the domestic and foreign price times the quantity of imports. A quota creates a similar difference between the domestic demand price and the foreign supply price for the quota-constrained quantity. This difference in prices times the import quantity is termed “quota rent”. The degree to which a tariff and quota are equivalent depends on who gets the quota rents and this, in turn, depends on how the quota is instituted. Licenses auctioned by the domestic government are closely equivalent to a tariff, at least in a competitive market. Licenses distributed to favored domestic individuals or firms enrich the licensed party at the expense of the government treasury. Licenses given to the foreign country or exporters are termed “voluntary export restraints” and the quota rent is then transferred abroad. Finally, lengthy bureaucratic procedures, lobbying and so forth can mean that the quota rents are dissipated in the use of real resources to get the licenses, the worse of all outcomes (filling out forms, standing in line, lobbying instead of actually working to produce things). This has been termed DUPS: directly unproductive activities.

We concluded with a discussion of a number of situations in which tariffs and quotas are not equivalent, even if set initially to achieve the same outcome. These include economic growth, world price fluctuations, domestic monopoly, foreign monopoly, and retaliatory trade wars. This list, plus the problem of the distribution of quota rents and DUPS, makes economists virtually unanimous in preferring tariffs to quotas when protection is a political inevitability.

REFERENCES

Baldwin, R. E. (1974), "Nontariff Distortions of International Trade." In R. E. Baldwin and J. D. Richardson (eds.), *International Trade and Finance.* Boston: Little, Brown.

Bhagwati, J. N. (1968), "More on the Equivalence of Tariffs and Quotas." *American Economic Review* 58, 142-146.

Bhagwati, J. N. (1982), "Directly-unproductive Profit-seeking (DUP) Activities." *Journal of Political Economy* 90, 988-1002.

Feenstra, R. C. (1992), "How Costly is Protectionism?" *Journal of Economic Perspectives* 6, 159-178.

Krueger, A. O. (1974), "The Political Economy of the Rent-Seeking Society." *American Economic Review* 69, 291-303.

Maskus, K. E. (1989), "Large Costs and Small Benefits of the American Sugar Programme." *The World Economy* 12, 85-104.

Figure 19.1

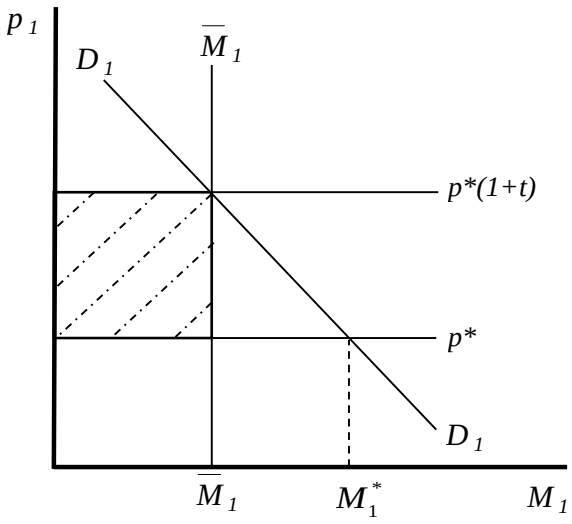
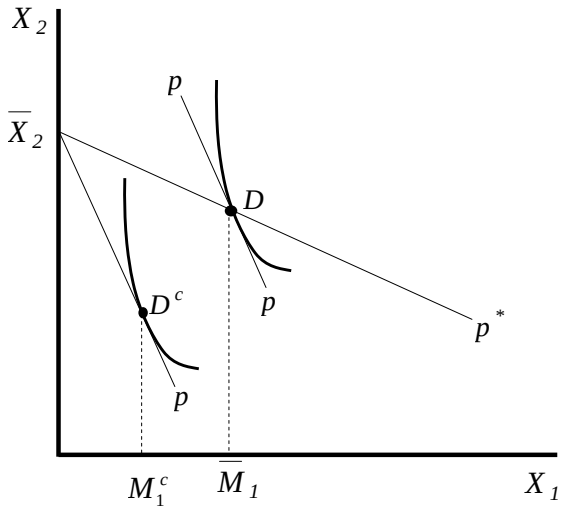


Figure 19.2



Chapter 20

STRATEGIC TRADE POLICY

20.1 Trade policy with increasing return to scale and imperfect competition

An interesting and important phenomenon occurs when governments act as agents in support of large domestic firms in the international marketplace. Certain actions by the governments in question are designed to give domestic firms important advantages over foreign rivals in competing for international business. These actions often involve some direct or indirect form of subsidy, which lowers the costs of the domestic firms relative to their competitors.

Although such subsidies or other forms of support may increase profits of the national firms, analysis in previous chapters (e.g., Chapter 10) should warn us that this in no way implies that such subsidies are welfare improving for the country. On the other hand, the fact that we know that large, imperfectly-competitive firms are "under producing" (producing outputs for which price exceeds marginal cost) might make us suspect that measures to stimulate the production and sales of large domestic firms might possibly be welfare improving.

From the mid-1980s to the mid-1990s, a large number of scientific papers were produced which examined trade policies, such as tariffs and quotas, under conditions of imperfect competition and increasing returns to scale. This has become known as the "strategic trade policy" literature, for better or for worse. Many interesting models produced non-traditional results, such as a finding that a production subside or an export subsidy can improve welfare. This result never holds in a competitive, distortion-free environment. Other authors changed the basic assumptions of the strategic models and completely reversed the findings.

The resulting literature is thus full of special cases, models and results which do not generalize in any simple fashion. The strategic trade policy literature is not a useful guide to government policy at this time. Collectively, this portfolio of special cases may help us understand the world better, a world in which trade is increasingly dominated by large, multinational firms. In this chapter, we will examine a number of special cases in order to get a flavor of the strategic trade-policy literature, and hopefully obtain some understanding of what sorts of environments produce what results. In particular, we will consider the specific form of competition and whether or not firms can enter or exit an industry in response to changing profits. The final sections of the chapter consider a few other important dimensions of strategic trade policy.

20.2 Export rivalry I: Cournot competition

In this section and in the next, we will consider the simplest possible model. There are two countries, home and foreign, each having a single firm producing an identical product. We will assume that there are no domestic sales: each firm only sells the product to a third country. The consequence of this weird assumption is that it makes changes in profits of the nation's firm (minus subsidy payments) equivalent to changes in welfare. There is no conflict between a higher price increasing profits but reducing domestic consumers' welfare. The job of the government is very simple: assist the domestic firm

in the international market place, and help the firm earn the maximum possible profits (Spencer and Brander, 1985 Brander and Spencer 1983).

Here, we will present a diagrammatic analysis, with a formal algebraic analysis in Section 20.4. A Cournot model, in which firms can be thought of as choosing their best output given the output of the other firm, gives rise to a construction called a “reaction curve” or “best response” curve. This gives the firm's optimal output for every possible output of its rival.

The reaction function for the home firm is shown in Figure 20.1 as RC_h. The reaction function is downward sloping: as X_f increases, the market remaining for the home firm is essentially shrinking, and it is optimal for the home firm to reduce its output, X_h. Point A in Figure 20.1 is the output the home firm would produce if the foreign firm produced nothing, and point B is the level of X_f at which the home firm quits producing entirely.

Figure 20.1

With some further analysis, we could derive a set of iso-profit curves for the home firm, shown as $\pi_{h1} < \pi_{h2} < \pi_{h3}$. Each point on one iso-profit curve yields the home firm the same level of profits. The home firm's profits improve as its rival's output, X_f decreases, so the home firm moves to higher iso-profit curves as it moves *down* the reaction function. The reaction function is, in fact, the locus of points where the home firm's profits are maximized for a given level of X_f. Thus RC_h connects all of the "top" points on the iso-profits curves. The best possible point for the home firm in Figure 20.1 is point A, where it is a monopolist.

A similar reaction curve for the foreign firm is shown as RC_f in Figure 20.2. RC_h and RC_f are in fact mirror images of one another under assumptions used later in Section 20.4 (including the assumption that the firms have equal marginal costs). The Cournot equilibrium is given by point C in Figure 20.2, where the two reaction curves intersect: each firm is choosing its optimal output given the output of the other firm. The foreign firm has iso-profit curves corresponding to those of the home firm, with these curves being vertical on RC_f: RC_f is the locus of optimal choices of X_f for given levels of X_h. The iso-profit curves for the home and foreign firm at the Cournot equilibrium are shown as π_{hc} and π_{fc} respectively. These iso-profit curves are perpendicular to one another at the Cournot equilibrium, indicating that each firm is choosing its optimal output given the output of the other firm. The hatched area between these two iso-profit curves gives a region in which both firms could be better off by colluding to mutually reduce outputs.

Figure 20.2

Now consider an activist government in country h. The foreign firm's reaction curve RC_f could be thought of as a constraint: the home government can assist the home firm in choosing an output such that the home firm's profits are maximized, subject to being on the foreign reaction curve RC_f. In Figure 20.2, we see that the best possible point for the home firm is point S, where iso-profit curve π_{hs} is just tangent to the foreign reaction function RC_f. How can the home government influence the situation such that the Cournot equilibrium is shifted to *S* in Figure 20.2? One way would be for the home government to give the home firm an output subsidy. This would reduce the home’s firm's marginal cost, and make it willing to supply a larger quantity of X_h for any given level of X_f. In other words, the output subsidy would shift the home firm's reaction function to the right in Figure 20.2. If the home government chooses just the right subsidy, the home firm's reaction function will be shifted just enough so that it passes

through point *S* in Figure 20.2. This new reaction function is shown by the dashed line in this diagram. The home firm earns higher profits in international markets, and these profits are a component of home's national income.[1]

Note that the foreign firm is worse off after the home subsidy. The foreign firm is on a lower iso-profit curve than π_{fc} in Figure 20.2, and thus foreign national income is reduced. This result that the subsidy makes home better off and foreign worse off is generally called "rent shifting", in the sense that oligopolistic rents in this market are shifted from the foreign firm to the home firm.[2]

A final important point is that, if both governments play this game, the outcome is jointly sub-optimal and reduces both countries' welfare relative to the initial Cournot equilibrium. If both countries institute an equal subsidy, both reaction functions in Figure 20.2 shift out and their intersection point is to the northeast of C. Both firms earn lower profits (net of the subsidy) and considerable benefits are captured by third countries which buy the subsidized goods.

To sum up, Figure 20.2 establishes a case in which active government support for a domestic firm in the international market place can increase national welfare. However, it is a fragile argument as we shall show in the next section.

20.3 Export rivalry II: Bertrand competition

A plausible alternative to the Cournot assumption that firms pick quantities is that firms pick prices. Each firm picks its optimal price given the price of its rival. This is known as Bertrand competition. This might seem a relatively harmless change, but in fact it reverses the result of the previous section. The home government's optimal policy is to *tax* the home firm, not subsidize it (Eaton and Grossman 1985).

Under the assumptions of linear demand and constant marginal cost, the outcome of Bertrand competition is more competitive than Cournot. Bertrand competitors produce more than Cournot competitors, and the equilibrium price and firm profits are lower than under Cournot. A tax by the home government can restraint Bertrand competition, and thereby reduce profits of the Home firm by less than the resulting tax revenue. Home welfare increases.

Another way economists think about this is that the firm's outputs are "strategic substitutes" in the Cournot case: an increase in the home-country subsidy will lead the foreign firm to reduce output, giving a benefit to the home firm. In Bertrand competition, the firm's prices are "strategic complements": an increase in the home-country's tax leads the foreign firm to increase its price, giving a benefit to the home firm.

Figure 20.3 shows a Bertrand reaction function for the Home firm, labelled RC_h. In the Bertrand case, the prices of the two firms are on the axis since prices rather than quantities are now the strategic variables. If the Foreign firm raises its price p_f, that reduces the demand for X_f and increases the demand for X_h. Because of the increased demand, it is optimal for the Home firm to increase its price p_h (generally by less than the increase in p_f), and hence the Bertrand reaction curve is upward sloping, reflecting this idea of strategic complements. A set of iso-profit curves for the home firm can be derived here in the same fashion as in Figure 20.1. Profits of the home firm increase as the foreign firm increases its price because, when p_f increases, more demand is shifted to the home firm. Figure 20.3 shows three

iso-profit curves with $\pi_{h3} > \pi_{h2} > \pi_{h1}$. The reaction curve RC_h is the locus of "bottom" points on the iso-profit curves; that is, the reaction curve gives the optimal (profit maximizing) level of p_h for each level of p_f. Opposite to the Cournot case, profits for the home firm now increase as the firm moves up its reaction curve, benefitting when its rival (foreign) raises its price.

Figure 20.3

Figure 20.4 adds the corresponding reaction curve for the Foreign firm, RC_f. Bertrand equilibrium is given by the intersection of the reaction curves at point B in Figure 20.4. At the Bertrand equilibrium, the two iso-profit curves π_{hb} and π_{fb} are perpendicular: each firm is choosing its optimal price given the price of the other firm. The hatched area between the two iso-profit curves is a set of mutually preferred points, in which both firms could achieve higher profits if they colluded to increase prices.

Figure 20.4

Now consider an activist home government as we did in the previous section. The foreign reaction curve RC_f can be thought of as a constraint and the home government's task is to help guide its firm to the highest profit level consistent with this constraint. The best point for home on RC_f is point *T*, with corresponding profit level π_{ht} in Figure 20.4. The appropriate policy for the home government is to use tax/subsidy policy to shift the home firm's reaction function to the right, to the position of the dashed line through T in Figure 20.4. The appropriate policy is now an output tax, the reverse of the optimal policy in the Cournot case. When the reaction curve is shifted to the right, we are saying that the Home firm now wishes to charge a higher price p_h for any given level of p_f. This is consistent with a higher level of marginal cost for the home firm, and hence is accomplished by a tax, not a subsidy.

It is interesting to note that home's tax also serves to raise the profits of the foreign firm, whereas the optimal subsidy in the Cournot case lowered the profits of the foreign firm. The tax imposed by the Home government acts to restraint competition, raising the pre-tax profits of both firms, though the home firm is likely worse off after paying the tax and won't like this outcome at all.[3] Finally, if both countries institute a tax, both countries benefit and the rest of the world loses. All results of the Cournot case are reversed.

The contrast between these results and that of the previous section serves to illustrate the caveat offered in the introduction to this chapter. The theory of strategic trade policy is very sensitive to the underlying assumptions and cannot really be taken as a guide to government policy at this time.

20.4 Cournot with and without entry, adding domestic consumption

We now add a more formal treatment of Cournot competition, and do this with and without free entry and exit of firms in response to the imposition of a subsidy. We will also add domestic consumption. The assumption used above that all sales go to third countries is a rather odd assumption indeed, especially for the firms often cited as applications of strategic-trade-policy theory (e.g., Boeing and Airbus).

We will use the quasi-linear utility function of Chapter 11, and assume two identical countries, each having a single firm. The countries are denoted *h* and *f*, with country *h* instituting a per-unit subsidy

s and country *f* remaining passive. Firm h's profits are given as follows, with an equivalent expression for firm *f*, except there is no subsidy term in the profit equation of firm f.

$$\pi_h = pX_h - cX_h + sX_h = [\alpha - \beta(X_h + X_f)]X_h - cX_h + sX_h \tag{20.1}$$

Each firm picks its optimal output given the output of the other firm. The first-order conditions for profit maximization are then:

$$\frac{d\pi_h}{dX_h} = \alpha - 2\beta X_h - \beta X_f - c + s = 0 \tag{20.2}$$

$$\frac{d\pi_f}{dX_f} = \alpha - 2\beta X_f - \beta X_h - c = 0 \tag{20.3}$$

Solving these equations explicitly give us the reaction or best-response functions that we used in Figures 20.1 and 20.2.

$$X_h = \frac{\alpha - c + s}{2\beta} - \frac{1}{2}X_f \qquad X_f = \frac{\alpha - c}{2\beta} - \frac{1}{2}X_h \tag{20.4}$$

Solving these two simultaneous equations gives us the Cournot equilibrium outputs.

$$X_h = \frac{\alpha - c + 2s}{3\beta} \qquad X_f = \frac{\alpha - c - s}{3\beta} \tag{20.5}$$

If we substitute these outputs back into the profit equations for the two firms, we will find that profits have a simple expression in this model, they are just β times the equilibrium output squared.

$$\pi_h = \left[\alpha - 2\beta\frac{\alpha - c + s/2}{3\beta} - c + s\right]\frac{\alpha - c + 2s}{3\beta} = \beta X_h^2 \tag{20.6}$$

$$\pi_f = \beta X_f^2 \tag{20.7}$$

From this, we can calculate the optimal subsidy which was illustrated graphically in Figure 20.2. The government's objective it to maximize the local firm's profits minus the subsidy payments. Denoting X_h as a function of *s*, the firm's objective is to maximize its profits, given by

$$\max \pi_h - sX_h \equiv \beta X_h(s)^2 - sX_h(s) \tag{20.8}$$

Using the first equation of (20.5) to substitute for X_h, we can differentiate (20.8), setting the derivative to zero to get the optimal value of the subsidy.

$$\frac{d\pi_h}{ds} = 2\beta X_h\frac{dX_h}{ds} - s\frac{dX_h}{ds} - X_h = 0 \qquad \frac{dX_h}{ds} = \frac{2}{3\beta} \tag{20.9}$$

where the second equation follows from (20.5). When this is substituted into (20.9), we have a formula for the optimal subsidy.

$$s = \frac{\alpha - c}{4} \tag{20.10}$$

An odd assumption that was used in the early literature is that there is no domestic consumption. Thus now we return to our more standard case of two countries, each with one domestic firm selling to its home market and exporting to its rival's market. Adding in domestic consumption requires us to consider the effect of the subsidy on consumer surplus as well as on profit income. Using the utility function which generates the linear demand curve in (20.1), we derived the expression for consumer surplus in Chapter 13, ending with equation (13.14). Let X_{hh} and X_{hf} denote firm h's sales to its home and rival's market respectively, and X_{fh} denotes the sales (exports) of firm *f* to h's market. We showed that consumer surplus in country *h* is equal to

$$CS_h = (\beta/2)(X_{hh} + X_{fh})^2 \tag{20.11}$$

We then showed in Chapter 13 that welfare of country *h* consists of consumer surplus, plus profit income, and now we modify that only by subtracting subsidy payments. Welfare for country *h* is given by consumer surplus plus profits minus the subsidy payment.

$$U_h = (\beta/2)(X_{hh} + X_{fh})^2 + \beta X_{hh}^2 + \beta X_{hf}^2 - s(X_{hh} + X_{hf}) \tag{20.12}$$

where the home sales, export sales, and foreign imports are given from (20.5) by

$$X_{hh} = X_{hf} = \frac{\alpha - c + 2s}{3\beta} \qquad X_{fh} = \frac{\alpha - c - s}{3\beta} \tag{20.13}$$

In addition to the profit-shifting motive, a subsidy increases domestic consumption and consumer surplus: the sum of the two equations in (20.13) is positive. This also applies to country *f*: consumer surplus increases since imports from *h* go up by more than domestic sales decrease. However, this must be outweighed by the loss of profit income in country *f*, and so the latter must be worse off by country h's subsidy.

These results are shown in Figures 20.5 and 20.6, the former is for the subsidy-imposing country *h* and the latter is for the passive country f. The countries are identical and point A in each Figure represents the Cournot equilibrium; there is no net trade in the symmetric free-market equilibrium. The subsidy moves country *h* to a production point B and to consumption point D in Figure 20.5. Country *h* becomes a net exporter of *X* and welfare improves in spite of a fall in the world price of *X*. Country f's firm reduces output to point B in Figure 20.6, and welfare is given by the indifference curve through the consumption point D.

Figure 20.5 **Figure 20.6**

It is interesting to remark how different the result is here from the case of a production subsidy by one of two identical countries that we considered in Chapter 10 in an environment of perfect competition. In Figure 10.3 we noted that this must make the subsidizing country worse off and the passive country

better off. Now we reverse both of these results. The basic underlying intuition behind this is that we are now in an initially-distorted world, where the price of *X* exceeds marginal cost and *X* is under-produced in equilibrium. The subsidy acts as an offsetting distortion for the subsidizing country. However, for country *f,* the subsidy in country *h* now exacerbates the initial distortion, leading to a smaller output of a good that was already under-produced.

A second unrealistic feature (in addition to no domestic consumption) of the basic model of Section 20.2 above is that there is no entry or exit of firms in either country in response to the subsidy. In this section, we make the polar opposite assumption, as in Chapter 11, and assume that the numbers of firms in both countries adjusts to the subsidy in country *h* until there are zero profits for all remaining firms (Venables 1985, Horstmann and Markusen 1986, Markusen and Venables 1988). The identical country *h* firms produce goods denoted by X_h , X_{hi} being the output of the ith firm while foreign goods are denoted by X_f , with X_{fj} being the output of the jth firm in country h.

The identical *X* goods are freely traded so there is a single world price *p* as before, given by

$$p = \alpha - \beta\left[\sum_i X_{hi}\right] - \beta\left[\sum_j X_{fj}\right] \tag{20.14}$$

with profits of firm X_h given by

$$\pi_h = pX_{hi} = \left[\alpha - \beta\left[\sum_j X_{hj} + \sum_j X_{fj}\right]\right]X_{hi} - cX_{hi} - sX_{hi} \tag{20.15}$$

Marginal revenue minus marginal cost for firm i in country *h* is given by differentiating (20.15) with respect to X_i, holding all other X_h outputs and all X_f outputs constant.

$$MR - MC = \alpha - 2\beta X_{hi} - \beta\sum_{j\neq i} X_{hj} - \beta\sum_j X_{fj} - c + s = 0 \tag{20.16}$$

Now impose symmetry. X_h and X_f will denote the outputs of a representative firm X_{hi} and X_{fj} firm respectively, and n_h and n_f the number of firms. All firms that are active in equilibrium will produce the same amount.

$$MR - MC = \alpha - \beta(n_h + 1)X_h - \beta n_f X_f - c + s = 0 \tag{20.17}$$

Add *f* to denote a fixed cost. The zero profit condition is that the profits of the representative firm are exactly zero.

$$\alpha X - \beta n_h X_h^2 - \beta n_f X_f X_h - cX_h + sX_h - F = 0 \tag{20.18}$$

Multiply *MR* - *MC* in (20.17) through by X_h. Solve the two equations for X_h.

$$X_h = \left[\frac{F}{\beta}\right]^{1/2} \tag{20.19}$$

The important thing to note is that the output quantity of a representative X_h firm does not depend on marginal costs or on the subsidy. Thus a small specific subsidy to X_h does not change output per X_h firm or change average cost. Further, if we were to solve equations (20.17) and (20.18) without the subsidy term we would get exactly the same result and the output of a country *f* firm would be given by (20.19). However, if the representative X_h and X_f are equal, then we will arrive at a contradiction. If the X_h firm breaks even in equation (20.18) with the subsidy term, then a representative X_f firm without the subsidy must be making losses. The effect of country h's subsidy is to drive country *f* out of production.

