'Sutton blazes his own trail. Distilling many years of careful observation of how firms actually behave, Sutton begins with simple premises but draws out surprising implications. This book artfully brings together two strands of economic thinking that have too often proceeded independently, the industrial-organization literature on market structure and the trade literature on the quality choices of heterogeneous firms. In placing firms' investments in capabilities at the center of the story, this book points the way to the future of research on globalization and development.'

Eric Verhoogen, *Associate Professor, Department of Economics and School of International and Public Affairs, Columbia University*

'Sutton's concept of a firm's capability - a two-dimensional measure of productivity and quality - promises to be a compact and fruitful way of describing and explaining many diverse phenomena in industrial organization and international trade. This book, with its rich mix of theory and facts viewed through this lens, is a valuable primer and toolkit for researchers in these fields.'

Avinash K. Dixit, *John J. F. Sherrerd '52 University Professor of Economics, Emeritus, Princeton University*

This book offers a new perspective on the economics of globalization, based on the concepts of firms' capabilities as the immediate cause of countries' wealth. It presents new ways of looking at the way China, India, and Africa have been drawn into the global economy over the past two decades. It offers new perspectives on some of the most central questions in the current debate: What effects does the rise of China have for the advanced industrial economies? Why have some industries adapted quickly and effectively to the changing global scene, while others have not? How were the 'Transition Economies' of Eastern Europe affected by trade liberalization? How have the economic prospects of sub-Saharan African countries changed over the past decade? This analysis contributes to the recent literature on quality and trade, which is providing a new and different approach to the analysis of globalization, and which focuses on those economic mechanisms that are central to the current wave of this centuries-old phenomenon.

OXFORD
UNIVERSITY PRESS

www.oup.com

ISBN 978-0-19-927453-6

9 780199 274536

Sutton COMPETING IN CAPABILITIES OXFORD

Clarendon Lectures in Economics

John Sutton

COMPETING IN CAPABILITIES

The Globalization Process

COMPETING IN CAPABILITIES

OUP UNCORRECTED PROOF – REVISES, 3/10/2012, SPi

Competing in Capabilities

The Globalization Process

JOHN SUTTON

OXFORD
UNIVERSITY PRESS

OXFORD
UNIVERSITY PRESS

Great Clarendon Street, Oxford, OX2 6DP,
United Kingdom

Oxford University Press is a department of the University of Oxford.
If furthers the University's objective of excellence in research, scholarship,
and education by publishing worldwide. Oxford is a registered trade mark of
Oxford University Press in the UK and in certain other countries

© John Sutton 2012

The moral rights of the author have been asserted

First Edition published in 2012
Impression: 1

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, without the prior permission in writing of Oxford University Press, or as expressly permitted by law, by licence or under terms agreed with the appropriate reprographics rights organization. Enquiries concerning reproduction outside the scope of the above should be sent to the Rights Department, Oxford University Press, at the address above

You must not circulate this work in any other form
and you must impose this same condition on any acquirer

British Library Cataloguing in Publication Data
Data available

ISBN 978–0–19–927453–6

Printed in Great Britain by
~~CPI Group (UK) Ltd, Croydon, CR0 4YY~~

For Jules, Nick, and Will

OUP UNCORRECTED PROOF – REVISES, 3/10/2012, SPi

OUP UNCORRECTED PROOF – REVISES, 3/10/2012, SPi

Acknowledgments

The origins of this volume began with my Keynes Lecture of 2000 to the British Academy (Sutton 2001) which I developed further as the Clarendon Lectures of 2004. My Presidential Address to the Royal Economic Society, which forms the basis of Chapters 1 and 3 below, and my recent joint paper with Dan Trefler which forms the basis of Chapter 2, allowed me to crystallize and develop some ideas set out in the Lectures.

I would like to thank Ana Valero, Clemens Von Oertzen, Qi Zhang, and the students in my 'Globalization and Strategy' course of 2011–12 for their careful reading of the draft version of this volume.

And one final thank you to all my colleagues in the 'Capabilities' lunch group, for the best conversations I have had on this subject.

OUP UNCORRECTED PROOF – REVISES, 3/10/2012, SPi

Contents

List of Figures

List of Boxes

OUP UNCORRECTED PROOF – REVISES, 1/10/2012, SPi

History and Theory

The globalization process we have witnessed over the past twenty years is just the latest act in a long-term drama. The first great wave of globalization occurred in the latter half of the nineteenth century; the second took off in the years following the Second World War; the current act began with the opening up of China and India and the collapse of the Soviet Union.

All three phases have much in common: the growth of trade volumes, the flow of international capital, and the ramifications that these bring. Yet each act has been distinct in terms of the most salient features of the process – and each has led, in its turn, to a new body of economic analysis concerned with its interpretation. Great events cast a long shadow: history shapes theory.

GLOBALIZATION IN HISTORY

The second half of the nineteenth century witnessed a massive opening up of world markets. The spread of the railways and improvements in shipping led to falling transport costs. Tariff barriers were falling in the industrialized world. Huge flows of capital moved across the globe, while improvements in shipping and the developments of rail networks drastically reduced transport costs, leading to a boom in the trade of commodities and manufactures.

The great drivers of trade in this period, as contemporaries well understood, were 'differences in factor endowments': the land-rich Americas exported agricultural produce to Europe, which in turn exported manufactured goods to the Americas. This process of trade based on comparative advantage between parts of the world in which the ratio of land to labour (or fixed capital stock) differed substantially, was familiar since the days of David Ricardo. Such trade tends to bring about a convergence in factor prices: wage rates in the new world, as compared with the old, would rise as a result of the increased demand for labour driven by the new trade flows.

The codification of the economics underlying this process did not come about until much later. The standard formulation is due to Heckscher and Ohlin, both of whom wrote their key works in the 1920s. Their analysis of the links from factor endowments to commodity and factor prices provided testable predictions of the theory: but data was lacking. Only very recently, in the work of O'Rourke and Williamson (2001), has a wide-ranging and systematic analysis of the predicted 'factor price convergence' been attempted. The results, subject to caveats about limitations imposed by data availability, suggest strongly that the classical Heckscher–Ohlin models capture the primary mechanisms driving the process in this period: it was indeed an era of factor-endowment driven trade.

THE POST SECOND WORLD WAR ERA

The wave of globalization that followed the Second World War began with the establishment of new international institutions (the IMF, the World Bank, and GATT) and the formation of the European Economic Community. It was marked by a major lowering of tariff barriers and a huge increase in the volume of international trade. But what struck many economists at the time as the most remarkable feature of the process was the role of intra-industry trade between very similar industrial economies. In other words, the paradigm case involved the export of German cars to France, and French cars to Germany – the trade was 'within the industry', in contrast to the pattern captured in the Heckscher–Ohlin framework. Differences in factor intensity between France and Germany were extremely small: the driver of this 'intra-industry' trade was clearly different in kind.

The dominance of this 'intra-industry' type of trade within the newly established European Economic Community was demonstrated by Balassa (1967). By the 1970s some models of this type of trade had been proposed Grubel and Lloyd (1975), but it was the publication of

Paul Krugman's article of 1979 that sparked off a new literature based on 'monopolistic competition' models. In these models, competing products offered by different firms were 'horizontally differentiated', i.e. if they were offered at equal prices, some consumers would prefer one variety while others would choose a different variety.[1] The production of any product ('variety') involved both a fixed cost, and a constant marginal cost ('increasing returns'). Under free entry, a number of firms offering different products would be present in the market; their equilibrium prices would exceed marginal cost, and their gross profit margins would suffice to cover their fixed (or sunk) costs.

Now if two economies of this kind are joined via free trade, all consumers gain access to the varieties produced in both countries. This means lower prices, as a larger number of firms come into competition with each other. The new long-run equilibrium number of firms will be less than the total number existing across both countries in the preliberalization set-up; each firm will have a higher sales volume, however, so its average cost (including the contribution to fixed costs) will be correspondingly less – and it is this 'exploitation of scale economies' that constitutes both the driver of the trade flows, and the source of welfare gains from trade in this setting.

This, then, was the story economists came to tell about the drivers of trade in the post-war era. As with the Heckscher–Ohlin model, this view took some time to become codified and accepted; it was not until the 1980s that it became the standard way of looking at these issues.

One feature of these 'intra-industry' trade models that cast a long shadow in the trade literature, was their reliance on a 'monopolistic competition' framework. In such a setting, any fixed or sunk costs of entry are exogenous (i.e. their size is determined by technological considerations, outside the control of firms). A central implication of this, is that as the global market gets larger, more and more firms enter each industry, so all industries can become fragmented (i.e. occupied by a large number of firms, each of whom has a very small market share). This is precisely the feature that seems a poor representation of many global industries – such as aircraft, pharmaceutical, or video games – and it is this that will form our point of departure in Chapter 1.

In what follows, we turn to the third of our 'waves of globalization', which began with the opening up of the Chinese and Indian

[1] This is in contrast to 'vertical product differentiation', i.e. differentiation by quality. In this latter setting, if two goods whose quality differs are offered at the same price, then all consumers will choose the same ('high-quality') good.

economies, and with the collapse of the Soviet Union. What appears to be the dominant feature of this recent era, is neither 'factor differences' driven trade (though this certainly forms an important element in the story developed below), nor is it 'trade between equals'. Rather, what seems most striking here, is the opening up of trade between parts of the world in which the initial levels of industrial capabilities were widely different. This suggests taking a different point of departure in modelling the process.

