Essays in the Theory of Economic Growth

by

Ashley Lester

B.Ec.(Hons I), University of Sydney (1997)

Submitted to the Department of Economics
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2005

Signature of Author ....
Department of Economics
17 May 2005

Certified by ....
Daron Acemoglu
Charles P. Kindleberger Professor of Applied Economics
Thesis Supervisor

Certified by ....
Abhijit Banerjee
Ford International Professor of Economics
Thesis Supervisor

Accepted by ....
Peter Temin
Elisha Gray II Professor of Economics
Chairman, Departmental Committee on Graduate Studies

Essays in the Theory of Economic Growth

by

Ashley Lester

Submitted to the Department of Economics
on 17 May 2005, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

This thesis is a collection of three theoretical essays on institutions and economic growth.

Chapter 1 considers a particular institution: ethnicity. Ethnic, religious and tribal divisions are empirically associated with economic underdevelopment. I construct a model in which groups form endogenously to enable cooperation between their members in a prisoner's dilemma. Groups sustain trust through monitoring, whereas only the Nash equilibrium, trade, is possible in an anonymous market. Optimal group size trades off the benefits of increased scale and the costs of reduced ability to detect cheating. Inter-group hostility can enable each group to enforce more trusting behavior between its own members. Even if groups may form optimally, in equilibrium they may persist inefficiently.

Chapters 2 and 3 consider some distributional implications of technical change in a model with human capital. Both chapters distinguish between general skills, that are equally useful with any vintage of technology, and specific skills, that are associated with a particular vintage.

In Chapter 2, I construct a model of slow technology diffusion. In developing countries, diffusion takes the form of a "dual economy", in which a gradually increasing fraction of workers use modern technology, while the remainder use traditional technology. Intermediate technologies are never used. During the transition, wages of specific-skill workers fall, as workers with general skills disproportionately join the modern sector. The model can also be applied to technology diffusion in developed countries.

Chapter 3 asks why, early in the modern era, technical change was primarily deskilling, while in the modern era it is skill-biased. Whereas previous explanations have focused on changes in technology, this paper suggests that changes in skills themselves were important. High-skill workers invest in specific skills if technical change is slow, and in general skills if it is rapid. This generates a U-shaped relationship between the rate of technical change and the skill-premium. Moreover, with low rates of technical change the modern sector is unskill-intensive, whereas the reverse is true when technical change is faster. The predictions of the model are compared with the historical experience.

Thesis Supervisor: Daron Acemoglu
Title: Charles P. Kindleberger Professor of Applied Economics

Thesis Supervisor: Abhijit Banerjee
Title: Ford International Professor of Economics

Acknowledgements

Writing a doctorate can sometimes be a lonely activity, but it is not one that can be undertaken alone. I have therefore been fortunate over my time at MIT to have benefitted from much help and support both from people at MIT and from further afield.

I would like to thank the Reserve Bank of Australia for financial support that enabled me to pursue this PhD. I owe a debt of gratitude to the RBA that I hope some day to repay. I would also like to thank the Department of Economics at MIT for awarding me a fellowship.

At MIT, I would like to thank the staff and faculty at the Department of Economics for making this the great place to learn that it is. I owe a special debt to my advisers. Ivan Werning asked a thousand sceptical and difficult questions, and expanded my horizons beyond Cambridge. Abhijit Banerjee helped me see the interesting questions and often pointed me towards more interesting answers. Daron Acemoglu was never less than an exacting taskmaster, for which I particularly thank him. His intellectual energy and curiosity were an inspiration throughout my time at MIT.

I would also like to thank many classmates and friends. Justin Wolfers, John Romalis and Jonathan Kearns helped me make it to MIT in the first place and then to settle in when I got here. My office-mates for the past three years, Bill Kerr and Byron Lutz, have shared good offices and bad, days when the phone rang and days when it didn't, and many conversations at least some of which were about economics. For their friendship, support and advice I would also like to thank Matilde Bombardini, Guy Debelle, the Fellas, Andrew Hertzberg, Geraint Jones, Daniel Paravisini, Bruce Preston, Veronica Rappoport, James Vickery and many others.

Many thanks of course to Mum and Dad and to Nana, who have always believed in me and supported me. Thanks also to Amelia, whose unexpected proximity for the past four years has been a tremendous good fortune for me.

Finally, my heart-felt thanks go to Nancy for her patience, understanding and love.

Contents

List of Figures

Chapter 1

Trade and the End of Trust: Endogenous Group Formation and Inter-Group Hostility

Summary 1 *Ethnic, religious and tribal divisions are empirically associated with economic underdevelopment. To investigate this relationship, I construct a model in which groups form endogenously as an alternative to an anonymous market. Groups sustain trust, the good outcome in a Prisoner's Dilemma, through monitoring, whereas only the Nash equilibrium, trade, is possible in the market. Optimal group size trades off the benefits of increased scale and the costs of reduced ability to detect cheating. I derive conditions under which groups are socially optimal, and conversely, groups form but are socially inefficient. Hostility between groups can arise as a by-product of group formation because it enables the group to enforce more trusting behavior between its own members. More infrequent contact between groups and technology that increases the extent of scale economies each lead to greater inter-group hostility. They also raise the probability that groups are socially inefficient. A non-stationary extension highlights that even groups that were efficient when they formed may persist after they become highly inefficient.*

1.1 Introduction

Ethnic, religious and tribal divisions are empirically associated with economic underdevelopment. Starting with Easterly and Levine (1997), a series of articles has highlighted the negative correlation between economic outcomes and various measures of fractionalization.[1] One obvious source of this association is ethnic conflict, which has come to considerable prominence since the end of the Cold War. However, even when actual violence is absent, fractionalization appears to have a range of negative effects, including inadequate provision of public goods and poor economic growth.[2] Most of the economics literature examining this relationship has focused on inter-group hostility, while treating ethnic groups themselves as exogenous.[3] In contrast, most social scientists - sociologists, political scientists, and anthropologists - do not view ethnic groups as exogenous or "primordial." Rather, they are "constructivists": they see ethnic and other social groups as constructed by their members.[4]

What can we learn about the economic effects of ethnic and other groups if we view them, not as exogenous, but as social constructs? Can this approach help us to understand the sources of inter-group conflict? I model group formation as a response to economic opportunities, such as the extent of returns to specialization and the efficiency of access to markets. In this setting, groups can become hostile to outsiders as a way of ensuring the loyalty of their own members. Hostility is a strategy that only arises either when it is needed to maintain group loyalty or when the cost of hostility is low, because, for example, the group rarely interacts with outsiders. This solves Fearon and Laitin's (1996) "puzzle" of inter-ethnic cooperation: their observation that, although inter-ethnic hostility occurs frequently, lack of conflict between groups is probably even more widespread. The model suggests that optimal policies towards groups depend on the setting which created the group, and that it may be more productive to try to change the economic setting than to focus on group behaviour directly.

[1] See also Alesina et.al. (2003), Aghion et.al. (2004), and Alesina and La Ferrara (2003).

[2] Some recent evidence on this is provided in Banerjee and Somanathan (2004).

[3] In addition to the empirical work discussed above, theoretical contributions include Esteban and Ray (1999), Acemoglu et.al. (2004) and Padro-i-Miquel (2004).

[4] For discussions of ethnicity and constructivism, see Chandra (2001) and Bates (forthcoming). Eltringham (2004) highlights the malleability of ethnicity by listing eleven associated characteristics, none of which are necessary or sufficient in themselves. Even kinship or shared ancestry is a constructed concept: Bates refers to a "mythology of consanguinity (quoted in Horowitz (1985), p60)," echoing Weber (1968), who stressed the importance of "a *subjective* belief in ... common descent" (p389).

In modelling groups, I draw on the well-established view that groups sustain trust amongst their members in economic transactions. Horowitz (1985), in his study of ethnic conflict, argues that "In the West, most tasks outside the home are performed by organizations not based on kinship. The same is simply not true in Asia and Africa... One reason for the difference is capacity: extended families are able to help their members in more transactions than nuclear families are (p63)." Trust may take many forms depending on the nature of the economic interaction involved. For example, Banerjee and Munshi (2004) find evidence that the Gounders who dominate the knitted garment industry in Tirupur, India, extend credit more readily to other Gounders than to outsiders. Moreover, it seems clear that groups sustain trust amongst their members through extensive circulation of information. For instance, Fafchamps (2004) finds evidence of information sharing and client referral amongst four distinct groups of Kenyan Asians involved in manufacturing.

The model reflects the idea that ethnic groups try to sustain trust amongst as many people as possible by circulating information. The production structure draws on Dixit (2003a). Agents are matched randomly each period and then play a prisoner's dilemma. Cooperation is sustained because unilateral deviations can be detected and the rest of the group informed. The probability with which a deviation is detected declines as the group expands. The productivity of a particular match depends on the characteristics of the partners, represented as distance from each other. In larger populations, agents can be further apart, and therefore the best possible matches are more valuable. However, because matching is random, the best matches do not always occur. It follows that, as population expands, the value of the best, or marginal, match increases faster than the value of the average match. Consequently, in sufficiently large populations, trust cannot be sustained in the most valuable transactions. Moreover, further expansion of the group reduces the quality of information and therefore the extent of trust. Accordingly, if there is an optimal group, it is finite.

Agents can also interact in an anonymous marketplace. Since it is anonymous, no trust can be sustained in the market. The market has two roles. First, even agents who prefer to trade within their group sometimes have to interact with members of other groups in the market. Second, it provides the autarky option for group members who have been caught cheating within their group.

One of the most important results of the paper is that inter-group hostility can arise because reducing the value of the market increases the ability of the group to sustain trust amongst its members. Therefore, if the marginal member is sufficiently valuable to the group, the group will try to reduce the value of market trades, a strategy I interpret as inter-group hostility. Groups can do this because the game includes multiple, Pareto-ranked Nash equilibria, as in a co-ordination game.[5]

This interpretation of inter-group conflict supports views that emphasize rational actors constructing ethnic identity "on the ground." This contrasts with most existing papers, which feature exogenous groups and argue that hostility is fomented primarily by elites to serve their own purposes.[6] One example of violence constructed on the ground is given by Fearon and Laitin (2000). Basque separatists, concerned that their fellow Basques were assimilating into the Spanish population, provoked the central Spanish authorities into taking action that they knew would antagonize previously assimilationist Basques. Ganguly (1997) has made similar arguments about Sikh revivalism in Punjab and Muslim fundamentalism in Kashmir.

In addition to suggesting a new interpretation of inter-group hostility, the model generates a number of comparative static predictions of interest. These relate in particular to changes in returns to scale and efficiency of matching, on the one hand, and changes in the probability of interaction with the market on the other. Increases in returns to scale and the efficiency of matching (that is, a greater probability of meeting distant trading partners) place pressure on groups because they raise the return to being in a larger population, such as the market. If the increases are large enough, groups may disintegrate entirely, and the economy can efficiently shift to a market equilibrium. Smaller increases raise the probability that, if groups persist, they are socially inefficient relative to the market equilibrium. In these circumstances, the group is particularly eager to sustain trust as broadly as possible, and accordingly inter-group hostility increases. Increases in the probability of interaction with the market do not affect the value of the market equilibrium, but they do make the survival of socially inefficient groups more likely. Moreover, they also raise inter-group hostility. Intuitively, a given degree of

[5] The mechanism is in some ways similar to Tirole (1996), in that it relies on group members being imperfectly distinguished by outsiders. Outsiders must therefore rely on the group's collective reputation when choosing their strategy.

[6] For analyses of ethnic hostility fomented by an elite, see Padro-i-Miquel (2004) and Acemoglu et.al. (2004).

hostility reduces the autarky value of group members by the same amount regardless of how segregated the group is, but is cheaper to inflict for groups that interact with the outside world less frequently.

The model therefore yields insights into the circumstances under which groups persist or dissolve. The non-stationary section of the paper highlights that even if groups originally form optimally, they can persist inefficiently when circumstances change, and this persistence itself leads to additional inter-group hostility. Nevertheless, ethnic groups can in fact change quite rapidly.[7] For example, although most modern European countries appear to be relatively homogeneous, in some cases this is a surprisingly recent development. France, for example, became dramatically more homogeneous during the late nineteenth century. Conversely, many of the chronic ethnic problems of modern Africa also turn out to be of surprisingly recent provenance, concerning groups that took their modern form only during or after the colonial period.

This is consistent with the prediction of the model that change is most likely when technology or infrastructure changes substantially, for example during the Industrial Revolution in Europe or the colonial period in Africa. Moreover, groups are more likely to be replaced by the market when these changes are accompanied by policies that reduce the ability of groups to distribute information, and by integrationist policies more broadly. In much of Europe, these conditions were present and groups largely dissolved. By contrast, in much of Africa, introduction of new technology and infrastructure was more limited, and policies often explicitly favored ethnic division. Groups therefore persisted and became more hostile. In confronting modern ethnic divisions, the model highlights that new technology and infrastructure can be as important in causing a shift to the market as policies more obviously directed at groups. For example, building a road and a market place where exchange can take place may be a more productive way of reducing group conflict than encouraging negotiations between group leaders.

This paper is related to a number of existing strands of the literature. The role of groups in sustaining trust has been extensively investigated. In Dixit (2003a), although no groups can form, small populations can sustain trust by informal enforcement: an agent whose partner has cheated informs other agents with a probability that declines with distance. Larger

[7] The following examples are discussed more extensively in Section 6.

populations do better by relying on a costly external enforcement mechanism, such as a formal legal system. Putnam (2000) is a well known sociological analysis of groups and trust. Avner Greif (1994) and Greif, Milgrom and Weingast (1994) consider medieval merchant traders, who sustained trust by forming guilds, or through an ethnic group, the Maghribi. Fafchamps (2004) provides an extensive discussion of various forms of trust in economic transactions in developing countries. Theoretical investigations of groups and trust include Ellison (1994) and Ghosh and Ray (1996). Kranton (1997), who compares anonymous market exchange with ongoing non-market interactions, is closer in spirit to the present paper. But since all exchange in her model is between ongoing bilateral trading partners, neither the scope of the group nor interaction between groups can be considered.