The fact that the subsidy has no effect on the output of an *individual* X_h firm does not mean that either country is unaffected. Since X_h is constant, the average cost of X_h must be constant. But if the producer price is unchanged by the subsidy, this in turn means that the world price of X_h must fall. The world price is less than average cost, the latter being equal to $p + s$: raising s with $(p + s) = ac$ constant means p must fall.

The situation for countries' *h* and *f* is shown in Figures 20.7 and 20.8 respectively. In each Figure, point A represents the identical free-market equilibrium in the two identical countries. The effect of the subsidy is to cause entry in country *h*, shifting the linear segment of the production frontier inward from $F\bar{X}$ to $F^s\bar{X}$. Total fixed costs used in X_h production rise from $\bar{Y}F$ to $\bar{Y}F^s$. With average cost constant because output per firm X_h is constant, the new production point in Figure 20.7 must be at point B. With a lower price ratio *p*, the consumption point must be something like point D in Figure 20.7. In country *h*, there are more firms each producing at the old scale of production. Country *h* exports good *X*, but this is not beneficial and, once again, we learn not to confuse exports with welfare.

Figure 20.7 **Figure 20.8**

The effect of country h's subsidy on country *f* is shown in Figure 20.8. Country f's firm are driven out of *X* production and so the country specializes by producing at $\bar{Y}$. The world price of *X* has fallen and so country *f* reaps a welfare gain, consuming at point D in Figure 20.8. This reverses the result of Section 20.2 and indeed re-establishes the result of competitive trade theory: the subsidizing country is worse off and the passive country is better off.

20.5 Other issues and further reading

Strategic trade policy was an important and fascinating topic for researchers from the mid-1980s to the mid-1990s as we indicated earlier. Many modifications of the basic structure and assumptions were proposed and analyzed. We cannot cover these here, but rather just outline a few of the modifications and extensions for those interested.

Foreign ownership.

Dick (1993) points out that the basic argument of Section 20.2 relied on the implicit assumption that each domestic firm is domestically owned, so its profits are all included in the national income stream. But empirically, there is widely-spread international equity ownership of large, oligopolistic firms and so part of their profit income goes to foreigners. Dick shows that at a rather modest share of foreign ownership, the argument for a subsidy collapses, even with Cournot competition and no entry or exit.

Segmented versus integrated markets

A common assumption in strategic trade-policy literature is termed market segmentation or price discrimination: firms can set prices differently in home and export markets. The alternative assumption is termed integrated markets, also known as factory-gate prices or mill prices. Firms cannot charge different prices depending on the destination of the goods. It turns out that this assumption is quite important once there are tariffs or transport costs between markets, as shown by somewhat contradictory results in Horstmann and Markusen (1986) (integrated markets) and Venables (1985) (segmented markets). The two are reconciled in Markusen and Venables (1988), who demonstrate that strategic trade-policy benefits are significantly weakened in integrated markets, relative to segmented markets.

The ability of firm to price discriminate across international markets comes up in many other trade-policy controversies, such as whether or not parallel trade is allowed. Parallel trade basically allows independent firms to arbitrage goods from where they are cheap to where they are expensive without the permission of the original manufacturing firm (e.g., US pharmaceutical products are cheaper in Mexico than in the US). Firms want to segment markets to prevent this from happening, and fiercely lobby their governments to make parallel trade illegal.

Import protection as export promotion

Another interesting example of the strategic use of export subsidies occurs when the *marginal cost* of production, as well as average cost, is decreasing (Krugman, 1984) with higher output. Suppose that a home firm and a foreign rival are each selling in their own markets, exporting to their rival's market, and perhaps selling in third countries. Assume that marginal cost decreases with the amount produced. Now assume that the foreign country erects a trade barrier (a tariff or quota) to imports of the home-country product.

The effect of the foreign trade barrier is to restrict the foreign market for its domestic firm, or at least to give that firm a larger share of the domestic market. Output of the foreign firm increases and that of the home firm falls, given constant levels of sales by each firm in other markets. But that is not the end of the story. This change in output levels implies a decrease in the marginal cost of production for the foreign firm and an increase in marginal cost for the home firm. Thus there will be further ramifications of the foreign trade barrier in that the foreign firm will now be more competitive in the home country's market and in third country markets. Conversely, the home firm will be less competitive in its own market and third-country markets.

The equilibrium condition of marginal revenue equals marginal cost now implies that the foreign firm will expand sales in the home country's market and in third countries, and that home firm's sales will shrink. Rents in these markets will be shifted to the foreign firm as described in an earlier section and the prices charged could even fall in the foreign market, providing a benefit to consumers in that country. Import protection by the foreign country thus becomes export promotion.

Krugman also shows that this argument holds up when the scale economies are in the form of dynamic learning-by-doing, in which marginal costs of production fall with accumulated sales experience. Protecting or reserving the home market for the domestic firm allows that firm to lower its costs more quickly and therefore, to compete more effectively in export markets. By reserving their domestic market for domestic firms, the protecting country can develope a more competitive industry than would occur strictly through market forces (Baldwin and Krugman 1988).

As in the case of the production subsidy discussed earlier, this argument may be weakened by the inefficient entry of new firms in response to the introduction of protection. Nevertheless, it again illustrates why the existence of scale economies and imperfect competition may imply radically different policies from those of traditional theory as discussed in the introduction to this chapter.

Quotas and VERs as “facilitating practices”

Articles by Harris (1985) and Krishna (1989) raise interesting points about how non-tariff instruments of trade policy can lead to some unexpected outcomes in the presence of imperfect competition and scale economies. These authors both raise the possibility that a voluntary export restraint imposed by home on the foreign exporter of some good competing in a duopoly with a home firm may lead to an increase in the profits of both firms. A Bertrand duopoly model is used, much like that developed in Section 20.3 above. The voluntary export restraint puts a check on the amount the foreign firm can export to the home country, but leaves the rents generated by the supply restriction in the hands of the foreign firm. Both authors show that the VER can lead to an increase in the profits of both firms by restraining the "excessive" competition associated with Bertrand pricing. In Krishna's terminology, the VER facilitates collusion. In Harris' terminology, that is why the voluntary export restraint is voluntary.

Trade Policy in monopolistic-competition models

Strategic trade policy questions have also been examined in monopolistic-competition models of the type introduced in Chapter 12. Flam and Helpman (1987) show that a tariff on imported, differentiated goods improves welfare, both due to the usual optimal-tariff argument (the tariff improves the terms of trade) and because consumer expenditure is shifted to domestic substitute products, produced with increasing returns to scale. This leads to a beneficial expansion in the domestic increasing-returns sector.

Markusen (1989, 1990) later shows that this result depends on the domestic and imported differentiated products being substitutes. While it may be reasonable to assume that differentiated consumer goods are generally substitutes, this is not so obvious with differentiated producer goods. An imported machine and a domestic computer may be complements in production, for example. Markusen shows that if the imported and domestic differentiated products are general-equilibrium complements, then an import tariff may reduce welfare. The tariff generates a favorable terms-of-trade effect, but the increased prices of the imported (producer) goods may generate a fall in demand for and production of the domestic inputs produced with increasing returns to scale. This can generate a negative "production expansion effect" that outweighs the favorable terms-of-trade effect. Once again, we see that appropriate policy depends very much on the underlying structure of the economy.

20.6 Summary

This Chapter has presented some examples of why scale economies and imperfect competition may give governments an incentive to assist strategically their domestic firms in international markets. By "assist," we mean actions such as direct or indirect subsidies that lower the costs to domestic firms of doing business abroad.

We noted that such subsidies are not optimal in models based on constant returns to scale and perfect competition. Domestic welfare deteriorates as a consequence of production or export subsidies

and the only beneficiaries are foreign consumers who can buy cheaper imports. Scale economies and imperfect competition imply an excess of price (the value of a good to consumers) over marginal cost (the value of resources needed to produce an additional unit). In such a situation, there is an incentive for governments to stimulate production. Furthermore, the existence of monopoly profits or rents in world markets may imply that a government can transfer more of those rents to domestic firms by subsidizing domestic production.

These are important findings, but it must be remembered that their validity rests on fairly specific assumptions. We demonstrated in two cases (Bertrand competition, free entry and exit) that pro-subsidy arguments are not valid. Arguments in favor of subsidies must, therefore, be very carefully evaluated in light of industry structure and other variables and assumptions. At the present time, the theory offers little support for a comprehensive government program to support its domestic firms in the international marketplace. Some specific findings are as follows.

When a home firm is competing with a foreign rival for sales to third markets, the home government can shift oligopoly rents in favor of the home firm with a production or export subsidy. This increases home country welfare if the home and foreign firms are Cournot competitors. Such a subsidy would reduce home welfare if the firms are Bertrand competitors. The optimal policy is actually a production or export tax in the Bertrand case, which increases home welfare as well as the profits of the foreign firm.

We add domestic consumption to the basic Cournot model, and exploit the type of diagrammatic analysis introduced in Chapter 11. Here, we see quite clearly why the traditional argument against production or export subsidies following from a competitive model may be reversed. With price in excess of marginal cost, an expansion in the output of the imperfectly-competitive, increasing-returns sector is beneficial and conversely a contraction is harmful. A small production subsidy can improve domestic welfare.

However, this reversal of the conventional competitive result can be reversed back again if firms are allowed to enter or exit in response to policy changes. Under special assumptions (linear demand, constant marginal cost), it can be shown that a production subsidy in fact leads only to the entry of new firms, not to an expansion in the output of the existing firms. No beneficial capture of scale economies occurs and, as in the case of the competitive model, the production subsidy hurts the subsidizing country and helps the other country.

In some cases, domestic and foreign markets are linked in a way that produces interesting implications for strategic trade policy. If marginal costs are declining or if firms are initially operating near zero profits, then the imposition of import protection in the foreign country can lead to ramifications back in the home country. This protection reserves the foreign market for foreign firms, giving them advantages in terms of lower marginal costs or forcing the exit of home firms from the market. In either case, the home country can suffer losses both through the loss of sales to the foreign market, but also through loss of sales and/or higher prices in its domestic market.

It has been shown that some policies, notably voluntary export restraints, can facilitate reduced competition between domestic and foreign firms. It is possible that such policies can then lead to increased profits for both domestic and foreign firms. Consumers are of course the losers.

There is generally some presumption that tariffs are beneficial in monopolistically-competitive

industries. They generate a favorable terms-of-trade effect, and lead to a beneficial expansion by domestic increasing-returns firms. This conclusion can be reversed however, if the imported goods are general-equilibrium complements for the domestic increasing-returns goods. In such a case, the higher import prices may lead to reduced production of the domestic goods, generating a welfare-reducing contraction in domestic increasing-returns production.

REFERENCES

Baldwin and Krugman, P. R. (1988), "Market Access and International Competition: A Simulation Study of 16K Random Access Memories", in Robert Feenstra (editor), *Empirical Methods for International Trade*, Cambridge: MIT Press.

Brander, J. A., and B. J. Spencer (1985)., "Export Subsidies and International Market Share Rivalry." *Journal of International Economics* 18, 227-242.

Deardorff, A. (1987), "Why do Governments Prefer Nontariff Barriers?", *Carnegie-Rochester Conference Series on Public Policy* 26m 191-216.

Dick, Andrew R. (1993), "Strategic Trade Policy and Welfare: the Empirical Consequences of Cross-Ownership", *Journal of International Economics* 35, 227-249.

Eaton, J., and G. M. Grossman (1985), "Optimal Trade and Industrial Policy Under Oligopoly." *Quarterly Journal of Economics* 101, 383-406.

Flam, H. and E. Helpman (1987), "Industrial Policy under Monopolistic Competition", *Journal of International Economics* 22, 79-102.

Harris, R. (1985), "Why Voluntary Export Restraints are 'Voluntary'", *Canadian Journal of Economics* 18, 799-809.

Helpman, E. and P. R. Krugman (1989), *Trade Policy and Market Structure*, Cambridge: MIT Press.

Horstmann, I., and J. R. Markusen (1986), "Up the Average Cost Curve: Inefficient Entry and the New Protectionism." *Journal of International Economics* 20, 225-248.

Krishna, K. (1989), "Trade Restrictions as Facilitating Practices", *Journal of International Economics* 26, 251-270.

Krugman, P. R. (1984). "Import Protection as Export Promotion: International Competition in the Presence of Oligopoly and Economies of Scale." In H. Kierzkowski (ed.), *Monopolistic Competition in International Trade*. Oxford: Oxford University Press.

Markusen, J. R. (1989), "Trade in Producer Services and in other Specialized, Intermediate Inputs", *American Economic Review* 79, 85-95.

Markusen, J. R. (1990), "Derationalizing Tariffs with Specialized Intermediate Inputs and Differentiated Final Goods", *Journal of International Economics* 28, 375-384.

Markusen James R. and Anthony J. Venables (1988), "Trade Policy with Increasing Returns and Imperfect Competition: Contradictory Results from Competing Assumptions," *Journal of International Economics* 24, 299-316.

Spencer, B. J., and J. A. Brander (1983), "International R & D Rivalry and Industrial Strategy." *Review of Economics Studies* 50, 707-722.

Venables, A. J. (1975). "Trade and Trade Policy with Imperfect Competition: The Case of Identical Products and Free Entry." *Journal of International Economics* 19, 1-20.

ENDNOTES

1. When the subsidy is applied, the home firm's actual profits at S in Figure 20.2 will be higher than π_{hs} by the amount of the subsidy payment, sX_h if s is the specific (per unit) subsidy rate. This amount of higher profits is exactly offset from a welfare point of view by the fact that consumers (taxpayers) pay the subsidy sX_h. Thus the *net* contribution of the subsidy to increased national income in the home country is the difference between π_{hc} and π_{hs} in Figure 20.2. An alternative way of saying the same thing is that π_{hs} in Figure 20.2 equals the firm's profits minus the value of subsidy payments.

2. The term rent shifting is a bit of a misnomer, or rather the phenomenon is more general. The same result is found in a model of external economies in which there are no monopoly rents for example. The result arises from the fact that price exceeds marginal cost (and not necessarily from the existence of profits) and thus the country whose firm expands production captures the excess of price over marginal cost on added output, while the country whose firm contracts suffers the loss of price over marginal cost on reduced output.

3. Similar to our discussion in the previous section, the home firm's actual profits after the tax are not given by π_{ht} in Figure 20.4, but are reduced below that amount by the tax. This reduction is exactly offset from a welfare point of view by the tax collection, so the difference between π_{ht} and π_{hb} is the *net* benefit of the tax to the home country. The tax increases the home country's welfare (measured as before in pre-tax or pre-subsidy profits), not the private profits of the home firm.

Figure 20.1

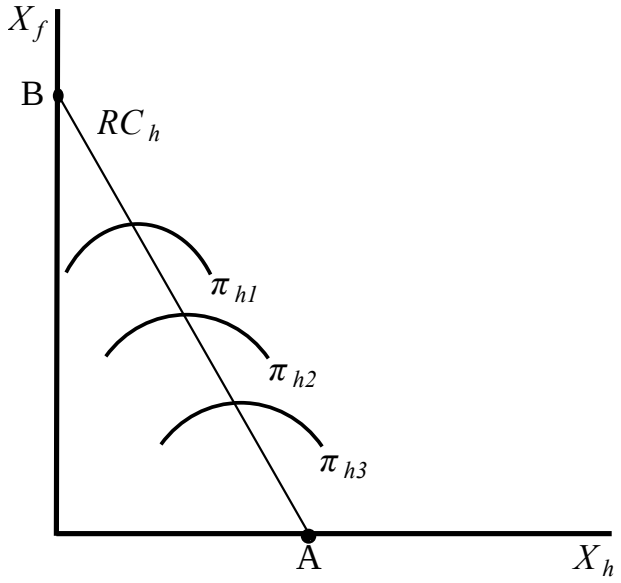


Figure 20.2

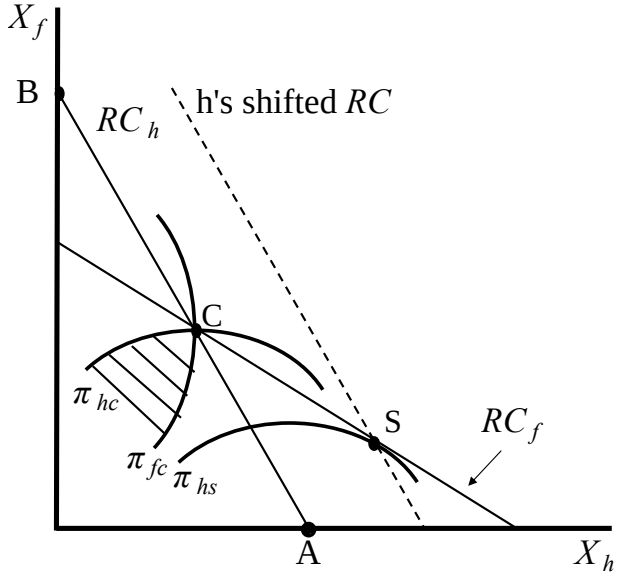


Figure 20.3

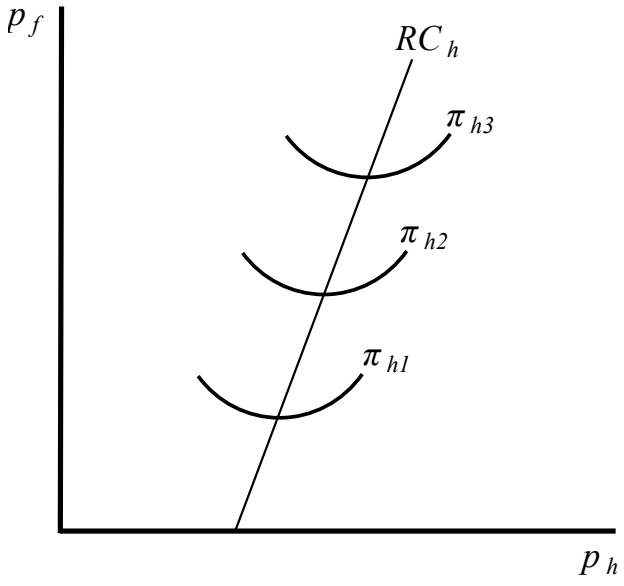


Figure 20.4

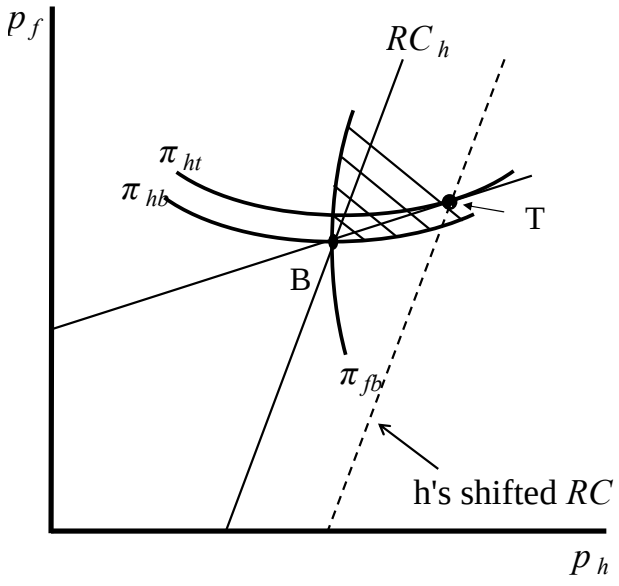


Figure 20.5

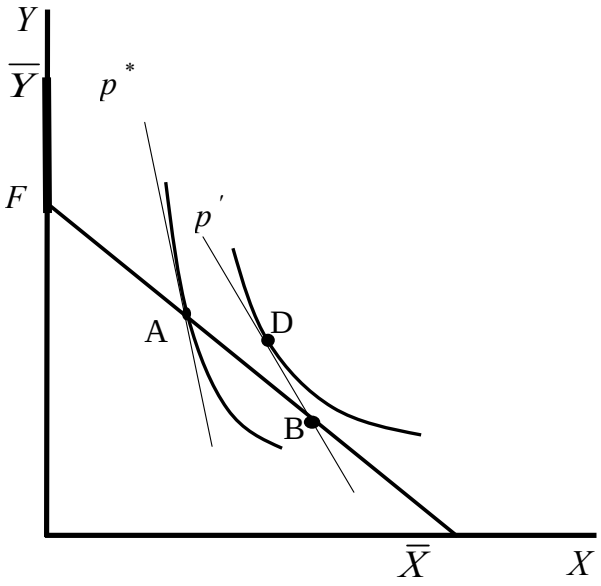


Figure 20.6

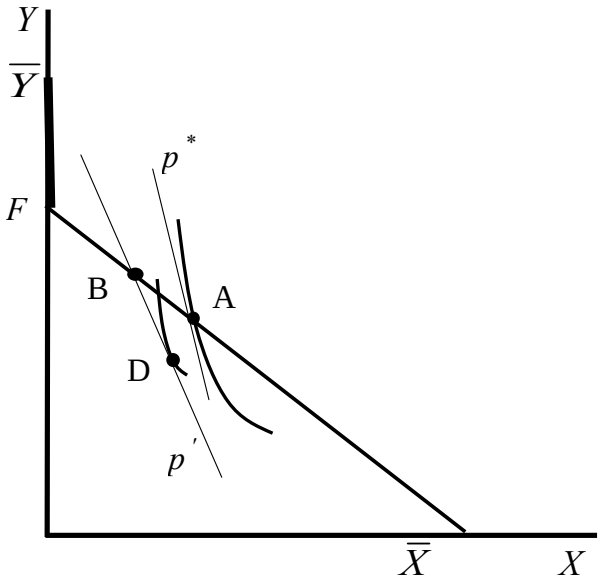


Figure 20.7

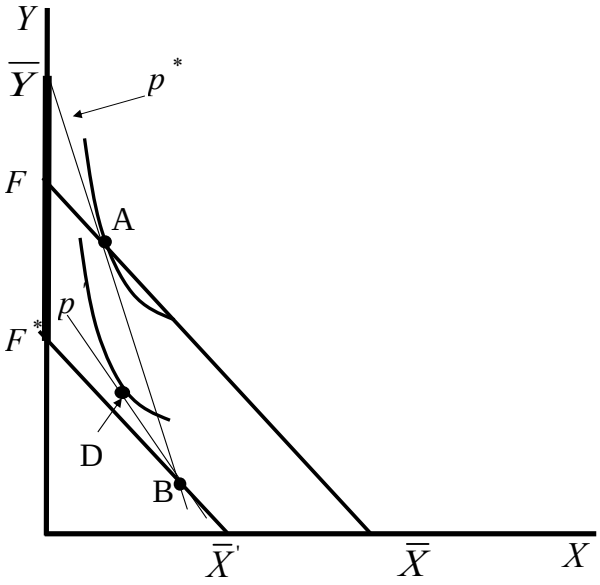
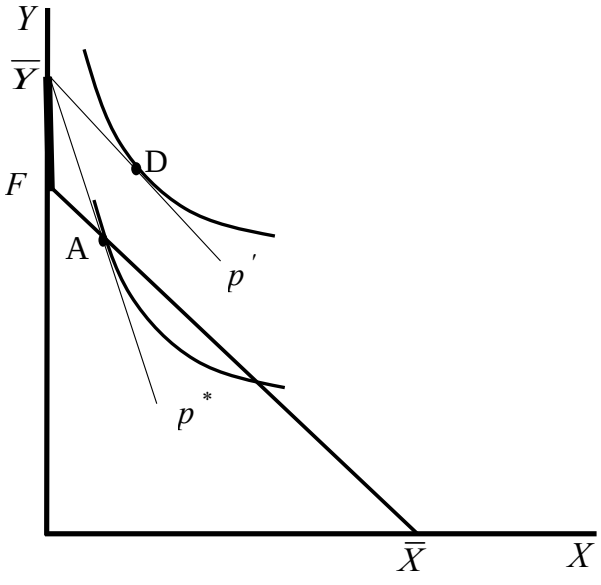


Figure 20.8



Chapter 21

MULTILATERAL TRADE AGREEMENTS: THE WORLD TRADE ORGANIZATION

21.1 Introduction

The text to this point has analyzed government policies that affect trade, investment and welfare largely as decisions made by individual countries, without much regard for their impacts on foreign nations. Clearly, however, such policies often have important regional or global effects and we may expect foreign governments to react with their own market interventions. Consider one basic example. Both the United States and the European Union traditionally have supported the incomes of their wheat and corn farmers (and the profits of their agribusiness companies) with a variety of price supports, direct payments, government purchase programs, import barriers and export subsidies. These policies reduce imports into these major markets, while also artificially expanding exports. A primary result is that global grain prices are generally lower than they would be otherwise, which diminishes the incomes of farmers in other countries and may stimulate trade barriers there. This issue has long been a source of policy conflict at the global level.