THE CURRENT PHASE

The most salient feature of the process we have seen at work over the past twenty years is the transformation of productivity and quality levels, and the widening of the range of products offered on international markets, by China, India, and the economies of Eastern Europe. The roots of this process are various. In some cases, the widened opportunities available to firms in these countries stimulated local firms to respond. More typically, partnerships of one kind or another with foreign firms, or the entry of foreign firms operating plants alone, was the key vehicle. But whatever the mechanism, the outcome was a striking transformation from a population of firms that hugely lagged behind global standards of productivity and quality, to one which could compete effectively in global markets. Central to this story was the role played by 'quality', an element long relegated to the footnotes of International Economics. But over the past decade, a new literature has emerged on 'Quality and Trade', which is deeply concerned with these issues. Meanwhile, economists have begun to build bridges towards their Business School colleagues, who have long grappled with the closely intertwined question of firms' 'capabilities'. It is these lines of analysis that I want to explore in what follows.[2]

[2] A large literature on trade and quality has developed in recent years: see for example Feenstra (1984), Schott (2004), Hummels and Klenow (2005), Hallak (2006), Verhoogen (2008), Hallak and Schott (2011), Kugler and Verhoogen (2012), Khandelwal (2010). These authors as well as Hummels and Skiba (2004), Schott (2008), Choi, Hummels, and Xiang (2009), Johnson (2012), and Baldwin and Harrigan (2011) examine the relationship between trade flows and quality. Verhoogen (2008) explores the relationship between quality, trade, and inequality, as do Goldberg and Pavcnik (2007). Amiti and Khandelwal (2009) explore the impact of trade restrictions on quality upgrading. Hallak and Sivadasan (2009) and Bastos and Silva (2010) provide

The novelty of the present treatment, relative to the 'Quality and Trade' literature, is that, by developing a framework that has proved useful in the Industrial Organization literature, the 'Cournot Model with Quality', it builds a bridge from the IO literature both to the Management literature on Capabilities, and to the Trade literature. One key feature of this framework is that it *endogenizes* the notion of scarce (and so valuable) capabilities, showing *why* it must always be the case that some capabilities remain relatively scarce at equilibrium. This idea, central to the 'market structure' literature in Industrial Organization, will be our point of departure in Chapter 1.

Each of the three chapters that follow begins by introducing a single key assumption, and what follows is a working out of its implications. These three key assumptions, taken in isolation, may seem innocuous. The first says that 'bad products cannot drive out good', i.e. a high-quality product will hold onto some minimal market share no matter how many low-quality products it competes with. This assumption leads, by the end of Chapter 1, to the conclusion that the capabilities required to produce high-quality products will inevitably be 'scarce': the global market will be dominated by a small number of firms.

The key assumption driving Chapter 2 is that scarce capabilities tend to be clustered (in terms of geography). This assumption is familiar from the Geography and Trade literature – but here the agenda is different. By combining it with the key assumption of Chapter 1, we arrive at a theory of the currently popular 'product mix diagrams' that link a country's wealth to the mix of goods it produces and exports.

In Chapter 3 we introduce the last, and least controversial, of our key assumptions: you can't make something out of nothing. In other words, not all costs are labour costs. All manufactured exports require for their production some material inputs, whether in the form of raw material, components, or bought-in sub-assemblies. The central result of Chapter 3 is that this simple feature is enough to drive a wedge between the economics of 'productivity' and the economics of 'quality' . . . and in so doing, it makes clear why the new literature on Quality and Trade actually matters.

As the analysis of globalization continues through Chapters 4 and 5, we will be drawn into issues that require a deeper discussion of

insights into the role of quality and productivity heterogeneity. Finally, Grossman and Helpman's (1991) quality-ladder model provides a dynamic link between quality, trade, and growth.

'capabilities' than that set out in Chapter 1, and it is at this point that we ask: what are the basic determinants of firms' capabilities? How can these capabilities be developed? And how can they be transferred across firms and countries? And what do these answers imply about Industrial Development in a globalizing world?

1

Capabilities

1.1. THE WEALTH OF NATIONS

Ever since Adam Smith wrote the Wealth of Nations, one of the central questions of economics has been, why are some countries richer or poorer than others? Answers abound: in the 1960s, the standard response was that rich countries had more capital stock (i.e. 'machinery') than poor countries; and this larger capital stock had its origin in the savings made by earlier generations. In the 1990s, the standard response was that rich countries had a better set of legal and social institutions, which provided an environment within which businesses could thrive, and the roots of these institutions ran deep into a country's history. The popularity of different answers sometimes reflected the salience of recent experience: the 1990s focus on institutions grew out of the apparent failure of the former Soviet Union countries to respond to the overnight liberalization of (capital) markets introduced at the beginning of the decade. Some answers were better than others: the 'capital stock' story is not without merit, but if that was the whole story, the poverty of Sub-Saharan Africa could have been solved three times over by devoting half a century of aid money to the purchase of machines. One of the most poignant moments on British TV came in the middle of Peter Jay's series on 'The Road to Riches'; see Jay (2001). Walking through a Tanzanian cashew nut factory built with World Bank money, so that home-grown produce could be processed locally, he found himself alone, among the perfectly functioning machinery.

So what of all the explanations? Clearly, we're in complex terrain here. Many things matter, and their separate influences may be deeply intertwined. But there is one way forward that allows us to separate out two stages of the analysis in a useful way.

My point of departure lies in distinguishing between *proximate* and *ultimate* causes of differences in wealth. The proximate cause lies, for the most part,[1] in the capabilities of firms. The ultimate causes, whether they be human capital, institutions, or otherwise, lead to economic growth through one central channel: by raising the volume and quality of output produced by the country's firms from the inputs available to them. This is almost a truism, since all that's involved in this claim is a restatement of the old adage: productivity determines GDP per capita. So let's start there …

1.2. THE WIDGET MAKERS

Imagine that the world consists of a bunch of different countries, in each of which people spend some of their time making widgets. Everyone in the world loves widgets; it's the only thing they spend money on. Every country has its own colour, blue or green, and its citizens can make only that colour. Half of all the widgets made in the world are blue, and half are green.

But when it comes to 'consuming' widgets, everyone likes to have an even mix of colours. They'll swap a blue for a green, or vice versa, on a one-for-one basis, to get an even mix. But they don't mind all that much, and if they're asked to swap at a less favourable rate, they will say no.

The outcome is that widgets get swapped across country boundaries, until everyone has the mix they want. This constitutes 'trade'. Now for 'productivity': people in some countries are quicker or more industrious than others, and they produce more widgets in a year: their 'productivity' is higher. So is their wealth. The number of widgets they end up with after doing all their swaps is the same as they produced. For individuals, and for countries, productivity determines wealth.

Of course you could say I'm only talking about proximate causes. Why Mr Smith, or Lithuania, has a higher or lower level of productivity goes back to Mr Smith's difficult childhood, or Lithuania's cultural heritage. Yes, it does; and we'll talk about these deeper questions

[1] The only major qualifications are (i) that this statement takes as given each country's endowment of human capital and natural resources, and (ii) that the size and productivity of the non-market (public) sector of the economy also affects outcomes.

underlying productivity and quality in Chapters 4 and 5. But for now, productivity determines GDP/capita: we're talking 'only' about the proximate causes of wealth. The reason this is worth focussing on, is that whatever are the deep and 'true' causes of wealth differences, they can only affect wealth insofar as they affect productivity. Productivity is the channel through which these things affect wealth.

So what's missing in the world of widgets? A lot, obviously. But the bits I want to focus on are quite simple and straightforward: I want to expand the set of goods from the simple widgets to all kinds of goods. And I want to show that the features of goods that will matter to my present story are simple in the extreme. By the end of this chapter, we'll have arrived at a restatement of the link from productivity to wealth, within the richer framework of 'capabilities'. With that aim in mind, let's start with a firm that makes some superior class of widget …

1.3. THE FIRM

In a market economy, a firm's viability depends on its earning a flow of profit that is at least as valuable as the costs it has incurred in establishing its business. Those costs are of many kinds: building a factory, inventing or developing a product, even building up a brand image. All that is in the past; we take a snapshot of the firm as a functioning entity, competing in the marketplace with its very superior widgets. What determines the flow of profit it can generate? Only two things can matter, once we've reduced things to this simple scenario: one is the number of labour hours it takes to make a widget; we'll assume, realistically, that that's a constant, independently of the number of units produced: call it c. Then, if w denotes the hourly wage rate, the production cost of a widget is wc. It is useful to take the number of widgets produced per hour, which is $1/c$, as our measure of the firm's productivity level.

How much can you sell a very superior widget for? The higher the price you quote, the fewer you sell. Raising the quality of the widget means that at any given price, you'll sell more units. By the same token, for any sales volume you're aiming at, you can support a higher price. Quality, in other words, is a 'demand shifter'. Using the same language, productivity is a 'cost shifter'.

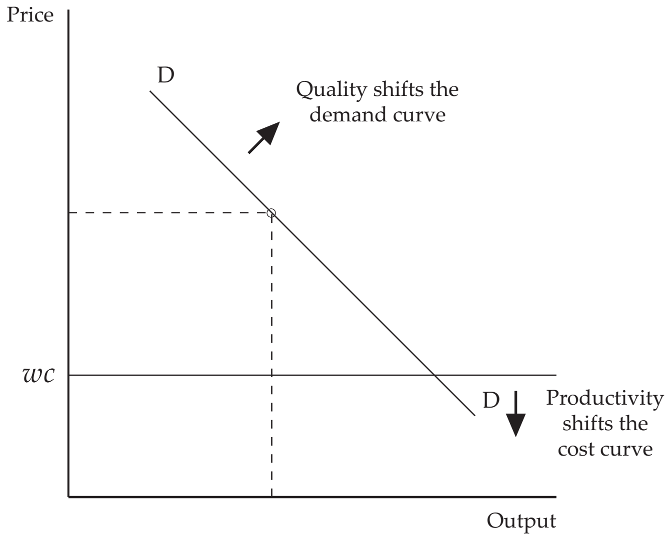

Fig. 1.1. Demand Shifters and Cost Shifters

It sounds simple, but it's a powerful formalism. It immediately throws a very broad meaning onto the word 'quality'. In all that follows, 'quality' refers to anything that shifts the demand schedule outwards: technical characteristics, after sales service, brand image, and so on … as we'll see in Box 1.2 below.