Although economists have recently become more interested in ethnic groups, only Caselli and Coleman (2003) endogenize ethnic group formation. However, in their paper, groups exist purely to extract rents, so society as a whole would always strictly benefit from the elimination of groups. Moreover, the number of ethnic groups, although not the size of each, is exogenous. Alesina and Spolaore (2003) describe optimal country formation subject to exogenous ethnic divisions, but their work assumes differences in preferences across groups.

Another related literature considers club goods, although in that literature the optimal size of a group is determined directly by congestion costs.[8] However, some of the club good literature does raise the possibility of apparently self-destructive actions as a way of committing to the group.[9] Finally, the observation that, in problems of limited commitment, improvements in the autarky option can reduce welfare has previously been made both in the macroeconomic literature on insurance with limited commitment and in the industrial organization literature on relational contracting.[10]

Section 2 of the paper introduces the stationary model of group formation. Section 3 shows how inter-group hostility arises naturally out of this model. Section 4 considers the social welfare implications of group formation. Section 5 considers the non-stationary model. This setting magnifies the potential for inefficiency and hostility seen in the stationary model.

[8]For an overview of club theory, see Sandler and Tschirhart (1997).

[9]See Berman (2000) for a discussion of this behaviour in the context of Israeli ultra-orthdox Jews.

[10]See Ligon, Thomas and Worrall (2002) and Attanasio and Rios-Rull (2000) for the insurance problem, and Baker, Gibbons and Murphy (2002) and Levin (2003) on relational contracting.

Section 6 discusses the model in the context of several case studies of ethnic groups, focusing particularly on the homogenization of France in the late nineteenth century and the development of the divide between Hutus and Tutsis in Rwanda at a similar time. Section 7 concludes.

1.2 Group Formation

1.2.1 The Environment

Production Locations

There is an *ex ante* homogeneous set of agents of mass $2L$. Despite dealing with a continuum of agents, in the body of the paper I will refer to decisions made by "individuals" or "agents" for the sake of readability. Agents die and are replaced at a constant rate, δ_0 and also discount the future, so that their effective discount factor is δ. Agents maximize expected lifetime utility,

$$E\sum_t \delta^t u\left(c_t\right).$$

All payoffs are specified in units of utility. At birth, agents can choose to join a group. This choice is irrevocable, in the sense that there is an infinite cost of switching groups.

Agents interact with each other in one of two types of locations. These need not be geographically distinct from each other, although they may be. Interactions can be considered as either a form of trade or a form of joint production. The first location, the market, is potentially inhabited by all agents. Actions within the market are anonymous except that the group to which an agent belongs can be observed. The second location is associated with the particular group to which an agent belongs. An agent who is a member of a group can seek interactions in the market or within the group. If they seek interactions in the market, they find the market with probability one.[11] If they seek interactions in the group, they find the group with probability p and the market with probability $1-p$.

Although more general imperfect search processes could be specified, this simple specification makes two key points. First, although the group is effective at excluding outsiders,

[11] Actually, probability $1-\epsilon$, although except in the section dealing with equilibrium group formation I consider only the limiting case as $\epsilon \rightarrow 0$.

it cannot entirely avoid contact with outsiders in other settings. This assumption is needed so that there is some interaction between members of different groups: otherwise inter-group relations are moot. Second, members of the group have some control over whether they find the group or the market in a given period. This generates a limited commitment problem for members of the group.

Matching and Production

Regardless of where an agent finds themself in a given period, production takes the same physical form, drawn largely from Dixit (2003a). Agents in the locale are arranged uniformly around a circle of circumference $2P_{lt}$, where $2P_{lt}$ is the mass of agents in locale l in period t. (For convenience, I will refer to this circle as being of "size" P.) The distance between any two agents is the shorter of the two arcs between them on the circle. This distance represents differences between agents that can be exploited in production or trade. One interpretation is that, although all agents are *ex ante* homogeneous, they can specialize in some quasi-fixed characteristic such as occupation, or some endowment of human and/or natural resources. Agents who are further apart from each other are therefore more different from each other in some relevant dimension, although all agents are *ex ante* alike.

Figure 1.1 depicts an equilibrium in which all agents have joined one of four groups, and seek the groups each period. The size of group A each period is G. The market consists solely of failed searches for the group, and is therefore of size $(1-p)\,L$.

Each period, each agent is randomly matched with another agent around the circle, and can, if they wish to do so, engage in an interaction, the product of which is immediately consumed. (Recall that utility is linear in payoffs.) An agent's utility in the period is normalized to zero if either they are not matched with anyone, or they are matched but one or both parties to the match choose not to interact.

The probability of meeting an agent at distance x is

$$\frac{\alpha e^{-\alpha x}}{2},$$

where $\alpha/2$ is a rescaling constant to ensure that the aggregate meeting probability is less than

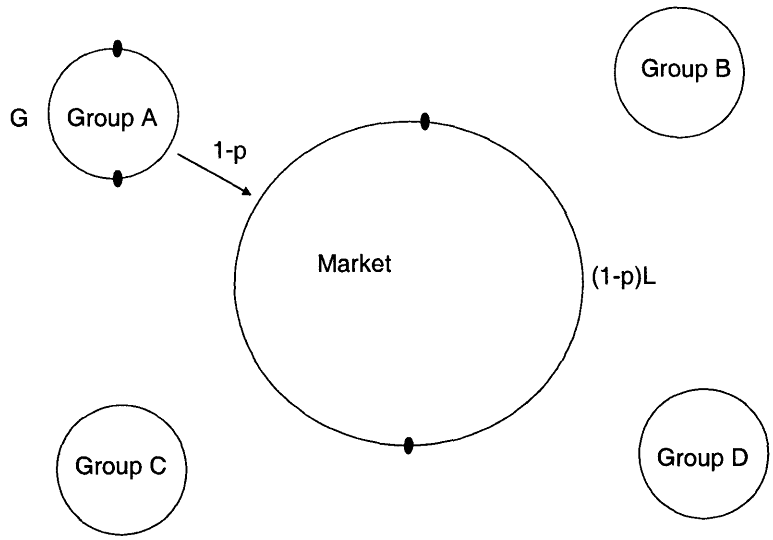

Figure 1-1: Groups and the Market

one, and $\alpha > 0$. α is a localization parameter: a greater value of α raises the probability of nearby matches relative to distant matches. Lower values of α are associated with more effective search. This might occur for a variety of reasons: information about favorable trading opportunities may disseminate more effectively, or roads and other physical infrastructure might bring otherwise distant traders effectively closer together.

There is no crowding out: adding extra agents to the population does not change the probability of meeting an agent at a given distance. This simplification necessarily implies that in any finite population the probability of a match in each period is less than one. In fact, the aggregate meeting probability is

$$2\int_0^{P_{lt}} \frac{\alpha e^{-\alpha x}}{2} dx = 1 - e^{-\alpha P_{lt}}.$$

A greater value of α therefore also raises the probability of matching in a given period for any finite population. Because α has the dual role of controlling localization and aggregate meeting probability, some of the comparative statics with respect to α will be ambiguous.[12] For sufficiently large populations, however, the effect of α on localization is more important than its effect on aggregate meeting probabilities.[13]

If they choose to interact, agents play a game in which the payoff to player j is

$$\pi_j\left(a_j, a_{-j}|\Omega_1\right) e^{\theta x},$$

where $\pi_j\left(a_j, a_{-j}|\Omega_1\right)$ is a constant determined by the actions of player j, a_j, and player $-j$, a_{-j}, which may be conditional on any information the two players have about each other, Ω_1, including the distance from each other, x. $e^{\theta x}$ is a multiplicative factor that depends on the distance between the two players. Assume that $\theta > 0$, which implies that more distant trades are potentially more valuable, so that there are increasing returns to scale in group size.

Increasing returns to scale are derived from the *ex post* heterogeneity previously discussed. The exact source of the returns to scale depends on the nature of the interaction between the players. In a trade setting, two agents growing different strains of wheat might stand to gain little from meeting, while a wheat farmer and a cattle farmer might expect large gains from trade. In a production setting, owners of different complementary factors will gain more from meeting than two owners of the same factor. Similarly, the return to meeting someone in a dissimilar occupation is probably greater than that to meeting someone in the same occupation: an architect might gain more from meeting an engineer than from meeting another architect. Although at some point it might be possible to imagine two trading partners who are actually too different from each other, all that is needed for the results is that interactions are increasing in value up to a sufficiently great distance. Following Dixit, assume that $\alpha > \theta$, so that expected utility does not increase infinitely as the population grows infinitely large.

The constant $\pi_j\left(a_j, a_{-j}\right)$ is determined by the players' choice of actions according to the

[12] Note that these roles are contradictory rather than reinforcing: greater α raises aggregate meeting probability but reduces the distance of those matches that occur.

[13] Normalising meeting probabilities along the lines of Dixit (2003a), so that the probability of a match at distance x is

$$\frac{\alpha e^{-\alpha x}}{2\left(1 - e^{-\alpha P_{lt}}\right)}$$

would greatly complicate the solution without adding much intuition. Such a modification would actually encourge finite group formation along the lines to be discussed.

table below. The players' choice of strategies is the set $a_j \in \{[0,\bar{A}],C\}$. This represents two types of strategies. The choice of C, for co-operation, corresponds to fully trusting the other player. It is the good outcome in a prisoner's dilemma (so $T^+ > T$), and the social optimum is for both players to play C. Any action $s_j \in [0,\bar{A}]$ is a variety of trade. Within trade, there is a pure co-ordination game, so that any value of trade up to $\bar{A} < T$ can be sustained as a Nash equilibrium. While the choice of 0 is an arbitrary normalization in the co-ordination game, it is important that T^- is less than the minimum payoff to *Trade* , so that *Trust* can never be a Nash equilibrium. An equilibrium in which agents play less than $\bar{A}$ will be taken to represent inter-group hostility. For now, assume that all agents co-ordinate their strategy on a fixed value of $A \leq \bar{A}$ when not playing trust.

	Trust (C)	Trade (s_2)
Trust (C)	T,T	T^-,T^+
Trade (s_1)	T^+,T^-	$\left\{ \begin{array}{c} s_j \text{ if } s_j \leq s_{-j} \leq \bar{A} \\ 0 \text{ if } s_j > s_{-j} \end{array} \right\} \forall j \in \{1,2\}$

There are two key parts to this game. The first is the prisoner's dilemma, while the second is the coordination game, which creates multiple, Pareto-ranked one-shot equilibria. Intuitively, groups form to enable their members to reach the good outcome in the prisoner's dilemma, while the coordination game allows the group to pursue a range of strategies with respect to the outside world.[14]

Both of these elements seem reasonable in the types of economies under consideration. The desirability of maintaining trust in economic interactions is well understood and fairly intuitive. For instance, in either credit markets or joint production, returns are higher if a partner can be trusted to perform without monitoring. Similarly, specifying a complete contract is costly, and parties who trust each other may prefer to leave the contract unspecified. Economic relations involving ethnic groups or kinship groups often take this form. For example, Geertz (1963) describes how in Bali kinship groups assemble from time to time to produce large *gamelan* (percussion instruments). Although some payments are made, capital is shared

[14]Formally, Section 3 will show that the group uses the multiplicity of Pareto-ranked equilibria to sustain different degrees of trust (equilibrium payoffs) inside the group. The strategy is therefore similar to that considered by Benoit and Krishna (1985) in finitely repeated games.

and participation is seen as a moral responsibility of group members (pp93-6). This may be presumed to be a more efficient arrangement than hiring workers whose tasks need to be contracted upon and performance monitored.

Continuing with the *gamelan* example, if a worker is hired on a spot market, so that trust cannot be sustained, there may be many tasks they could be assigned. Both their wage and their task will depend on their employer's expectation of their capability and honesty. Each of these combinations of tasks and wages could correspond to a different coordination equilibrium. If the employer has a low opinion of their new employee's likely capability, they may use them purely in menial tasks. If they have a higher opinion of their employee, they may trust them with more important tasks relating to the manufacturing itself. Given that the employee is hired only once, there may be no opportunity to discover that a poor opinion of their abilities was misplaced. Similarly, in credit markets, if a creditor has a low opinion of a lender's trustworthiness, they may limit lending and impose onerous conditions on the lender to ensure repayment.

Information

Trading partners in the market have no individual information about each other. They can observe which group their trading partner joined at birth, although not their current status within the group. The only other information available to market participants is an announcement made by each group at the start of the period as to which strategy its members will pursue in the market that period.

The group can sustain trust because it can transmit information. Within the group, the flow of information is costless, so that effectively any necessary announcement can be made by a central authority at the end of each period. The group gathers information only about transactions within its location. For simplicity, all actions associated with monitoring are costless. The quality of monitoring deteriorates when more members are added to the group, so that the probability of being caught after having cheated is

$$\bar{m}e^{-\beta G}$$

where $\bar{m}$ is just a multiplicative constant, $\beta > 0$ and G is the total size of the group. The probability of an erroneous adverse finding is zero.

If an agent is caught cheating, they are immediately given a record that is publicly available to all group members in the future. This is denoted

$$R_j \in \{0, 1\}$$

where $R_j = 1$ if a member has been caught cheating, and $R_j = 0$ otherwise.

An investigation is launched only if one of the agents complains. Complainants are assumed to be exactly indifferent about complaining: no cost is involved, but nor is there any compensation if the complaint is upheld. If an agent complains, the other agent may launch a counter-complaint. Hence, an agent will not complain unless a deviation has been unilateral. This is important, since under some circumstances groups will wish to include some interactions in which neither agent plays *Trust*. Without this mechanism, such interactions would involve the risk of expulsion from the group.

There are many reasons why information might deteriorate with group size. For example, a single agent, such as a chief, will find monitoring more and more diverse specialities increasingly difficult. On the other hand, if agents can specialise in monitoring a narrow set of transactions, information need not deteriorate as assumed. This form of constant returns to scale monitoring seems, however, a better characterisation of a nation state or modern bureacracy than of relatively informal groups. Dixit (2003a) considers a costly monitoring structure with more of this flavour, while Dixit (2003b) considers, in a slightly different setting, profit-maximizing intermediaries. While leaving the exact micro-foundations of the current assumption outside the present model, it therefore seems the most appropriate assumption when considering groups.