In analytical terms, the essential difficulty is that, in pursuing policies designed to maximize some measure of national or sectoral domestic benefit, governments may impose significant costs on trading partners. Thus, national trade policies and commercial regulations can embody *cross-border externalities*, which may be costly and invite retaliation. This issue arises across the spectrum of trade and regulatory policies, as suggested by the following examples. First, if a large country imposes an optimal tariff or import quota it will drive down the price received by foreign exporters and reduce welfare abroad. Second, production subsidies offered to oligopolistic domestic firms grant them an export advantage at the expense of foreign competitors. Third, if some countries have relatively weak environmental rules, permitting extensive emissions of greenhouse gases or the release of effluents into rivers and oceans, there are at least two negative international externalities. One is that the resulting pollution damages are felt abroad, causing health problems in foreign countries. The other is that lax pollution regulations can generate export advantages for domestic firms, to the detriment of competing international enterprises.

A closely related problem is that there are *global public goods*, such as reductions in greenhouse gas emissions, improvements in public health, increases in innovation and rules supporting open markets in trade and investment, that are insufficiently provided by private international markets and governments pursuing individual regulatory policies. For example, it is highly unlikely that private pharmaceutical companies would make costly investments in researching and developing medicines for diseases that exist only in impoverished countries, because those markets are not profitable. Rather, dealing with such problems requires a cooperative international solution among governments, international organizations and private firms.

Whenever such externalities arise, or global public goods are underdeveloped, there is a role for regional or global agreements and institutions to manage conflicts and facilitate cooperation. Thus, various forms of *international governance* exist in nearly all major areas of policy, including public health, environmental protection, labor standards, banking regulations, and technical specifications for

producing certain goods.[1]

In this chapter we focus on multilateral agreements and institutions governing international trade policies, leaving preferential trade agreements to the next chapter. Our emphasis is largely on analytical models and evidence that help explain the logic and impacts of such agreements, but we provide a descriptive overview as well.

21.2 The Logic of Trade Agreements

Why do countries have to negotiate with each other to reduce trade barriers when, as described earlier in the text, there are significant gains from unilateral open trade? The answer that economists focus on primarily is the fact that large countries can gain welfare at the expense of their trading partners by imposing a tariff. As described in Chapter 18, a country with significant monopsony power in trade faces a welfare tradeoff when it raises a tax on imports: the volume of trade is diminished, reducing economic efficiency and well-being, but the foreign export price is pushed down, improving the country's terms of trade. The country's optimal tariff maximizes the difference between the terms-of-trade (TT) gains and the volume-of-trade (VT) losses.

The difficulty, of course, is that the exporting country or countries are made worse off because the tariff reduces the price their goods receive in trade. As a result, these countries may choose to retaliate with their own tariffs, since doing so would tend to re-balance the terms of trade back in their favor. The net result would be lower trade volumes but a largely unchanged international price ratio, which would likely make all countries worse off compared to global free trade.

This basic logic may be seen in Figure 21.1, which extends the analysis in Figure 18.4. With no tariffs in place the home and foreign economies share price ratio p^{*F} in free trade and the volume of trade in good x_1 is OF. Now let home impose a tariff, shifting its import-demand curve in toward the origin. The result is a lower trade volume of OT, while the foreign economy must accept a lower relative export price at p^{*t}. Home is better off (if this is the optimal tariff) but foreign is unambiguously worse off, for both its volume of trade (efficiency) and terms of trade are lower. In response, foreign might impose an "optimal" retaliatory tariff (subject to home's existing tariff) of its own in good x_2. Rather than draw a trade diagram for that good, however, it is enough to recall that a foreign tariff on x_2 is equivalent to a foreign export tax on x_1. Thus, if the foreign economy did retaliate its export-supply curve would shift in toward the origin, as shown. This retaliatory equilibrium would involve an even-lower level of trade, at OR, while the relative price ratio of p^{*r} would be close to the original terms of trade. In brief, this kind of reaction could significantly reduce the amount of trade while having little net impact on the terms of trade. Global welfare would be smaller and both countries likely would lose.

Figure 21.1

Figure 21.1 does not depict a full policy equilibrium for two reasons. First, it is evident that the home country may react to foreign's tax by changing its tariff yet again, and foreign may respond to that change, and so on in a multiple-stage game. Second, it is difficult to depict welfare-maximizing policy reactions in goods space. Thus, we show how equilibrium retaliatory tariffs are determined in Figure

[1]A good introduction to this complex subject is the volume by Kaul, Grunberg and Stern (1999).

21.2. The home tariff rate is listed on the horizontal axis and the foreign rate on the vertical axis. The origin of the diagram represents free trade. Because no tariffs exist at this point both countries face the same price ratio (not shown) and the situation is Pareto-efficient.

Consider the home country's best-response tariff function, which shows the "optimal" tariff for home for any level of the foreign tax. If foreign has no tariff, home would choose the optimal tariff shown at point t^{hO}, which therefore must be on home's best-response tariff function. Suppose now that foreign has a small positive tariff. Recalling the analysis in Section 18.5, home's best response would depend on the VT losses versus TT gains it would achieve by altering its own import tax, assuming the foreign tariff is fixed.[2] For a small foreign tariff, home is likely to raise its own tax because the TT gains would apply to a large volume of imports, while the efficiency loss would be relatively small. Thus, it is possible (though not necessary) that the home best-response function (BR^h) initially slopes upward. However, as the foreign tariff grows larger the available TT gains diminish and the VT losses grow. As a result, at some point it will be best for home to begin reducing its tariff and BR^h begins to slope backward toward the vertical axis. Similarly, foreign's best-response function (BR^f) begins from its optimal-tariff (with a zero home tariff) and may initially slope upward, but at some point that economy will reduce its tariff and the curve is negatively sloped.

Figure 21.2

To depict welfare along these curves, recall the discussion of firm-level reaction curves in Section 20.2. In Figure 20.1, the home firm's reaction curve was the locus of sales levels at which its iso-profit contours became horizontal, meaning its profits were the highest possible given any level of foreign sales. Analogous to that analysis, a country's best-response function must show tariff rates that generate the highest national welfare for any given foreign tariff. Thus, B^{rh} is the locus of home tariff rates for which all iso-welfare contours become horizontal and B^{rf} is the locus of foreign rates at which its contours become vertical. In this context, a home iso-welfare contour refers to combinations of home and foreign tariff rates that keep home welfare (the sum of VT and TT effects) constant. For example, the contour at point B shows that if both tariff rates were on that curve but to the left and below the horizontal point, if foreign were to raise its tariff, then home could do so also and retain the same welfare. A final observation is that home's welfare rises as it moves down and to the right on its best-response function (the optimal tariff is its maximum welfare). Similarly, foreign welfare rises as it moves up and to the left on its function.

Now we may analyze the outcome of a retaliatory game. The Nash equilibrium lies at point N, which shows the tariff rates that would emerge. Note that these equilibrium import taxes may be higher or lower than the individually optimal tariffs. The primary point, however, is that reaching the Nash-equilibrium tariffs through retaliation makes both countries worse off than in free trade, for they are both on lower welfare contours than they would be at point T.[3] Note also that the two welfare contours

---

[2]We do not go through the formal mathematics here, which would be similar to those in Chapter 18. The difference is that in achieving an expression like (18.26) for the change in home's welfare from altering its tariff, we must account for the effect of different levels of the foreign tariff on export supply and world price. There is a similar expression for the change in foreign's welfare, holding constant home's tax. The best-response functions are calculated by setting these welfare changes equal to zero.

[3]Strictly speaking, this is not a necessary outcome. Depending on various parameters it is possible for one country to gain and the other to lose welfare in Nash equilibrium relative to free trade.

intersect at point N, which is, therefore, not Pareto-efficient. The fact that both countries are worse off at N than in free trade implies that unilateral tariff-setting can easily generate a *Prisoner's Dilemma* outcome: while zero tariffs are mutually beneficial, each country on its own would choose higher taxes, greatly diminishing trade and efficiency.

Clearly, both countries may gain by jointly agreeing to reduce tariffs. For example, any negotiated agreement to lower trade taxes such that both countries move to the southeast, inside the lens marked by the welfare contours through point N, would mutually raise welfare. To maximize global (two-country) welfare, negotiations should achieve a point on the contract curve CC, along which the iso-welfare contours for both countries are tangent to each other.[4] In particular, the outcome at reciprocal free trade (point T) is both globally welfare-improving (compared to N) and Pareto-efficient. We depict this fact by showing tangency at point T, which arises because home and foreign share the same (free-trade) price ratio.

To summarize, the existence of this *terms-of-trade externality*, under which one large country attempts to gain welfare through a tariff that reduces its partner's welfare, leading to higher tariffs abroad in response, offers a sound reason for negotiating agreements to reduce taxes on trade. Doing so binds each country so that it may not unilaterally deviate from its commitments by raising tariffs. This is the primary analytical argument for reciprocal trade-liberalization agreements, such as the World Trade Organization, described in the next section.

Before leaving this section, however, we might ask what other reasons might induce governments choose unilaterally to impose high trade taxes, generally to the detriment of their economies. For example, the standard terms-of-trade model offers no explanation for why small countries, which cannot affect world prices, often have high trade barriers.

Rather than work through the fundamental theories we simply describe these arguments, which seek to explain the endogenous formation of trade policy, rather than treating tariffs as exogenous variables. One possibility is that governments are not focused solely on maximizing national welfare as assumed by the previous analysis. Rather, they may also pursue *political-economic* objectives that lead them to favor high tariffs (or export subsidies) in some sectors over others, regardless of impacts on world prices. One primary reason is that industry interests lobby for special protection to preserve economic rents, as noted in Chapter 9. In the standard *protection-for-sale* model the levels of trade taxes across sectors depend on a variety of factors, such as industry size and the ability to organize firms and workers to lobby.[5] Another is that citizens may care about the effects of international trade on the distribution of

---

However, the losses to one nation dominate the gains to the other and the world is worse off. It is also possible for these nonlinear best-response functions to intersect more than once, suggesting multiple Nash equilibria are possible. For details, see Johnson (1953-1954), Kennan and Riezman (1988) and Bagwell and Staiger (2002).

[4]CC is not necessarily linear. Mayer (1981) showed that tariff rates along this locus must satisfy the relationship $(1 + t^h) = 1/(1 + t^f)$, which ensures that relative prices are the same in both countries, absent any other price distortions.

[5]The basic logic of this argument for tariffs may be traced back to Olson (1965) and was statistically tested by Baldwin (1985) among many others. Grossman and Helpman (2002) offer a theoretical analysis.

incomes across factors (types of labor, land and capital) and geographic regions of the country. Thus, in the classic *median-voter* model, policymakers will select trade taxes that appeal to the majority of voters.[6] At the most basic level we might anticipate that the median voter would prefer free trade since it should maximize national income. However, it is quite possible that she would favor protection if there are multiple ballot issues to consider and the intensity of preferences varies across voters and issues, or if the distributional outcomes of trade policy are uncertain.

While political-economy reasons for trade policy are certainly important in practice, they do not fundamentally change the logic for negotiating trade agreements, as stressed by Bagwell and Staiger (2002). Politically motivated tariffs generate inefficiencies in production and consumption in the importing country, tending to reduce welfare even if part of those costs are shifted onto foreign trade partners through lower export prices. Put differently, it is possible to imagine Nash-equilibrium tariffs that are generated by political incentives but these would also be inefficient in comparison with free trade. In turn, mutually agreed trade liberalization can raise national and global welfare, with the same logic as above.

Political-economic pressures do raise at least two interesting additional policy complications, however. First, because trade liberalization exposes import-competing industries to competition it may be difficult simply to move quickly and sustainably from the protected equilibrium to free trade. Rather, trade agreements generally phase in their tariff cuts gradually to give firms and workers an opportunity to anticipate the needs to adjust.[7] Moreover, as we will describe in the next section, agreements generally offer various safety valves that may permit temporary reimposition of trade protection under certain circumstances.

The second complication is that there will always be political pressures in each nation to restore tariffs or add new trade barriers once they are liberalized. Future governments may give in to these pressures, which would raise inefficiency costs and generate uncertainty for investors. In this context, trade liberalization is subject to what economists call *time-inconsistency*: while it may be beneficial to lower tariffs and eliminate quotas now, there will inevitably be pressures to raise them again in the future. Thus, governments may prefer to enter trade agreements, such as the WTO and NAFTA, which make it quite costly to raise trade restrictions in the future and thereby raise the credibility of their actions. This analysis points to the fact that major trade agreements serve as a strong *commitment mechanism* that overcomes the time-inconsistency problem. For example, in negotiating NAFTA Mexican trade officials often pointed to the credibility and certainty that the agreement would generate about sustaining open markets, thereby attracting foreign direct investment.

21.3 The World Trade Organization

In contrast to a preferential agreement, in which a small number of countries grant free trade to each other and the pact is not open for further members to join, a multilateral accord attempts to cover a large number of countries – in the extreme the entire world – and is open to new accessions by nations

---

[6]An analytical treatment is in Magee, Brock and Young (1989).

[7]Furusawa and Lai (1999) show in a dynamic model that if workers face adjustment costs then negotiated tariff cuts generally need to be implemented gradually for countries to keep them in place. See also Chisik (2003) for a similar argument where there are costly adjustments to capital stocks.

that agree to meet its terms. The World Trade Organization (WTO), founded in 1995 with 123 member countries, is the premier example of a multilateral trade agreement. As of early 2011 there were 152 members, which accounted for the vast majority of international trade flows.[8] The European Union also belongs as a single institution, along with its individual members. Some of the more notable new participants include China, Saudi Arabia, Vietnam and nearly all countries of the former Soviet Bloc. An additional 30 nations have observer status, which means they were actively negotiating the terms of their accession. Included in the latter group was the Russian Federation, the largest nation that had not yet joined. Clearly, the WTO has attracted nearly universal participation by the world's economies.

From GATT to the WTO

The WTO is a successor to an accord called the General Agreement on Tariffs and Trade (GATT). The GATT was founded in the years following World War II, along with other international economic institutions, including the International Monetary Fund, the World Bank, and various United Nations agencies. In particular, the GATT was an agreement among 23 countries, both developed and developing, to negotiate over time reciprocal reductions in tariffs. It also contained a series of articles setting out general principles and obligations that countries had to meet in fashioning their own trade policy.

The GATT had considerable success in its objective of reducing trade barriers among its contracting parties. Before it gave way to the WTO, there were eight "GATT Rounds" of multilateral trade negotiations (MTN). The first five were devoted to reciprocal tariff cuts among a relatively small number of countries. The Kennedy Round, from 1964-67, was the first to attract fairly wide participation and it resulted in significant tariff cuts, by about 1/3 from their postwar rates, among the industrialized nations. Large numbers of developing countries agreed to join the GATT after the Kennedy Round because the agreement was expanded to recognize their particular needs. Thus, developed countries agreed not to demand much reciprocity from developing countries and began to offer preferential market access to their products. These provisions introduced the notion of *special and differential treatment* for poor nations into the GATT.

The seventh MTN was the Tokyo Round, from 1974-79. The Tokyo Round further slashed tariffs on manufactured goods, again by about 1/3 for the rich nations. Further, for the first time numerous developing countries agreed to significant trade liberalization through the GATT. The most notable aspect of the Tokyo Round was to bring into the GATT rules covering a range of non-tariff barriers (NTBs). Several specific agreements were negotiated among subsets of countries on policies covering government subsidies, safeguards, antidumping tariffs, public procurement policies, customs procedures and technical standards for protecting health and safety. The Tokyo Round attracted further participation by developing countries in cutting tariffs.

The final, and most ambitious, GATT negotiation was the Uruguay Round, from 1986 to 1994. While a major focus of this Round continued to be reducing tariffs and establishing disciplines on the use of NTBs, it was aimed largely at resolving major perceived problems in the structural architecture of the GATT, as discussed in the next subsection. Its primary outcome was to launch the WTO, which, as an organization rather than agreement, has considerably stronger mechanisms for settling disputes among member countries than did the GATT, on which it was founded. The agreement began to extend liberalization commitments and policy rules to new and controversial areas, including services, foreign

[8]For details see World Trade Organization, *Annual Report, 2011*, at www.wto.org.

direct investment, and intellectual property rights. There was also a considerable emphasis on bringing developing countries fully into both the negotiations and in meeting the full range of WTO obligations, including tariff cuts.

The WTO came into effect on January 1, 1995. There has been just one MTN since that time, the Doha Development Round, which began in 2001 and was still ongoing in the middle of 2011. As its name suggests, this Round was supposed to make global trade policy more integrated with the needs of economic development. However, the sheer complexities of reaching agreements across 153 disparate countries to cut tariffs, open service markets and reduce differences in commercial regulations stalled the Doha Round.

Membership in the GATT and WTO has reduced tariffs sharply over time, as shown in Table 21.1. The first column lists estimates of the average percentage tariff rates in a selection of countries, weighted by the share of goods imported into those countries in 22 industries in 1987. Thus, these data measure the effective height of import taxes after the large reductions achieved (at least for the developed economies) in the Kennedy and Tokyo Rounds. These weighted-average tariffs were very low in the United States and Switzerland and somewhat higher in the 12 countries then in the European Union. At that time Australia still had significantly high tariffs. Among the emerging economies, Hong Kong and Singapore have had very low tariffs for a long time, while Chinese Taipei (Taiwan) and South Korea had average rates over ten percent.[9] Import duties were generally much higher in the developing countries, even after the Tokyo Round.

Table 21.1

As may be seen in the next column, import-weighted average tariffs had declined considerably in most countries by 2008, due to the Uruguay Round and other cuts.[10] Australia and Japan saw significant reductions among developed countries, while the EU ended up with much lower average tariffs despite expanding from 12 to 27 nations. The most dramatic declines came in such developing countries as Brazil, India and Turkey. For its part, China's estimated average tariff rate in 1994 was 45 percent. China joined the WTO in 2001 and immediately embarked on major trade-liberalization policy, with weighted-average tariffs now just over four percent. Clearly, world trade is much less burdened by trade taxes currently than it was in the 1980s and earlier.

The remaining columns show interesting features of current tariffs. First, there is a difference between the tariff rates countries actually charge and what they negotiate at the WTO. Thus, the third column shows the "most favored nation" (MFN) simple-average applied tariff rates.[11] These are the taxes charged on imports from all WTO members that do not receive some kind of preferred lower rates, such as those in regional free trade agreements. Thus, for example, the average U.S. tax is 3.5 percent, but

[9]In 1987 Hong Kong was an independent protectorate of the United Kingdom and is now formally part of China. But Hong Kong retains its own authority to set trade and immigration policy.

[10]These columns are not strictly comparable because the weights in 2008 are based on more disaggregated product trade than the 22 industries in 1987.

[11]Note that these rates averaged equally across all goods classes are generally higher than the rates that use import weights, since imports are often higher in products with lower tariffs.

Canada and Mexico actually face lower rates due to their membership in NAFTA.