The great advantage of wrapping up these disparate elements of product quality into a 'demand shifter', is that (a) for many purposes, they all have the same effect, and lumping them together is harmless; and (b) we now get a simple statement: no matter how complex a market we insert this firm into, the only things that can affect its profit flow, and so the viability of this business, are the two parameters representing the cost shifter and the demand shifter. We'll refer to the cost shifter as productivity, and we'll represent it by $1/c$. (Sometimes we'll use the label c, rather than $1/c$, in the interests of keeping the notation simple.) The demand shifter will be referred to as 'quality', and we'll label it as u.

It's this pair of numbers that we'll refer to as the firm's 'capability' in the market for widgets. More generally, we'll refer to the firm's capability by asking what values u and c take in each of the markets in which it operates, or might operate. (This richer description in terms of a *set* of (u, c) values is explored in the next chapter.) We'll think of these markets as being narrowly defined ('widgets' rather than 'engineering components') so that the u and c can be thought of as referring to a specific product (or product line). But that's all that's to it for the moment.

When I use the word 'capability' in this way, I am using shorthand for what I might more properly refer to as the firm's 'revealed capability', i.e. the observable outcome of the firm's underlying capability. In the Managment literature, the term capability is used to refer to these deeper attributes of the firm, which are (often) not directly observable or measurable (Box 1.1).

Of course, it's tempting to ask at once about these deeper attributes: where do the u and c come from? We will reach this level of discussion in Chapter 4, but we begin in this chapter with the idea that it's the firm's past efforts and investments that have left it with its current values of u and c…and we ask, given some set of firms in today's market, with their various levels of u and c, how will competition between them pan out?

Box 1.1. **The Roots of Capabilities**

The term 'capability' is used in the main text as shorthand for what might better be called 'revealed capability', i.e. the performance-relevant outcome of the firm's underlying capabilities. This begs the question: what is the nature of these underlying 'capabilities'?

The primary point to be made here is that the 'interesting' elements of this underlying capability are those elements that can't be bought 'off the shelf', i.e. those to which all firms have free access on the market. This idea lies at the core of the 'Resource Based Theory of the Firm' literature, which traces its roots to Penrose (1959), via Rumelt (1984), Wernerfelt (1984) and Lockett, Thompson, and Morgenstern (2009).

More recently, the literature of 'organizational capital' has attempted to unearth the factors that lead one firm to achieve higher levels of productivity and quality than its rivals. (For a review, see Gibbons and Henderson (2011).)

These are difficult questions. One of the central problems in addressing the issues involved, is that differences in the performance (profitability) of firms will usually be driven both by factors internal to the firm (our focus of interest here), and by external factors, i.e. the different market environment in which different firms operate. Unravelling these two kinds of influence is notoriously difficult. To address this issue, some researchers have focussed on groups of closely comparable firms, between which differences in the external environment are both small, and identifiable (so that they can be controlled for in an econometric study). Thus, interest has focussed on such groups as restaurants belonging to the same chain (Gibbons and Henderson (2011)), or the set of all ready-mix concrete plants in a city (Hortaçsu and Syverson (2007), Syverson (2008)).

978–0–19–927453–6 02-John-Sutton-c01-drv John (Typeset by SPi) 12 of 129 October 1, 2012 10:3

OUP UNCORRECTED PROOF – REVISES, 1/10/2012, SPi

Box 1.1. **(continued)**

The latter setting is of particular interest here, as it illustrates one influence that can be central to performance differences. Ready-mix concrete plants supply concrete, shipped in trucks, to construction sites in a local area (usually a city). The concrete, once prepared, must be delivered quickly. The main attribute that distinguishes high-performing from low-performing depots is the effectiveness with which they manage the complex and demanding set of decisions about scheduling deliveries. As Syverson (2008) remarks: 'Ready-mix concrete producers are not just manufacturers, they are logisticians: they deliver, typically on short notice, a perishable product to time-sensitive buyers in multiple locations.' This is the task of a team comprising a handful of individuals, and the skill of one individual (team leader) is a crucial determinant of performance. So, in this instance, the capability of the 'firm' rests to a high degree on the skills of a single individual.

Moving one step up in terms of organizational complexity, we have the law firms studied by Garicano and Hubbard (2005, 2007, 2009). Here, lawyers of differing levels of ability are sorted into different roles, and a firm's capability rests on the 'architecture' of its organization.

It is the extension and elaboration of such ideas, in more complex settings, that underlies the recent literature on 'Organizational Capital' (Prescott and Visscher (1980), Atkeson and Kehoe (2005)).

This literature identifies one key source of a firm's value in the organizational structure of the firm, as opposed to its proprietory knowledge, or its market position. (In other words, with the assembling of a team of people who work effectively together, within some framework of rules, routines and tacit understandings that have been put in place or have evolved over time.) To see what this implies, consider, for example, the Aquafresh company in Ghana (Sutton and Kpenty (2012)). This company began life in the clothing and textiles sector, but when this sector came under intense competition from Chinese imports, the firm reinvented itself as a maker of soft drinks. Its expertise in clothing and textiles was secondary to the fact that it was a well-functioning midsize firm, that could reorient itself in the product market as market circumstances changed.

A more general perspective on this issue emerges from the work of Peter Schott, Andrew Bernard, Steven Redding, and Bradford Jensen, who have studied the way in which US industries adapt to competition from low-wage countries. Key to their results is that a primary survival mechanism lies in switching the balance of a firm's activities towards new product lines. See Bernard, Jensen, and Schott (2006), Bernard, Redding, and Schott (2010, 2011).

1.4. COMPETING IN CAPABILITIES

Suppose we have a number of competing firms, each characterized by its levels of c and u. When competition occurs between these differently placed firms, what happens?

Some will be 'active': they will produce some level of output, and have a price that exceeds their level of unit cost (and so have positive profit margins). The remaining firms will be 'inactive' in the sense that they will be unable to sell any quantity of output at a price sufficient to cover their unit costs of production. How does this happen? Consumers will buy a product only if it offers a price – quality combination that is as good as that offered by rival firms. So there will be some threshold in terms of the price – quality combination that firms must reach in order to survive. If a firm's quality is low, or its productivity is poor, then, even if it cuts its price to the level of its unit cost, pushing its profit margin to zero, it may still offer an unacceptable price – quality combination. Such a firm will remain inactive.

So we are left with the outcome where some sets of firms will be active, and their prices will vary with their respective quality levels, while their profit margins will vary both with their quality levels and their productivity (and so cost) levels.

How good a 'capability level' does a firm need in order to be active? Given the equilibrium prices ground out by competition between the active firms, there will be a curve in the space of quality, u, and productivity, $1/c$, above which a potential entrant to the market needs to be. An illustration of this threshold level is shown in Figure 1.2.

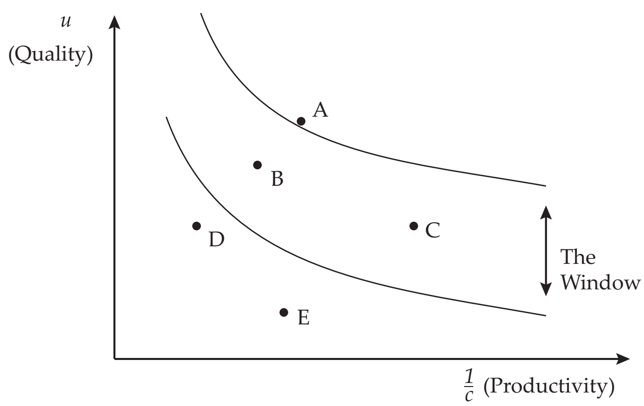

Fig. 1.2. The Capability 'Window'. The leading firm, labelled A, defines the top of the window. Its active rivals are B and C, while firms D and E are inactive. The process of competition between A, B and C determines equilibrium prices, and so fixes the bottom of the window, i.e. a curve in (u, $1/c$) space below which a firm will sell zero at equilibrium ('be inactive').

978–0–19–927453–6 02-John-Sutton-c01-drv John (Typeset by SPi) 14 of 129 October 1, 2012 10:3

OUP UNCORRECTED PROOF – REVISES, 1/10/2012, SPi

All active firms have a capability level that lies between the highest level in the market and threshold level needed for entry. I will refer to the band lying between these two curves as the 'window': to be an active firm, you need to get your capability level into this window. (Figure 1.2.)

1.5. THE FIRST KEY ASSUMPTION: BAD PRODUCTS CANNOT DRIVE OUT GOOD

Consider an industry comprising a number of firms, indexed by i, where c_i and u_i denote, respectively, the productivity and quality parameters of firm i. We describe the demand side of the market as follows: all consumers have the same utility function, of the form

$$U = (ux)^{\delta} z^{1-\delta} \tag{1.1}$$

Here, z denotes the individual's consumption of some 'outside good'; we focus on the good quantity purchased in our market of interest; here, the quantity purchased is x and the quality offered by the firm supplying the good is u. It follows from the form of the Cobb – Douglas utility function that consumers spend a fixed fraction $1 - \delta$ of their income on the outside good, independently of prices and qualities, and so the remaining fraction in the market is what concerns us here. It is convenient to denote total consumer expenditure in this market by a constant, S. The second feature of this utility function is that the consumer will choose the product, or one of the several products, that offer the best quality – price ratio.[2] It follows that, at equilibrium, all firms that are active, in the sense of having strictly positive sales revenue, must have the same, i.e. the equal-highest, quality – price ratio. This demand system also has a second, less obvious feature that will play a crucial role in all that follows.