1.2.2 Equilibrium

The equilibrium concept is perfect public equilibrium (Fudenberg et.al., 1994).[15] I restrict attention to perfect public equilibria in which groups attain privately optimal expected payoffs for their members, conditional on group structure and grim punishment strategies for group

[15] The equilibrium is not simply sub-game perfect because actions are conditioned, not on the trading partner's past actions, but on a public signal of their past actions.

members caught cheating. Agents face a choice of strategies over which group to join, if any, and, in all subsequent phases, what action to play in the stage game, conditional on distance and information about their trading partner. Therefore, agent j's strategy space is defined over

$$\Sigma_j = \sigma_j(l_j) \times \bigcup_N \{\{F_j : (V_g(G_n, R_j), V_m) \in \Omega_0\} \times \{a_j : (x, D, R_j, R_{-j}) \in \Omega_1\}\}. \qquad (1.1)$$

Consider the three terms in the strategy space in turn. First, $\sigma_j(l_j)$ is a mixed strategy profile for agent j over each location in N, the set of possible groups. After the agent's group, $n \in N$, is known, each period the agent has two more sets of choices. First, $F_j \in \{h, o\}$ is agent j's choice of search strategy. If the agent plays h, they seek their "home", the group to which they belong. If the agent plays o, they seek the outside, that is, the market. The agent's choice is conditioned on their information set at the start of the period, Ω_0. This includes the expected payoff of finding the group, n, to which agent j belongs, $V_g(G_n, R_j)$, and the expected payoff to the market, V_m. The payoff to finding the group is a function of the size of the group, G_n, and whether the agent has previously been caught cheating, R_j. The expected payoff of the market is also a function of its size, although for simplicity I suppress this in the notation. Finally, $a_j \in \{[0, \bar{A}], C\}$ is agent j's action profile in the subgame conditional on their information set before playing. This contains x, the distance of the match, and their location, $D \in \{n, m\}$. If $D = n$, the agent also has information on any past record of cheating by themselves (trivially) and by their partner, R_j and R_{-j}.

Agents choose their group membership strategies according to a Nash concept, by calculating expected pay-offs conditional on the strategies of other agents. Group formation is non-cooperative, so the group is unable to exclude prospective members.[16]

To simplify the discussion, I solve for equilibrium recursively by first describing payoffs to groups as a function of optimal equilibrium strategies, and then consider the variety of possible

[16]This accords with many descriptions of ethnicity, which focus on self-identification. In his study of the Lue people of Thailand, for instance, Moerman writes that "someone is Lue by virtue of believing and calling himself Lue and acting in ways that validate his Lueness." (quoted in Eltringham (2004), 183). Nevertheless, an interesting alternative approach to ethnicity would emphasise the need for acceptance by other group members. Some discussion of this theme can be found in Migdal (ed) (2004). An economic model taking this approach might use an equilibrium concept such as Bernheim, Peleg and Winston's (1987) Coalition Proof Nash equilibrium.

group structures in equilibrium. The latter discussion involves several complications that follow from considering a continuum of agents and a discontinuity in group payoffs.

Sub-Game Strategies

Each period, an agent makes a choice F_j, whether to seek the group, and a choice a_j, conditional on the outcome of their first decision and the process of random matching, of strategy in the subgame. Groups only exist if an agent who has not cheated in the past prefers to find the group rather than the market. Moreover, the grim punishment strategy and the coordination game together ensure that agents who have been caught cheating must prefer the market. Hence, if groups exist it must be true that

$$F_j(R_j) = \begin{cases} h & \text{if} \quad R_j = 0 \\ o & \text{if} \quad R_j = 1 \end{cases}$$

In the market, there is no punishment for not playing trust, so for given expectations of other players' strategies,

$$a_{jm} = A$$

where a_{jm} is j's choice of action in the market (that is, given location $D = m$).

In the group, I assert (for now) that the optimal choice of strategy is to play:

$$a_{jn}(x|R_j = 0) = \begin{cases} C & \text{if} \quad R_{-j} = 0 \quad \text{and} \quad x \leq G_T \\ \bar{A} & \text{if} \quad R_{-j} = 0 \quad \text{and} \quad x > G_T \\ 0 & \text{if} \quad R_{-j} = 1 \end{cases}$$

So long as an agent's trading partner has not been reported cheating in the past, an agent, when located in the group, plays *Trust* in all trades that are not too distant, and the maximum attainable value of *Trade* in all other transactions. If the trading partner has been reported cheating in the past, then agent j pursues a grim punishment strategy. This choice of actions derives from an incentive compatibility constraint, which is considered in the next section.

Finally, it is necessary for completeness to specify an agent's choice of actions within the group conditional on their having been caught cheating in the past. Note that this will not actually occur in equilibrium. However, given that the agent knows that the grim punishment strategy is being pursued in the group, their best response is to play

$$a_{jn} = 0 \text{ if } R_j = 1.$$

1.2.3 Optimal Group Size

Incentive Compatibility

Suppose that all agents expect all other agents to form groups and search for them in equilibrium. Then the size of the market is $(1-p)L$. All transactions in the market, being anonymous, will involve the agents playing A, which for now is exogenous. I have argued that, in equilibrium, relatively nearby transactions within the group will involve both players playing *Trust*. Transactions at greater distance will be unable to sustain trust, so assume that the group can coordinate on the greatest possible value of trade, $\bar{A}$, which is clearly optimal.

Associated with these strategies are a set of expected values that define the expected value of transactions associated with being a member of a group of size G, with extent of trust G_T. The first two of these, equations 1.2 and 1.4 , are the expected value of trades in a particular location, conditional on finding that location, and mirror those in equation 1.1. The first, $V_g(G|R_j = 0)$ is the expected value of trades within the group, n, to which agent j belongs, assuming they have not been caught cheating in the past. For simplicity, I will denote this as simply $V_g(G)$, omitting the condition on the report unless otherwise specified. The second, V_m, is the expected value of trade in the market. Finally, V_h is the expected value of being a group member in good standing who seeks trades within the group. That is, it is the expected value of playing h in the search phase.

$$\begin{aligned}
V_g(G) &= 2\int_0^{G_T} Te^{\theta x}\frac{\alpha e^{-\alpha x}}{2}dx + 2\int_{G_T}^{G} \bar{A}e^{\theta x}\frac{\alpha e^{-\alpha x}}{2}dx && (1.2)\\
&= \lambda T\left(1 - e^{(\theta-\alpha)G_T}\right) + \lambda\bar{A}\left(e^{(\theta-\alpha)G_T} - e^{(\theta-\alpha)G}\right) && (1.3)\\
V_m &= 2\int_0^{(1-p)L} Ae^{\theta x}\frac{\alpha e^{-\alpha x}}{2}dx && (1.4)\\
&= A\lambda\left(1 - e^{(\theta-\alpha)(1-p)L}\right) && (1.5)\\
V_h &= pV_n(G) + (1-p)V_m && (1.6)
\end{aligned}$$

Note that $\lambda = \frac{\alpha}{\alpha-\theta}$ is a constant that parameterises net returns to scale. It is increasing in θ and decreasing in α. The mass of the group is normalised to $2G/p$, so that each period the group is of size G. Naturally, no group forms if the returns to scale lost by forming a group exceed the gains from trusting relationships. Therefore, groups only form if in equilibrium

$$T\left(1-e^{(\theta-\alpha)G_T}\right)+\bar{A}\left(e^{(\theta-\alpha)G_T}-e^{(\theta-\alpha)G}\right)>A\left(1-e^{(\theta-\alpha)(1-p)L}\right). \tag{1.7}$$

The group would like to expand as much as possible, to capture any possible increasing returns to scale. However, the group needs to overcome the incentive compatibility problem in the prisoner's dilemma if it wants to guarantee that players continue to play trust. The group is able to sustain trust among at least some of its members by the threat of future punishment of members who do not comply. For sufficiently distant trades, however, the temptation to cheat becomes so great that the group is unable to sustain trust. The point at which the incentive compatibility constraint binds therefore defines the limits of trust, G_T. The incentive compatibility constraint for a group of size G is

$$\begin{aligned} Te^{\theta G_T}+\delta V_h &\geq T^+e^{\theta G_T}+\delta\left(\bar{m}e^{-\beta G}V_m+\left(1-\bar{m}e^{-\beta G}\right)V_h\right) \\ V_h-V_m &\geq \frac{\left(T^+-T\right)e^{\beta G+\theta G_T}}{\delta\bar{m}} \end{aligned}$$

or

$$\lambda T\left(1-e^{(\theta-\alpha)G_T}\right)+\lambda\bar{A}\left(e^{(\theta-\alpha)G_T}-e^{(\theta-\alpha)G}\right)\geq\lambda A\left(1-e^{(\theta-\alpha)(1-p)L}\right)+ce^{\beta G+\theta G_T} \tag{1.8}$$

where $c=\frac{(1-\delta)\left(T^+-T\right)}{\delta\bar{m}p}$ and of course $G\geq G_T$. The left hand side of this equation defines the benefits of group membership as group size expands. It is concave in both G_T and G, and approaches a maximum of T, reflecting the bounded returns to scale in expectation. The right hand side defines the temptation to cheat on a trade at distance G_T as the group gets larger. It is convex in both G_T and G, reflecting the strongly increasing returns to scale of distant interactions.

Optimality and the Extent of Trust

The incentive compatibility constraint can be used to find the optimal group size for a given market. This problem is equivalent to allowing coalitions to form in the pre-play phase, subject to the expectation that all other agents also join groups. Because of this expectation, the size of the group does not affect the size of the market. Hence, the group need only be concerned with maximizing the payoff to within-group interactions.

The program is

$$\max_{G,G_T} \lambda T\left(1-e^{(\theta-\alpha)G_T}\right)+\lambda\bar{A}\left(e^{(\theta-\alpha)G_T}-e^{(\theta-\alpha)G}\right) \tag{1.9}$$

s.t.

$$\lambda T\left(1-e^{(\theta-\alpha)G_T}\right)+\lambda\bar{A}\left(e^{(\theta-\alpha)G_T}-e^{(\theta-\alpha)G}\right)\geq\lambda A\left(1-e^{(\theta-\alpha)(1-p)L}\right)+ce^{\beta G+\theta G_T} \tag{1.10}$$

$$G\geq G_T \tag{1.11}$$

This problem yields the following first order conditions (in addition to the constraints):

$$\alpha\left(T-\bar{A}\right)e^{(\theta-\alpha)G_T^*}(1+\mu_1)-c\theta e^{\beta G^*+\theta G_T^*}\mu_1-\mu_2 \;=\; 0 \tag{1.12}$$

$$\alpha\bar{A}e^{(\theta-\alpha)G^*}(1+\mu_1)-c\beta e^{\beta G^*+\theta G_T^*}\mu_1+\mu_2 \;=\; 0 \tag{1.13}$$

where μ_1 is the multiplier on the IC constraint, and μ_2 is the multiplier on the feasibility constraint. The problem can be solved for two distinct sets of parameter values.

Case 1: Semi-Trust Groups.

Assume that the feasibility constraint is slack, so $\mu_2=0$. Rearranging the first-order conditions, equations 1.12 and 1.13, yields the solution

$$G^*=G_T^*+\frac{1}{\alpha-\theta}\ln\left(\frac{\theta}{\beta}\frac{\bar{A}}{T-\bar{A}}\right) \tag{1.14}$$

which is the solution so long as

$$\theta\bar{A}>\beta\left(T-\bar{A}\right). \tag{1.15}$$

The left-hand side of equation 1.15 represents the marginal return to the group of adding an additional member whose most distant trade cannot be trusted (and for whom the payoff is therefore proportional to $\bar{A}$). The right-hand side represents the cost of adding a marginal group member, which is that the marginal *Trust* interaction now cannot be sustained, and inolves payoffs of $\bar{A}$ instead. Optimal additions to the group are finite (from equation 1.14) since both $\alpha - \theta > 0$ and $\beta > 0$. Beyond the margin, additional group members yield lower expected returns as a result of imperfect matching (the $\alpha - \theta$ term), while continuing to cost additional *Trust* trades (the β term).

This solution for G yields an implicit solution for G_T^*, which is the value of G_T that solves (from equation 1.10) the equation

$$\left[T - A\left(1 - e^{(\theta-\alpha)(1-p)L}\right)\right] - e^{(\theta-\alpha)G_T^*}\left[T - \bar{A}\left(1 - \frac{\beta}{\theta}\frac{T-\bar{A}}{\bar{A}}\right)\right] = \frac{c}{\lambda}\left(\frac{\theta}{\beta}\frac{\bar{A}}{T-\bar{A}}\right)^{\frac{\beta}{\alpha-\theta}} e^{(\beta+\theta)G_T^*}$$

The left-hand side of this equation is concave in G_T, while the right-hand side is convex in G_T, so the solution is unique so long as the left-hand side is greater than zero for $G_T = 0$.[17] The second order conditions confirm that the solution to the problem is indeed the maximum.

Case 2: Full Trust Groups.

If the feasibility constraint binds, $G_T = G$, so the problem is simply

$$\max_{G} \lambda T\left(1 - e^{(\theta-\alpha)G}\right) \tag{1.16}$$

$$\text{s.t.} \quad \lambda T\left(1 - e^{(\theta-\alpha)G}\right) \geq \lambda A\left(1 - e^{(\theta-\alpha)(1-p)L}\right) + ce^{(\beta+\theta)G} \tag{1.17}$$

which is maximized at the greatest solution (if one exists) of the IC constraint. This leads to the first proposition of the paper.

Proposition 1 *In a population of size L, if an optimal group exists, then:*

[17] This condition need not be true. In that case, there are either no solutions (and no groups) or two solutions, in which case the greatest value of G_T^* is relevant.