The next column shows the average "bound" tariffs, which are the maximum rates agreed to at the WTO. Countries are free to charge MFN taxes below these bindings but cannot exceed them except in rare circumstances. In the developed world bound tariffs are generally not much higher than the applied rates, meaning that these countries have committed themselves not to raising import charges. South Korea's average bound tariff is considerably higher than its applied rate, however, indicating that it has reserved some ability to increase taxes. Argentina, Brazil, India, Mexico and Turkey all have applied rates around 10 percent on average but their negotiated bindings are considerably higher. China is remarkable for having both a relatively low average tax rate of 9.6 percent and an average bound rate that is not much larger, at 10.0 percent. In it 2001 accession agreement, China committed to its new WTO partners that it would sustain low bound rates.

A final interesting observation is that import tariffs are generally considerably higher in agricultural goods than in manufacturing and other non-agricultural goods. For example, in the EU the average tariff in agriculture is 13.5 percent in comparison with 4.0 percent otherwise. Current agricultural tariffs average 21.0 percent in Japan, 48.6 percent in Korea, and 31.8 percent in India. Clearly, most countries of the world prefer to protect domestic farmers from import competition through high import barriers, a clear outcome of political-economic considerations. Indeed, an inability to find agreement among WTO members on cutting tariffs in agriculture has been a primary reason for the failure of the Doha Round negotiations.

Basic Principles and Functions of the WTO

It is one thing to describe the economic logic for cutting taxes on trade on a global basis. It is another to develop a set of principles and an organizational architecture for actually achieving and sustaining this goal. We now describe briefly the fundamental principles of the WTO (and the GATT before it) and its primary features.[12]

The title "World Trade Organization" refers to two interrelated processes. First is a set of negotiated agreements to liberalize international trade and establish obligations preventing member governments from violating their commitments to open markets. These agreements are based on fundamental principles and mutually agreed rules of behavior. Indeed, the WTO is often called a rules-based approach to multilateral trade cooperation. Second is an organizational secretariat, located in Geneva, Switzerland, that facilitates enforcement of these obligations and supports ongoing multilateral trade negotiations among members. It is important to note that the organization's policies and procedures are determined by mutual agreement among countries, not by the WTO's staff.

Nondiscrimination

The first guiding principle is nondiscrimination, which embodies two obligations. First is *national treatment* (NT), by which each nation must treat imports, once they have cleared the border, at least equal treatment to that given domestic competing goods. This principle extends to FDI in that domestic affiliates of multinational firms cannot be treated less favorably than competing domestic production. Thus, domestic taxes and regulations cannot be biased against foreign goods. The second

[12]Students interested in a detailed discussion of these issues should read Hoekman and Kostecki (2009).

form of nondiscrimination is that each WTO member must unconditionally give to the products of all other members the most favorable treatment it offers to products from any one country. The *most favored nation* (MFN) requirement means that imports from all WTO members are charged the lowest tariffs. It extends also to NTBs and regulations, implying that a favorable tax offered to firms from one nation must be offered to firms from all others. In essence, MFN ensures that all international competitors are treated equally at the border and NT guarantees that there is no favoritism shown domestic firms inside the border.

The MFN principle carries strong economic advantages over a policy of bilateral discrimination. It permits comparative advantage to be the primary determinant of trade patterns, raising global efficiency. It makes negotiations on reducing trade restrictions easier because it requires that each country set just a single schedule of MFN import taxes rather than a series of bilateral tariff rates. MFN provides an important protection for small, developing countries, which individually have little or no bargaining leverage. Under MFN these countries are immediately given the most favorable access to markets in large nations. Without this principle, countries would be free to negotiate bilateral tariff agreements and exercise their market power, resulting in less liberalized and more fragmented world trade.

There are four major exceptions to the MFN rule. The most prominent is that the WTO permits groups of countries to form preferential trading arrangements, such as free trade agreements and customs unions, which we analyze in the next chapter. The primary requirement is that such agreements do not impose tariffs and regulations on the goods from excluded countries that are higher or more restrictive than those already existing. This essentially means that, even though members of FTAs offer each other zero tariffs, the overall external level of trade protection is not raised. This restriction has rarely been invoked by the GATT or WTO to deny the formation of a new trade agreement.[13]

A second exception is that developed countries are encouraged to provide favorable tariff treatment of goods from the developing countries by adopting a *Generalized System of Preferences* (GSP). As poor countries achieve higher income levels and increase their trade volumes they are often graduated out of the GSP and then face MFN tariffs. The third departure from MFN is that groups of like-minded countries have negotiated certain side agreements, commonly called Codes, which obligate only the Code signatories to treat each other without discrimination. For example, there is a Code on Government Procurement, applying mostly to developed economies, which establishes rules for governments to accept bids from domestic and foreign firms for various public contracts. These plurilateral agreements are based on "conditional MFN," meaning that only signatories are entitled to nondiscriminatory treatment. Finally, as discussed later in this section, countries may temporarily raise tariffs against imports from certain other nations that are found to be unfairly dumping or subsidizing exports or are not meeting their policy obligations.

Reciprocity

The second fundamental principle of the WTO is reciprocity. In particular, multilateral trade negotiations are reciprocal; countries agree to lower their tariffs or relax NTBs in return for other

[13]All these provisions mean that each WTO member actually has up to five tariff schedules, which list tariff rates for each product. First is the MFN bound schedule, then the MFN applied schedule. After that come any tariffs agreed to in FTAs and the GSP rates. Finally, there are the tariffs applied to goods from the few countries that are not in the WTO and are not awarded MFN treatment.

countries doing the same. Purely in economic terms we might question the need for reciprocity, since nations generally gain welfare by liberalizing trade. However, as analyzed above, it may be necessary for large countries to agree jointly to reduce barriers in order to move away from the Nash equilibrium. Reciprocity is also necessary to induce concessions by all participants. Without that expectation, some countries may choose not to liberalize trade while enjoying the export gains from their MFN treatment as others reduce tariffs. Through the Tokyo Round, many developing countries were seen by some observers as "free riders" in this sense and a large emphasis in the Uruguay Round was gaining full participation. Similarly, when a new country joins the WTO it will immediately gain MFN status. Existing members therefore require, through negotiations, that a new entrant engage in considerable market liberalization as part of its "entry fee".

Most fundamentally, reciprocity is necessary to achieve domestic political support for opening markets to trade. That policy will be opposed by import-competing industries and scarce factors, both of which must adjust to greater competition. However, if foreign countries are also cutting tariffs there will be greater opportunities for export expansion and export-oriented industries will lobby for trade agreements, offsetting the status-quo political pressures. Note that for this argument to be effective, exporting firms must expect significant gains from multilateral tariff cutting. One of the interesting current features of global trade is that the developed economies, and many of the largest developing economies, now have generally low tariffs, except in agriculture. This leaves WTO members little room to bargain unless they turn to difficult and controversial issues of liberalizing service markets and changing commercial regulations. This is another major reason for the continuing failure of the Doha Round.

Enforceability

The third basic WTO principle is that commitments and obligations must be enforceable. Trade agreements would have little value or staying power if countries could easily go back on their tariff cuts or violate essential rules. For example, we noted above that member countries actually negotiate *tariff bindings*, or ceilings above which they commit not to raise import duties. It is possible for a country to exceed these bindings, however, if it agrees to offer compensation, generally in the form of reductions in other tariffs, to exporting countries that are the principal suppliers of the goods in question. These negotiated cuts then must be offered to other members under the MFN principle, which raises the political costs of choosing to violate the bindings.

The more familiar enforcement element is the *dispute settlement* process of the WTO. Keep in mind that the WTO agreements embody far more than just tariff bindings. There are basic obligations (eg, MFN and NT), limits on government behavior (eg, bans on export subsidies and rules governing antidumping duties), policy standards (eg, intellectual property protection and public-health regulations applied to imports) and other kinds of market-access commitments (eg, establishment rights in services). If one member country thinks that another has violated any of these commitments, thereby nullifying or impairing its trade benefits, it can essentially file a lawsuit at the WTO. The two countries first must engage in consultation, which may resolve the conflict. If that process fails, a WTO dispute-settlement panel is established to hear the case. These disputes are highly complex and both sides present legal arguments about whether the offending behavior is, in fact, a violation of some commitment and results in economic damages to the complainant nation. After hearing these arguments, the panel issues a ruling, which may be appealed. If the ruling finds that the defendant nation has violated its commitments, the panel can order that country to modify or stop the offense. If the defendant finds, usually for political reasons, that it cannot end the violation it must offer some form of compensation or accept retaliation in

the form of higher tariffs in the complainant.[14]

One of the major reasons the GATT members established the WTO was to beef up this dispute settlement process. Under the GATT it was relatively ineffective at changing behavior for a number of reasons. For example, a country that "lost" a panel ruling could simply block it from being released and did not really have to comply with recommended policy modifications. That is, the GATT had no authority to require a change in policy, since the agreement relied solely on voluntary compliance. Under the WTO, in contrast, countries cannot block publication of a panel ruling and must take steps to modify their policies or pay compensation. This enforcement authority, rather rare among international organizations, is a primary feature of the WTO. In particular, it gives developing countries more leverage to petition for removal of trade barriers in the industrialized world.

Relief mechanisms

A final major principle of the WTO is that countries must have some safety valves to manage pressures that are inevitable in the system. For example, under WTO rules countries may use trade measures under certain circumstances to protect public health, deal with monopolies, and conserve exhaustible natural resources, among other "non-economic" objectives. Such trade measures cannot be discriminatory or disguised restrictions on international trade. In order to protect national security, countries may also adopt trade restrictions, which are not subject to these limitations. The linkages between trade policy and such non-trade objectives are complex and difficult to make, but critical in the globalized economy. We discuss and analyze this question further in Section 21.5.

A second type of safety valve permits WTO members to deploy import quotas if they encounter significant problems in financing balance-of-payments deficits or raise tariffs beyond bound levels to support the development of an infant industry. These provisions were established in the original GATT in 1947 when there was a considerable concern among countries about their ability to sustain their monetary reserves in an era of immobile capital and a strongly perceived need to use trade policy to foster industrial development. In practice such trade restrictions were used largely by developing countries in the late decades of the 20th century. These provisions are rarely used in today's era of rapid private capital flows and general doubts about the ability of tariffs to nurture competitive industries.

The third kind of pressure release is a set of actions countries may take to offset the impacts of rapid import surges or unfair competitive conditions. Regarding import surges, WTO rules permit a member to protect a domestic industry when it is threatened with serious injury from a rapid rise in imports. This is typically done by nondiscriminatory increases in tariffs on specific products beyond their bound rates. These actions, referred to as *safeguards*, have two important features. First, the rules state that such protection is to be temporary, lasting only as long as needed to remedy the injury. For example, safeguards tariffs used in the United States have often have been subject to a declining schedule over time. Second, because such actions violate the its commitment not to raise tariffs, the importing country is required to pay some compensation to foreign nations that stand to lose business. This may happen either through an agreement to lower tariffs on other industries or by allowing foreign governments to retaliate with higher tariffs of their own.

---

[14]An excellent review of the dispute settlement process is in Bown (2009). Interested readers may consult all of the panel reports at the WTO website: http://www.wto.org/english/tratop_e/dispu_e/dispu_e.htm.

These safeguards procedures try to achieve a balance between two important goals. On the one hand, countries may need at times to protect domestic industries and their workers from rapid and unanticipated increases in foreign competition. If there were no safety valve of this kind it would be much harder to convince domestic interests to support membership in the WTO. Thus, safeguards help meet the political objective of overcoming opposition to global trade agreements. On the other, the system needs to make it costly for individual governments to depart from their trade obligations. Otherwise, foreign trading partners would face uncertainty about their firms' access to export markets, thereby reducing their willingness to offer binding reciprocal commitments. Safeguards are an important building block of the WTO system.

These principles can be illustrated by the most recent instance of US safeguards policy. In 2002, the Bush Administration imposed import duties of between eight and thirty percent on imported steel products, after a finding by the US International Trade Commission of a detrimental import surge. These tariffs were scheduled to be phased out in 2005, with the temporary protection affording some breathing room to domestic steel producers. Immediately, several countries asked for consultation over removing the tariffs, which the US government refused to do. Subsequently, the European Union, Japan, South Korea, China and other countries filed complaints at the WTO, which issued a panel ruling in late 2003, showing that there was no evidence of an import surge. The panel directed the US Government to eliminate the tariffs and authorized foreign retaliation that could raise tariff on up to $2 billion in US exports, the largest such order ever issued. Initially the President vowed to keep the duties in place. Only after the EU threatened retaliation against important American export products did the United States agree to eliminate them, which it did in December 2003.

Safeguards permit protection against unanticipated import surges that result from ordinary market activity, or "fair competition". Largely because of the compensation and retaliation features, such tariffs are rarely imposed. Another key reason is that it is easier to raise duties against imports considered to be "unfair competition". There are two categories of unfairly traded imports. First are goods that are "dumped," or sold in a market by a foreign firm at an unfairly low price. The importing government can respond to these prices with *antidumping duties* (ADD). Second are products that have been given export or production subsidies by a foreign government, leading to low-cost imports. The subsidies may be offset with *countervailing duties* (CVD). The WTO agreements set out guidelines for national policies covering ADD and CVD. In essence, trade authorities in the importing country must demonstrate that dumping or export subsidization is taking place in a particular product and causing material injury for domestic producers making similar products. Note that "material injury" is a considerably weaker legal standard for showing damages than is "serious injury" under safeguards.

Evidence of dumping exists if the import price (not including any existing tariff) is less than the "fair value" of the good, which may be either the home price or the average production and shipping cost. In either case, if both dumping and injury are found to have occurred, an ADD is imposed to equal the *dumping margin*, or the difference between the import price and the fair value. In principle, it completely offsets the unfair pricing and thereby protects domestic competing firms. from The tariff is imposed until the dumping ceases or exporters agree to set a higher price.

This process seems unobjectionable on its face, if the dumping really is "unfair', though some economists wonder why importing countries would require foreign firms to raise their prices if they are willing to sell cheaply to domestic users. Greater concern arises over the specific legal procedures under which antidumping cases are prosecuted. For example, US trade authorities primarily use a comprehensive "constructed costs" to determine an exporter's average costs, a standard that tends to bias

fair value computations upward from underlying figures and readily results in findings of large margins. It is also important that ADD actions are taken against private firms rather than foreign governments, giving domestic enterprises an opportunity to ask for tailored protection. Analysts have convincingly demonstrated that in both the United States and the European Union domestic industries often use antidumping laws in a strategic way to protect markets and transfer profits from foreign competitors. These policies also tend to distort trade patterns from third countries.[15]

Countervailing duties, which are tariffs raised to protect domestic firms from the competitive effects of foreign export subsidies, operate in a similar fashion. Trade authorities in the importing nation must demonstrate that the subsidy induces one or more exporting firms abroad to charge a price below fair value and that domestic firms suffer material injury. In such cases a CVD equaling the subsidy margin is imposed on the imports of the subsidized firms until the subsidy is removed.

The WTO provisions on antidumping and countervailing duties establish rules within which importing countries must operate in order to avoid their use as unwarranted barriers to fairly traded imports or means of harassing foreign exporters. In addition, WTO members recognize that subsidizing production and exports can be damaging to world trade. At the same time, as we have discussed earlier in this text, there are circumstances in which subsidies to output, consumption, employment or other variables can be beneficial policies to attack market failures or encourage economic development. Drawing the line between beneficial subsidies and policies that unfairly distort trade is difficult. The WTO agreements embody a *subsidies code*, which tries to find this balance. In essence, the code bans direct export subsidies and fiscal production supports that encourage exports. So-called "pre-competitive subsidies", which are given to R&D, wage support, training programs, and the like, are generally permissible.

Policies regarding antidumping and export subsidies raise interesting analytical questions, which we address in Section 21.4.

Structure of the WTO Agreements

As the prior discussion suggests, the WTO encompasses far more than just commitments to reduce tariffs. We finish this section with a brief overview of the major agreements that make up the set of rules and provisions to which WTO members are supposed to adhere.[16]

In 1994 the Uruguay Round negotiations resulted in an umbrella agreement establishing the WTO itself. Attached to that are six major components, called Annexes, which make up the set of WTO accords. First, Annex 1A covers many complex agreements on trade in goods. These include, among others, the antidumping framework, the agreement on subsidies and countervailing duties, the safeguards provisions and rules governing how customs procedures operate. There are also agreements covering how governments may limit trade based on sanitary measures and technical product standards, plus special commitments regarding trade policy in agriculture and clothing.

Annexes 1B and 1C are important because they introduced new areas of competition into the

[15]Blonigen and Prusa (2003) describe a number of questionable practices in antidumping law and their effects on trade.

[16]Interested students can read the WTO text online at http://www.wto.org.

WTO. Specifically, 1B is the General Agreement on Trade in Services, or GATS, which offers a framework for willing countries to negotiate liberalized market access in designated service sectors. Annex 1C is the Agreement on Trade-Related Aspects of Intellectual Property Rights, or TRIPS, which sets out minimum protection standards that countries must observe in patents, copyrights and trademarks. The TRIPS Agreement has proved highly controversial since it is the first multilateral accord that requires nations to reform an important set of domestic regulations to accommodate international trade. For example, it requires developing countries to offer stronger patents for new pharmaceutical products, which are overwhelmingly innovated in the developed world. As a result, patients and public-health ministries poor countries may end up paying higher prices for imported medicines, while there may be little benefit in terms of growth in domestic drug innovation.[17]

Annex 2 is the Dispute Settlement Understanding, or DSU. It describes the procedures for consultation, panel investigations, appellate procedures and the scope for compensation and retaliation among members. Again, the DSU is a major form of relief valve for WTO members. It also plays the vital role of establishing a body of WTO precedents and law over time as more panel reports are issued. These reports are essential for interpreting the legal reach of the various Annexes and the obligations countries must meet in their trade policies.

Annex 3 establishes a Trade Policy Review Mechanism (TPRM), which calls for the WTO staff to issue periodic reports about the policies pursued by member countries. These reports are a vital part of ensuring policy transparency and adherence to WTO rules. Finally, Annex 4 contains four plurilateral agreements among subsets of countries, such as the Government Procurement Code mentioned earlier.

21.4 The Theory of Contingent Protection

We return here to an analytical treatment of ADD and CVD. Because these two types of tariffs are imposed in response to particular forms of trade behavior, they are jointly referred to as *contingent protection*.

An interesting analytical question is why firms engage in dumping. Again, there are two legal definitions of dumping. The first entails a firm selling its product in an export market at a price below the price of the same good in its home market. At a surface level this behavior seems odd since we would expect foreign sales to be priced sufficiently to cover the additional transport costs. However, examples of such pricing are common. The second definition pertains where a product is sold in an export market at a price below the average cost of producing it. Importing countries generally equate average cost with the "fair value" of the product and dumping constitutes sales at *less than fair value* (LTFV). Selling at a price below average cost also seems unprofitable on its face, so why might firms do this?

One possibility is that firms could experience an unexpected reduction in demand in their home market and, rather than cut production, choose to export their surplus output to foreign markets at a discount. This strategy could emerge in countries experiencing an economic recession, for example, which is why it is sometimes called *cyclical dumping*. If importers experience lower output and employment as a result, countering it with temporary AD duties is sensible, both economically and politically. However, available evidence indicates that such dumping is rare. For example, it seems there

[17]Maskus (2000) explains the global economics of intellectual property rights in detail.

was not much increase in antidumping (or safeguards) complaints or actions taken in the wake of the global recession of 2008-10 (Bown, 2011). It remains unclear why this was the case. One possibility is that the high costs of establishing distribution systems and marketing in foreign countries discourages cyclical dumping. Another is that multinational firms are now both significant exporters and importers across multiple stages of production, suggesting that they would oppose widespread increases in antidumping tariffs. Or perhaps adherence to the WTO itself diminishes the likelihood of raising tariffs even in a global downturn.

A second possibility is that a foreign firm may engage in *predatory dumping*, which means selling its goods in an export market at a price low enough to drive competing firms out of business or deter entry by other firms. This would require setting a price so low that competitors could not cover variable costs and would ultimately leave the market. Such fears offer the greatest political motivation for ADD, as domestic firms use them to lobby for protectionist antidumping laws. If predatory behavior exists it harms competition and importing governments are justified in levying an offsetting tariff. Again, however, there is little evidence of predatory dumping, largely because of the costs of sustaining it. Even if a foreign firm succeeded in establishing a monopoly it would have to continue charging a sufficiently low price to deter future entry. Moreover, predation is illegal under the antitrust laws of most countries and firms would likely be prosecuted if they attempted it.

This leaves the third, and primary, reason for dumping, which can be explained by the theory of international trade under imperfect competition. Suppose that firms produce differentiated goods, so each has some monopoly power in each country in which it sells. Recall from chapter 12 that a monopolist would maximize profits by setting a price in each market *j* (including her own home market) as a markup over marginal costs:

$$p_j = mc_j / (1 - 1/\sigma_j) \tag{21.1}$$

In this expression, marginal cost includes the cost of transporting an additional good to market *j*. If these transport costs are similar to each country, the firm would clearly set different prices across markets. In particular, the less elastic is the demand for the good in a specific location (that is, the lower is σ_j), the higher will be the price because of a higher markup over marginal cost. This practice is called *international price discrimination* and is quite likely to occur in many differentiated products and markets.[18]

Note for this outcome to hold there must be two further market features beyond imperfect competition. The first is that countries have preferences that generate different demand elasticities. While there are many reasons for such differences, two primary factors were mentioned in Chapter 14. Countries generally exhibit *home bias* in consumption, which would support more inelastic demand, and higher prices, for domestic goods. Further, the price sensitivity of consumers for certain goods may depend on incomes. Survey evidence finds, for example, that people in high per-capita-income economies are more willing to pay for prescription medicines, luxury automobiles, fashionable clothing

[18]Students who have studied microeconomic theory may recognize this outcome as "second degree" price discrimination, since the firm sets one price for groups of consumers (here, one price per country), rather than a different price for each individual consumer.

and digital products, permitting firms to charge them higher prices.[19] The second critical feature is that different markets must be partially or fully segmented (that is, isolated from one another) so that consumers or distributors in one country will not buy the good at a low price and resell it in other markets at higher prices. This kind of arbitrage, called *parallel trade* or *gray-market trade*, would tend to eliminate price differences. Where it is legal, specialized distributors often engage in parallel trade to profit from international price differences. However, in many countries this trade is illegal, which is the most important reason that markets for differentiated goods are segmented.[20]

How do these features generate dumping incentives for firms? First, if the home market has higher per-capita income than the export market, demand could be more elastic in the latter, inducing a lower price there. Next, home bias suggests that prices are likely to be higher in home markets than abroad, itself a feature of dumping. Interestingly, in a two-country world if home bias existed in both locations, then imperfectly competitive firms in each, possibly producing quite similar products, would sell in the other at a lower price. Because this activity results from profit-maximizing responses by monopolistically competitive firms facing international differences in demand, it might be called *equilibrium dumping*. It generally would not call for an ADD response since there is no distortion to offset with the tariff.