The easiest way to introduce this feature is to consider a setting in which all the firms have the same productivity parameter, and so the same unit cost of production. Quality levels, however, are different, in that one firm has a 'higher' quality u while all the other firms

[2] To see this, note that spending S on firm i's product yields S/p_i units of quality u_i and so the bracketed term in (1.1) equals Su_i/p_i.

978–0–19–927453–6 02-John-Sutton-c01-drv John (Typeset by SPi) 15 of 129 October 1, 2012 10:3

OUP UNCORRECTED PROOF – REVISES, 1/10/2012, SPi

have a 'lower' quality $v < u$. Now imagine some form of competition between the firms. (We will be looking at Cournot competition, but for the moment any other form, such as Bertrand competition, for example, will do. The feature we're about to encounter is a feature of the demand system per se, and does not reflect the use of any particular form of competition between the firms.) The one thing we can say, for any form of competition in which firms are standard 'profit-maximizers', is that firms will not price below their unit cost of production, and so earn negative profits (as they will always have an available action that leaves them with zero output and zero profit). This in turn means that the 'worst' scenario that can be faced by the 'higher' quality producer is that all its 'lower' quality rivals sell at a price that exactly coincides with the level of unit costs (unit marginal cost), which is, as we assumed, the same for all firms. But this in turn means that our 'higher quality' firm will always enjoy a price strictly greater than this common level of marginal cost, and correspondingly (positive) levels of output, sales revenue, and market share (defined as its sales revenue divided by total industry sales revenue, which in this setting equals S, as we noted earlier).

Now let's step back from the 'worst case scenario', and ask, how could such a scenario be reached? One story of interest here is one in which the number of 'lower quality' rivals increases over time; and as they enter the market in greater numbers, competition between them forces their common price down to the level of their common marginal cost. This process of entry may damage the profits, and reduce the market share, of the 'higher quality' firm – but as our discussion of 'worst case scenario' shows, it cannot drive that firm's market share to zero. There is a minimal market share that the 'higher quality' firm must retain, no matter how many 'lower quality' rivals it faces.

In other words, a 'higher quality' producer cannot be squeezed out of the market by 'lower quality' rivals, no matter how many such rivals appear. Now this may seem obvious – and indeed it is so plausible and natural an account of how things work in practice that this whole discussion may seem belaboured. But viewed against the context of the modern literature, this point really is of central importance. To see why, suppose we started instead with the standard 'constant elasticity of substitution' (CES) utility function. Whether we use it in its original form, as introduced by Dixit and Stiglitz and adapted to the International Trade literature by Krugman, or its extended various forms; it carries the central property that *all* firms' market shares will be eroded to zero as the number of rival firms increases indefinitely.[3]

Now the property that good products cannot be squeezed out by bad is not special to the utility function (1.1) above. Rather, it is a general feature of the wide class of models that provide a basis for the modern theory of market structure (Sutton (2007a)). It is essential to any account of why many industries around the world remain dominated by a small number of firms, despite the huge size of the global market – as we will see below. It is also an extremely plausible, indeed compelling, property of actual markets. But on a purely theoretical level, it is a substantial assumption.

The central theme of this chapter lies in exploring the consequences of this assumption that 'bad products cannot drive out good' … and the central result of the chapter is that this leads to the first of our key implications: some capabilities will always, necessarily, be 'scarce'.

It is time to move on. Having put in place the demand system defined by (1.1), we need to complete our description of the model by specifying the form of competition. Here, I'll use the most conventional form, namely Cournot competition (i.e. a Nash equilibrium in quantities).[4] Each firm sets a quantity of output. The quantity may be zero, in which case we say the firm is 'inactive'. All active firms have prices proportional to their qualities, i.e. p_i/u_i is the same for all active firms. We use the label λ to denote this common value of p_i/u_i. Given the quantities set by the firms, the value of λ is set to equate supply and demand in the market – and it is this that fixes the level of prices in the market. Given this mechanism to fix prices, each firm takes rivals' quantities as given, and chooses its own quantity to maximize profit. (In other words, this is the standard story we tell about how a Cournot equilibrium works, except for the complication that firms differ in the quality of their products.)

So what happens?

[3] This model dominated the International Trade literature throughout the 1980s and 1990s. The central feature of Sutton (2007b) is that it departs from this framework.

[4] Combining (1.1) with Cournot Competition define the 'Cournot Model with Quality' introduced by Sutton (1991), on which all the analysis that follows is based.

Box 1.2. **Demand Shifters**

In 'short-run' analysis (Sections 1.2 to 1.4), we take as given the capabilities of the firm, as represented by its demand and cost shifter for each narrowly defined market (in which it may or may not operate). In 'long run' analysis (Section 1.9), we look at the investments the firm makes in moving its demand and cost shifters.

So long as we are confining attention to the short run then, the demand and cost shifters of all firms active in the market suffice in themselves to fully determine the flow of profit accruing to each firm. This offers a powerful general framework – since all we now need to know is the $(u, 1/c)$ pairs. On the other hand, it hides a multitude: for we have deliberately defined the cost and demand shifters in such a general way that their values may reflect a very wide range of influences.

While the cost shifter is relatively easy to conceptualize, the demand shifter is more complex. It is often convenient to refer to the demand shift variable u as 'quality', but this is just shorthand for perceived quality, i.e. consumers' willingness-to-pay. Thus it includes not just 'quality' in the usual narrow sense (a feature of the product's physical characteristics), but also a range of characteristics that include, for example:

- brand advertising: for example, a major Indian tractor manufacturer is currently working to establish its tractors in the US market. Success will hinge just as much on customer perceptions, as on technical performance.
- services: for a machine tool maker, the network of service engineers available to repair the machines may be of similar importance to buyers as the machine's engineering characteristics.
- logistics: for a maker of clothing in the Far East selling to a UK department store, the ability to alter designs and deliver consignments at very short notice is just as important as the physical characteristics of the garment itself.

1.6. THE WINDOW

To see what happens, we begin by calculating the equilibrium level of output, price, and profit of some particular firm, labelled firm i. Let the unit wage level[5] faced by firm i be denoted w_i, and the number

[5] If all firms operate in the same (competitive) labour market, then w_i will be the same for all, and we may drop the subscript i. However, we will be interested, in later chapters, in markets in which the competing firms operate in different countries, with different wage rates.

978–0–19–927453–6 02-John-Sutton-c01-drv John (Typeset by SPi) 18 of 129 October 1, 2012 10:3

OUP UNCORRECTED PROOF – REVISES, 1/10/2012, SPi

of units of labour (the only variable input) per unit of output be c_i, so that marginal cost is $w_i c_i$ and firm i's profit[6] at output level x_i is

$$\Pi_i = p_i x_i - w_i c_i x_i$$

Now since

$$p_j = \lambda u_j \text{ for all firms}, j = 1 \text{ to } n, \tag{1.2}$$

and since the value of output in the market, $\sum p_j x_j$, equals consumer expenditure in the market S, it follows that, using (1.2),

$$\sum p_j x_j = \lambda \sum u_j x_j = S$$

or

$$\lambda = \frac{S}{\sum u_j x_j}$$

whence

$$p_i = \lambda u_i = \frac{u_i S}{\sum u_j x_j} \tag{1.3}$$

and so firm i's profit can be written as a function of its output x_i.

Firm i sets x_i to maximize

$$\begin{aligned}\Pi_i &= p_i x_i - w_i c_i x_i \\ &= \left(u_i S / \sum u_j x_j - w_i c_i\right) x_i\end{aligned}$$

taking $x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n$ as given.

A routine calculation, which is set out in Appendix 1.1, yields the Nash equilibrium solution for the firms' outputs and prices, given their capabilities. It is convenient to adopt the shorthand notation k_j to represent the ('effective cost') indicator, $w_j c_j / u_j$, and to express the solution in terms of quality-adjusted prices and outputs as follows: for any firm i with positive output at equilibrium (an 'active firm'):

$$\frac{p_i}{u_i} = \frac{1}{n-1} \sum_{j \text{ st } x_j > 0} k_j \tag{1.4}$$

and its (quality-adjusted) output is

[6] All fixed costs incurred in entering the market are sunk, and do not enter $\prod_i$.

978–0–19–927453–6 02-John-Sutton-c01-drv John (Typeset by SPi) 19 of 129 October 1, 2012 10:3

OUP UNCORRECTED PROOF – REVISES, 1/10/2012, SPi

$$u_i x_i = S \frac{n-1}{\sum\limits_{j\ st\ x_j>0} k_j} \left[1-(n-1) \frac{k_i}{\sum\limits_{j\ st\ x_j>0} k_j} \right] \tag{1.5}$$

so long as the term in [·] is non-negative, and zero otherwise. In the former case, firm *i*'s equilibrium profit equals

$$\Pi_i = S\pi_i = \left[1-(n-1) \frac{k_i}{\sum\limits_{j\ st\ x_j>0} k_j} \right]^2 S \tag{1.6}$$

and n denotes the number of firms that are active at equilibrium, and where we have introduced the notation π_i to represent firm *i*'s profit in a market of size $S = 1$.

The boxed output and profit equations (1.5) and (1.6) are of central and continuing importance in all that follows; indeed, the central analytical results of later chapters will require little more than the repeated use, in new contexts, of the output equation (1.5).

The shape of the profit function (1.6) is illustrated in Figure 1.3.

Consider first the case where firm *i*'s effective cost level k_i is equal to the mean level for all firms $\sum k_j/n$. Inserting this into the profit formula (1.6) yields $\Pi_i = S/n^2$. This is shown as point A in Figure 1.3.