1. *if* $\frac{\theta}{\beta}\frac{\bar{A}}{T-\bar{A}} > 1$*, an optimal group is of size* G^* *where* G^* *is equal to*

$$G^* = G_T^* + \frac{1}{\alpha - \theta}\ln\left(\frac{\theta}{\beta}\frac{\bar{A}}{T-\bar{A}}\right).$$

G_T^* *is the greatest value of* G_T *for which the inequality*

$$T - e^{(\theta-\alpha)G_T}\left[T - \bar{A}\left(1 - \frac{\beta}{\theta}\frac{T-\bar{A}}{\bar{A}}\right)\right] \geq A\left(1 - e^{(\theta-\alpha)(1-p)L}\right) + \frac{c}{\lambda}\left(\frac{\theta}{\beta}\frac{\bar{A}}{T-\bar{A}}\right)^{\frac{\beta}{\alpha-\theta}} e^{(\beta+\theta)G_T}$$

holds. In these groups, each agent plays $a_j = T$ *if* $x \leq G_T^*$ *and* $a_j = \bar{A}$ *otherwise.*

2. *if* $\frac{\theta}{\beta}\frac{\bar{A}}{T-\bar{A}} \leq 1$*, an optimal group is of size* G^* *where* G^* *is the greatest value of* G *for which the inequality*

$$\lambda T\left(1 - e^{(\theta-\alpha)G}\right) \geq \lambda A\left(1 - e^{(\theta-\alpha)(1-p)L}\right) + ce^{(\beta+\theta)G}$$

holds. In these groups, each agent plays $a_j = T$ $\forall x \leq G^*$*.*

The distinction between full-trust groups and semi-trust groups, and the central role of informational crowding out, highlights why, compared with a constant returns matching function, the matching function as specified, with its limited increasing returns, is actually disadvantageous for finite group formation. If instead all agents were guaranteed to meet someone each period, there would be a direct congestion externality in addition to the information externality modelled here. That is, adding an extra (non-trust) member of the group would reduce the probability of meeting a trust member of the group. For certain functional forms, this force could in itself be strong enough to motivate the formation of groups of finite size, even without informational crowding out.

Since none of the substantive results depend on whether optimal groups are full-trust or semi-trust groups, for convenience in presenting further results, I make the following assumption.

Assumption A1: *Assume that* $\frac{\theta}{\beta}\frac{A}{T-A} \leq 1$*, so that optimal groups are full-trust groups.*

The results from proposition one yield an implicit function for G^*, the value of G_T (and G) at which the incentive compatibility constraint exactly binds. G^* is implicitly determined as the greatest solution (if one exists) to the implicit funtion, $F(G^*)$:

$$\begin{aligned} F(G(z);z) &= \lambda T\left(1-e^{(\theta-\alpha)G}\right)-\lambda A\left(1-e^{(\theta-\alpha)(1-p)L}\right)-ce^{(\beta+\theta)G}=0 \\ \text{where } z &= \{\alpha,\theta,L,\delta,m,p,T^+,T,A\} \end{aligned}$$

Using the implicit function theorem, I can investigate the properties of G^* with respect to the vector of parameters, z. The partial derivative of $F(\cdot)$ with respect to G is

$$F_G=\alpha Te^{(\theta-\alpha)G}-(\beta+\theta)\,ce^{(\beta+\theta)G}.$$

It is straightforward to show that $F(G(x);x)$ is a concave function with a maximum at

$$\tilde{G}=\frac{1}{\alpha+\beta}\ln\left(\frac{\alpha}{\beta+\theta}\frac{\delta}{1-\delta}\frac{T}{T^+-T}\bar{m}p\right).$$

If a group exists, $F(\tilde{G})\geq 0$ and $G^*\geq\tilde{G}$, with the second inequality holding strictly if and only if the first inequality holds strictly. Since $\tilde{G}$ is the maximum of a strictly concave function and $G^*>\tilde{G}$, it follows that $F_{G^*}<0$.

The incentive compatibility constraint is illustrated in Figure 1.2. The positive part of $F(G)$ is the Group Return, the expected value of a group of size G. The negative part is the Cheat Return, the sum of the expected value of the market and the appropriately discounted value of cheating in a match at distance G. Group size is determined by the upper intersection of the two curves. A group equilibrium need not exist: this is the situation depicted in CR'. The fact that the Cheat Return intersects the Group Return from below illustrates intuitively why $F_{G^*}<0$: $F(G)$ is intuitively the difference between the return to trusting and the return to cheating, which is zero (and falling) at G^*.

Using this result, the comparative statics for optimal group size (and the extent of trust) are summarised in the proposition below.

Proposition 2 *The most distant trade over which the group can maintain trust, G_T, is increasing in the discount factor, δ, the effectiveness of monitoring, m, the effectiveness of group search, p, and the return to trust, T. It is decreasing in the attractiveness of cheating, T^+, the productivity of outside opportunity, A, the size of the total population, L, and the rate of*

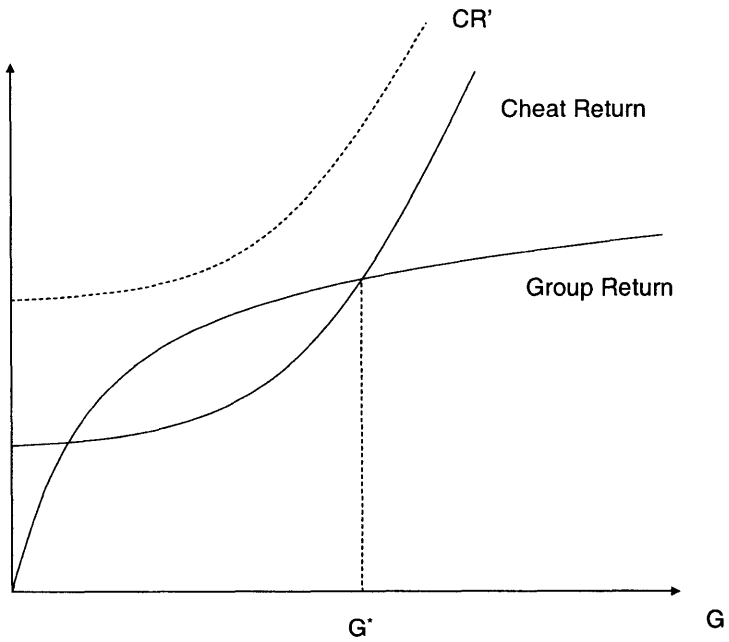

Figure 1-2: Incentive Compatibility and Group Size

decline of information, β. For sufficiently large groups, G_T is increasing in localization, α, and decreasing in returns to scale, θ.

Proof. *The proof follows immediately from differentiating $F(G)$.* ■

1.2.4 Equilibrium Group Formation

The equilibrium formation concept for groups is Nash. In a pre-play phase, each individual makes their decision to join a group, taking as given the decisions of all other individuals. Therefore, all groups will be symmetric (since otherwise some agents would join a different group) and all agents will be indifferent between joining their group and joining any other group.

A particular problem arises because, when there is a continuum of agents, any move by a single agent is infinitesimally small, and hence has no effect on overall group payoffs. Accordingly, without further refinement, *any* symmetric distribution of agents into groups is a possible equilibrium. I therefore consider moves only by small but measurable sets of agents, and take the limit of this process as the measure of the set shrinks towards zero. Similar problems are dealt with in Dubey and Shapley (1994) and Peters (1991). A related problem is that, if the population is segmented into very small groups, agents will choose to seek the market, and hence, if they can perfectly avoid their groups, their expected payoff is locally invariant to group size. An ϵ probability of mistakenly finding the group resolves this issue, since it breaks the indifference of agents to the size of their group.

Consider a partition of the population into J subsets $\{S_1, ..., S_J\}$, each having equal measure, $\frac{L}{J}$. Each S_j has a strategic choice over group location, in that it can choose among the set of locations, N. Any allocation of subsets to groups is an equilibrium if no subset can change groups and attain higher expected utility from the alternative group. Clearly, this implies that in equilibrium, all groups must provide equal expected utility.

The expected utility of a group is a continuous function of group size with a unique maximum, so long as the group is large enough so that its members seek it rather than the market. The per-period utility from joining group n is therefore

$$\begin{aligned} (1-\epsilon) V_m + \epsilon V_g (G_n) \text{ if } G_n &< G_{\min} \\ (1-p) V_m + p V_g (G_n) \text{ if } G_n &\geq G_{\min} \end{aligned}$$

where $G_{\min}$ is the minimum group size such that agents seek the group, given the size of the market. If a group equilibrium exists, $G_{\min} < G^*$, so $V_g' (G) > 0 \; \forall \; G < G_{\min}$. In order to rule out potential co-ordination problems over group location, assume that each potential location is ordered, so that, for example, if equilibrium involves $\tilde{n}$ groups, these are formed in the first $\tilde{n}$ potential locations. Furthermore, restrict attention to symmetric mixed-strategy equilibria, so that, when there are $\tilde{n}$ groups in equilibrium, each S_j sets

$$\begin{aligned} \sigma_j (l_i) &= \frac{1}{\tilde{n}} \text{ if } i \leq \tilde{n} \\ &= 0 \text{ if } i > \tilde{n} \end{aligned}$$

Given the actions of all other agents, the relevant component of the payoff is effectively simply the expected value of group membership. Since the expected value of group membership is continuous in group size, a symmetric Nash equilibrium satisfies the following properties:

Definition 3 *In a group Nash equilibrium:*

1. *each subset of the population, S_j, plays a mixed strategy profile in the pre-play phase,*

$$\begin{aligned}\sigma_j(l) &= \frac{1}{\tilde{n}} \forall i \leq \tilde{n} \\ &= 0 \forall i > \tilde{n}\end{aligned}$$

   Each group has the same size, $\tilde{G} = p\frac{L}{\tilde{n}}$, and payoff, $V_g(\tilde{G})$.

2. *All agents seek their group in the play phase, playing $F_j = n$. The size of the market is $(1-p)L$.*

3. *$V_g(\tilde{G}) \geq V_g(\tilde{G} + S_j) \forall j \in J$.*

The first property is simply a convenient way to sort agents into n symmetric groups without relying too much on co-ordination amongst the agents. Any arbitrary set of pure or mixed strategies that resulted in symmetric groups would also be an equilibrium. The second property establishes that by definition a group equilibrium involves active groups. This distinguishes the concept from the alternative equilibrium of the game, in which agents seek the market each period. The final property requires that each of the subsets of players, S_j, operating as a single agent, at least weakly prefers the group in which they find themselves to any other group. The following definition extends the concept of Nash equilibrium to a continuum of agents.

Lemma 1 *With a continuum of agents, a Nash equilibrium requires that $V_g'\left(\tilde{G}\right) < 0$.*

Proof. *The second property of the Nash equilibrium requires that $V_g(\tilde{G}) \geq V_g(\tilde{G} + S_j)$. Therefore, for any measurable S_j,*

$$\frac{V_g(\tilde{G} + S_j) - V_g(\tilde{G})}{S_j} \leq 0.$$

Now take the limit as $J \to \infty$, so that $S_j \to 0$, and apply the definition of a derivative:

$$\lim_{S_j \to 0} \frac{V_g(\tilde{G} + S_j) - V_g(\tilde{G})}{S_j} = V_g'(\tilde{G}) \leq 0.$$

■

The lemma formalises the intuition that in equilibrium, the value associated with a group cannot be increasing in group membership, since if it were, for some sufficiently fine partition of the population, some S_j would find it in their interest to move to the smaller group.

Lemma 2 *Any Nash equilibrium can have at most n^* groups, where n^* is the integer for which it is true that $n^* (G^*/p) < L < (n^* + 1)(G^*/p)$.*

Proof. *Consider a candidate equilibrium with $k > n^*$ groups. Each group has size $G^k = \frac{pL}{k}$. Then $G^k < G^*$. But $V_g(G)$ is a strictly convex function that attains a maximum at G^*. Hence, $V_g'(G^k) > 0$. But equilibrium requires $V_g'(G) < 0$. Hence, the candidate equilibrium cannot be an equilibrium.* ■

Although the equilibrium concept ensures there cannot be too many groups in equilibrium, it does not rule out the possibility that fewer than n^* groups may form, since this would require coordination amongst a larger coalition of players. The following lemma proves that the optimal group Nash equilibrium is the equilibrium involving n^* groups.

Lemma 3 *The optimal group Nash equilibrium contains n^* groups.*

Proof. *For a given market size, the optimal group size is G^*, with total membership G^*/p. From Definition 1, $V_g'(G) < 0 \forall G > G^*$. With $k' \leq n^*$ groups, group size is $\frac{pL}{k'}$, of value $V_g\left(\frac{pL}{k'}\right)$. It follows immediately that $V_g(G)$ is maximized at the minimum feasible value of G, $G = \frac{pL}{n^*} \geq G^*$. Therefore, the optimal group Nash equilibrium involves the maximum feasible number of groups, n^*.* ■

The preceding lemma establishes that the optimal group Nash equilibrium will involve groups slightly larger than G^*, where G^* is the optimal size of a group given market size. This makes it difficult to establish comparative statics for changes in parameters in general. A small change in a parameter, for example, may move actual group size much closer to optimal group size than before. In comparing welfare, therefore, I abstract from this problem, and compare group outcomes at G^*.[18]

[18] This approach is also often used in the club goods literature (see Sandler and Tschirhart (1997) for references).

1.3 Defensive Group Strategies

Proposition 2 established that sustainable group size falls in response to improvements in the outside opportunity, A, which has so far been exogenous. However, smaller groups come with the cost of fewer gains from increasing returns to scale. It is therefore possible for group members to be *worse off* as a result of improvements in the productivity of trade interactions. This result is established in the following proposition.

Proposition 4 *If the value of trade is exogenous and equal to A, improvements in A can reduce the welfare of group members.*

Proof. *The expected return to group membership each period is*

$$\begin{aligned} V(t) &= pV(g) + (1-p)V(a) \\ &= pT\left(1 - e^{(\theta-\alpha)G}\right) + (1-p)A\left(1 - e^{(\theta-\alpha)(1-p)L}\right) \end{aligned}$$

Accordingly, the change in value as a result of an exogenous change in A is

$$\frac{dV(t)}{dA} = pT\frac{\partial\left(1 - e^{(\theta-\alpha)G}\right)}{\partial G}\frac{dG}{dA} + (1-p)\left(1 - e^{(\theta-\alpha)(1-p)L}\right).$$

By the implicit function theorem,

$$\frac{dG}{dA} = -\frac{F_A}{F_G} = \frac{-\left(1 - e^{(\theta-\alpha)(1-p)L}\right)}{c(\beta+\theta)e^{(\beta+\theta)G} - (\alpha-\theta)Te^{(\theta-\alpha)G}} < 0$$

Substituting through and solving yields

$$\frac{dV}{dA} < 0 \Leftrightarrow G^* < \frac{1}{\alpha+\beta}\ln\left(\frac{T}{T^+ - T}\frac{p}{1-p}\frac{\delta}{1-\delta}\frac{\alpha}{\beta+\theta}\bar{m}\right).$$

Note that the right hand side is approximately equal to $\tilde{G} + \frac{p}{\alpha+\beta}$. ■

Recall that, at the start of each round, each group can announce the value of trade that its members will play, s_i. If all groups are symmetric, then this is the value of trade. If there is a disagreement among the groups of which players are members (or if one player is not a member of an active group), then assume for convenience that each player plays $\min\{s_1, s_2\}$ where s_i

is the preannounced strategy, if any, and $\bar{A}$ otherwise.