Interestingly, an extreme version of this possibility can exist even without home bias and product differentiation. Suppose, as in Chapter 11, two oligopolistic firms produce a homogeneous good but they perceive the home and foreign markets to be segmented, rather than fully integrated. In this case, with Cournot competition, each firm will actually export its good to the other market in order to take advantage of the demand there. This phenomenon is called *reciprocal dumping* and involves an interesting welfare tradeoff: the pro-competitive effect reduces prices and benefits consumers in both markets but any resources used in "cross-hauling" homogeneous goods are a pure economic waste (Brander and Krugman, 1983). Recalling our earlier analysis of strategic trade policy, the importing government might intervene with an ADD to capture for its treasury some of the profits made in its market by the foreign firm (Helpman and Krugman, 1989).

While doubts exist about the wisdom of current ADD policies, there is a strong analytical justification for using countervailing duties against foreign export (and output) subsidies, which are generally harmful in welfare terms. This may be seen in Figure 21.3, which shows a free-trade equilibrium between home and foreign at point A with trade quantity of OF and price at point p^{*F}, assuming competitive markets. Now suppose foreign pays an ad valorem subsidy to exports of good x_1, which shifts its export-supply curve to the right, generating an equilibrium at point Z, with a higher quantity of trade OS. The effect is a worsening of foreign's terms of trade: world (and home's) price falls to $p^{*s.}$ Further, foreign's domestic relative price rises to p^{fs} as the export subsidy generates scarcity in that market and drives up the producer and consumer prices of the good. The resulting distortions in economic decisions, along with the worsened terms of trade, imply that foreign suffers a decline in economic welfare. Note that because the subsidy must be paid on the full amount of exports, the fiscal

---

[19]The logic in Chapter 14 was that income elasticities of demand for such goods rise with wealth. The idea here extends that to the possibility that preferences of higher-income consumers (countries) also are less price-elastic.

[20]The United States, for example, prohibits parallel imports of products that are protected by domestic patents and copyrights. The EU also prohibits such imports from outside its region but they are permitted among the member countries. See Ganslandt and Maskus (2008) for details.

cost is area $p^{fs}p^{*s}ZV$ (measured in units of good x_2). In short, foreign's subsidy expands trade artificially at high cost.

Figure 21.3

The home country, however, actually gains from this subsidization because of the fall in the relative price of its import good. This gain in home's terms of trade outweighs the distortionary effects of the lower price. Thus, this policy amounts to a decision by the exporting government to transfer income to the importing nation. Overall, global welfare declines because of the departure from free trade.

Suppose that home is authorized to retaliate against this subsidy with a CVD calculated to restore the initial volume of trade. Home may choose to retaliate, despite its welfare gain from foreign's subsidy, in order to protect its competing firms from the reduced price. A tariff sufficient to shift home's import-demand curve down to M^t_{h1} would achieve the free-trade volume of OF but would worsen foreign's terms of trade even more by driving down the price it receives to p^{*st}. Note that because outputs and prices have returned to their pre-subsidy levels, there is no remaining distortionary impact of the export subsidy: it has been completely offset by the CVD. The only impact is a transfer of income from taxpayers in foreign to the government in home (the subsidy is paid to exporters who then surrender its proceeds in the form of tariff revenues in the amount $p^{*F}p^{*st}BA$). Presumably, if taxpayers in foreign understood this fact they would lobby for an end to the export subsidy or oppose it in the first place. Thus, the essential economic justification for CVDs is that they can restore optimality and deter the introduction of subsidies.

21.5 Global Trade Policy and Market Externalities

As we noted in the introduction, individual countries choosing their own trade policies raises a form of cross-border externality through a terms-of-trade impact. Equally, national regulatory policies, even though aimed primarily at addressing domestic market failures, generate related international problems. For example, if one country prefers to regulate its greenhouse gas emissions relatively lightly, there is a primary international externality in the form of diminished air quality in neighboring countries or faster global warming. There is also the secondary problem that limited regulation can reduce production costs and give domestic firms a competitive advantage over foreign firms. The secondary problem is manifested in altered trade patterns, with companies in the country with weaker regulations exporting more (and facing less import competition) than would be the case if they were required to pay the environmental costs of these emissions. Similar arguments pertain to differences in workplace standards protecting labor rights, the rigor of sanitary and technical regulations governing the production processes of goods, the strength of patent and copyright enforcement, the degree of antitrust competition, and a host of other critical regulatory issues. Each of these involve "behind the border" regulations with potentially important impacts on international trade and investment.

There are deep and intricate relationships between trade policy and the regulation of market failures with international cost spillovers. These complexities raise several analytical challenges for economists. First, while basic trade theory points to global efficiency and welfare gains from mutually free trade, this is not necessarily the case in other areas of regulation. For example, there is no meaningful definition of "free trade" in intellectual property rights, such as patents. Rather, those countries that are net exporters of new technologies prefer a global regime of strongly enforced and long-lasting patents to protect the ability of their enterprises to sell high-technology goods and ideas without

fear of losing profits to cheaper imitated versions. Net importers would prefer a regime of weaker and shorter patents to avoid paying higher prices for new goods and technologies, such as medicines and drought-resistant plant varieties, and to encourage local firms to copy those items. An internationally harmonized patent policy that would maximize global welfare is essentially impossible to define in this context.[21]

Second, because of the second-best economic environment raised by multiple distortions, the ways in which trade policies interact with regulatory market failures depend on the particular cases considered. It may not be possible to predict confidently whether restricting or liberalizing trade would be welfare-improving. For example, as we will describe in the next section, some evidence suggests that reducing trade barriers can offset the distortionary effects of poor environmental regulations sufficiently to improve air quality. Similarly, where market failures interact with international trade there inevitably are calls for using trade policies, such as countervailing tariffs and special import or export taxes, to counteract the impacts of inadequate regulation. However, tariffs are second-best approaches to market failures and may not be effective. More fundamentally, trade policies may address the indirect trade effects of a regulatory problem but simultaneously introduce additional costs and possibly reduce welfare.

The WTO and Market Access

The analytical complexities of these interactions make it difficult to state any general theorems about optimal mixes of trade and regulatory policies. However, there are important points to be made about the WTO and multilateral trade agreements in this area. We start by asking how a multilateral trade agreement might be helpful given its primary mandate of encouraging mutual increases in market access among its member states. Here an extension of the basic theory in Section 2 is useful.

Imagine that two governments, home and foreign, have two policy instruments at their disposal. The first is an *ad valorem* tariff *t*, which may be used to limit access to their import market. The second is a regulatory standard *s* of some kind designed to offset a domestic market distortion, which we assume to lie in the import-competing sector. For instance, this could be an effluent charge on firms emitting pollution, a minimum age at which people are permitted to work full time (to eliminate child labor), a minimum wage, a requirement that a good be produced with sanitary processes, stronger patent protection, or policies aimed at many other regulatory goals. Each government uses its two instruments to maximize national welfare. Note that the tariff and standard may interact. An increase in either nation's standard (eg, a higher minimum working age) raises costs in its import-competing sector. The effect would be to increase import demand, worsening the country's terms of trade. We immediately see that these policies are interrelated: a cut in the import tariff expands market access, which could be offset by a reduction in the standard.

As usual, let home import good x_1, foreign import good x_2, and the price ratio reflect the relative price of x_1. Home's import demand depends on domestic price, which is a function of the tariff and world price, and on its standard, and similarly for foreign. Balanced trade between the countries determines world price, an increase (decrease) in which raises (reduces) home's (foreign's) welfare. This means that, while each country sets its own tariff and standard, its welfare depends indirectly on the policy choices set by the other.

[21]Grossman and Lai (2004) present an important theoretical model of these tradeoffs.

How is an efficient policy equilibrium determined when there are two instruments in two countries? In essence three Pareto-efficiency conditions must be satisfied by the four policy choices.[22] The first is a condition in the home market: at the margin, any small changes in *t* and *s* together that leave the world price unchanged must leave home welfare the same. That is, holding constant the world price, these marginal changes in instruments just offset each other in home welfare. Note that because the terms of trade are unchanged there is no impact the foreign economy. This condition therefore states that, given constant foreign welfare the home government cannot further alter its policy mix to raise domestic welfare. The economic intuition is important. This condition states that, so long as there is no change in trade volume, and therefore foreign's market access in home remains intact, the home government should be free to choose the policy mix that maximizes its welfare. The second is a parallel condition for foreign-market efficiency. The third condition calls for international Pareto efficiency: policies chosen must make sure that neither country can be made better off without the other being made worse off.

The next question is whether, if countries set their policies unilaterally, the non-cooperative Nash choices would meet these efficiency criteria. Each government would choose policies to maximize domestic well-being, assuming the other country's policies are held constant. As might be expected, this maximization results in two reaction functions in each country, one for the tariff rate and one for the standard. As was the case analyzed in Section 21.2, the Nash tariffs will be set at higher levels than global efficiency requires. The new element is that, because the standard also affects world price, each country has an additional incentive to choose lower standards than needed for international efficiency. In short, both tariffs and standards are globally sub-optimal in the Nash equilibrium.

It follows that a trade agreement may be used, in principle, to move the world toward an improved, but perhaps not sustainable, welfare outcome. To understand this, consider the analysis in Figure 21.4. In panel A we depict a globally efficient set of tariff rates $(\bar{t}^{hE}, \bar{t}^{fE})$, and subsidies $(\bar{s}^{hE}, \bar{s}^{fE})$, which jointly place home and foreign on some contract curve at point E. In particular, given the efficient set of standards, the countries might choose through bargaining within a trade agreement to bind their tariffs at the levels depicted. These policies establish a given amount of market access in each market, reflected in the amount of imports each country accepts from the other.

However, as seen in panel B, depicting domestic policy tradeoffs between tariffs and standards, this outcome is not sustainable on its own. Home's welfare contours improve as the economy moves down and to the right, since for any tariff rate a lower standard reduces import demand and improves the terms of trade. The line labeled p^{*E} depicts combinations of home's tariffs and standard levels that keep the world price constant, as required by efficiency. Point E' depicts the same initial welfare outcome for home. Clearly, the home country can do better with a unilateral reduction in its standard, moving to point B and restricting the negotiated market access it had offered to foreign, which becomes worse off. A similar analysis holds for the foreign country. This analysis supports the theoretical concept of a "race to the bottom" in regulatory standards as countries liberalize their trade barriers.[23] To the degree that lower tariffs expose domestic import-competing firms to competition it can be offset by lower regulation. However, the reduced standard by home (or foreign) violates our conditions for full optimality in the

[22]Bagwell and Staiger (2002, Chapter 8) provide mathematical details.

[23]The papers in Bhagwati and Hudec (1996) offer a variety of cases where such problems could exist.

trade agreement.

This analysis raises the question of what an agreement like the WTO might do to deter this defection in standards policy. One answer is that the WTO rules already do require some disciplines in this regard. Recall the efficiency criterion: the mutually negotiated degree of market access should not be altered when a country alters its standards. In the dispute-settlement mechanisms of the WTO, any country that believes another is reducing its trade opportunities through domestic regulation may lodge a "non-violation complaint". This means that, even if the standard is not covered by WTO rules, any changes in the standard that "nullify or impair" the agreed-upon access benefits may be challenged. The resulting dispute could result in an order to restore the prior standards or offer additional market openings. In essence, this provision is designed to safeguard access commitments against subsequent standards revisions taken for commercial advantage.

A second approach is that some standards do fall directly under the WTO rules. Specifically, product standards come under the agreement on sanitary and phytosanitary measures (SPS) and the agreement on technical barriers to trade (TBT) in Annex 1A. These are complex legal accords. However, they share the basic principles that, while countries are free to take measures to protect their health status, environment, and other regulatory goals, the standards erected cannot amount to "...arbitrary or unjustifiable discrimination" nor can they be "...disguised restrictions on international trade". Measures that fail these provisions, according to certain technical rules, are subject to disputes.

These elements of the WTO raise concerns in the minds of many observers about the individual sovereignty of member nations to manage their own economies. This is perhaps the largest objection many have to global trading rules. As written, this interface between trade policy and domestic regulation is powerful yet relatively small in scope. The WTO Agreements have no direct control over government policies except for tariffs, nontariff barriers, trade subsidies and intellectual property rights. That is, member nations have chosen not to include disciplines on most standards and regulations, except to the extent they unjustifiably interfere with negotiated market access. There are no rules governing environmental standards, banking regulations, labor rights, or antitrust policy, to name several critical areas.

Thus, a final question to pose is whether the WTO agreement should be expanded to establish enforceable disciplines in such areas (Maskus, 2002). This is a fundamental question about the nature and objectives of global governance and has no ready answers. Multilateral trade agreements offer one advantage here. It is conceivable that negotiations could be undertaken across two or more disparate areas, with this "cross-issue bargaining" potentially generating wider economic gains. This possibility was one reason given for bringing intellectual property rights into the WTO. Most trade economists, however, agree that the WTO, as a compact designed to facilitate global commerce in goods, does not have sufficient competency to cover these broader elements of domestic regulation.

21.6 Empirical Evidence

A number of empirical questions have been posed about global trade liberalization. For example, trade economists study the potential gains from multilateral trade negotiations, such as the ongoing Doha Round, through computational general equilibrium models of the global economy. For example, Hertel, Hoekman and Martin (2002) calculate that a 40-percent reduction in the agricultural tariffs and export subsidies existing in the late 1990s would raise global income by around $60 billion per year, while a 40-

percent cut in industrial tariffs in all WTO members would expand global trade by around $380 billion per year.

In this section we review two other questions. First, is it really true that membership in the GATT and WTO has increased trade volumes significantly? Second, how does trade liberalization seem to affect environmental quality?

WTO membership and trade volumes

Given the importance of multilateral trade, it seems self-evident that countries that have been members of the GATT and WTO must trade more than countries that have not. However, a controversial study by Rose (2004) could not find such evidence. He estimated a standard gravity model (see Chapter 13) in which bilateral trade flows between pairs of countries were explained by mutual distance, GDP levels, and a variety of indicator variables for common language, common borders, common currency, joint membership in a preferential trade agreement, and other potential determinants of trade. The primary dummy variables of interest indicated whether both countries were in the GATT or WTO, or whether just one was. The sample covered 178 countries between 1948 and 1999, a period during which many countries were not in the GATT in the earlier decades. Some countries, such as China and Russia, were not in the GATT or WTO during this period, leaving them as the "control group" in the regressions.

The basic OLS regressions found that the standard gravity determinants of trade worked well to explain bilateral trade volumes. However, the coefficients on joint or single membership in the GATT/WTO were insignificantly different from zero and often negative. In only one case was there a significantly positive coefficient, which implied that two countries both in the multilateral system would increase their trade by about 16 percent compared to two that were not. While a 16-percent rise may seem important, the impact of WTO membership was considerably smaller than the other determinants of trade. A series of additional regressions were run to check how robust these findings were. Rose added country-pair fixed effects, estimated the regression for different periods, and broke down the coefficients by region and income class, among other tests. In general, his basic finding persisted. The author concluded that the simple act of belonging to the GATT/WTO had little impact on trade flows.

A finding this surprising rarely goes unchallenged in economics and several papers have extended the analysis in important ways. The most telling criticism is Subramanian and Wei (2007). They pointed out that simply asking whether membership itself affects trade was not a good test of the impacts of the actual commitments made by countries. Those member countries that actively engaged in reciprocal tariff cuts and other market-access commitments during the GATT/WTO years presumably should have experienced more trade growth than those that did not. Similarly, countries with a large production share in goods that were not subject to much trade liberalization, specifically agriculture and textiles and apparel, should have seen less trade growth. In fact, during the period studied most of the mutual tariff cuts were made by developed countries in industrial goods, while developing economies were often exempt from such cuts due to special and differential treatment. For example, prior to the Uruguay Round, 85 percent of industrial products had bound tariffs in the developed countries but only 32 percent did in the developing economies. Moreover, the export patterns of developing countries were biased towards agricultural products and apparel, precisely the goods that saw relatively little liberalization until after the WTO was founded. For these reasons the authors expected to see more trade increases among the developed countries than in other cases.

To this end, Subramanian and Wei estimated a similar gravity model using trade data very similar to those employed by Rose. Their initial specification, in which the basic variable was simply GATT/WTO membership, found a negative and significant coefficient, offering an even more pessimistic conclusion. However, when the importing countries in any pair were separated into developed and developing countries a major difference emerged. Where the importer was an industrialized country the gravity coefficients on the WTO dummy were around 0.9 to 1.0 and highly significant, suggesting that for these nations membership by around 250 percent in comparison with non-members. In contrast, the coefficients for developing countries were generally negative but small and insignificant. In the authors' words, then, membership in the WTO "...promotes trade strongly, but unevenly". Simply joining the global system, without actively engaging in mutual tariff liberalization, did not raise the trade volumes of developing nations. It is important to note that one of the key principles of the Uruguay Round, which founded the WTO, was to phase out the special treatment for developing countries and hold them to effective trade liberalization.

Is free trade good for the environment?

In Section 21.5 we pointed out that the interrelationships between trade policy and environmental regulation – or any significant regulatory issue – are complex and multifaceted, making it difficult to state any general theoretical results. It is possible to ask a number of empirical questions, however, starting with an obvious one: does trade liberalization around the world tend to improve or worsen environmental quality?

To organize this question, consider the framework set out in Antweiler, Copeland and Taylor (2001), which we depict graphically in Figure 21.5. Suppose that the home country is small, produces two final goods under constant returns, and has a comparative advantage in good x_1, while it has a tariff on imports of x_2. Production of the export good is "dirty", in that each unit of output generates emissions, which we label *G*. The home government imposes a pollution tax, inducing firms in this sector to adopt a technology $e(\eta)$, where η is the "abatement intensity", or the number of units of x_1 sacrificed to cleanup, divided by total output.[24] A higher tax generates a higher intensity, implying that less net output (total production less abatement) is available. Then total emissions are $G = e(\eta)x_1$. Good x_2 imposes no such externality. Then in Figure 21.5 the production frontier shows the net output of x_1 for all feasible outputs of x_2. The downward-sloping lines in the bottom panel reflect the relationship between emissions and output. Because of the pollution tax and import tariff the relative producer price of x_1 in home is below that in the rest of the world and the initial output equilibrium is at point A, with $p^* > p^h$. Output is Ox_1^A and the resulting total emissions are given by OG^A. Note also that at the world price ratio, home's national income is OI^A, measured in units of x_1 by the point where line p* through point a cuts the horizontal axis. We do not show consumption and simply note that home would import x_2 somewhere along the world price line northwest of A.

Figure 21.5

Now suppose that home cuts its tariff on imports, which raises the home relative producer price of x_1 to p^{h1} and induces output to shift to point C. We may now break down the resulting impacts on pollution into three fundamental components. First, suppose we hold national income constant along line

[24]This technology is assumed to have constant returns to scale and the same capital-labor ratio as output.

AI^A and do not change the pollution tax. Then the effect of trade is to shift the composition of output to point B, toward more of the dirty good. This *composition effect* of trade liberalization on emissions is the movement in the bottom panel from OG^A to OG^B and it necessarily increases pollution by expanding output of x_1.[25] Second, the economy is now bigger, as indicated by the movement from B to C. If we again hold the pollution policy constant, this *scale effect* is the shift from OG^B to OG^S. Again, this impact must raise emissions because we produce more of the dirty good at constant factor intensities.

The third component, called the *technique effect*, arises from an income-induced change in policy that reduces emissions intensity. Because real national income is higher, it is likely that the government would raise the pollution tax. One reason is that voter preferences for a cleaner environment tend to increase with higher income, an assumption that is well supported in the empirical literature (Grossman and Krueger, 1993). Thus, if the government raises the tax to increase the abatement intensity to η^C the new emissions line comes into play. The technique effect is then the emissions fall from to OG^S to OG^C, which cuts pollution.

While we have shown an overall cut in pollution, it is clear that the total impact of tariff cuts, even in a country that exports the dirty commodity, is ambiguous. The composition and scale effects expand pollution, while the technique effect, arising indirectly from more rigorous regulation as income expands, brings emissions down. Thus, empirical evidence should shed light on this question, as studied by Antweiler, Copeland and Taylor (2001). They compiled data on the airborne concentrations of sulphur dioxide (SO_2), a noxious gas emitted both by natural forces and industrial production, in 108 cities spanning 43 countries from 1971-1996. Concentrations are not the same as emissions, since the former depend on weather conditions and other factors. The authors assumed that concentrations were linearly related to underlying emissions. This permitted them to regress the log of SO_2 concentrations on variables explaining emissions from the theory, along with several interaction terms. Central explanatory variables included SCALE (local GDP per square kilometer in each city), KL (the national capital-to-labor endowment ratio), INC (lagged per-capita GNP in the country), and TI (trade intensity, the ratio of exports plus imports to GDP in each nation). The final variable was taken as a proxy variable for trade restrictions, assuming that more closed economies would have lower trade ratios. Further, because the sensitivity of concentrations to trade intensity would vary across countries, depending on levels of income and factor endowments, interactions were included between TI and these other variables. Several control variables were included to account for weather variations, type of government (communist or non-communist) and other factors.

The basic results are given in Table 21.2 for the authors' preferred specification, which included fixed effects for each location where SO_2 was measured. The first variable, SCALE, captures the effect of city size, holding constant trade openness and the country capital-labor ratios. It has a significantly positive effect on concentrations (emissions), as anticipated. The scale elasticity suggests that a one-percent increase in local economic size raises pollution by about 0.27 percent. The next variable is the capital-labor ratio, which directly expands concentrations. Thus, countries with higher capital-labor ratios, and therefore a larger proportion of output in capital-intensive goods, emit more concentrated SO_2. However, note that there is a negative coefficient on the interaction of KL with income (INC), which means that even capital-abundant economies emit less pollution as they become richer. Further, there is a negative interaction between the capital-labor ratio (relative to the global average) and trade intensity (TI), indicating also that more open economies emit relatively less gases. The composition elasticity

[25]Of course, if the economy exported the clean good the composition effect of trade liberalization would be to cut emissions.

listed takes account of both the direct impact and all these interactions (computed at sample means). Overall, then, an increase of one percent in the average country's capital abundance, holding scale and income constant, generates an increase in SO_2 concentrations of one percent.