As firm *i*'s effective cost level falls indefinitely, its profit rises asymptotically to S as its market share approaches unity. (Point B in Figure 1.3.)

The key feature is shown as point C in the figure: as k_i rises to a critical level, Π_i falls to zero. This is the 'lower threshold' mentioned in the preceding section. We can calculate the value of k_i to which this corresponds by simple inspection of the profit formula (1.6) (or the output formula (1.5)). The term in the bracket [·] falls to zero when

$$(n-1)k_i = \sum k_j = k_i + \sum_{j \neq i} k_j$$

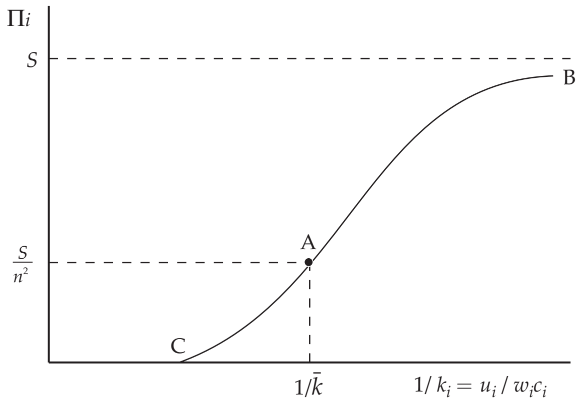

Fig. 1.3. Equilibrium profit Π_i as a function of $1/k_i$. The parameter $k_i = w_ic_i/u_i$ represents firm i's effective cost level. The diagram is drawn in terms of $1/k_i$ so that a rightward movement corresponds to a rise in quality u_i, or a rise in productivity, $1/c_i$.

or when

$$(n-2)k_i = \sum_{j\neq i} k_j \tag{1.7}$$

where the sums are taken over j such that $x_j > 0$. We can interpret this more easily by noting that the $n-1$ (active) rival firms have a mean value of effective cost k_j equal to

$$\left(\sum_{j\neq i} k_j\right)/(n-1).$$

Writing this mean value of rivals' effective costs as $\bar{k}$, (1.7) can be written as

$$k_i = \frac{n-1}{n-2}\bar{k} \tag{1.7$'$}$$

or

$$\frac{1}{k_i} = \frac{n-2}{n-1}\frac{1}{\bar{k}} \tag{1.7$''$}$$

which says that i's effective cost k_i can lie above $\bar{k}$ by at most this ratio $\frac{n-1}{n-2}$ (>1). This constitutes the first of two key propositions:

Proposition 1 (The Short Run Proposition)

For any set of n firms with effective cost levels k_j, there is a threshold level of effective cost above which a firm cannot achieve positive sales revenue at equilibrium.

1.7. THE VIABILITY CONDITION EXPLORED

In view of the central role played by the viability condition (1.7) in later chapters, it is worth pausing at this point to examine the condition more closely. In writing down (1.7), we began by looking at a context in which there were n firms, all of which had strictly positive output at equilibrium, and we focussed on one of these firms, labelled firm i, and asked: at what threshold value of k_i would x_i fall to zero? This threshold value is defined by the equality (1.7).

An alternative way of representing the threshold for viability is to imagine a potential entrant to the market. Denote by n_0 the number of incumbent firms, all with strictly positive output levels, and label the potential entrant as firm $n_0 + 1$. The threshold level of k_{n_0+1} below which this potential entrant can achieve positive sales revenue is given by equation (1.7): note that we can identify firm i in (1.7) as our potential entrant, remembering that the sum on the r.h.s. is taken over firm i's rivals, which in this context are the n_0 incumbent firms, while the total number of firms has become $n_0 + 1$. Hence, the viability threshold for our potential entrant is

$$k_{n_0+1} = \frac{1}{n_0 - 1} \sum_{j=1}^{n_0} k_j \tag{1.8}$$

This form of the viability condition for a potential entrant will be used in later chapters.

The preceding discussion may raise the following question: when we write down the viability condition, which firms are to be included in the summation? We have chosen to define the condition (1.7) by reference to the sum over firms whose output is *strictly* positive. It is obvious that the capability values of these firms affect the threshold. It is also obvious that the presence of firms whose capability values lie *strictly* above the threshold defined by (1.7) cannot affect the value of the threshold itself; these firms produce zero output, and have no effect on equilibrium prices. But what about firms that are exactly on the margin of viability, in the sense that their capability coincides exactly with the threshold value defined by (1.7)? It seems obvious, intuitively, that since their equilibrium value of output is zero, they cannot affect the value of the threshold. A simple calculation, set out in Appendix 1.3, shows that this is indeed the case.

This observation justifies our use of the strict inequality $x_j > 0$ in defining the number of firms to be counted in the formula for the viability threshold. It also shows that including firms that are exactly on the margin of viability is harmless: we obtain the same answer whether these firms are included in the summation or not.

1.8. THE OUTPUT EQUATION REVISITED

Now that we have explored the viability condition, it is helpful to digress briefly in order to point out some properties of the key output equation (1.5), on which much of our later analysis will rest.

We summarize the relevant properties in Lemma 1 below, the proof of which is given in Appendix 1.2. The key point to note about these results is that they refer to the way in which a firm's effective cost level $k_j = w_j c_j / u_j$ affects its equilibrium output *over a particular domain*. This domain is defined by the viability condition. Specifically, we are concerned here with a firm whose effective cost level lies between the (high) threshold level it must go below in order to attain viability, and the average effective cost level of the incumbent firms in the market. (This will be of interest in the next chapter, where we will be looking at a 'low quality' firm attempting to enter the market.)

Now what Lemma 1 says is that, over this domain, a reduction in a firm's effective cost level raises its equilibrium output (properties 1 and 2), while – less obviously[7] – a rise in quality that is fully offset by a proportional increase in (wage) costs will lead to a fall in output.

It will be useful, for further reference, to note some properties of the function.

[7] The proportionate changes in quality and the wage rate leave the effective cost level unchanged. From equation (1.5), this leaves profit unchanged, as π_i depends on w_i and u_i only via their ratio, $k_i = w_i c_i / u_i$. But inspection of the output equation (1.5) shows that the quality-adjusted output level $u_i x_i$ is a function of k_i – but x_i is *not* invariant to changes in u_i and w_i that leaves k_i fixed.

Lemma 1 *On the domain where the inequality*

$$\frac{w_i c_i}{u_i} < \frac{1}{n-1}\sum_j \frac{w_j c_j}{u_j} = \frac{n}{n-1}\left(\overline{\frac{w_j c_j}{u_j}}\right)$$

holds, so that $x_i > 0$:

1. $\dfrac{\partial x_i}{\partial w_i} < 0.$
2. $\dfrac{\partial x_i}{\partial u_i} > 0.$
3. *A rise in u_i and w_i that leaves u_i/w_i unchanged implies a fall in x_i.*

These results follow immediately from the form of the output function. (The proofs are given in Appendix 1.2.)

So why is the domain restriction needed? Why are the rather obvious-looking properties 1 and 2 not valid everywhere? To see the intuition behind this, think of a new entrant to the market, whose quality rises over time (so that its effective cost level falls). Now initially, this will indeed lead to a rise in its equilibrium output level. But what will happen as its quality level rises far above that of any rival firm? Its rivals become less and less effective competitors; the ratio of their prices to that of the now high-quality entrant falls to zero. As this happens, the high-quality firm's position will become close to that of a monopolist; and in the present model, the monopoly solution involves setting an arbitrarily high price and an arbitrarily low output level. This suggests the intuition that underlies the domain restriction in Lemma 1: It is *not* the case, in general, that rising capability implies higher output. But over the domain defined in Lemma 1, which is the relevant domain to consider in analysing a low-quality/high-cost entrant that is engaged in 'catching up' on its rivals, it is indeed the case that rising capability implies rising output. This result will play an important role in Chapter 2.

1.9. THE LONG RUN

Up to this point, we have taken each firm's capability to be fixed, i.e. determined by investments or efforts undertaken by the firm in the past. The second of the two basic propositions relates to the 'long run'. It addresses the question: suppose firms were in a position to invest in improving their capabilities. Then, foreseeing the competitive environment we have just described, how many firms would invest in capability building, and to what level?

The 'long run' question is central in what follows, and it raises a number of complex issues. In order to set the scene, however, it is useful to begin with a rather abstract discussion, which sets aside all the details of the capability building process that will occupy us in the next section. With that in mind, suppose that the firm has access to some 'R&D' process that will allow it to improve the quality u_i of its product, or to raise its productivity $1/c_i$, for some fixed financial outlay. A higher level of u, or a lower level of c, just requires a correspondingly higher fixed outlay on 'R&D activity'.