It is important to remember that the reason for reducing the value of trade is to reduce the outside option of group members. If, for example, ostracised group members could be perfectly recognised by non-group members willing to play $\bar{A}$, there would be no reason for groups to reduce the value of trade. Inter-group hostility, in this interpretation, is designed to increase loyalty to one's own group: reduced payoffs to other groups are merely a by-product.

Allowing the group to choose A yields the following program for an optimal group:

$$\begin{aligned}
\max_{G,A}\mathcal{L} &= \lambda\left[pT\left(1-e^{(\theta-\alpha)G}\right)+(1-p)\,A\left(1-e^{(\theta-\alpha)(1-p)L}\right)\right] && (1.18)\\
&+\mu_1\left(\lambda T\left(1-e^{(\theta-\alpha)G}\right)-\lambda A\left(1-e^{(\theta-\alpha)(1-p)L}\right)-ce^{(\beta+\theta)G}\right) && (1.19)\\
&+\mu_2\left(\bar{A}-A\right) && (1.20)\\
&+\mu_3 A && (1.21)
\end{aligned}$$

Assuming an interior solution for A, then from the first order conditions $\mu_1 = 1-p$. Accordingly, for interior solutions, the value of relaxing the IC constraint is proportional to how often group members find the market. The optimal size of the group is

$$\begin{aligned}
G^{**} &= \frac{1}{\alpha+\beta}\ln\left(\frac{T}{T^{+}-T}\frac{p}{1-p}\frac{\delta}{1-\delta}\frac{\alpha}{\beta+\theta}\bar{m}\right) && (1.22)\\
&= \frac{\ln\kappa}{\alpha+\beta} && (1.23)
\end{aligned}$$

while the optimal value of trade is

$$\begin{aligned}
A^{**} &= \frac{T}{1-e^{(\theta-\alpha)(1-p)L}}\left[1-\kappa^{\frac{\theta-\alpha}{\alpha+\beta}}\left(1+\frac{\alpha-\theta}{(\beta+\theta)(1-p)}\right)\right] && (1.24)\\
&= \frac{T}{1-e^{(\theta-\alpha)(1-p)L}}\left[1-e^{(\theta-\alpha)G^{**}}\left(1+\frac{\alpha-\theta}{(\beta+\theta)(1-p)}\right)\right]. && (1.25)
\end{aligned}$$

This approach yields the following analytic comparative statics.

Proposition 5 *When the value of outside trade can be limited, G^{**} is increasing in T, p, δ and m. It is decreasing in β, θ and T^{+}. It is decreasing in α if $\ln\kappa > \frac{\alpha+\beta}{\alpha}$ and increasing otherwise. Population size, L, has no effect on optimal group size. A^{**} is increasing in δ,m*

and T. It is decreasing in T^+ and L.

Proof. *The proof follows immediately from differentiating equations 1.22 and 1.24.* ■

These comparative statics are relatively straightforward except that, as before, group size increases in α only for sufficiently large groups.[19] Note that, when the group can choose the value of trade, population size no longer affects optimal group size, but it does raise optimal hostility. This is another channel whereby modernization (here in the form of a population expansion) can increase ethnic hostility.

Further results with respect to the optimal value of trade are more difficult to sign. The reason for this can be found in equation 1.25 . Rearranging the equation slightly,

$$A^{**}\left(1 - e^{(\theta-\alpha)(1-p)L}\right) = T\left\{1 - e^{(\theta-\alpha)G^{**}}\left[1+Y\right]\right\}, \tag{1.26}$$

where $Y = \frac{\alpha-\theta}{(\beta+\theta)(1-p)}$. The left hand side is the expected value of market transactions, the right hand side the expected value of group transactions, up to an adjustment factor, $1+Y$. Larger values of Y reduce the optimal value of trade, other things being equal. Y is the product of $\frac{1}{1-p}$ and $\frac{\alpha-\theta}{\beta+\theta}$. Both these terms can be interpreted in terms of the opportunity cost of hostility to the market. The first is larger when the group is more segregated, reflecting the lower cost of hostility in this case. The second term increases when localization increases or returns to scale fall. It also increases when information dissipates less quickly, reflecting the smaller marginal cost to the group of expanding trust.

While analytic expressions for the comparative statics with respect to α, β, θ and p can be derived, it is difficult to sign them, because each parameter has several offsetting effects. In addition to the direct effect on the optimal value of trade through Y, there is a further effect through optimal group size, and still a further effect through the size of the market.

[19]This condition seems likely to be met. Assuming conservatively that $p = .9$, $\delta = .9$, $\bar{m} = .9$ and $T^+ = 2T$, and rearranging, the condition requires

$$\ln\left(\frac{73}{e}\right) > \frac{\beta}{\alpha} - \ln\alpha + \ln\left(\beta+\theta\right).$$

By assumption, $\theta < \alpha$. Therefore, replacing θ with α and noting that $\ln(1+X) < X$ for $X > 0$, the condition is guaranteed if

$$\frac{\beta}{\alpha} < \frac{1}{2}\ln\left(\frac{73}{e}\right) \approx 1.65.$$

This simply states that information must not decline too rapidly relative to meeting probabilities.

The most helpful approach turns out to be to assume that parameters are such that $A^{**} = 0$, and then consider the effect of small changes in the parameters on the sign of A^{**}. Note that

$$A^{**} > 0 \Leftrightarrow 1 - e^{(\theta-\alpha)G^{**}} [1 + Y] > 0$$

For Y sufficiently small, this approximately implies

$$G^{**} > \frac{1}{(1-p)(\beta+\theta)}.$$

Taking the derivative of both sides shows that (locally), A^{**} is unambiguously decreasing in p and θ. It is increasing in α so long as G^{**} is sufficiently large, that is, so long as G^{**} itself is increasing in localization. Finally the effect of β depends on the parameters as follows:

$$\frac{dA^{**}}{d\beta} < 0 \Leftrightarrow \frac{1}{\alpha+\beta}\left(G^{**} + \frac{1}{\beta+\theta}\right) > \frac{1}{(1-p)(\beta+\theta)^2}.$$

Most of these results are intuitively appealing. As p increases, for example, reducing the value of trade becomes less costly but remains effective in expanding the range of trust. This strongly supports an interpretation of ethnic group conflict that emphasizes the effect of colonial policies that divided subject populations, whether into Hindus and Muslims in India, or Hutus and Tutsis in Rwanda. Similarly, a higher value of θ raises the value to the group of expanding. Finally, when matches are more localized (α high), high value but distant trades contribute relatively less to utility, and hence reducing the value of trade is less worthwhile.

1.4 Social Welfare

This section considers the circumstances under which it is optimal for groups to form, and the circumstances under which it is possible for groups to form even if they are not socially optimal. For convenience, it is assumed that the optimal group Nash equilibrium is the equilibrium of interest, and that the population can be partitioned into precisely optimal groups.

Consider first groups that take the value of outside trade as exogenous. Given that an agent expects everyone else to join a group, they also do so if

$$T\left(1-e^{(\theta-\alpha)G^*}\right) > A\left(1-e^{(\theta-\alpha)(1-p)L}\right). \tag{1.27}$$

That is, the agent joins the group so long as it provides greater expected utility than the market, given that the market is of size $(1-p)L$. The condition for social optimality, however, is that

$$pT\left(1-e^{(\theta-\alpha)G^*}\right) + (1-p)A\left(1-e^{(\theta-\alpha)(1-p)L}\right) > A\left(1-e^{(\theta-\alpha)L}\right)$$

which can be rearranged as

$$T\left(1-e^{(\theta-\alpha)G^*}\right) > A\left(1-e^{(\theta-\alpha)(1-p)L}\right) + \frac{A}{p}\left[e^{(\theta-\alpha)(1-p)L} - e^{(\theta-\alpha)L}\right]. \tag{1.28}$$

Taken together, equations 1.27 and 1.28 show that there is a wedge between the private and social conditions for group membership, equal to

$$\frac{A}{p}\left[e^{(\theta-\alpha)(1-p)L} - e^{(\theta-\alpha)L}\right].$$

Graphically, Figure 1.3 shows how the social optimum varies depending on the degree of segregation of the groups. For values of $p < p^*$, the value of the group is undefined, since no agents actually go to the group, and of course the group would be unable to sustain trust even if they did. At p^*, a full-trust group can be sustained. Once the group becomes privately optimal, its value increases, because it is able to sustain trust among its members. The value of the group can be convex over some ranges of p, since an increase in p not merely increases the value of a group of given size, but also increases the potential size of the group. If groups are sufficiently segregated, so that $p > p^{**}$, the group equilibrium dominates the market equilibrium. Note that the expected value of group transactions must strictly exceed the expected value of the market equilibrium in order for the group equilibrium to be optimal, since total welfare includes both group and (small) market transactions. For intermediate values of p such that $p^* < p < p^{**}$, the optimal group equilibrium is dominated by the market equilibrium.

Attempts to encourage integration by different groups therefore have an ambiguous effect on social welfare. If the different groups are already somewhat integrated, they may be suboptimal,

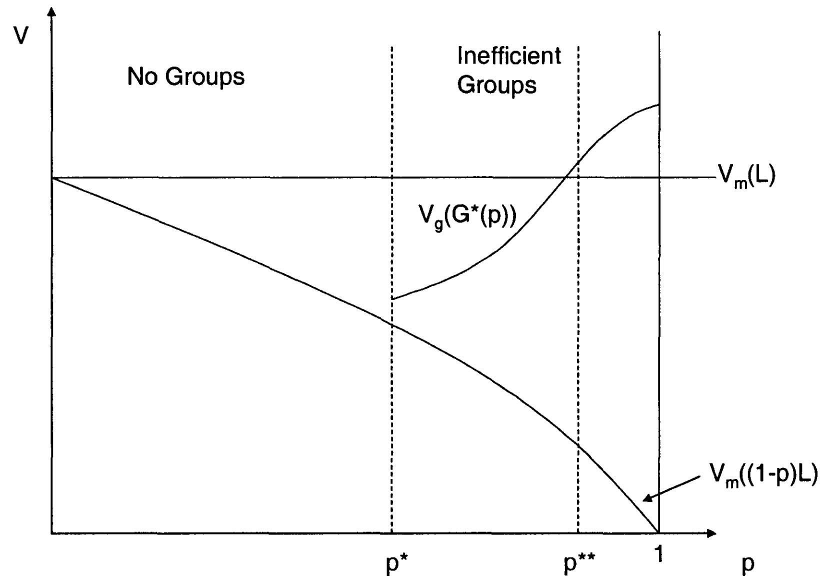

Figure 1-3: Social and Private Efficiency of Groups

and forcing p to fall may increase social welfare, although for this, p must fall below p^*. If, on the other hand, groups are quite distinct, attempts to force integration are likely to prove counter-productive.

Consider now groups that can also set the value of trade optimally, and assume an interior solution. In this case, it is straightforward to use the envelope theorem to examine changes in welfare. These can be compared with changes in welfare for a population of identical size that is in the market equilibrium.

Some parameters, of course, affect only the group equilibrium. Increases in p and falls in β raise welfare in the optimal group equilibrium without affecting the market equilibrium. Increases in θ and reductions in α, however, disproportionately favor populations that are in the market equilibrium. This result is proved with respect to θ in the following proposition. The result with respect to α can be proved similarly.

Proposition 6 *The elasticity of income in location D with respect to an increase in returns to* *scale, θ, is denoted $\varepsilon_{V_D,\theta}$. It is greater (1) for a population in the market equilibrium than for the same population in a group equilibrium and (2) for a population in a market equilibrium than for an optimal group of the same size.*

Proof. *The elasticity of value in a location of size P with equilibrium payoffs proportional to π with respect to a change in θ is*

$$\varepsilon_{V_D,\theta} = \frac{dV_D(P)}{d\theta}\frac{\theta}{V_D(P)} = \frac{\theta}{\alpha-\theta}\frac{1-e^{(\theta-\alpha)P}\left(1+(\alpha-\theta)P\right)}{1-e^{(\theta-\alpha)P}} \tag{1.29}$$

Taking the derivative with respect to P yields

$$\frac{d\varepsilon_{V_D,\theta}}{dP} = \frac{\theta e^{(\theta-\alpha)P}}{\left(1-e^{(\theta-\alpha)P}\right)^2}\left[e^{(\theta-\alpha)P}+(\alpha-\theta)P-1\right] > 0.$$

Note that the elasticity is a function of population size but not equilibrium payoffs. Accordingly, the elasticity of value in the market equilibrium is simply $\varepsilon_{V,\theta}(L)$. The elasticity of value in the group equilibrium is a convex combination of the elasticity of value in the group, $\varepsilon_{V_g,\theta}(G)$ and in the small market, $\varepsilon_{V_m,\theta}((1-p)L)$. Group size falls when θ increases, so $\varepsilon_{V_g,\theta}(G)$ overstates the true elasticity of returns to the group. Since both G and $(1-p)L$ are smaller than L, both $\varepsilon_{V_g,\theta}(G)$ and $\varepsilon_{V_m,\theta}((1-p)L)$ are less than $\varepsilon_{V_m,\theta}(L)$.