Table 21.2

The effects of income (lagged GNP per capita) enter in a variety of direct and indirect means but the overall effect, called the technique elasticity, is significantly negative and indeed rather large. This finding supports the view that policymakers in higher-income economies encourage cleaner technologies. Finally, consider the impacts of trade intensity (a questionable proxy for openness to trade). The direct coefficient is significantly negative. Taking account of all channels, the trade-induced elasticity is clearly negative. Based on this evidence, then, economies that are more open to trade generate significantly lower concentrations of sulphur dioxide than average. Whether this result would hold in other pollution databases remains to be seen.

REFERENCES

Antweiler, Werner, Brian R. Copeland, and M. Scott Taylor (2001), "Is Free Trade Good for the Environment?" *American Economic Review* 91, 877-908.

Bagwell, Kyle and Robert W. Staiger (2002), *The Economics of the World Trading System*, Cambridge, MA: MIT Press.

Baldwin, Robert E. (1985), *The Political Economy of U.S. Import Policy*. Cambridge, MA: MIT Press.

Bhagwati, Jagdish N. and Robert E. Hudec, editors (1996), *Fair Trade and Harmonization: Prerequisites for Free Trade?* Cambridge, MA: MIT Press.

Blonigen, Bruce A. and Thomas J. Prusa, (2003), "Antidumping," in E. Kwan Choi and James Harrigan, eds., *Handbook of International Trade*. Malden, MA: Blackwell Publishing.

Bown, Chad P. (2009), *Self-Enforcing Trade: Developing Countries and WTO Dispute Settlement*. Washington DC: Brookings Institution.

Bown, Chad P., editor (2011), *The Great Recession and Import Protection*. Washington DC: The World Bank.

Brander, James and Paul Krugman (1983), "A 'Reciprocal Dumping' Model of International Trade," *Journal of International Economics* 15, 313-321.

Chisik, Richard (2003), "Gradualism in Free Trade Agreements: A Theoretical Justification," *Journal of International Economics* 59, 367-397.

Furusawa, Taiji and Edwin L.-C. Lai (1998), "Adjustment Costs and Gradual Trade

Liberalization," *Journal of International Economics* 49, 333-361.

Ganslandt, Mattias and Keith E. Maskus (2008), "Intellectual Property Rights, Parallel Imports and Strategic Behavior," in K.E. Maskus, editor, *Intellectual Property, Growth and Trade*. Amsterdam: Elsevier Publishing.

Grossman, Gene M. and Elhanan Helpman (2002), *Interest Groups and Trade Policy*. Princeton: Princeton University Press.

Grossman, Gene M. and Alan B. Krueger (1993), "Environmental Impacts of a North American Free Trade Agreement," in Peter Garber, editor, *The US-Mexico Free Trade Agreement*. Cambridge, MA: MIT Press.

Grossman, Gene M. and Edwin L.-C. Lai (2004), "International Protection of Intellectual Property," *American Economic Review* 94, 1635-1653.

Helpman, Elhanan and Paul R. Krugman (1989), *Trade Policy and Market Structure*. Cambridge MA: MIT Press.

Hertel, Thomas W., Bernard M. Hoekman and Will Martin (2002), "Developing Countries and a New Round of WTO Negotiations," *The World Bank Research Observer* 17, 113-140.

Hoekman, Bernard M. and Michel M. Kostecki (2009), *The Political Economy of the World Trading System: the WTO and Beyond, Third Edition*. Oxford: Oxford University Press.

Johnson, Harry G. (1953-54), "Optimum Tariffs and Retaliation," *Review of Economic Studies* 21,142-153.

Kaul, Inge, Isabelle Grunberg and Marc A. Stern, editors (1999), *Global Public Goods: International Cooperation in the 21st Century*. Oxford: Oxford University Press.

Kennan, John and Raymond G. Riezman (1988), "Do Big Countries Win Tariff Wars?" *International Economic Review* 29, 81-85.

Magee, Stephen P., William A. Brock and Leslie Young (1989), *Black Hole Tariffs and Endogenous Policy Theory: Political Economy in General Equilibrium*. Cambridge: Cambridge University Press.

Maskus, Keith E. (2000), *Intellectual Property Rights in the Global Economy*. Washington DC: Peterson Institute for International Economics.

Maskus, Keith E. (2002), "Regulatory Standards in the WTO: Comparing Intellectual Property Rights with Competition Policy, Environmental Protection, and Core Labor Standards," *World Trade Review* 1, 135-152.

Mayer, Wolfgang (1981), "Theoretical Considerations on Negotiated Tariff Adjustments," *Oxford Economic Papers* 33, 135-153.

Olson, Mancur (1965), *The Logic of Collective Action*. Cambridge, MA: Harvard University Press.

Rose, Andrew K. (2004), "Do We Really Know that the WTO Increases Trade?" *American Economic Review* 94, 98-114.

Subramanian, Arvind and Shang-Jin Wei (2007), "The WTO Promotes Trade, Strongly but Unevenly," *Journal of International Economics* 72, 151-175.

ENDNOTES

1. A good introduction to this complex subject is the volume by Kaul, Grunberg and Stern (1999).

2. We do not go through the formal mathematics here, which would be similar to those in Chapter 18. The difference is that in achieving an expression like (18.26) for the change in home's welfare from altering its tariff, we must account for the effect of different levels of the foreign tariff on export supply and world price. There is a similar expression for the change in foreign's welfare, holding constant home's tax. The best-response functions are calculated by setting these welfare changes equal to zero.

3. Strictly speaking, this is not a necessary outcome. Depending on various parameters it is possible for one country to gain and the other to lose welfare in Nash equilibrium relative to free trade. However, the losses to one nation dominate the gains to the other and the world is worse off. It is also possible for these nonlinear best-response functions to intersect more than once, suggesting multiple Nash equilibria are possible. For details, see Johnson (1953-1954), Kennan and Riezman (1988) and Bagwell and Staiger (2002).

4. CC is not necessarily linear. Mayer (1981) showed that tariff rates along this locus must satisfy the relationship $(1 + t^h) = 1/(1 + t^f)$, which ensures that relative prices are the same in both countries, absent any other price distortions.

5. The basic logic of this argument for tariffs may be traced back to Olson (1965) and was statistically tested by Baldwin (1985) among many others. Grossman and Helpman (2002) offer a theoretical analysis.

6. An analytical treatment is in Magee, Brock and Young (1989).

7. Furusawa and Lai (1999) show in a dynamic model that if workers face adjustment costs then negotiated tariff cuts generally need to be implemented gradually for countries to keep them in place. See also Chisik (2003) for a similar argument where there are costly adjustments to capital stocks.

8.For details see World Trade Organization, *Annual Report, 2011*, at www.wto.org.

9.In 1987 Hong Kong was an independent protectorate of the United Kingdom and is now formally part of China. But Hong Kong retains its own authority to set trade and immigration policy.

10. These columns are not strictly comparable because the weights in 2008 are based on more disaggregated product trade than the 22 industries in 1987.

11. Note that these rates averaged equally across all goods classes are generally higher than the rates that use import weights, since imports are often higher in products with lower tariffs.

12. Students interested in a detailed discussion of these issues should read Hoekman and Kostecki (2009).

13. All these provisions mean that each WTO member actually has up to five tariff schedules, which list tariff rates for each product. First is the MFN bound schedule, then the MFN applied schedule. After that come any tariffs agreed to in FTAs and the GSP rates. Finally, there are the tariffs applied to goods from the few countries that are not in the WTO and are not awarded MFN treatment.

14. An excellent review of the dispute settlement process is in Bown (2009). Interested readers may consult all of the panel reports at the WTO website: http://www.wto.org/english/tratop_e/dispu_e/dispu_e.htm.

15. Blonigen and Prusa (2003) describe a number of questionable practices in antidumping law and their effects on trade.

16. Interested students can read the WTO text online at http://www.wto.org.

17. Maskus (2000) explains the global economics of intellectual property rights in detail.

18. Students who have studied microeconomic theory may recognize this outcome as "second degree" price discrimination, since the firm sets one price for groups of consumers (here, one price per country), rather than a different price for each individual consumer.

19. The logic in Chapter 14 was that income elasticities of demand for such goods rise with wealth. The idea here extends that to the possibility that preferences of higher-income consumers (countries) also are less price-elastic.

20. The United States, for example, prohibits parallel imports of products that are protected by domestic patents and copyrights. The EU also prohibits such imports from outside its region but they are permitted among the member countries. See Ganslandt and Maskus (2008) for details.

21. Grossman and Lai (2004) present an important theoretical model of these tradeoffs.

22. Bagwell and Staiger (2002, Chapter 8) provide mathematical details.

23. The papers in Bhagwati and Hudec (1996) offer a variety of cases where such problems could exist.

24. This technology is assumed to have constant returns to scale and the same capital-labor ratio as output.

25. Of course, if the economy exported the clean good the composition effect of trade liberalization would be to cut emissions.

Table 21.1 Average Tariff Rates by Selected Country(%)

	1987	2008	2009	2009	2009	2009
Developed Countries	Imp weighted	Imp weighted	MFN applied	MFN Bound	MFN Ag	MFN Other
Australia	14.8	5.6	3.5	10.0	1.3	3.8
Canada	4.6	3.3	4.5	6.7	10.7	3.5
EU-12	6.6	na	na	na	na	na
EU-27	na	2.9	5.3	5.3	13.5	4.0
Japan	6.2	2.0	4.9	5.1	21.0	2.5
Switzerland	3.5	3.4	6.5	8.0	36.9	1.9
United States	3.3	2.0	3.5	3.5	4.7	3.3
Newly Industrialized Countries						
Chinese Taipei	13.4	1.9	6.1	6.4	16.6	4.5
Hong Kong, China	0.0	0.0	0.0	0.0	0.0	0.0
Singapore	0.9	0.0	0.0	10.4	0.2	0.0
South Korea	10.8	8.3	12.1	16.6	48.6	6.6
Developing Countries						
Argentina	13.0	13.3	12.6	31.9	10.3	13.0
Brazil	20.8	8.8	13.6	31.4	10.2	14.1
China	45.0*	4.3	9.6	10.0	15.6	8.7
India	47.0	6.0	12.9	48.5	31.8	10.1
Mexico	10.6	11.1	11.5	36.1	22.1	9.9
Turkey	16.5	4.0	9.7	28.6	42.9	4.8

Note: *Simple average for 1994

Sources: 1987 data computed from Deardorff and Stern (1989); China 1994 figure from Harvard University, Global Trade Negotiations webpage; remaining data from WTO, Tariff Profiles on line

Table 21.2 Estimated Impacts of Trade Intensity on SO_2 Concentrations

Variable	*Coefficient*	*Elasticity*	*Estimate*
Intercept	-4.32***		
SCALE	0.058***	Scale	0.266***
KL	0.461**	Composition	1.006**
KL Squared	0.006		
INC	-0.096	Technique	-1.153**
INC Squared	0.559***		
KL*INC	-0.381***		
TI	-3.142**	Trade-induced	-0.864***
TI*RELKL	-2.252*		
TI*(RELKLSQ)	-0.123		
TI*RELINC	2.687***		
TI*(RELINCSQ)	-0.595**		
TI*RELKL*RELINC	0.900**		

Note: *Significant at 95% confidence
**Significant at 99% confidence
***Significant at 99.9% confidence
Source: Antweiler, et al (2001, Table 1)