It's worth noting what we are *not* dealing with here. Suppose a seller of the felt-tip pen with which I'm writing this paragraph decided to double the amount of ink in each pen. This is a quality increase, and will shift the demand curve outwards, as I discover these pens now last longer. But the firm needs to incur double the old level of cost on the ink, and so the firm's marginal cost schedule rises too: this type of quality change does *not* fall within the scope of the result I'm about to describe. But suppose my pen-maker employed a design house to advise him, and they suggested a redesign of the tip that stopped me ending each writing session with inky fingers: there is no change in the unit cost of production, a one-off payment to the design house, and all pens are better … in this instance, the rise in capability involves a rise in *fixed* (and sunk) outlays.[8]

Now within this latter setting, a fundamental result holds; and the practical importance of this result derives in part from the fact that we have said nothing about *how* the rise in capability can be brought about. All that matters is that it involves a fixed (and sunk) cost, i.e. one that has nothing to do with the firm's current size (or level of output). Rather, it represents something like the development of a new product design, or the discovery of a new and better arrangement of

[8] But of course things are never this simple in practice: every design change carries *some* implications for unit costs, and it may involve either a rise or a fall. So it's worth noting that the simple case I'm describing here turns out to be of broad applicability. It will apply to cases where either (a) the fixed outlay leads to a fall in unit cost of production, with no change on the demand side; or (b) the fixed outlay leads to a rise in demand, with at most a *small* rise in unit costs (readers who are interested in pursuing what exactly 'small' means here may wish to refer to Sutton (1991), pp. 70–1 for the details).

the production process. Once learned, it can be implemented across any number of units of output that the firm produces. This is true of all the ways of 'raising capabilities' that we will explore in later chapters: the differences and details of how the change in capability is achieved, are immaterial, beyond the simple requirement that it involves only some 'one-off' expenditure of money and/or effort on the part of the firm to bring it about.[9]

Now think of a group of firms contemplating how much to spend on this kind of 'capability building' exercise. The right amount to spend will clearly depend on how much rival firms are spending. A certain number of firms will play the game, and invest in building capability. Others will hold back, and drop out. Our focus is on the number of capability-builders, who will constitute the 'active firms' of Figure 1.2.

The central result we develop below has a paradoxical flavour: the harder it is to build capability, the greater the number of firms in the market. Why?

The reason behind this result goes to the heart of the underlying economics. Suppose the cost outlay required to get from a lower-quality level v to a higher-quality level u is low. Then it might seem that more firms will be willing to pay the price and move up from v to u. But by the same token, it may then be worth moving not just to u, but to some higher quality (call it u^*). So there's a tension here: as it becomes cheaper to move up the quality spectrum, more firms would like to make any given move – but for the same reason, any single firm would want to move up higher (to u^*, or u^{**}, or $u^{***}\ldots$), whatever the level chosen by its rivals. And it turns out that this second effect wins out. Instead of inducing more firms to move up a little, the access to cheaper ways of building capability has the effect of encouraging a *smaller* number of firms to stay in the game – but this smaller number of players will each spend *more*. Making capability-building cheaper means a *smaller* number of firms each spends *more* on their capability building efforts. That's the paradox.

The second thing that emerges in this setting follows immediately from the story we've just told. It relates to the question: suppose the size of the (global) market got bigger (as happened when India and China brought down the trade barriers that had largely isolated them from global markets during the 1960s, 1970s, and early 1980s)–what

[9] Of course, over time, there will in general be a series of such fixed outlays, bringing us to successively higher levels of capability – but on each occasion, the firm pays some fixed fee in terms of money and/or effort to bring about the current improvement.

happens? The intuition we've just seen provides the right clue. As the size of the market gets bigger, we might expect it to support more and more producers. But as the market gets bigger, the extra profit available to the firm with the best design, or the top-performance product, rises – and this induces all firms to invest more in capability building. So instead of an ever-rising number of players in an ever-growing global economy, we have instead a group of players – who in some industries can be few in number – each spending more and more on its capability-building efforts as the global market expands.

The effect of this is most dramatic in those industries where there is a narrowly defined market in which all consumers want the same thing. A good example is the market for wide-body commercial jet aircraft. Here, the buyers are airlines, and their aim is simple: how to achieve the lowest carrying cost per passenger-mile. Throughout the first decades of jet age, from the 1960s to the 1980s, the best way to achieve this was to expand aircraft size. And so the major players – Lockheed, McDonnell, Douglas, Boeing, and Airbus – were involved in a decades-long game that ended with the survival of only two global players: Boeing and Airbus (Sutton (1998), Chapter 15).

At the other end of the spectrum are two kinds of industry in which the 'window' can accommodate, even in the long run, a very large number of players. The first of these corresponds to the case where 'capability building' is expensive. But in what sense? Think of the sugar industry. There's a given, readily available, technology for turning sugar beet or sugar cane into granulated sugar. How could a sugar maker improve its capability? There are ways of organizing production that will save labour – but there's fairly limited scope for this. And you might come up with some technical innovation that squeezes a bit more sugar out of a beet, or cane. Again, the scope is limited. And what of the demand side? Sugar is a pure chemical (sucrose). The (white) sugar on your table is 99 per cent pure. Purifying it to be 99.9999 per cent sucrose might be an interesting task for industrial chemists, but consumers don't care. So you could spend a great deal of 'fixed outlays' here for a minimal shift in your cost or demand schedules. The result: the global sugar industry supports a huge number of refiners. No one refiner can steal much market from others by outspending them in 'capability building'.

There's another kind of industry that also permits a huge number of firms to co-exit in the window. In this instance, the key lies in the fact that the product offerings in any market differ from each other in two quite distinct ways. So far we've focussed on the fact that producers differ in 'quality' – and everyone likes better 'quality'. But what of other features, like the body styling of a car, or the lyrics of a song? Some buyers will prefer one firm's offering, while others prefer a rival's ... even though both are offered at the same price. This is called 'horizontal' differentiation, as opposed to the 'vertical' (quality-improving) differentiation we've been considering up to now, and it arises in its simplest form in the context of geographical location. Given that the evening newspaper sells for the same price in all stores, I'll patronize the one in my own neighbourhood, just as you'll patronize the one closer to you.[10]

Now the practical setting where this 'horizontal' differentiation becomes most important is well illustrated by the case of scientific instruments. Take, for instance, flow-meters. These instruments are used to measure the flow of liquids in industrial plants. Many types exist, based on different scientific principles (electromagnetic, ultrasonic, and so on). Different types lend themselves better to different applications. In general chemical plants, the electromagnetic and ultrasonic types are good substitutes. In the oil industry, the ultrasonic is preferred at any price (oil doesn't conduct electricity, so the electromagnetic type just doesn't work). In a market like this, there's room for a very large number of players in the window: no process of capability building by makers of electromagnetic flow-meters can ever drive ultrasonic specialists out of the oil industry segment. So the window can be wide, and can accommodate a multitude of firms, in industries of this kind.[11]

[10] The contrast with 'quality', also known as 'vertical' differentiation, is that if prices are equal, *all* consumers will choose the better of the two products: its quality is unambiguously 'higher' – as with, say, the operating speed of a computer: if two equal price machines differ in nothing except operating speed, no one picks the slow one.

[11] There is one more possibility that is worth noting, for the sake of completeness. In the model that follows, we will assume, to keep things simple, that all individuals have the same income. If we allow for a range of incomes across consumers, then the window of viable qualities widens, as there is more room for low-quality products sold to low-income consumers. At the extreme, if we allow the range of incomes to extend downwards to zero, it is possible to obtain an equilibrium market structure in which a small number of 'high capability' firms have an (arbitrarily) large combined market share, while a large number of (arbitrarily) small firms remain viable, albeit with a very small combined market share. (See, for example, Shaked and Sutton (1983).)

978–0–19–927453–6 02-John-Sutton-c01-drv John (Typeset by SPi) 28 of 129 October 1, 2012 10:3

OUP UNCORRECTED PROOF – REVISES, 1/10/2012, SPi

1.10. WHY (SOME) CAPABILITIES ARE (RELATIVELY) SCARCE

In what follows, we develop this result within the special setting of the Cournot model with quality.[12] The easiest way to proceed is by setting up a 3-stage game, as illustrated in Figure 1.4.

The third stage of this game is simply the Cournot model with quality that we've seen already. The quality and productivity level of the firm, which are determined by decisions made at earlier stages, will fix each firm's (gross) profit level, as given by our profit function (1.9), i.e.

$$\Pi_i = S\pi_i = \left[1 - (n-1) \frac{k_i}{\sum_{j\ st\ x_j>0} k_j} \right]^2 \cdot S \qquad (1.9)$$

Now this function, as we've seen, is symmetric in u_i and $1/c_i$, so we'll lose nothing at this point by just setting the values of c_i and w_i equal to 1 for all firms, $k_i = w_i c_i / u_i$ becomes $1/u_i$. The profit function can now be written

$$\Pi_i = \left[1 - (n-1) \frac{1/u_i}{\sum_{j\ st\ x_j>0} 1/u_j} \right]^2 \cdot S \qquad (1.10)$$

Now we introduce a function $F(u)$ to represent the fixed and sunk outlays that must be incurred by any firm to achieve quality level u.

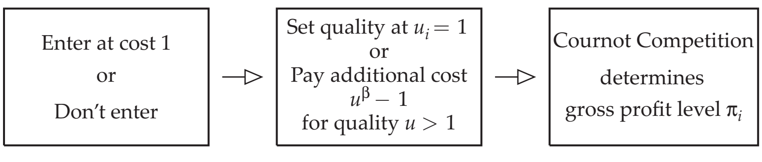

Fig. 1.4. The 3-stage game

[12] For a broad treatment, that allows for the presence of horizontal as well as vertical differentiation, see Sutton (1991), Chapter 3.

We'll take an iso-elastic function here, to keep things simple, setting

$$F(u) = u^{\beta}$$

so that

$$F(1) = 1$$

and

$$\frac{u}{F}\frac{dF}{du} = \beta$$

so that β represents the elasticity of F with respect to u (and $1/\beta$ is the elasticity of u with respect to F). A high value of β means that quality is relatively unresponsive to these cost outlays (which we may think of, for the moment, as R&D outlays – though we'll return to this point in Chapter 5, Box 5.1). A low value of β means that R&D spending is very effective in raising u – a relatively small outlay leads to a big rise in quality. (Figure 1.5.)

With this in hand, we can now come back to the structure of the 3-stage game of Figure 1.4.