Consider now the elasticity of value for a group of optimal size that chooses hostility optimally. Applying the envelope theorem yields:

$$\varepsilon_{V_g,\theta} = \frac{\theta}{\alpha-\theta}\frac{1-e^{(\theta-\alpha)G^*}\left(1+(\alpha-\theta)G^*\right)-\frac{c(1-p)}{T\lambda}e^{(\beta+\theta)G^*}(\alpha-\theta)G^*}{1-e^{(\theta-\alpha)G^*}-\frac{c(1-p)}{T\lambda}e^{(\beta+\theta)G^*}}. \tag{1.30}$$

Comparing equations 1.29 and 1.30 yields, after considerable simplification, that

$$\varepsilon_{V_D,\theta}(P) > \varepsilon_{V_g,\theta}(P) \Leftrightarrow e^{(\theta-\alpha)P} > 1-(\alpha-\theta)P$$

which is true for all values of P. ■

This result provides one interpretation of Easterly and Levine's (1997) African growth tragedy: economies in which groups are important gain less from modern technologies that increase returns to scale than those in which the population is united. This is both because of a pure fragmentation effect, reflecting returns to scale, and because some of the returns to scale are lost by the need for increased inter-group hostility.

1.5 Non-Stationarity and Shocks

Within the stationary model, multiple equilibria are possible, and even the best group outcome may be inferior to the market equilibrium. Is it possible to say anything about a non-stationary model, for example when there is a shock to the value of some parameters?

Recall that the population is replaced at rate δ_0, so that each period $\delta_0 L$ new agents arrive in the population and face a choice of group. Suppose that the number of optimal groups has increased as a result of, say, a one-time unanticipated increase in θ. It is in the social interest to reduce the size of each group and increase the number of groups. But each existing member of the population is already in their group and it is infinitely costly to switch groups. Therefore, a new group will form only if it is in the private interest of new members of the population.

To consider new group formation in this context, agents require a rule for when they will form a new group, since otherwise any arbitrary beliefs might support some kind of group formation. A natural rule to consider is to allow incoming agents to form a new group if, assuming all other incoming agents make the same calculation, it is privately worthwhile to form a new group. Assume for convenience that the flow of agents is large enough so that, if all entering agents form a new group, they prefer to find the group rather than the market. This assumption keeps the size of the market constant throughout any transition.

If the flow of the population is large enough, so that $\delta_0 L > G^*$, an entering cohort can form a group immediately. If, however, the population flow is small, then the population may be stuck with too few groups for some time, or even indefinitely. New members of the population form a new group only if the returns to doing so in the transition exceed the returns to joining an existing, inefficiently large group. Eventually, the new group will grow until it reaches the same size as the old groups. Note that this will be exact, since otherwise new agents would reallocate themselves between groups. Denoting the new group as G^- (since it is smaller than optimal) and the old group as G^+, the new group forms if

$$\sum_{t=0}^{T} \delta^t V_g\left(G_t^-\right) \geq \sum_{t=0}^{T} \delta^t V_g\left(G_t^+\right),$$

where T is the period in which G^- and G^+ attain the same size. Although a complete characterization of these dynamics is difficult, intuitively new groups will tend to form only when the inefficiencies of the old groups have become large. Moreover, the larger the optimal group size relative to the rate of population flow, the more difficult it will be to start new groups. It follows that greater segregation of groups, which raises optimal group size, also makes it more difficult to form new groups if it becomes socially optimal to do so.

The following lemma is useful in discussing non-stationarity.

Lemma 4 *The social return to joining a group smaller than the optimum exceeds the private return.*

Proof. *Consider a static choice of groups for a small but measurable quantity of agents, S_j. Suppose that there is at least one group of less than optimal size, G^-, and one of greater than optimal size, G^+. When all agents are group members, the size of the market is fixed at $(1-p)L$. If there are no further changes in group structure, agents join the small group if*

$$V_g\left(G^- + S_j\right) - V_g\left(G^+ + S_j\right) > 0.$$

Taking the limit as $S_j \to 0$, agents join the small group if

$$V_g\left(G^-\right) - V_g\left(G^+\right) > 0 \tag{1.31}$$

The utilitarian social criterion, however, is

$$\begin{aligned} \left(G^- + S_j\right) V_g\left(G^- + S_j\right) + G^+ V_g\left(G^+\right) &> G^- V_g\left(G^-\right) + \left(G^+ + S_j\right) V_g\left(G^+ + S_j\right) \\ V_g\left(G^- + S_j\right) - V_g\left(G^+ + S_j\right) &> G^+ \left[\frac{V_g\left(G^+ + S_j\right) - V_g\left(G^+\right)}{S_j}\right] \\ &\quad - G^- \left[\frac{V_g\left(G^- + S_j\right) - V_g\left(G^-\right)}{S_j}\right]. \end{aligned}$$

Again, taking the limit as $S_j \to 0$,

$$V_g\left(G^-\right) - V_g\left(G^+\right) > G^+ V_g'\left(G^+\right) - G^- V_g'\left(G^-\right). \tag{1.32}$$

Since by construction, $V_g'\left(G^+\right) < 0$ and $V_g'\left(G^-\right) > 0$, the right-hand side is strictly negative, and therefore, comparing equation 1.32 with equation 1.31, agents are too reluctant to join the small group. ■

The following result shows that persistence of overly large groups has important negative consequences for welfare. In particular, it raises the degree of hostility that each group optimally pursues.

Proposition 7 *Increases in group size above the optimal group size, G^{**}, reduce the constrained optimal value of trade, $A^\dagger$.*

Proof. *For given group size, the group solves the following program:*

$$\max_{G_T, A} \lambda p \left\{ T\left(1 - e^{(\theta-\alpha)G_T}\right) + \bar{A}\left(e^{(\theta-\alpha)G_T} - e^{(\theta-\alpha)G}\right)\right\} + \lambda(1-p) A\zeta$$

$$\begin{aligned} s.t.\ \lambda\left[T\left(1 - e^{(\theta-\alpha)G_T}\right) + \bar{A}\left(e^{(\theta-\alpha)G_T} - e^{(\theta-\alpha)G}\right)\right] &\geq \lambda A\zeta + c e^{\beta G + \theta G_T} \\ \bar{A} &\geq A \end{aligned}$$

where $\zeta = 1 - e^{(\theta-\alpha)(1-p)L}$.

Assuming the second constraint is slack, the first order conditions yield the usual result that the Lagrange multiplier $\mu_1 = 1 - p$ and substituting through yields

$$\begin{aligned} G_T^\dagger &= \frac{1}{\alpha}\ln\left(\frac{\left(T - \bar{A}\right)\alpha}{(1-p)c\theta}\right) - \frac{\beta}{\alpha}G \\ A^\dagger &= \frac{1}{\zeta}\left[T\left(1 - e^{(\theta-\alpha)G_T^\dagger}\right) + \bar{A}\left(e^{(\theta-\alpha)G_T^\dagger} - e^{(\theta-\alpha)G}\right) - \frac{c}{\lambda}e^{\beta G + \theta G_T^\dagger}\right] \end{aligned}$$

Take the total derivative,

$$\frac{dA^\dagger}{dG} = \frac{\partial A^\dagger}{\partial G} + \frac{\partial A^\dagger}{\partial G_T}\frac{dG_T^\dagger}{dG}$$

These terms are:

$$\begin{aligned}
\frac{\partial A^{\dagger}}{\partial G} &= \frac{1}{\zeta}\left[\bar{A}\left(\alpha-\theta\right)e^{(\theta-\alpha)G}-\frac{c\beta}{\lambda}e^{\beta G+\theta G_T}\right]\\
\frac{\partial A^{\dagger}}{\partial G_T} &= \frac{1}{\zeta}\left[\left(T-\bar{A}\right)\left(\alpha-\theta\right)e^{(\theta-\alpha)G_T}-\frac{c\theta}{\lambda}e^{\beta G+\theta G_T}\right]\\
\frac{dG_T^{\dagger}}{dG} &= -\frac{\beta}{\alpha}
\end{aligned}$$

This yields the result that $\frac{dA^{\dagger}}{dG}<0$ *if*

$$\alpha\bar{A}e^{(\theta-\alpha)G}<\beta e^{\theta G_T}\left[\left(T-\bar{A}\right)e^{-\alpha G_T}-\frac{c}{\lambda}e^{\beta G}\right]. \tag{1.33}$$

Now, taking the derivative of $V(G)$ *the same way (still treating total group size as given) yields the result that* $V'(G)<0$ *if*

$$\alpha\bar{A}e^{(\theta-\alpha)G}<\beta e^{\theta G_T}\left[\left(T-\bar{A}\right)e^{-\alpha G_T}-\frac{(1-p)\,c}{\lambda}e^{\beta G}\right]. \tag{1.34}$$

Comparing the right hand-size of equations 1.33 and 1.34, it is clear that a group being larger than the optimum size is a sufficient condition for it to choose to reduce the value of trade below the value it would have chosen at the optimum. ■

Hence, for a fixed number of groups, the externality to creating a new group implies not merely that new groups form inefficiently slowly, but also that existing groups become increasingly hostile to each other until and unless a new group eventually forms.

1.6 Discussion: France and Rwanda

This paper has suggested that understanding the endogeneity of fractionalization can yield novel insights into both the welfare implications of groups and the sources of inter-group hostility. It has also proposed a model in which hostility arises between groups in certain situations, not as a result of competition for resources or power but as a way of strengthening the group itself. This section considers the development experience of two countries, France and Rwanda, during the late 19th century and early 20th century, in some detail, to illustrate the mechanisms at work in the model. The reason for considering these two very different countries is that they represent relatively clear cases of the disappearance of ethnic groups (France) and their creation (Rwanda).

A final section considers several more disparate strands of literature and their implications for the model.

1.6.1 France, 1870-1914

Today, by most measures, France is both highly homogeneous and economically very successful. According to Alesina et.al.'s (2003) data, France has among the world's lowest rates of ethnic and linguistic fractionalization, and also a relatively low rate of religious fractionalization. This accords with most people's conception of France as a united nation with a strong sense of national identity, particularly proud of its culture and especially its language. But this apparent modern unity is in fact largely the product of a brief spurt of nation-building during the Third Republic (that is, between the end of the Franco-Prussian War and the start of the First World War).

Linguistically, France was highly diverse even late in the 19th Century. According to Eugen Weber (1976) (on whose book, *Peasants into Frenchmen*, most of this account is based), for "almost half the children who would reach adulthood in the last quarter of the [nineteenth] century" French was "a foreign language" (p67). In 1835, French was spoken by only the north-west corner of the country, its western border marked by Brittany, its eastern extremity little past Paris, and its southern border scarcely extending to Bourges. In other regions, French was supplanted by a variety of other languages: Catalan, German, Flemish, Breton, Basque, langue d'Oc, and other, now forgotten, varieties of *patois*. These were not mere dialects: French was entirely incomprehensible to many speakers of *patois*.

Fractionalization was not confined to the merely linguistic. Many people outside the valleys of the Loire and the Seine regarded France as a foreign country, and felt a "deep-seated sense of difference" from the French (p98). Around the Pyrenees as late as 1907, there is evidence that the loyalty of most lay with the Basques or with Spain, but not with France. The Breton forces in the Franco-Prussian War had to be reminded by their leader "I beg you to forget that you are Bretons, and to remember only that you are French", while a Breton delegation met

Woodrow Wilson at Versailles to plead for national self-determination.

Weber lists a number of factors that were instrumental in the successful creation of a national French identity during the Third Republic. At the top of his list is the creation of an extensive road and rail network, most importantly extending beyond highways to local roads that brought previously isolated villages into closer contact with the outside world. Weber writes

> So, roads and rail lines that were secondary in name but primary in fact brought the isolated patches of the countryside out of their autarky - cultural as well as economic - into the market economy and the modern world... The conjunction ... resulted in a crash program of national integration of unparalleled scope and effectiveness; a program that could operate on this scale only because, for the first time, economic and technological conditions offered the possibility for radical cultural change. (206)

The coming of rail and road had the effect both of bringing people from outlying areas into contact with other French, and of reducing the scope for artisanal handicraft compared to machines and factories. Thus, they can be thought of as reducing p and increasing θ, as well, perhaps, as reducing α, the localization of matches.

Another important factor in the spread of French language and national feeling was the extension of conscription, which can reasonably be considered a straightforward reduction in p. Conscription, particularly late in the century when it became almost universal, played an important role in expanding the use of French, although units, being regionally based, sometimes persisted with the use of *patois*. Resistance to the army also died out as conditions for soldiers improved to the point where they even exceeded the average conditions of rural life in many areas - an increase in A relative to T that the group was powerless to prevent.

Finally, many commentators have highlighted the importance of universal, compulsory and free public schooling, in French. Weber emphasises that schools became useful only once the world had already changed - in terms of the model, once productive parameters such as θ and α had already changed - but it is equally true that, these parameters having changed, the schools themselves were important instruments of unification.

Thus, the model suggests that the policies of integration pursued in France reduced sustainable group size until groups were no longer sustainable. While increasing θ and reducing α might be expected to lead to inter-group hostility, nation-building policies that reduced p limited hostility. Why urban French pursued these costly integrationist policies is admittedly an important question that lies largely outside the scope of the model. One interpretation is that they had simply already advanced to the stage where they functioned as a market, perhaps because of greater density of interactions and earlier access to industrialization. In that case, it was clearly in their interest to expand the scope of the market as much as possible.

1.6.2 Rwanda, 1860 to 1961

Rwanda, it need hardly be pointed out, is a very different country to France along many dimensions. But the contemporary importance of Rwanda's ethnic divide can lead to impressions of the role of ethnicity in Rwanda that are just as misleading as contemporary impressions of French ethnicity. The genocide of 1994 was the most recent and dramatic instance of the conflict between the country's two main ethnic groups, the Hutu and the Tutsi. The divide between these groups also lay at the heart of Rwanda's movement to independence from Belgium in the years 1959-1961. It may, therefore, be surprising to find that the divide between Hutus and Tutsis is largely a relatively modern creation.

The Hutu and Tutsi divide is highly unusual even in Africa. Many of the usual indicators of distinct ethnicity are missing. There is no language barrier between Hutus and Tutsis, and there are only minor linguistic variations in the pre-colonial language, Kinyarwanda, across regions (Newbury (1988)). Nor are there systematic religious differences between Hutus and Tutsis. On Ijwi Island in the 1970s, in neighbouring Zaire (as it was then known), Hutus and Tutsis were typically known not as Hutus or Tutsis, but either by their clan identity or simply as "Badusi" (a collective term for Rwandans in general, or sometimes for Tutsis in particular).