Figure 21.1

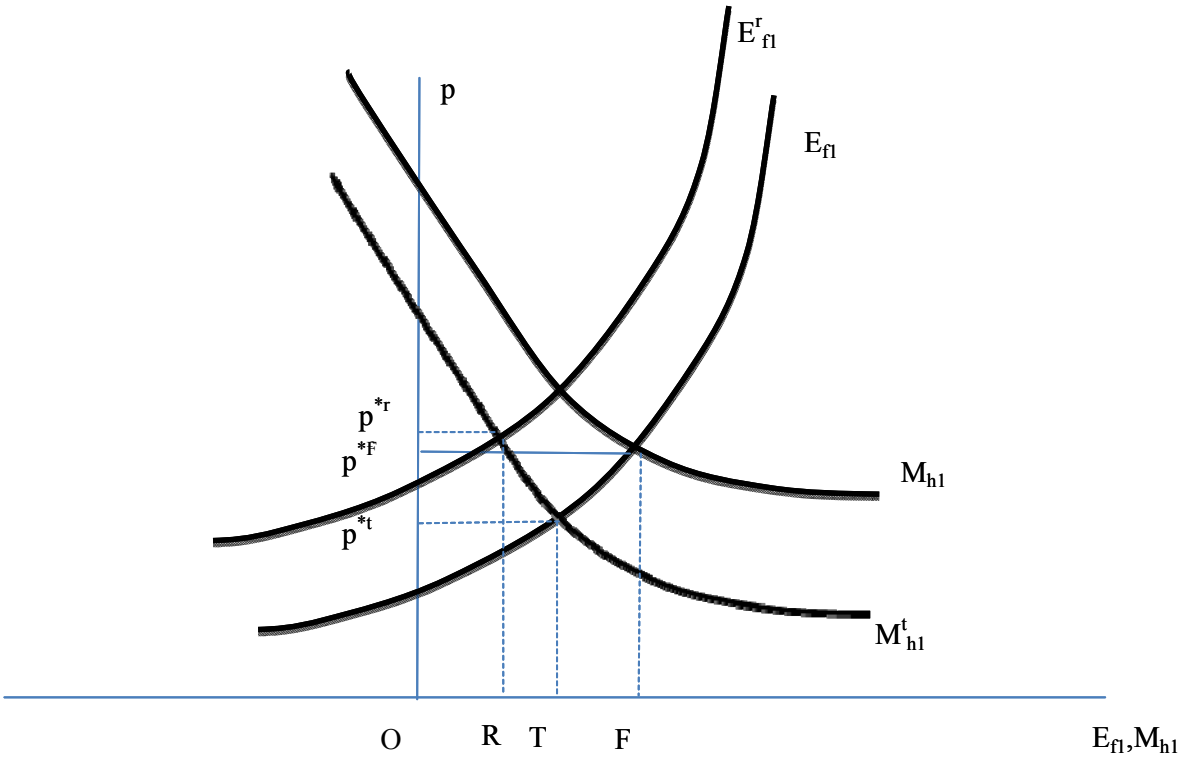


Figure 21.2

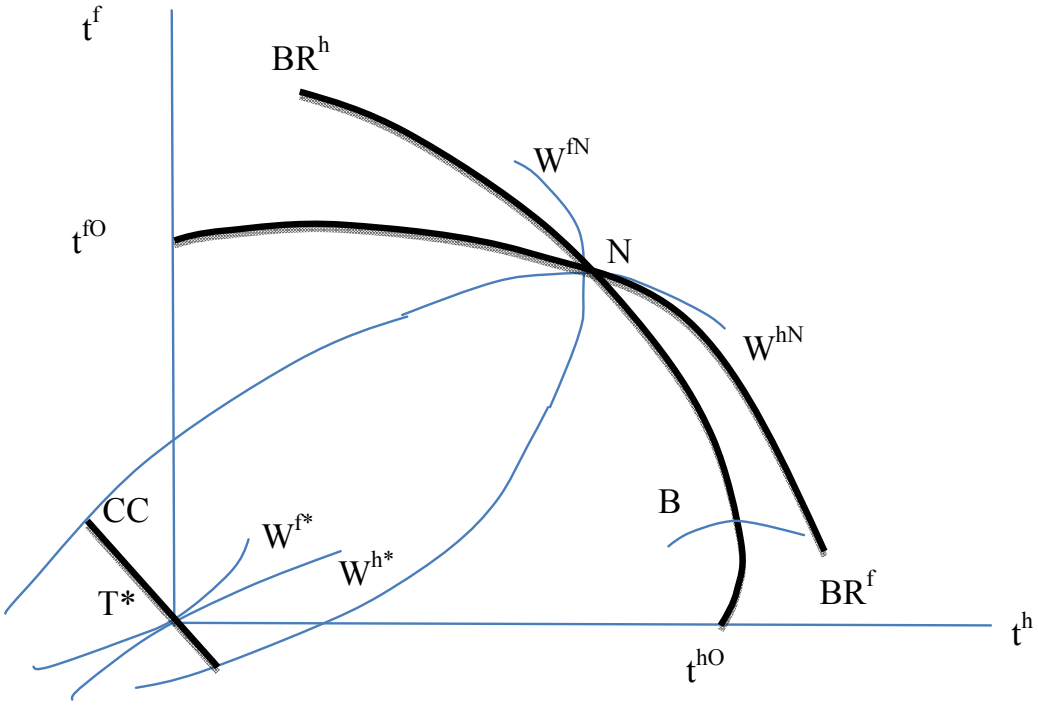


Figure 21.3

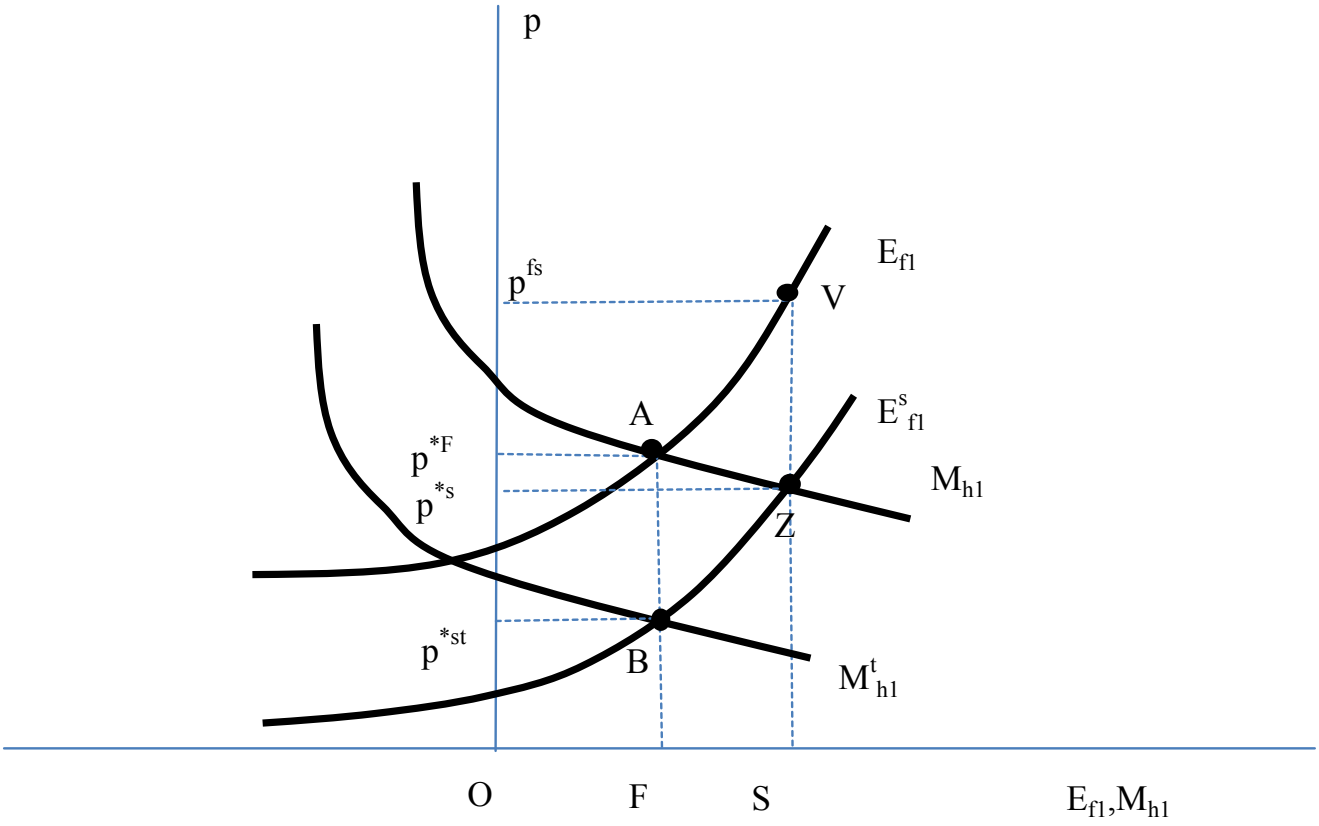


Figure 21.4a

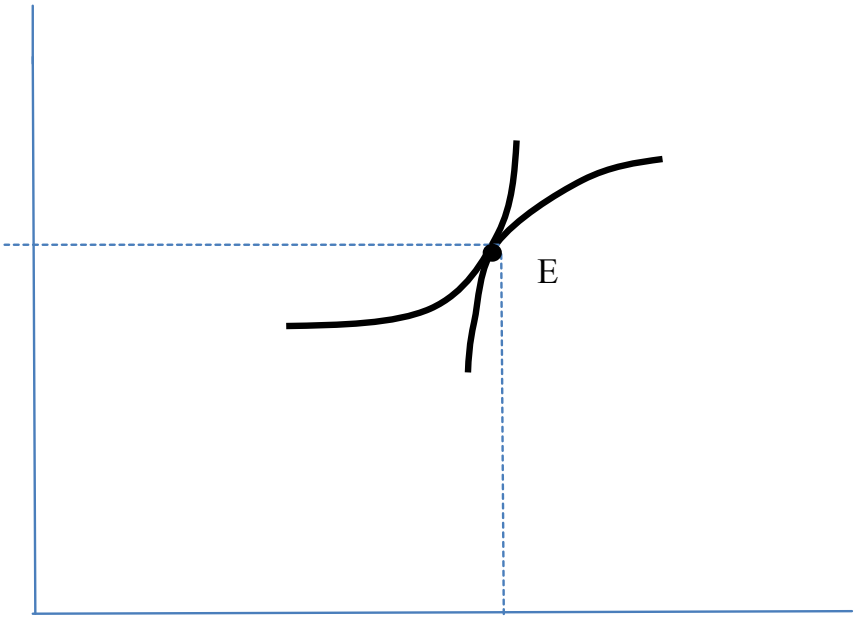


Figure 21.4b

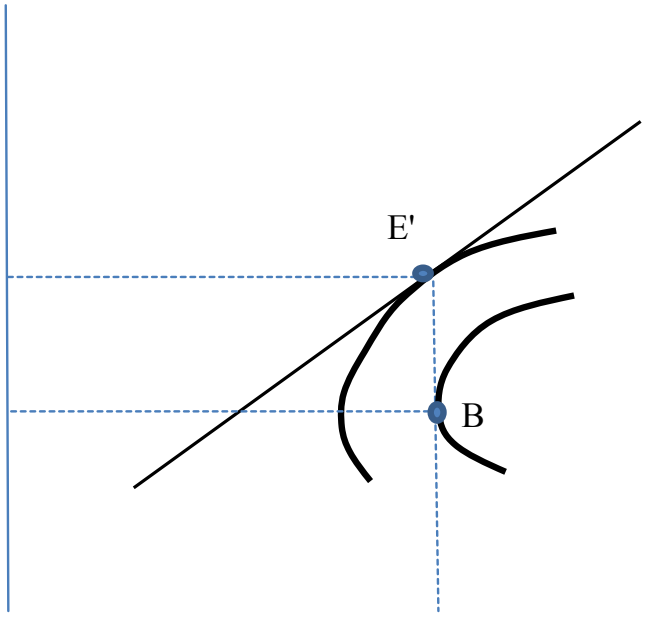
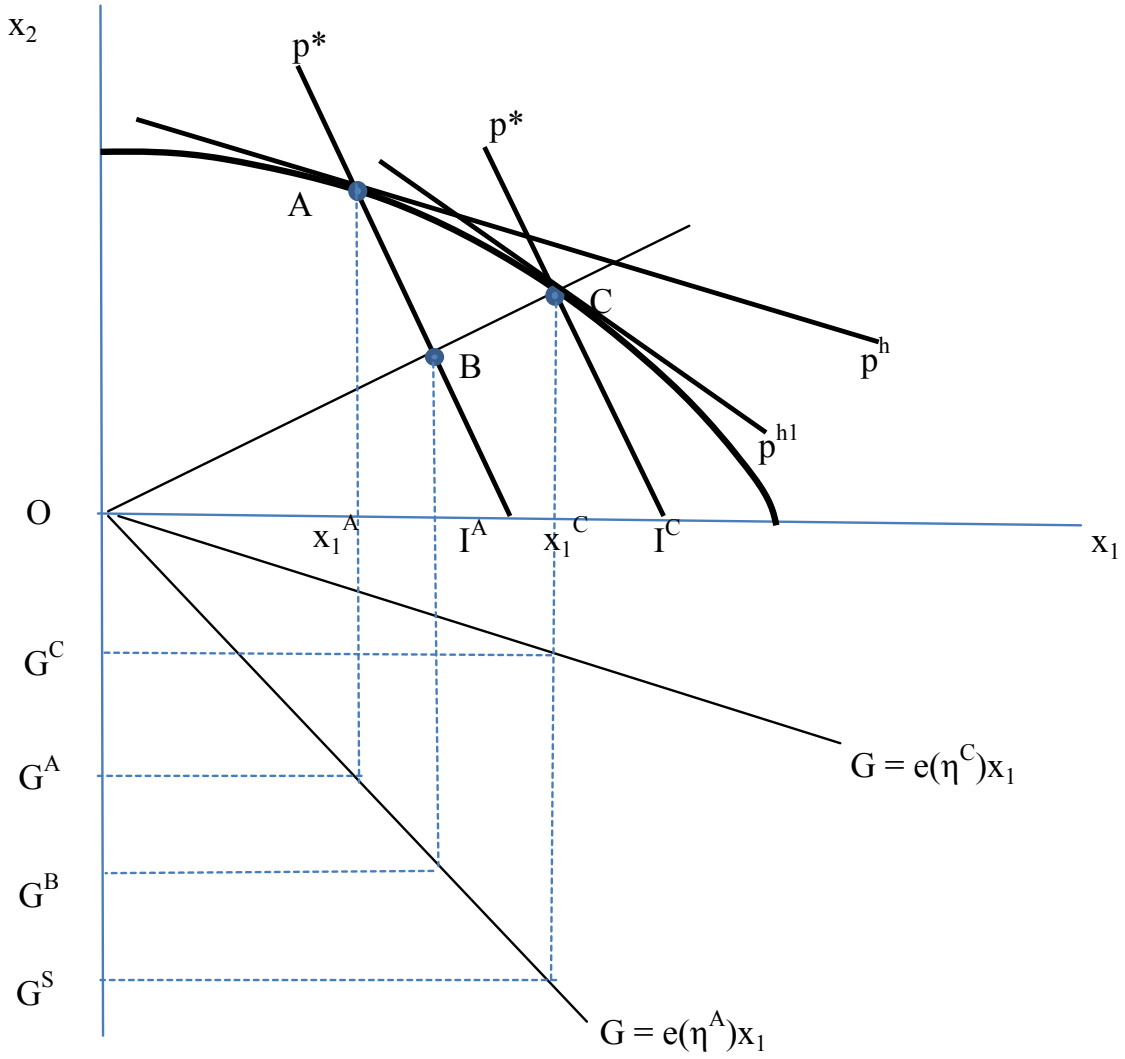


Figure 21.5



CHAPTER 22

PREFERENTIAL TRADE AGREEMENTS

22.1 Introduction

While the WTO is the central institution for multilateral trade liberalization, many countries also embrace the concept of preferential trade agreements (PTAs), in which a subset of countries mutually offer lower trade barriers on the products coming from each participant. Because such arrangements are often made on a regional basis, such as the North American Free Trade Agreement (NAFTA), they are often referred to as regional trade agreements. The number of PTAs in the world has grown markedly in recent years. As of May 15, 2011 there were 297 PTAs in force, according to the WTO.

Preferential trade agreements may be usefully classified in terms of their scope. Many are *partial PTAs*, in which countries simply reduce tariffs, perhaps to zero, on a limited schedule of goods. More common are *free trade agreements (FTAs)*, which feature the elimination of trade barriers among all member nations. NAFTA is an example of an FTA, for it has zero tariffs on virtually all trade among Mexico, Canada and the United States. However, each member sustains its own tariff schedule, which it applies to products from non-member countries. One important implication is that countries in an FTA must negotiate *rules of origin*, which define terms under which a good is certified to have been produced in the region and therefore eligible for free trade. Otherwise non-member countries would simply export their goods to the lowest-tax member and then transship the goods to their ultimate destinations. These rules can be highly complex and sometimes may be an impediment to trade themselves (Hoekman and Kostecki, 2009).

A tighter degree of integration is a *customs union*, in which member countries have free trade among themselves and share a common tariff schedule against goods from nom-members. One advantage is that it makes rules of origin less significant, which reduces bureaucratic trade costs. MERCOSUR, an agreement among Argentina, Brazil, Paraguay and Uruguay, is a customs union. Finally, some countries may choose to go beyond a customs union to the mutual elimination of barriers to capital and labor movements. This form of trade policy coordination is called a *common market*. The European Union, also called the European Economic Community, is the primary example of a common market. Its members share a common external tariff (CET) and are required to permit the free flow of labor and capital.

In this section we discuss the basic economic effects that arise specifically in the context of PTAs. Students should recognize, however, that because PTAs expand regional markets they would generally garner the same kinds of gains discussed in earlier chapters. First, there could be gains from trade associated with specialization that takes advantage of member country differences in endowments or tastes. Second, a PTA may allow firms within the region to enjoy greater economies of scale. Third, domestic industries will face increased competition, so there may be pro-competitive gains as monopoly power is diminished, along with benefits from gaining access to a larger number of varieties.

22.2 Welfare Basics: Trade Creation and Trade Diversion

It seems natural to think that PTAs are a movement toward free trade and therefore must raise

economic welfare. However, this view is false. The central feature of PTAs is that they set tariffs to zero among member countries but remain positive against non-members. This means they are a form of *partial* or *piecemeal* tariff reforms: rather than go to fully free trade, some trade restrictions are cut and others are not. This is an example of a second-best policy described in Chapter 10. PTAs replace one distortionary policy – a uniform set of positive tariffs – with another distortion – some positive and some zero tariffs. Put differently, the preferential tariffs artificially discriminate against some suppliers in favor of others. The immediate implication is that a participant in a PTA could gain or lose welfare.

This simple insight is illustrated in Figure 22.1, where we depict country A to be a small importer of good x_1. Because A is small, the export-supply curves from a potential partner B and R, the rest of the world, are perfectly elastic. We draw R to be the more competitive supplier, with price p^R, and B to have a higher price at p^B. In the initial situation, A has a non-discriminatory tariff, which shifts its import-demand curve downward to M^t_{A1}. Country A then buys the good from R because it is the cheaper supplier. The volume of imports is OT, which arrive from R at price p^R but are sold on A's market at p^A. Note that this equilibrium generates tax revenue (in terms of good x_2) of area $p^A\, p^R ED$. Now suppose country A enters into a free trade agreement with B, which means the tariff is eliminated on good x_1 from the partner but remains in place for the good from R. The good now becomes available to A's consumers at price p^B, who will therefore import from B in the amount ON and cease importing from R. In this new equilibrium no tariff revenue is generated in A.

There are several ways to think about the effects of this FTA. First, note that the tariff cut expands the volume of trade from OT to ON, with the additional imports coming from partner B. This is an example of *trade creation*, which we may define as additional trade coming from a partner country as a result of the preferential trade liberalization. At the same time, the prior volume of imports had come from R whereas now they are purchased from B. This is the result of *trade diversion*, which is the reduction of imports from a lower-cost supplier outside the preferential region in favor of imports from a higher-cost supplier inside it.[1]

Figure 22.1

A second approach is to use the language of Chapter 18. Note that the tariff cut generates both a volume-of-trade effect and a terms-of-trade effect. The first effect is associated with higher welfare as consumers benefit from the lower price. The second effect reduces welfare as country A's policy worsens the price it must pay for imports from p^R to p^B. This difference is equivalent to rewriting equation (18.26) as follows:

$$dW = (p^A - p^B)dE_1^B - E_1^R dp^* \qquad (22.1)$$

where dp^* refers to the change from the price in R to the price in B. This expression holds for small changes in trade and prices. In Figure 22.1 the VT gains may be represented heuristically by the additional area under the import-demand curve but above price p^B, while the TT losses are given by the elimination of the tariff revenue. The net welfare impact would be area DJG less area $p^B\, p^R EJ$, which could be positive or negative.

[1]Identification of the concepts of trade creation and trade diversion is generally attributed to Viner (1950).

Note that this analysis depicted a single homogeneous good imported by country A. Any PTA would involve tariff cuts on many imports and one could sum the effects to determine whether trade creation or trade diversion dominates overall. Because of this basic theory, some analysts argue that PTAs that result in net trade creation must increase welfare and those that generate net trade diversion must reduce well-being. This linkage can be misleading, however, as the next example shows.

Suppose instead of two goods we consider a more complex world of three goods, which we number x_1, x_2, and x_3, and three countries, A, B and R. Imagine that each country is completely specialized, with A producing x_1, B producing x_2, and R producing x_3. In this case, A must import the latter two goods and, because they are different products, they may face different tariffs prior to an FTA. We continue to assume A is a small importer, facing fixed prices from B and R. These prices are depicted in Figure 22.2 as p_B^2 and p_R^3, respectively.[2] Curve M_{A2} is the import-demand curve for good x_2 in country A. To avoid clutter in the diagram we do not show the tariff-distorted curve, which implicitly goes through point J. Thus, the pre-FTA situation in country A is to import x_2 in the amount OT, with domestic price p_A^2. If A now enters a FTA with B and eliminates this tariff the domestic price of x_2 falls to p_B^2 and there is a net welfare gain of area DJG. The volume of trade rises to ON and this outcome may reasonably be termed a gain from trade creation. Note that this effect in good x_2 simply reflects the fact that country A is small and therefore gains from eliminating this tariff.

The complication is that goods x_2 and x_3 are interrelated in demand. The fall in the price of x_2 will reduce the demand for imports of x_3 if the goods are net substitutes, as shown in the second panel by the shift to M_{A3}^*. Prior to the FTA imports of x_3 were OY but the agreement reduces this volume to OZ. Because the tariff remains in place there is no change in domestic price. However, tariff revenue falls by the amount HEFJ, which is an appropriate measure of the welfare cost of the FTA in this good and may reasonably be attributed to trade diversion.

Figure 22.2

Now imagine that goods x_2 and x_3 are net complements instead of substitutes. In this case the import-demand curve for the latter product would shift upward (not shown) from its original level. Then the import volume from R would expand, as would tariff revenue, implying a welfare gain for country A. Here the phrase "trade diversion" is misleading since trade with the external region actually expands and welfare increases. In general, demand-side complexities imply that the economic effects of FTAs depend on numerous market parameters.[3]

---

[2]These prices may be defined relative to good x_1, which would be the numeraire in this case. Thus, in both panels areas, such as tariff revenues, are measured in units of good x_1.

[3]The mathematics of the case with multiple differentiated goods are worked out in Konan and Maskus (2011). They show that key parameters include the proportion of trade that had existed among partners prior to the FTA, the relative size of tariffs cut within the FTA versus those sustained on non-members, and the degree of substitutability among goods. See also Riezman (1979) and Kowalczyk (2000).

There are numerous other possibilities that could be considered in a full accounting of the trade and welfare impacts of a preferential trade agreement involving many goods and countries. Among the more important are the effects on the internal terms of trade (price changes within the PTA) and the external terms of trade (price changes on goods from outside the region), the interrelationships of tariff cuts with other taxes in an economy, and how the formation of PTAs influences capital formation and FDI. Indeed, economists sometimes discuss the *investment creation* and *investment diversion* effects of a free trade area. Also important is the possibility that when PTAs are formed they change competition among imperfectly competitive firms subject to increasing returns to scale. In that case the analysis in Chapters 11 and 12 becomes relevant. Because these are complex issues the full impacts are generally analyzed with computable general equilibrium models, as described in the final section of this chapter.

22.3 A General Welfare Theorem

This considerable degree of complexity makes it seems impossible to say anything definitive about the welfare effects of particular preferential trade agreements. However, a particular formation was set out by Kemp and Wan (1976) that suggests gains are available to all members of a customs union. Recall that a customs union (CU) is an agreement among a set of countries to offer zero tariffs to each other but establish the same set of tariffs against non-members. These common external tariffs (CET) offer the key to the welfare result.

The Kemp-Wan Theorem may be stated as follows. Consider a customs union among any set of countries. There exists a CET schedule that keeps the vector of world prices constant, implying that the vector of net external purchases made by the CU as a whole is unchanged. This CET, combined with free trade among members of the CU, makes it possible for income transfers to be made among members such that, in each country, the consumption bundle in the CU is revealed preferred to that prior to the agreement.

The full proof of this theorem is complex but may be understood intuitively in terms of the analysis of the gains from trade in Chapter 18. Forming a CU involves, for each country, a comparison of two tariff regimes: one with an initial set of tariffs unique to each versus one in which a subset of countries share a CET. The particular CET chosen, however, keeps prices and trade volumes of each good constant for non-members, with two implications. First, the non-member nations are not harmed by the CU. Second, taking the CU as a whole there is no terms-of-trade effect. Equivalently, the aggregate amount of tariff revenues do not change when added up among all the CU members.

The difficulty is that trade-tax revenues will rise in some member countries and fall in others with the shift to the CET, depending on initial tariff levels and the changes in import volumes. However, the analysis in equation (18.13) covers this situation nicely. We reproduce it here as equation (22.2), with superscript *c* for member *i* indicating the situation with the customs union in place and superscript *r* referring to the pre-agreement case. Note first that the value of production at internal CU prices must be at least as large as the pre-CU production bundle valued also at those prices. Adding the appropriate balanced-trade conditions we get that the left-hand side exceeds the first term on the right-hand side, or that the CU consumption bundle is preferred to the pre-CU bundle.

$$\sum_i p_i^c D_i^c \geq \sum_i p_i^c D_i^r + \sum_i (p_i^r - p_i^c)(D_i^r - X_i^r) \qquad (22.2)$$

The final term may be positive or negative and countries where it is negative may be worse off with the CU. However, recall that the aggregate change in tariff revenues across the CU is zero. Therefore, it is possible for countries with positive revenue changes to transfer them to those with negative changes, implying that for each member the final term (post-transfer) of equation (22.2) is zero. It follows that each CU participant gains from the agreement, while the rest of the world experiences no change in welfare.[4]

Three additional important comments should be made. First, even though these transfers are feasible, they may not actually be made in practice. Nonetheless, economists are willing to claim that there is a welfare gain so long as such transfers are possible in principle. This is an example of the *compensation principle* widely used in welfare economics. Second, the Kemp-Wan criterion may be compared to the WTO principle permitting the formation of PTAs. The WTO rules state essentially that free trade agreements and customs unions are acceptable so long as they do not cause member states to raise average levels of trade protection against non-members. This is quite different from the Kemp-Wan proposal that a CET be set so that the custom union's net imports of each good are unchanged, which is the more appropriate means of ensuring that non-members are not disadvantaged. Finally, a CU organized according to Kemp-Wan principles can be meaningfully thought of as a partial movement toward free trade, since it does not change external trading terms while it does establish zero internal trade restrictions. Thus, the larger is a CU in terms of the number or size of its member nations the further is the movement toward global free trade.

22.4 Endogenous FTA Formation

The theory presented so far considers the effects on a single partner of a customs union or free trade area established among an exogenous set of countries. However, such agreements do not arise randomly among nations. Rather, they are the result of conscious choices made by governments. This observation raises the obvious question of which economic factors determine the endogenous decisions of countries to form CUs and FTAs. Put differently, which economic characteristics make it likely that a set of countries – or the world – would gain welfare as a result of forming a preferential trade agreement?

Because FTAs embody numerous elements of second-best economics there are no fully definitive theoretical answers to this question. However, economists have identified several basic factors that increase the likelihood of mutual gains among agreement partners. For example, Krugman (1991) set out a basic model in which an arbitrary number of countries exist. Consumers display "love of variety" preferences (as in Chapter 12) and each country has a single firm producing a monopolistically competitive good with increasing returns to scale and using just one factor, labor. The author calculated world welfare as these countries were combined into various configurations of equally sized FTAs, or trading blocs. Within these blocs countries have zero tariffs but each sets its optimal tariffs against goods from non-members. Krugman's computations found that as the number of FTAs went down (and, therefore, countries joined ever-larger blocs), world welfare would decline and be minimized at three.

---

[4]A similar result for free trade agreements was proven by Krishna and Panagariya (2002). In their case each country must reset its individual tariff schedule after entering an FTA so that its imports from, and exports to, non-member countries are unchanged for each good. The FTA therefore has no impact on prices in the rest of the world, while within each country an expression like equation (22.1) holds. It should be noted that this result ignores complications from such elements as rules of origin.

Global welfare would rise at two blocs and be maximized with just one, corresponding to multilateral free trade.

The result that three FTAs would minimize global well-being was met at the time with considerable concern in light of what seemed to be three emerging major trading blocs: North America, the European Union and East Asia. However, it is the outcome of highly specialized assumptions of product differentiation and love of variety. In this world, an FTA among two small countries beneficially reduces prices within the agreement on just the two goods produced there, while raising the relative prices of all goods from excluded countries. Here "trade diversion" dominates "trade creation" because the excluded number of countries (and varieties) is far larger than the number included. As the number of blocs falls to three there is still net trade diversion since two-thirds of the world's products are outside the FTAs and just one-third inside, while the larger economic sizes of the blocs encourage higher optimal tariffs. These factors explain why the "3-bloc" solution is worst from a welfare standpoint. Only when the world reaches two blocs, so that the number of included and excluded goods are the same, can trade creation gains begin to outweigh trade diversion losses.

Krugman's model was criticized on other important grounds and we mention two here. First, by assuming that all countries produce a single good with just labor, the model ignored the important possibility that countries could differ in factor endowments and the efficiency of producing goods with different factor intensities (Deardorff and Stern, 1994). However, the ability of countries within an FTA to specialize resources and reduce costs according to comparative advantage is a critical potential source of welfare gains. Put differently, trade creation within the FTA because of specialization should enhance welfare.

Second, Krugman's model assumed there were no trade costs, either inside an FTA region or from countries outside the agreement. Suppose, instead, that countries were located on three "continents", with low intra-continental transport costs and high inter-continental transport costs, which would make welfare calculations very different. In the starkest case of zero local trade costs and infinite inter-continental costs, the emergence of an FTA on each continent would effectively mirror global free trade and maximize welfare.[5] Indeed, the presence of trade costs suggests that economic geography matters greatly for the selection of FTA partners, which is consistent with the regional orientation of most trade agreements. In this regard, countries in close proximity to each other may be considered "natural" partners (Frankel, Stein, and Wei, 1995).

This kind of analysis suggests there are three major economic factors determining which countries might choose to form FTAs with each other. First, to the extent that consumers prefer variety among differentiated goods, which are produced under increasing returns, joint market size and similarity should be important, as discussed in Chapters 12 and 16. Second, within an FTA relatively large differences in capital-labor endowments should expand specialization and welfare. Third, countries in close proximity should have lower transport costs and enjoy larger net FTA gains. A further observation is that what really matters is the scale of these factors relative to those in the rest of the world. For example, in MERCOSUR there are variations in capital and labor abundance between Brazil and Uruguay, but these may be small compared to those between Brazil and the United States, tending to limit the intra-FTA specialization gains.

These factors were studied closely in an important paper by Baier and Bergstrand (2004). They

[5]Krugman himself noted this possibility in a later paper.

developed a simulation model that supports the following hypotheses. Net welfare gains among two FTA partners are more likely: (1) the lower the distance between them; (2) the further they are from the rest of the world (ROW); (3) the larger is the sum of their economic outputs (GDPs); (4) the more similar they are in economic size; (5) the smaller is the economic size of ROW; (6) the greater is the difference in factor endowments between them; and (7) the smaller is the difference between their relative endowments and those of ROW. In turn, countries with these characteristics should be more likely to be FTA partners. The authors studied all possible FTA pairings of 54 countries in 1996 and found strong empirical support for each of these hypotheses. Thus, the probability that two countries share an FTA is significantly more likely if they are closer together, jointly larger in size but close together in scale, and have markedly different capital-labor ratios.