Stage 1 is the entry stage, at which each of some ('large') number of potential entrants decides either to enter, and thereby incur a set-up cost of $F(1) = 1$, or not to enter. At stage 2, each firm decides whether

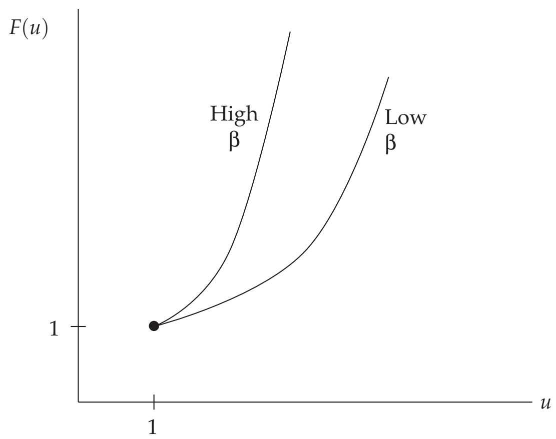

Fig. 1.5. The Fixed Cost Schedule $F(u) = u^{\beta}$ on $u \geq 1$.

OUP UNCORRECTED PROOF – REVISES, 1/10/2012, SPi

to have a 'basic' quality of $u = 1$, or a higher quality $u > 1$. If it decides the latter, then it pays the incremental fee of $F(u) - F(1) = u^\beta - 1$.

We characterize equilibrium as a (subgame) perfect (pure strategy) Nash equilibrium of the 3-stage game. In other words, we proceed by backward induction, asking at stage 2 the question: Given the number n of firms that have entered at stage 1, what quality will they choose? We'll then go back to stage 1 and ask, given this relation between the number of entrants and the quality they will choose at stage 2, how many of the (large) number of potential entrants will enter?

We can characterize the equilibrium outcome by noting that it must satisfy two conditions. The first condition is that the level of quality u chosen by each firm at stage 2 must be optimal. This optimal configuration of the quality levels can take two forms. The first is one in which all firms set $u_i = 1$. Here, we require that the cost of raising u to any level $u > 1$ exceeds the gross profit gained by doing so. This gives us the (necessary) condition

$$\left.\frac{dF}{du_i}\right|_{u_j=1,\forall j} \geq \left.\frac{d\Pi_i}{du_i}\right|_{u_j=1,\forall j} \tag{1.11}$$

In what follows, we will distinguish two regimes, according as (1.11) is satisfied as a strict inequality ('regime I') or as an equality ('regime II'). We will show that the first regime corresponds to the case where market size S is small.

We begin with the second regime, where S is large. In this setting, we have an interior solution, where $u > 1$, and a necessary condition for equilibrium is that

$$\frac{dF}{du_i} = \frac{d\Pi_i}{du_i} \quad \text{for all } i \tag{1.12}$$

We seek a symmetric equilibrium, i.e. one in which all the u_j are equal at equilibrium. With this in mind, set $u_j = \overline{u}$ for all $j \neq i$, differentiate the profit function (1.10) with respect to u_i, and finally set $u_i = \overline{u}$ to obtain

$$u_i \left.\frac{d\Pi_i}{du_i}\right|_{u_i=\overline{u}} = 2\frac{(n-1)^2}{n^3} S \tag{1.13}$$

This is the first condition for equilibrium, in the regime where the (common) level of quality in the (symmetric) equilibrium of the stage 2 subgame exceeds unity.

We saw above that in the symmetric setting where all firms have the same quality $\overline{u}$, the value of $\Pi(\overline{u})$ is S/n^2, so that (1.13) implies

$$\left.\frac{u_i}{\Pi_i}\frac{d\Pi_i}{du_i}\right|_{u_i=\overline{u}} = 2\frac{(n-1)^2}{n} \tag{1.14}$$

We now go back to stage 1, the entry stage. Here, it is convenient to begin by treating n as a continuous variable, so that the condition determining the equilibrium number n of entrants is the equality condition

$$\Pi(\overline{u}) = F(\overline{u}) \tag{1.15}$$

Combining (1.15) with equation (1.12) it follows that

$$\left.\frac{u_i}{\Pi_i}\frac{d\Pi_i}{du_i}\right|_{u_i=\overline{u}} = \left.\frac{u_i}{F}\frac{dF}{du_i}\right|_{u_i=\overline{u}} = \beta \tag{1.16}$$

so from (1.14) and (1.16) we have

$$\beta = 2\frac{(n-1)^2}{n}$$

or

$$n + \frac{1}{n} - 2 = \frac{\beta}{2} \tag{1.17}$$

Equation (1.17) is the key equilibrium describing market structure: it links the number of firms n that are active at equilibrium in this 'large market size' regime to the elasticity parameter β. We illustrate this in Figure 1.6.

So what we have established here is this: in a regime in which the firms choose a quality level strictly greater than 1, there will at equilibrium be a fixed number of entrants, independent of the size of the market: it is the absence of the market size parameter in Figure 1.6 which is the key point. The equilibrium number of firms depends on β alone, and not on S.

Before turning to the interpretation of this, we complete the analysis by pasting together the two possible regimes, and so describing the relation between market size and the equilibrium number of firms. (Figure 1.7.) In doing this, we'll replace the number of firms n with its reciprocal $1/n$, which is the conventional measure in the Industrial Organization literature. In the theoretical literature, the key summary measure of market structure is the share of industry sales accounted for by the largest firm, which is written as C_1 (the 'one-firm

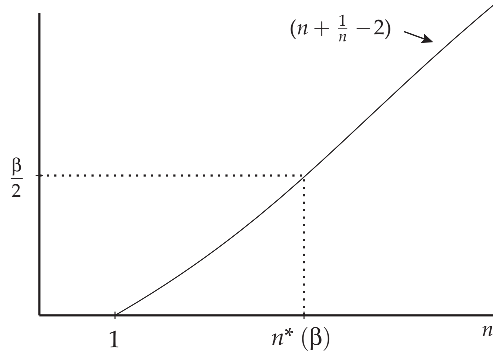

Fig. 1.6. The Equilibrium Number of Firms n^* as a Function of β.

concentration ratio'). In our present symmetric set-up, where all firms are of equal size at equilibrium, this is simply $1/n$.

Now for the market size vs market structure relationship. In Figure 1.7, we see that as market size S rises, the effect is to induce entry, as in the basic Cournot model. All firms set $u = 1$, and don't invest at stage 2. They each earn final stage profit S/n^2, and the number of entrants is fixed by the zero-profit condition, $S/n^2 = F(1) = 1$, whence $n = \sqrt{S}$. So as S rises, n rises, and $C_1 = 1/n$ falls. This is 'regime 1'.

The condition for being in regime 1 is that no firm wants to deviate by investing at stage 2 to achieve a quality level strictly greater than unity. This condition can be written as

$$\left.\frac{d\Pi_i}{du_i}\right|_{u_i=1} = S\left.\frac{d\pi_i}{du_i}\right|_{u_i=1} \leq \left.\frac{dF}{du_i}\right|_{u_i=1}$$

Using equation (1.13), this reduces to

$$u_i\frac{d\Pi_i}{du_i} = 2\frac{(n-1)^2}{n^3}S \leq u_i\left.\frac{dF}{du_i}\right|_{u_i=1} = \beta u_i^{\beta}\Big|_{u_i=1} = \beta$$

If we replace the inequality by an equality here, and insert the asymptotic value of n, which we labelled as $n^*(\beta)$ above, into the left-hand side expression, we obtain a critical value of S, which depends on β, at which we move from regime 1 to regime 2, as shown in Figure 1.7.

So what happens as S increases is that we reach a critical value of S, beyond which the effects of further increases in market size lead, not

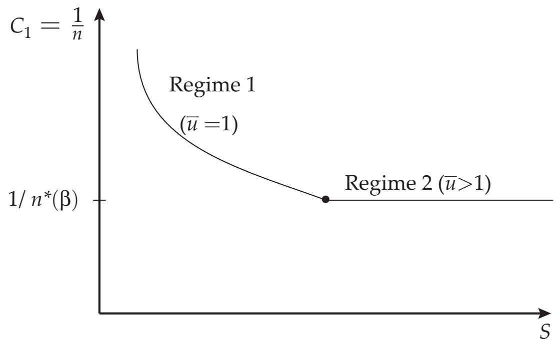

Fig. 1.7. Market Size and Concentration

to entry and falling concentration, but to an increasing escalation of expenditure on quality improvement. The bigger S, the higher is $\overline{u}$; but the number of firms remains unchanged.

So what of the dependence of n^*, and so $C_1 = 1/n^*$, on β? What we see from Figure 1.6 is that a fall in β leads to a fall in $n^*(\beta)$ and so to a rise in $C_1 = 1/n^*(\beta)$. This is the paradox: as it gets 'easier' to build quality, fewer firms do so. The resolution of the paradox lies in looking at it in a different way: a low value of β means that the quality improvement a firm gets for a given expenditure F is greater. Hence, the attractiveness to a deviant firm of outspending its rivals is greater. So escalation is more intense, and equilibrium spending is higher. It is an endogenous outcome of market forces that the industry becomes an increasingly expensive one to enter. In the language of our earlier discussion, a small number of firms build up capabilities that are relatively scarce.

This result constitutes the second of our two key propositions:

Proposition 2 (The Long Run Proposition)

Given any value of β, there is a corresponding lower bound to concentration, independently of the size of the market.

1.11. ROBUSTNESS

In the spirit of the present volume, the above discussion has been conducted in the special setting of the Cournot model. Its importance, however, lies in the fact that it illustrates a far more general result. All

OUP UNCORRECTED PROOF – REVISES, 1/10/2012, SPi

that is needed for this 'non-convergence' property to hold, is that our consumer utility (or demand) function conforms to the basis idea that 'bad qualities cannot drive out good'. Once this feature is present, we can extend and relax each of the special features of the example, and in particular the form of price competition, and the entry process. A full discussion of this is outside our scope, and the reader is referred to Sutton (1998) for the most general treatment (See also Symeonidis (2002)). There is, however, one point that is of fundamental interest in our present context: this relates to the way in which we can relax the form of the entry process.