What, then, distinguishes the Hutu, who comprise over 80 per cent of the population, from the Tutsi, who comprise most of the remainder? One answer is that the division in some ways resembles a class division. Before independence, the Tutsi dominated both property and the apparatus of the state. In traditional usage, being a Tutsi signified possession of wealth and power almost as much as it signified descent. In her study of the prefecture of Kinyaga, Newbury (1988) argues that the concept of being a "Hutu" had little political significance until at least the reign of King Rwabugiri from around the 1860s. Successful Hutus often "became" Tutsis.

After the 1860s, however, the division between Hutus and Tutsis solidified. The early Europeans in Rwanda constructed elaborate theories of the racial distinctness of Hutus and Tutsis from each other.[20] First the Germans and later the Belgians consciously fostered the Hutu/Tutsi division, largely to create a cooperative group of local proxies. For example, the Belgians united in one person what had previously in many cases been the roles of three overlapping chiefs: whereas at least one of these had previously often been a non-Tutsi, the new position was almost invariably filled by a Tutsi (Prunier (1995)). Moreover, education was provided almost exclusively by Catholic missionary schools, that taught only - with rare exceptions - the favored Tutsis. The development of a Hutu identity arose largely in response to the increased privileges and rigidification of identity that Tutsis underwent under Rwabugiri and the colonial powers.

Colonization probably affected all of the parameters of the model. Colonists brought with them some European productive technologies, thereby increasing θ. However, investment in infratructure, for instance, was considerably lower than in European countries, so α may not have fallen very far. Many of the colonists' policies consciously fostered divisions between the population, increasing p. Even the reliance on what were perceived to have been traditional power structures, such as chiefs, reinforced the role of groups, perhaps reducing β. In France, by contrast, the central authorities consciously sent outsiders to remote areas to serve as bureaucrats and magistrates. On balance, these forces probably produced three key changes. First, there was some increase in the potential attractiveness of the market (θ increased), although probably considerably less than in France or other developed nations. Second, other changes did not systematically push the population towards the market equilibrium. In particular, increased p and reduced β reinforced the ability to form groups. Finally, the potential for group hostility was greatly increased, through the increases in both p and θ. The model therefore qualitatively captures some of the main developments in Rwanda that created the

[20]Modern authors remain divided on the highly emotive subject of the extent to which the ancestors of modern Tutsis were a distinct group from the rest of the population, possibly having migrated to Rwanda from another area in the distant past. For example, Newbury and Destexhe both downplay the significance of any distinct origins along these lines, while Prunier (p16) cautiously accepts the possibility of distinct origins.

setting for the extreme ethnic division of more recent times.

1.6.3 Other Evidence

Although Rwanda is in many ways unusual even within Africa in the exact circumstances (not to mention the intensity) of its conflict, many of the themes of construction of ethnic identity are similar throughout the rest of Africa. For example, Nigeria is one of the world's most ethnically heterogeneous nations. From around the 1920s, tribal associations emerged, founded in cities but with members who came originally from a particular area of the countryside. A distinguished early scholar of these groups, James Coleman, writing in 1952, saw these as giving voice to "ethnic groups previously inarticulate". He saw the primary role of these associations as providing mutual aid and protection, in accelerating acculturation to the cities, and in assisting development of the villages of their rural origin, rather than engaging in rent-seeking or competition (Coleman (1994)).[21]

Moreover, although the Hutu/Tutsi divide is relatively unusual in Africa in the extent to which it resembles a class or caste divide, the pattern of colonial rulers creating or exacerbating "ethnic" divides is widespread. Migdal (1988) studies the creation of new groups during the period of British rule in Sierra Leone in some detail, and cites studies relating to Zimbabwe, and further afield, the Karen on the Thai/Burmese border. Migdal writes

> tribes, linguistic communities, and ethnic groups changed dramatically in the late nineteenth and early twentieth centuries, in large part because of the colonialists' support of certain forms of social organizations. The new tribes maintained a relationship to the past... [but] the new social fragmentation ... did not depend entirely on old tribal symbols and ties. (130)

This process of group reformation fits quite neatly with the stylised process of new group formation described in the non-stationary section of the paper.

[21] A recent example of rearrangement of ethnicity is Somalia, which until recently appeared to be largely homogeneous (over 80 per cent Somali). However, following a disastrous war against Ethiopia, from 1976, and protracted civil strife, it is now rife with separatist movements defined by clan loyalties. (See Alesina et.al. (2003) and Fafchamps (2004).)

More recently, Ted Miguel (2003) has studied the role of ethnic diversity in public goods provision in neighbouring areas of Kenya and Tanzania. Miguel argues that, whereas ethnic divisions have played a key role in Kenyan politics, Tanzania has pursued a self-consciously nation-building strategy since independence. In Kenya, the local degree of ethnic fragmentation is heavily associated with poor local provision of public goods such as schools and well maintenance, whereas in Tanzania, local ethnic fragmentation appears to have no effect on local provision of public goods. The most important nation-building strategies cited by Miguel are linguistic unification and a nationalistic public school curriculum, mirroring the policies that Weber argues were effective in France in the late nineteenth century. While such strategies appear to have been succesful in reducing the effect of fractionalization in Tanzania, Miguel cautions that they may not be universally appropriate.

This caution raises the broader question of when the appropriate response to ethnic fractionalization is nation-building and when it is separation through, for example, self-determination. This question is particularly pertinent in Africa, with its colonial era boundaries that often bear little relation to ethnic divisions or even geography.[22] The model presented here suggests a framework for thinking about this issue in particular cases. When ethnic groups are already very distinct, redrawing national boundaries to achieve greater separation of groups may be optimal.[23] On the other hand, very heterogeneous nations with intermingled ethnic groups in frequent contact may have a better chance of pursuing nation-building policies successfully. Although policies that aim explicitly at nation-building may succeed under the right circumstances, they carry with them the cost of disrupting existing productive arrangements. They are more likely to be succesful when accompanied by policies that increase the relative attractiveness of the market, such as openness to modern technology and investment in suitable infrastructure.

[22] See Herbst (2000) for an influential recent discussion of the problem of national borders in Africa.

[23] Unfortunately, as the example of Rwanda highlights, groups can be effectively quite separate without necessarily being geographically segregated.

1.7 Conclusion

This paper has adopted a constructivist approach to group formation as a way of gaining understanding of inter-group hostility. It has focused in particular on ethnic groups, since previous research by economists, which has largely treated ethnicity as exogenous, has found they are associated with poor economic outcomes, often because of inter-ethnic hostility. The model suggests that ethnicity, and even inter-ethnic hostility, can be an efficient response to economic parameters. Thus, the empirical association of ethnic diversity with economic underperformance may suffer from reverse causality.

In the model, both group formation and inter-group hostility are responses to parameters including returns to scale, efficiency of search, information flows and the ability to separate effectively from the rest of the population. Hostility, when it arises, need not be the product of elite manipulation. Instead, it arises as a way of ensuring the group's own viability by expanding the range of trust within the group.

Although ethnic group formation and hostility can be efficient, they need not be so. Even if groups arise efficiently, the non-stationary section of the paper shows that it may be difficult for them to adjust efficiently to new settings. This is likely to be particularly true when optimal groups are large relative to the population. Inefficiently slow adjustment tends to produce more hostility than would otherwise occur. Modernization, broadly construed, is likely to exacerbate these problems, so long as ethnic groups persist. Nation-building policies such as infrastructure investment or explicitly integrationist policies may ameliorate some of these problems. However, integrationist policies in particular come with a short-run cost of disrupting existing ethnic structures. Only if the population is already relatively integrated can such policies be expected to raise welfare.

This paper has deliberately adopted an extreme version of constructivism, in which everyone is exactly alike *ex ante*. Allowing individuals to differ in their ability to join particular ethnic groups could introduce a host of interesting issues. In particular, not all groups would be symmetric. Relations between groups of markedly different size, such as the Hutus and Tutsis, could then be addressed explicitly, as could the question of why small communities often dominate particular branches of trade, as for instance South Asians dominate manufacturing in Kenya.

More broadly, the model can potentially be applied to many types of groups aside from ethnic groups. Indeed, once ethnic groups are modeled explicitly it becomes clear that conceptually they are not entirely distinct from other types of groups, such as religious groups or social organizations. These could also be analyzed along the lines of the model in this paper. These forms of social capital have tended in the past to be discussed rather informally, without a clear accounting of their likely costs and benefits. This paper provides a starting point from which to derive a more systematic understanding of the relationship between non-market organizations and the market.

Bibliography

[1] Acemoglu, Daron, James Robinson and Thierry Verdier (2004). "Kleptocracy and Divide-and-Rule: A Model of Personal Rule." *Journal of the European Economic Association*, 2, 162-192.

[2] Aghion, Philippe, Alberto Alesina and Francesco Trebbi (2004). "Endogenous Political Institutions." *Quarterly Journal of Economics*, 119, 565-612.

[3] Akerlof, George and Rachel Kranton (2000). "Economics and Identity." *Quarterly Journal of Economics*, 115, 715-753.

[4] Akerlof, George and Rachel Kranton (2004). *Identity and the Economics of Organizations.* mimeo.

[5] Alesina, Alberto and Eliana La Ferrara (2003). "Ethnic Diversity and Economic Performance." Harvard Mimeo.

[6] Alesina, Alberto, Arnaud Devleeschauwer, William Easterly, Sergio Kurlat and Romain Wacziarg (2003). "Fractionalization." *Journal of Economic Growth*, 8, 155-194.

[7] Alesina, Alberto, Enrico Spolaore and Romain Wacziarg (2004). "Trade, Growth and the Size of Countries." Harvard Mimeo.

[8] Attanasio, Orazio and Ríos-Rull, José-Victor (2000). "Consumption Smoothing in Island Economies: Can Public Insurance Reduce Welfare?" *European Economic Review*, 44, 1225-1258.

[9] Baker, George, Robert Gibbons and Kevin Murphy (2002). "Relational Contracts and the Theory of the Firm." *Quarterly Journal of Economics*, 117, 39-84.

[10] Banerjee, Abhijit and Kaivan Munshi (2004). "How Efficiently is Capital Allocated? Evidecnce from the Knitted Garment Industry in Tirupur." *Review of Economic Studies*, 71, 19-42.

[11] Banerjee and Somanathan (2004). "History, Social Divisions and Public Goods in Rural India." MIT Mimeo.

[12] Bates, Robert (forthcoming), "Ethnicity" in *The Elgar Companion to Development Studies.*

[13] Benoit, Jean-Pierre and Vijay Krishna (1985). "Finitely Repeated Games." *Econometrica*, 53, 905-922.

[14] Berman, Eli (2000). "Sect, Subsidy and Sacrifice: An Economist's View of Ultra-Orthodox Jews." *Quarterly Journal of Economics*, 115, 905-953.

[15] Bernheim, B. Douglas, Bezalel Peleg and Michael Whinston (1987). "Coalition-Proof Nash Equilibria: I. Concepts." *Journal of Economic Theory*, 42, 1-12.

[16] Caselli, Francesco and Wilbur Coleman (2002). "On the Theory of Ethnic Conflict." Mimeo.

[17] Chandra, Kanchan (2001). "Ethnic Bargains, Group Instability, and Social Choice Theory." *Politics and Society*, 29, 337-362.

[18] Coleman, James (1994). *Nationalism and Development in Africa: Selected Essays.* University of California Press.

[19] Destexhe, Alain (1995). *Rwanda and Genocide in the Twentieth Century.* New York University Press.

[20] Dixit, Avinash (2003). "Trade Expansion and Contract Enforcement." *Journal of Political Economy*, 111, 1293-1316.

[21] Dixit, Avinash (2003). "On Modes of Economic Governance." *Econometrica*, 71, 449-481

[22] Dubey, Pradeep and Lloyd Shapley (1994). "Noncooperative General Exchange with a Continuum of Traders: Two Models." *Journal of Mathematical Economics*, 23, 253-293.

[23] Easterly, William and Ross Levine (1997). "Africa's Growth Tragedy: Policies and Ethnic Divisions." *Quarterly Journal of Economics*, 112, 1203-1250.

[24] Ellison, Glenn (1994). "Cooperation in the Prisoner's Dilemma with Anonymous Random Matching." *Review of Economic Studies*, 61, 567-588.

[25] Eltringham, Nigel (2004). *Accounting for Horror: Post-Genocide Debates in Rwanda.* Pluto Press.

[26] Esteban, Joan and Debraj Ray (1999). "Conflict and Distribution." *Journal of Economic Theory*, 87, 379-415.

[27] Fafchamps, Marcel (2004). *Market Institutions in Sub-Saharan Africa: Theory and Evidence.* MIT Press.

[28] Fearon, James and David Laitin (1996). "Explaining Interethnic Cooperation." *American Political Science Review*, 90, 715-735.

[29] Fearon, James and David Laitin (2000). "Violence and the Social Construction of Ethnic Identity." *International Organization*, 54, 845-877.

[30] Fudenberg, Drew, David Levine and Eric Maskin (1994). "The Folk Theorem with Imperfect Public Information." *Econometrica*, 62, 997-1039.

[31] Ganguly, Sumit (1997). *The Crisis in Kashmir: Portents of War, Hopes of Peace.* Cambridge University Press.

[32] Geertz, Clifford (1963). *Peddlers and Princes: Social Change and Economic Modernization in Two Indonesian Towns.* University of Chicago Press.

[33] Ghosh, Parikshit and Debraj Ray (1996). "Cooperation in Community Interaction without Information Flows." *Review of Economic Studies*, 63, 491-519.

[34] Greif, Avner (1994). "Cultural Beliefs and the Organization of Society: A Historical and Theoretical Reflection on Collectivist and Individualist Societies." *Journal of Political Economy*, 102, 912-950.

[35] Greif, Avner, Paul Milgrom and Barry Weingast (1994). "Coordination, Commitment, and Enforcement: The Case of the Merchant Guild." *Journal of Political Economy*, 102, 745-776.

[36] Herbst, Jeffrey (2000). *States and Power in Africa: Comparative Lessons in Authority and Control.* Princeton University Press.

[37] Horowitz, Donald (1985). *Ethnic Groups in Conflict.* University of California Press.

[38] Levin, Jonathan (2003). "Relational Incentive Contracts." *American Economic Review*, 93, 835-857.