Such evidence helps us understand why FTAs form. Consider NAFTA, for example, which is an agreement among Canada, the United States and Mexico. The potential gains to Canada and the United States stem largely from the large joint market size and similarity in incomes, both of which permit expansion of industrial production and economies of scale. The primary advantage to Mexico emerges from its relative labor abundance in the region, permitting more efficient production according to comparative advantage.

22.5 Empirical Evidence

A primary question about FTAs and CUs is whether they create more trade among members than they divert away from non-members. One common means of studyng this is to estimate a standard gravity model, as explained in Chapter 13, and enter dummy variables that indicate whether two countries both belong to a particular FTA. This was the approach in Frankel, Stein and Wei (1995), who studied the impact of the Asia-Pacific Economic Cooperation (APEC) grouping of countries, the European Community (EC) and NAFTA on bilateral trade flows. They found that membership in APEC and the EC significantly raised trade among members, but NAFTA did not, which is unsurprising since their data sample ended in 1995 and NAFTA came into being just one year earlier.

There are at least two significant problems with that approach. First, as noted above, membership in a FTA is not exogenously given, but rather an endogenous policy choice. This means that FTAs are likely to include countries that already have unusually high trade among themselves, perhaps due to close geographical proximity or strong historical ties. Thus, the error terms would be correlated with the dummy variables and the OLS coefficients of membership would be biased. Second, while the standard gravity model may give some indication about trade creation it is not suitable for computing trade diversion.

The most comprehensive analysis that took these factors into account is Magee (2008). He used data on bilateral trade flows among 133 countries over the period 1980 to 1998 to study both trade creation and trade diversion within a general gravity framework. For this purpose, he created two types of dummy variables. First, $RTA_{ijt}^{k} = 1$ if countries *i* and *j* both belonged to a preferential trade agreement in year *t*, while $TD_{ijt}^{k} = 1$ if one of the two countries was a member of agreement *k* and the other was not. For example, the United States and Canada are both in NAFTA and France and Germany are both in the EU but the United States and France do not share membership in a preferential trade arrangement. The coefficient on *RTA* should then be a measure of trade creation impacts and that on *TD*

should capture trade diversion effects. To manage the problem of endogenous FTA choice, Magee included fixed effects for each pair of countries. These "dyad" effects control for specific relationships among any two nations that do not change over time, such as distance, common language, adjacency and colonial or historical trading relationships.

Table 22.1 lists the primary results for the coefficients on RTA and TD.[6] The first column shows the basic equation, in which the RTA and TD variables are included only for the contemporaneous sample years. The coefficient on RTA is 0.597 and highly significant. This estimate implies that if two countries are in the same preferential trade agreement their joint bilateral trade increases by 82 percent.[7] In contrast, if one is in an agreement it reduces imports from a non-member slightly. However, these equations do not include the dyad effects, which are in the remaining regressions. In column two we see that the estimated trade creation coefficient falls to 0.352, suggesting a 42 percent increase in intra-bloc trade, while the trade diversion coefficient becomes virtually zero and insignificant. Thus, estimated trade creation is large and significant, while trade diversion is absent in the data.

Table 22.1

The remaining columns pose the interesting question of how long it takes these trade effects to occur. Because trade agreements tend to be negotiated for years it is possible that firms could expand their international commerce in anticipation of their enactment. Thus, the "before" coefficients are the sum of the impacts in the four years prior to implementation. Furthermore, it generally takes time for firms to adjust their trade fully after agreements are put into place. Thus, the "after" coefficients are the sums of these FTA and TD impacts over the 18 years from implementation date. Finally, the "cumulative" effects add together the before- and after-implementation effects. As may be seen, the dynamic trade creation of joint membership in the average trade agreement is powerful: the coefficient of 0.638 implies that the cumulative impact is to raise trade by 89 percent. This is the combination of a significant anticipatory increase prior to implementation and a larger increase after the agreement is enacted. In contrast, there are no significant trade diversion impacts, even in this dynamic specification. The final three columns repeat this analysis but break the agreements down into customs unions, free trade areas and partial preferential trade agreements. It seems that CUs establish the highest intra-bloc long-run cumulative trade creation (estimated at 129 percent). The effect of FTAs is also significantly positive, though surprisingly the trade growth comes largely in anticipation of their enactment. Partial PTAs have smaller and statistically insignificant effects. Again, all estimates of trade diversion are essentially zero.

These findings strongly support the view that FTAs and CUs are generally trade-creating and benign in their effects on non-member countries. Before wholly accepting that conclusion, however, some shortcomings of the analysis should be noted. First, trade diversion is measured only as a function of one country being in a PTA and the other being out, ignoring the potentially important effects of the size of a specific PTA on external countries. For example, the European Union is a very large CU and it

---

[6]To conserve space, we do not show coefficients for the standard gravity variables, such as GDP. Because there are additional problems with OLS estimation when the dependent variable could be the log of zero for some bilateral observations and also exhibits significant heteroskedasticity, the author used a nonlinear estimation technique called Poisson pseudo-maximum likelihood.

[7]This is computed from the formula $e^{0.597} = 1.82$, since the dependent variable is the log of bilateral trade.

is likely that the trade diversion effects of having many members with mutual preferences are not captured by simple bilateral dummy variables. Second, our earlier theoretical discussion pointed out that a critical determinant of trade impacts is the degree of substitution between goods produced within the bloc and goods from outside the bloc. Properly assessing that factor is impossible with aggregate trade data, rather than detailed product information.

Romalis (2007) developed a product-level model in order to study the trade impacts of NAFTA with detailed data. On the demand side, consumers perceive goods from each country to be differentiated. Thus, US, Japanese and German cars are differentiated goods that enter into a love-of-variety utility function with substitution elasticity $\sigma > 1$, with similar treatment for other industries. There are transport costs between nations for each good and countries impose *ad valorem* tariffs on imports. When a subset of countries decide to eliminate tariffs on imports from FTA partners, consumers substitute toward goods produced within the bloc and away from imports. On the production side, Romalis permits export-supply functions from non-members to have positive elasticities that must be estimated. This is important for assessing the terms-of-trade effects of NAFTA.

Romalis assembled data on around 4,000 product categories for the period 1980 to 2000, which covered years both before and after NAFTA was implemented in 1994. The first step in the analysis was to estimate the elasticity of substitution for each NAFTA member. This was possible by computing the ratio of, for example, Canada's imports in the United States to Brazil's imports in the United States for each product and doing the same for those shares in a comparison country, the EU. With NAFTA the United States offered Canada lower tariffs but not Brazil. By comparing the impact on the Canada-Brazil US import ratio with its effect on the Canada-Brazil EU import ratio, and doing so for every excluded country (like Brazil), Romalis could estimate the implied substitution elasticities across goods and countries.

Because such elasticities are of great interest in trade theory we report particular estimates in Table 2.2, Panel A. These estimates use the schedule of bound tariffs averaged across product lines between 1989 and 1999. The figure in the first column suggests that the elasticity of substitution in the United States between Canadian imports and goods from non-NAFTA countries was 6.68, while that between Mexican imports and other imports was 10.15. These high estimates suggest that American consumers substitute easily among sources of imported goods, which should have a significant impact on trade flows when tariffs are cut preferentially. In Canada the corresponding figures were 2.85 (from US) and 7.32 (from Mexico), while in Mexico they were 2.50 (from US) and 0.77 (from Canada). Thus, the smaller economies in NAFTA tend to have lower substitution parameters, particularly regarding imports from the United States.

Table 2.2

The second step was to estimate the supply elasticity of exports from outside NAFTA, which Romalis did using detailed US import data and a series of strong assumptions about production conditions abroad. Using US tariff rates as instrumental variables for production in each country, he computed a supply elasticity across all goods of approximately 1.9. This suggests that exports to the United States were relatively elastic during that period. However, 1.9 is far less than an infinite elasticity, suggesting that the preferential tariff cuts in NAFTA would reduce world supply prices through trade diversion and modestly improve the American terms of trade. Note, however, that his decision to apply this rate to all products from all non-NAFTA nations is questionable.

With these demand and supply characteristics in place, Romalis went on to compute the implied impacts on trade volumes and welfare in each NAFTA member and the rest of the world. These computations, listed in Panel B of Table 2.2, are far more pessimistic about the potential gains from the agreement.[8] Note first that substantial changes in bilateral trade volumes within the bloc were estimated, with US-Canada trade rising by 3.7 percent, US-Mexico trade by over 23 percent, and Canada-Mexico trade by over 24 percent. These computations suggest significant scope for trade creation within NAFTA, at least in terms of trade volumes and higher outputs. However, consistent with the relatively high substitution elasticities in Panel A, there was also trade diversion found, as indicated by the reductions in trade volumes with ROW. In turn, this trade diversion asserted itself in the welfare calculations from implied reductions in tariff revenues in each NAFTA member. As noted in the final portion of the table, US output gains were offset by revenue losses, leaving no net welfare gain despite the improvement in the country's terms of trade. Canada was estimated to experience a small welfare loss and Mexico an even larger loss at 0.3 percent of GDP. Note that the reductions in trade were not enough to change ROW output or welfare levels appreciably. Romalis concluded that, at least within his study period ending in 1999, NAFTA generated enough trade diversion that the United States observed little or no net gain, and its partners suffered small net losses, while not affecting ROW.

The differences between these two studies are stark and deserve comment. Both find that FTAs generate large positive effects on within-bloc trade volumes. This fact in itself is often used by some politicians to trumpet the benefits of regional trade agreements. However, as Romalis reminds us, larger trade volumes do not necessarily imply overall welfare gains when analysts also consider the price and revenue implications.[9] It is not possible to compute welfare impacts of CUs and FTAs in Magee's analysis.

A second important note is that Romalis' finding of very small welfare changes – less than 0.5 percent of GDP in all cases – is actually typical of traditional economic analyses of FTAs. Numerous studies, both using econometrics and computational equilibrium methods, generally have predicted that various regional trading agreements would have only minimal impacts on overall GDP levels and economic well-being. In that context, it is puzzling that trade policymakers in so many countries tout the benefits of FTAs and spend so much time negotiating them. We noted earlier the remarkable recent proliferation in the worldwide number of such agreements.

One potential answer to this puzzle is that a small overall impact disguises the possibility that FTAs have large effects on individual industries and the potential winners may have significant lobbying power to push negotiators. A more comprehensive answer is that governments find it easier to agree on trade concessions when negotiating with a small number of countries than in the multilateral WTO context. Thus, basic political-economy pressures may explain much of the enthusiasm for regional trade agreements.

A second answer is that the traditional analysis may be highly misleading. The primary reason is

---

[8]These impacts actually reflect the combined impacts of NAFTA (implemented in 1994) and the earlier bilateral deal between the United States and Canada, the Canada-US Free Trade Agreement, or CUSFTA (implemented in 1989).

[9]Interested students could read the analysis by Chang and Winters (2002), who directly studied the effects of MERCOSUR on internal prices and prices of goods from outside the bloc. They found that foreign exporters dropped their prices considerably in Brazil after MERCOSUR was formed.

that the vast majority of studies, including those by Romalis and Magee, assume that economies are largely built on perfect competition or large-numbers monopolistic competition and that market structures and factor endowments are static. If, instead, some industries are characterized by economies of scale and imperfect competition there could be significant rationalization and pro-competitive gains from FTAs. Further, if the implementation of trade agreements changes economic conditions sufficiently that trade partners attract new capital and firms engage in more innovation the dynamic effects could be far larger than the static impacts.

These points were made forcefully by Kehoe (2005), who compared the actual increases in trade within NAFTA to those predicted by prominent CGE models published as the agreement was being negotiated. Kehoe found that the actual increases in trade volumes between 1988 and 1999 were far larger than those predicted. For example, Brown, Deardorff and Stern (BDS; 1992) computed that NAFTA would induce US exports to rise by 2.3 percent of GDP but the actual rise in the data was 30 percent. Similar growth figures for Mexican exports over GDP were 51 percent (predicted) but 141 percent (actual) and for Canada 4.3 percent (predicted) and 53 percent actual. These differences were even greater at the sectoral level. The BDS model predicted that Mexico's exports to the United States of transport equipment, as a percentage of GDP, would rise by 6.2 percent while the actual ratio more than tripled.

From such comparisons we conclude that actual trade and production flows may be considerably more responsive to trade reforms than basic static models would suggest. In turn, potential welfare impacts may be larger. What factors could explain this difference? There are several hypotheses and all are essential subjects of current economic research. First, there may be pro-competitive gains from breaking down monopolies. These gains can make the broader economy more efficient and expand production, with large impacts on welfare (Konan and Maskus, 2006). Second, larger markets generated by FTAs can raise domestic investment and attract inflows of capital, which can expand productivity (Kehoe, 2005). Third, market expansion also encourages new firms to enter and existing firms to develop new products (Rutherford and Tarr, 2002). This expansion in the *extensive margin* of trade seems especially important in NAFTA (Kehoe, 2005). These issues are likely far more important than static trade creation and trade diversion in assessing the long-term impacts of trade agreements, a claim that needs much more research by economists.

REFERENCES

Baier, Scott L. and Jeffrey H. Bergtrand (2004), "Economic Determinants of Free Trade Agreements," *Journal of International Economics* 64, 29-63.

Brown, Drusilla K., Alan V. Deardorff, and Robert M. Stern (1992), "A North American Free Trade Agreement: Analytical Issues and a Computational Assessment," *The World Economy* 15, 11-30.

Chang, Won and L. Alan Winters (2002), "How Regional Blocs Affect Excluded Countries: The Price Effects of MERCOSUR," *American Economic Review* 92, 889-904.

Deardorff, Alan V. and Robert M. Stern (1994), "Multilateral Trade Negotiations and Preferential Trading Arrangements," in Alan V. Deardorff and Robert M. Stern, editors, *Analytical and Negotiating Issues in the Global Trading System*, Ann Arbor: University of Michigan Press.

Frankel, Jeffrey A., Ernesto Stein, and Shang-Jin Wei (1995), "Trading Blocs and the Americas: The Natural, the Unnatural and the Supernatural," *Journal of Development Economics* 47, 61-95.

Kehoe, Timothy J. (2005), "An Evaluation of the Performance of Applied General Equilibrium Models on the Impact of NAFTA," in Timothy J. Kehoe, T.N. Srinivasan and John Whalley, editors, *Frontiers in Applied General Equilibrium Modeling*. Cambridge: Cambridge University Press.

Kemp, Murray C. And Henry Wan Jr. (1976), "An Elementary Proposition Concerning the Formation of Customs Unions," in Murray C. Kemp, editor, *Three Topics in the Theory of International Trade: Distribution, Welfare, and Uncertainty*. Amsterdam: North-Holland.

Konan, Denise E. and Keith E. Maskus (2006), "Quantifying the Impact of Services Liberalization in a Developing Country," *Journal of Development Economics* 81, 142-162.

Konan, Denise E. and Keith E. Maskus (2011), "Preferential Trade and Welfare with Differentiated Products," *Review of International Economics*, forthcoming.

Kowalczyk, Carsten (2000), "Welfare and Integration," *International Economic Review* 41, 483-494.

Krishna, Pravin and Arvind Panagariya (2002), "On Necessarily Welfare-Enhancing Free Trade Areas," *Journal of International Economics* 7: 353-367.

Krugman, Paul (1991), "Is Bilateralism Bad?" in Elhanan Helpman and Asaf Razin, editors, *International Trade and Trade Policy*, Cambridge, MA: MIT Press.

Magee, Christopher S. P. (2008), "New Measures of Trade Creation and Trade Diversion,"

Journal of International Economics 75, 349-362.

Riezman, Raymond (1979), "A 3 X 3 Model of Customs Unions," *Journal of International Economics* 9, 341-354.

Romalis, John (2007), "NAFTA's and CUSFTA's Impact on International Trade," *Review of Economics and Statistics* 89, 416-435.

Rutherford, Thomas F. and David G. Tarr (2002), "Trade Liberalization, Product Variety and Growth in a Small Open Economy: A Quantitative Assessment," *Journal of International Economics* 56, 247-272.

Viner, Jacob (1950), *The Customs Union Issue*. New York: Carnegie Endowment for International Peace.

Table 22.1 Estimated Impacts of PTAs on Bilateral Trade

	Dependent variable: bilateral real trade flows, 1980-99					
Variable	*all PTA*	*all PTA*	*all PTA*	*CU*	*FTA*	*PTA*
RTA	0.597[a]	0.352[a]	0.010	0.108[a]	-0.032	-0.015
TD	-0.029[a]	-0.003	-0.002	-0.003	-0.004	0.000
Cumulative RTA			0.638[a]	0.829[a]	0.508[a]	0.169
Before RTA			0.231[a]	0.050	0.367[a]	-0.003
After RTA			0.407[a]	0.779[a]	0.141	0.172
Cumulative TD			-0.002	-0.010	-0.004	0.039
Before TD			0.007	0.002	0.027	0.019
After TD			-0.009	-0.012	-0.031	0.020
Dyad fixed effects	No	Yes	Yes	Yes	Yes	Yes

Note: [a]indicates a variable is significantly different from zero at the 1% level.
Source: Magee (2008, Table 1)

Table 22.2 Substitution Elasticities and Welfare Impacts of NAFTA

Panel A: Estimated Elasticities of Substitution

US-EU import shares		*Canada-EU import shares*		*Mexico-EU import shares*	
Canada	*Mexico*	*US*	*Mexico*	*US*	*Canada*
6.68	10.15	2.84	7.32	2.50	0.77

Panel B: Estimated Trade and Welfare Impacts of NAFTA [a]

Percentage Change in Bilateral Trade Volumes

US-Canada	*US-Mexico*	*US-ROW*	*Canada-Mexico*	*Canada-ROW*	*Mexico-ROW*
3.74%	23.18%	-0.30%	24.49%	-1.49%	-9.40%

Percentage Change in Welfare (% of GDP)

Source	*United States*	*Canada*	*Mexico*	*ROW*
Output	0.04%	0.28%	1.09%	0.00%
Tariff Revenue	-0.04%	-0.31%	-1.39%	NA
Welfare	0.00%	-0.03%	-0.30%	0.00%

Note: [a] Sum of CUSFTA and NAFTA impacts.
Source: Romalis (2007, Tables 3 and 5)

Figure 22.1

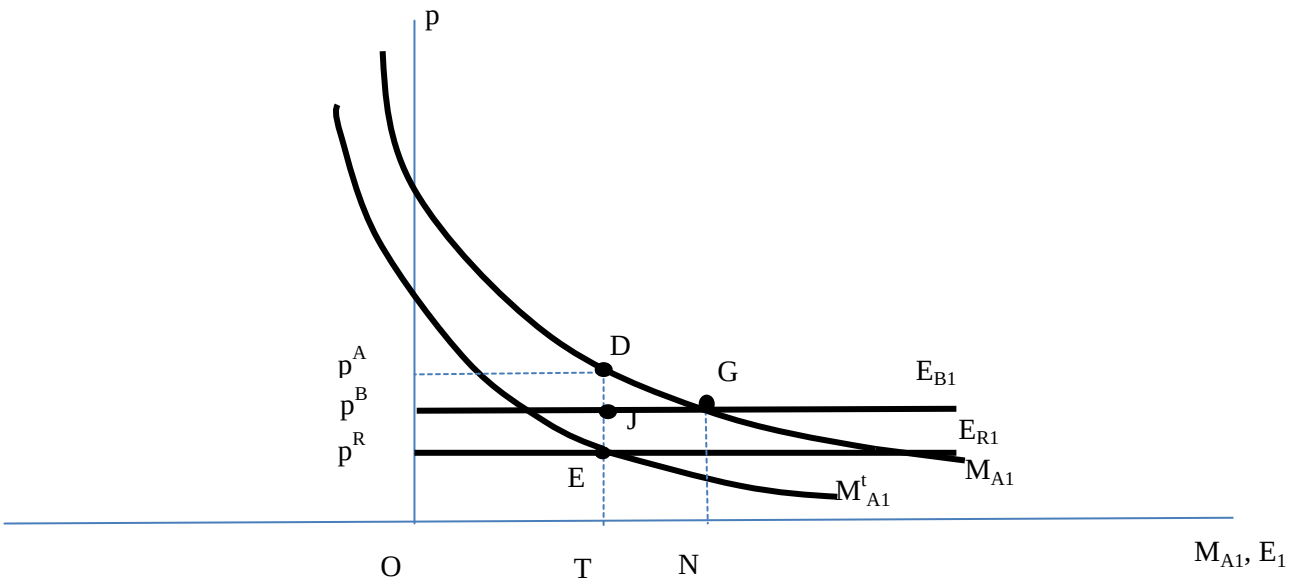


Figure 22.2 Panel A

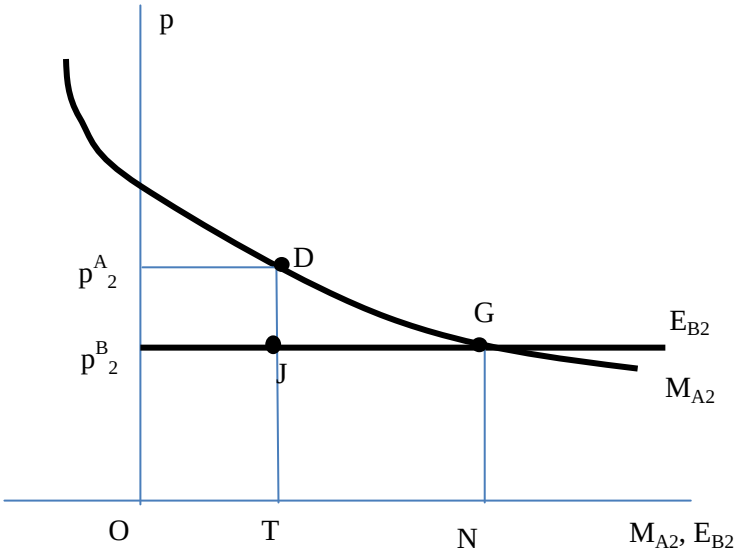


Figure 22.2 Panel B