Suppose, instead of our 3-stage game, we replace stages 1 and 2 with any finite sequence of stages; and then assign to each firm any stage at which it enters the game. As of that stage, the firm is free to enter, and make any investment it wishes. In other words, we can build in any 'historical' story we wish, giving any firm (or in the 'Trade' setting of later chapters, any country) an 'advantage' or 'disadvantage' of early or later entry into the (global) market, and the non-convergence result continues to hold. So this result captures something very robust and general about the outcome of a competitive market in which firms compete in qualities. More generally, while we have focussed on quality here, the argument relates more generally to capability: recall that, up to this point in our discussion, profit depends on quality u and productivity $1/c$ only via the ratio u/c, and so all that we have said about competition in quality carries over directly if we replace u with $1/c$, or with u/c. Competing in capabilities, in other words, involves a process in which, independently of the accidents of history, some capabilities will be 'scarce' – and it is this simple idea that will carry us forward to the next chapter, where we ask: where are these 'scarce' capabilities located, and why does it matter?

1.12. LOOKING AHEAD: A PREVIEW

With all this machinery in place, we can now sketch an informal outline of the way the globalization process will be represented in the chapters that follow.

If we flatten out the isoquants of Figure 1.2 by placing $\frac{u}{wc}$ on the vertical axis, and $1/wc$ on the horizontal axis, we obtain an alternative picture of the window, as shown in Figure 1.8, panel (i). In panel

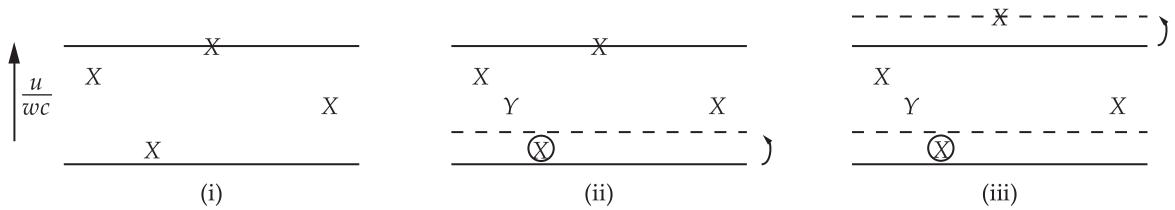

Fig. 1.8. The Moving Window

Notes: Panel (i) shows four viable producers, together with the u/wc 'window'. Panel (ii) shows the entry of a new firm Y into the global market, the consequent upward shift in the bottom of the window, and the exit of one of the original producers. Panel (iii) illustrates how changing incentives for long run investments move the top of the window upwards.

978–0–19–927453–6 02-John-Sutton-c01-drv John (Typeset by SPi) 36 of 129 October 1, 2012 10:3

OUP UNCORRECTED PROOF – REVISES, 1/10/2012, SPi

(ii) of the Figure, we illustrate the entry of a new producer from a country that has just opened up its borders to free trade. As it enters the window, two mechanisms come into play – and it is these mechanisms that drive the story in the chapters that follow.

First, the presence of the new firm makes price competition more intense, thus raising the threshold for viability (the bottom of the 'window') – and this may lead to the exit of some firms that were viable hitherto.

The second thing that will happen relates to firms' investments (in R&D or otherwise) that are aimed at improving their levels of quality (or productivity). As more firms enter the window, the market shares of all firms are reduced, and so the marginal returns to 'new' or 'incremental' investments in R&D aimed at raising their levels of quality (or productivity) will rise. Jumping ahead of the pack to obtain a large market share yields a greater incremental return. In the long run, as such investments are undertaken, the quality (or productivity) levels of some or all of the firms will rise – and so, in particular, the top level of $\frac{u}{wc}$ which serves to define the top of the window, will rise.[13] (The details of this argument, which brings us beyond the analysis of the present chapter, are discussed in Chapter 4.)

So the outcome of this 'globalization' process is a 'moving window' – as China's and India's producers enter the global market, they exert pressure on existing producers, leading to the exit of the least competitive, both in their home markets and in their export destinations. But they now find themselves in an increasingly competitive international environment, where the goalposts are constantly moving forwards . . .

It is this story that we set out to explore in the chapters that follow – but before we can do so, we need to move in the next chapter from our present, single market, 'partial equilibrium' framework to a multimarket 'general equilibrium' setting. To set the ground for this, we conclude this chapter by looking at a very simple general equilibrium setting.

[13] Incidentally, these shifts will also, in general, lead to a further upward shift in the bottom of the window (not illustrated in Figure 1.8).

1.13. CAPABILITIES AND WEALTH

We end this chapter by looking at the relation between firms' capabilities and welfare. To do this, we need to extend the previous analysis by looking at equilibrium in the labour market, and so endogenizing the wage w, which we took as a given parameter faced by the firm in Section 1.2.

With this in mind, we consider a (single) country, in which there is a single industry of the kind modelled in Section 1.2, and a single wage rate w faced by all firms. We ignore the 'outside good', and let all individuals devote all their income to the purchase of this good. We extend the individual's utility function to incorporate labour supply, writing it as

$$U = ux - \frac{1}{2}l^2$$

This form has the property that the marginal utility of leisure, $\frac{dU}{dl}$, equals l – so the individual labour supply schedule will have the form of a ray through the origin, i.e. the volume of labour supplied will rise in direct proportion to the wage rate w, for any given level of prices and qualities.

The individual (or 'worker') supplies l units of labour at wage rate w, and spends income wl on x units of good of quality u sold at price p. The constrained maximization problem

$$\max_{l} U = ux - \frac{1}{2}l^2 \tag{1.18}$$

$$\textit{subject to } px = wl$$

yields the solution

$$l = u \cdot \frac{w}{p} \tag{1.18a}$$

The consumer's budget constraint implies that the equilibrium level of consumption per capita is wl/p, and we will write this $\bar{x}$ (to distinguish it from the level of output per firm, which we continue to write as x). Substituting $\bar{x} = wl/p$ and $l = uw/p$ into (1.18) yields

$$U = \frac{1}{2}\left(u\frac{w}{p}\right)^2 \tag{1.19}$$

978–0–19–927453–6 02-John-Sutton-c01-drv John (Typeset by SPi) 38 of 129 October 1, 2012 10:3

OUP UNCORRECTED PROOF – REVISES, 1/10/2012, SPi

We now turn to the analysis of the firms. In this section, we focus on a symmetric set-up in which n firms offer the same quality u. Profits are distributed to a separate set of individuals ('non-workers'), who have the same utility function (1.18) but with l constrained to zero, i.e. $U = ux$. These firms face a market demand schedule of the form $p \sum_i x_i = S$, where S is total expenditure, as before.

In the present, symmetric, setting, the output per firm given by equation (1.5) reduces to (recall marginal cost in our present setting is wc):

$$x = \frac{n-1}{n^2}\frac{S}{wc} \tag{1.20}$$

and equilibrium price is (from equation (1.4))

$$p = \frac{n}{n-1}cw \tag{1.21}$$

Using (1.21) to substitute for w/p in (1.19) yields

$$U = \frac{1}{2}\left(\frac{n-1}{n}\right)^2\left(\frac{u}{c}\right)^2$$

which gives the link between capability u/c and the individual worker's utility, which we will take as our welfare indicator.

What is the level of the real wage in this economy? We can define the real wage as the number of units of quality-adjusted output that can be exchanged for one unit of labour input. This is given by

$$\frac{w}{p/u}$$

From equation (1.21) we have

$$\frac{w}{p/u} = \frac{n-1}{n} \cdot \frac{u}{c}$$

This links capability u/c to the real wage.[14]

It remains to calculate the level of activity in the economy, which we measure as the total labour input at equilibrium. Denoting the

[14] The factor $(n-1)/n$ reflects the level of the price–cost margin, which is linked to the equilibrium return to past investment in capability building, as we saw in Section 1.9.

total number of workers by N, and using (1.18a), total labour supply equals

$$L^S = Nl^S = N \cdot u\frac{w}{p} \tag{1.22}$$

which on substituting for w/p from (1.21) yields

$$L^S = N\frac{n-1}{n}\frac{u}{c} \tag{1.23}$$

while labour demand is given by

$$L^D = ncx \tag{1.24}$$

(remembering that x is per-firm output and one unit of output requires c units of labour output). Equating $L^S = L^D$ we find the value of per-firm output in terms of the primitives of the model,

$$x = N\frac{n-1}{n^2}\left(\frac{u}{c}\right)\frac{1}{c}$$

and equating this to the per-firm output equation (1.20) we obtain

$$\frac{1}{w} \cdot \frac{S}{N} = \frac{u}{c} \tag{1.25}$$

We can interpret the l.h.s. expression as gross national expenditure per capita, expressed in wage units (i.e. using the wage rate as our numeraire).

What the equation tells us is that as capability $\frac{u}{c}$ rises, the real wage rises and, given our upward sloping labour supply function, the level of employment (and so output) expands, and so real GDP per capita rises.

Incidentally, equation (1.25) can be derived more directly: readers who are interested will find details in Appendix 1.4.

1.14. SUMMING UP

We began this chapter by introducing a model in which bad products cannot drive out good. This led us to the notion of a 'window' of capability within which a firm must lie in order to be viable. Competition to establish oneself in this window implies a shakeout of firms: in a long run equilibrium, some kinds of market will be dominated by a small number of firms. In other words, some capabilities will necessarily be relatively scarce, and so valuable. Finally, capabilities determine wealth – and in the next chapter, we extend this notion to a multi-country setting, in which the capabilities of a country's firms form the proximate determinant of its GDP per capita.