[39] Ligon, Ethan, Jonathan Thomas and Tim Worrall (2002). "Informal Insurance Arrangements with Limited Commitment: Theory and Evidence from Village Economies." *Review of Economic Studies*, 69, 209-244.

[40] Migdal, Joel (1988). *Strong Societies and Weak States: State-Society Relations and State Capabilities in the Third World.* Princeton University Press.

[41] Migdal, Joel (ed.) (2004). *Boundaries and Belonging.* Cambridge University Press.

[42] Miguel, Edward (2003). *Tribe or Nation? Nation-Building and Public Goods in Kenya versus Tanzania.* Mimeo, UC Berkeley.

[43] Newbury, Catharine (1988). *The Cohesion of Oppression: Clientship and Ethnicity in Rwanda 1860-1960.* Columbia University Press.

[44] Padro i Miquel, Gerard (2004). *Captured by the Government: Ethnic Divisions and Political Accountability.* Mimeo, MIT.

[45] Peters, Michael (1991). "Ex Ante Price Offers in Matching Games Non Steady States." *Econometrica*, 59, 1425-1454.

[46] Prunier, Gerard (1995). *The Rwanda Crisis: History of a Genocide.* Columbia University Press.

[47] Putnam, Robert (2000). *Bowling Alone: the Collapse and Revival of American Community.* Simon and Schuster.

[48] Sandler, Todd and John Tschirhart (1997). "Club Theory: Thirty Years Later." *Public Choice*, 93, 335-355.

[49] Tirole, Jean (1996). "A Theory of Collective Reputations (with Applications to the Persistence of Corruption and to Firm Quality)." *Review of Economic Studies*, 63, 1-22.

[50] Weber, Eugen (1976). *Peasants into Frenchmen: The Modernization of Rural France 1870-1914*. Stanford University Press.

[51] Weber, Max (1968). *Economy and Society.* Bedminster Press.

Chapter 2

Inequality and the Dual Economy: Technology Adoption with Specific and General Skills

Summary 2 *Slow technology diffusion has important consequences for income inequality. Across countries, it can account for much of world inequality. Within countries, the arrival of new technologies often causes popular concern that the wages of some workers will fall. In developing countries, diffusion is often represented in the stylized form of a "dual economy," in which a gradually increasing fraction of workers use modern technology, while the remainder use traditional technology. This paper generates this pattern of technology adoption endogenously, and shows that it causes the wages of some workers to fall. Diffusion is slow because production requires both skills that are equally useful with any vintage of technology and skills that are imperfectly transferable across vintages. When these skills are sufficiently specific, a dual economy is optimal, even though intermediate technologies are also available. Workers with transferable skills disproportionately join the modern sector, leaving those with specific skills worse off. As in classical trade models, changes in factor supplies affect the allocation of labor across sectors, but not factor rewards. A dynamic extension shows that workers without transferable skills become worse off throughout the period of transition to the modern technology. A further dynamic extension applies the model to developed countries.*

2.1 Introduction

Slow technology diffusion has important consequences for income inequality. Most obviously, a large fraction of cross-country income differences can be explained by differences in technology.[1] The way in which developing countries adopt technology is therefore of considerable interest. This is often represented in the stylized form of a "dual economy," in which a gradually increasing fraction of workers use modern technology, while the remainder use traditional technology.[2] This process is a common justification for the Kuznets Curve, according to which inequality increases early in modernization and later decreases. The dual economy is therefore one example of the way in which technology diffusion can affect income inequality within countries, as well as across them. In fact, it is often claimed that new technologies not merely raise inequality, but can actually leave some workers worse off in absolute terms. This was notoriously believed to be the case, for instance, by the Luddites during the Industrial Revolution. Their claims had at least superficial plausibility: the handloom weavers of Lancashire, for example, lost three quarters of their earnings over a thirty year period from the 1790s to the 1820s.[3] More recently, technical change is often cited as a reason for increased inequality in the US since the 1970s, and sometimes as an explanation for the falling wages of many US workers over much of that time.

This paper presents a simple framework to unify these observations on technology diffusion through a dual economy and its effect on wages. Any vintage of technology requires for production both general skills that can easily be transferred across vintages and specific skills that are imperfectly transferable. Technologies diffuse slowly because, on arrival, new technologies are biased towards general skills. As a result, so long as specific skills are sufficiently specific, and there is a large enough supply of them, two vintages of technology are simultaneously in use. Because the modern technology favors general skills, the modern sector is relatively general-skill intensive. This reduces the absolute wage of specific-skill workers, who mainly work with the old technology but with fewer complementary inputs. Wages are set by relative productivity in the two sectors, independent of factor supplies. An economy is "developed" when it is optimal

[1] See, for example, Hall and Jones (1999).

[2] For a recent example, see Banerjee and Newman (1998). For the canonical discussions, see Lewis (1954) and Harris and Todaro (1969).

[3] For more on the plight of the handloom weavers, see Blythell (1969).

to close the traditional sector entirely, and use only modern technology instead. Much of the intuition for the model is exactly parallel to that of a small open economy in classical trade models.

While the static model illustrates why wages of some workers fall on arrival of new technology, it cannot describe the process of diffusion or changes in wages during modernization. The dynamic model considers these questions. In the dynamic model, external learning-by-doing yields a gradual increase in the specific skills required for the modern technology.[4] Eventually, enough skills are acquired in the modern sector so that the traditional sector closes entirely. Wages of workers with specific skills continue to fall throughout the transition, and only start to recover when the traditional sector closes. The model therefore generates a strong form of Kuznets Curve. The model is easily extended to developed countries. In developed countries, there is a balanced growth path involving the use of this period's frontier technology and last period's frontier technology. Economies endowed with more workers with transferable skills have larger modern sectors, and as a result equilibrium efficiency of specific-skill workers is higher in both sectors. An increased rate of technical progress and increased transferability of specific skills each have ambiguous effects on equilibrium inequality, although increased transferability certainly increases inequality initially.

The idea that some human capital is vintage specific has been explored in several previous papers. The key paper in the literature is Chari and Hopenhayn (1991). In Chari and Hopenhayn, all human capital is equally vintage-specific, and this leads to their interest in the response of the lifetime pattern of earnings to changes in technical progress. In particular, they are interested in the complementarity between experienced and inexperienced workers in the same vintage. Most other models of vintage human capital do not emphasise the complementarity between types that is the focus of interest here. These include Aghion, Howitt and Violante (2002), who consider the effect of GPTs, Krueger and Kumar (2004), and Galor and Tsiddon (1997), who develop a rich model of growth focusing on intergenerational mobility and ability.

Much of the extensive literature investigating why developing countries adopt new technol-

[4]Stokey (1991) has considered learning-by-doing with human capital. Famous empirical examples of the importance of learning-by-doing include David (1975) and Lucas (1993).

ogy slowly or not at all does not consider technology diffusion through a dual economy. Instead, it assumes that developing countries as a whole adopt a succession of better technologies. Most of the time, this is implicit, and slow adoption is the result of exogenous country differences such as productivity differences (Zeira (1998)) or barriers to adoption (Parente and Prescott (1994)). Keller (1996) is somewhat closer to the current paper in considering the relationship between technology opening and human capital.[5] In Basu and Weil (1998), this approach is explicit, as developing countries develop by using "appropriate technology," that is technology intermediate between the traditional technology and the frontier.

By contrast, most existing papers on a dual economy assume the existence of precisely two modes of production, rather than generating this as an optimal outcome, as here. An example is Beaudry and Francois (2004), who assume exactly two technologies and emphasize a complementarity between experienced and inexperienced workers different types of workers in learning about the new technology.[6] In other papers, the older technology survives because some non-competitive factor retards adoption of the new technology.[7]

A vast literature considers the interaction between technical change and inequality. The possibility that new technology can make some workers worse off in absolute terms has also been of persistent interest to economists at least since Ricardo's famous chapter on "Machinery" in the 1821 edition of his *Principles.* Only a few papers generate this prediction, and most of these rely on complementarity between different types of workers. In Acemoglu (1999), the complementarity arises because the economy is not perfectly competitive, but involves instead search and screening. Kremer (1993 and 1997) and Kremer and Maskin (1995) are closer in spirit to the present model, in that the O-ring production function can generate behaviour similar to that described here even with perfect competition. But the production function is highly non-standard, and they do not consider the transition to a modern economy. Beaudry and Green (2002) also generate falls in wages for some workers derived from complementary factors and two possible production technologies. However, their model focuses on explaining recent changes in the US wage distribution by considering a one-time introduction of a new,

[5] Another paper concerned with slow adoption caused by human capital is Kremer and Thomson (1998).

[6] In an extension, they consider multiple technology steps, that could lead to an endogenous dual economy. But their main interest is in comparing steady states rather than exploring this possibility.

[7] Banerjee and Newman (1998), for example, are concerned with imperfect credit markets. Acemoglu and Shimer (2002) generate technology dispersion from search, although they do not impose exactly two technologies.

permanently skill-biased technology to a developed economy.

The paper proceeds by developing static results in Section 2 , before turning to dynamic results and learning in Section 3. Some possible extensions to allow for supply responses and multi-good settings are discussed in Section 4. Section 5 concludes.

2.2 The Static Model

This section introduces the static model of technology adoption with vintage-specific human capital. Interpreted literally, the static model represents the effect on an economy in steady state of the exogenous introduction of a set of new, but not fully understood, technologies. Such a pure experiment may occasionally have occured, for instance after the fall of the Tokugawa Shogunate in Japan or the loss of the Opium Wars in China. More broadly, however, it approximates the many instances when developing countries have switched from inward-looking policy regimes to policies favoring trade and foreign investment. To establish the intuition for the model, I first use Lerner diagrams to describe a dual equilibrium geometrically, before describing the model algebraically. I establish the conditions under which the dual or separating equilibrium involves using the traditional technology and the frontier technology, rather than a single interior technology.

2.2.1 A Graphical Representation

As foreshadowed in the introduction, much of the intuition in the present model is similar to the intuition that comes from considering a small open economy in classical trade models. Adapting the graphical analysis from that literature highlights three key points about the dual equilibrium. First, it shows the circumstances under which a dual economy, rather than universal adoption of an intermediate technology, will indeed be optimal. Second, it highlights the importance of factor endowments and the cone of diversification. An economy insufficiently endowed with general skills never opens a modern sector, while an economy with enough workers with general skills operates only the modern sector. Finally, it highlights that wages in an economy within the cone of diversification are determined purely by the relative position of the isoquants, independent of relative factor supplies.

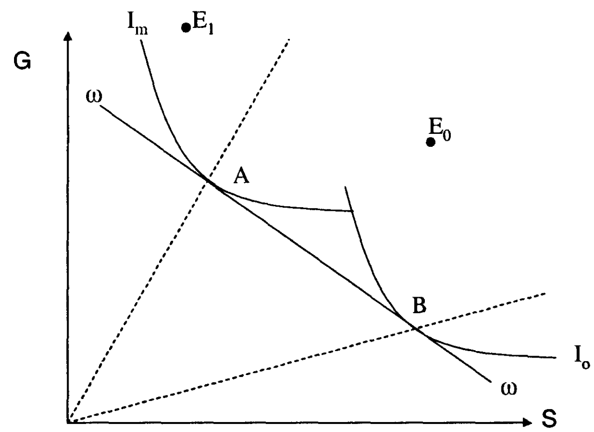

Figure 2-1: Equilibrium with Two Technologies

In the Lerner diagram in Figure 2.1, perfectly competitive firms choose their factor inputs and production technology, subject to factor prices. The quantity of general-skill labor appears on the vertical axis, that of specific-skill labor on the horizontal axis. The endowment of this economy, which consists of supplies of general and specific-skill labor, is at point E_0. The two available technologies are represented by the isoquants I_o and I_m. I_m, the modern technology, is general-skill biased, since at any given factor-price ratio it employs relatively more general-skill labor.[8]

In equilibrium, both technologies are used so long as the endowment of the economy lies within the cone of diversification, defined by the area between the rays OA and OB. In that case, factor prices are determined solely by the relative position of the two isoquants, as illustrated by the factor price line, $\omega\omega$: relative supplies do not affect relative prices. The factor price line can also be interpreted as an isocost line.

If the endowment is at a point such as E_1, the supply of general skill is so great that the

[8] In the trade literature, these are "unit value" isoquants, which can be drawn because output prices are fixed in a small open economy. Here, they are simple isoquants, because all firms produce the same good.

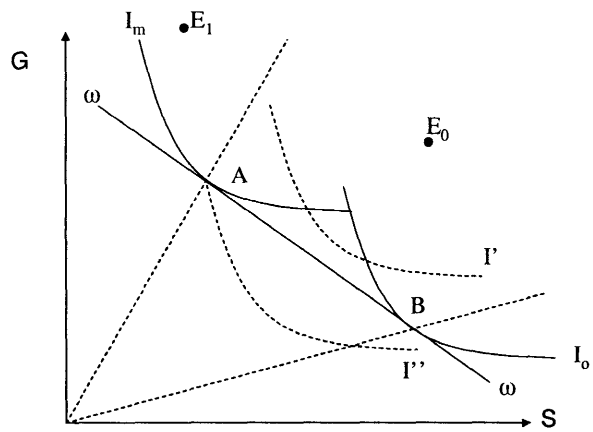

Figure 2-2: Intermediate Technologies

specific-skill intensive technology is not used. This is because no convex combination of the rays OA and OB can attain point E_1, so factor markets cannot clear. With endowment E_1, firms use only the technology represented by isoquant I_m, and relative supplies once again determine factor prices. The relative price of the abundant factor, G, must be lower than when both technologies are in use.

Now consider introducing other technologies whose factor bias lies somewhere between that of I_o and I_m. Facing the same factor prices, these technologies will be used if and only if some point on the new isoquants lies below the factor price line. In Figure 2.2, the technology associated with isoquant I' is unprofitable, whereas the technology represented by isoquant I'' will replace one or both of the previous technologies. Which one is replaced depends on the factor endowment of the economy. If there is a continuum of technologies such as I' or I'', none will be used so long as none of them have isoquants that cross the factor price line. In this case, the two extreme technologies, I_o and I_m, are used. This is the dual economy equilibrium.

Because isoquants are convex, the isoquant crosses the factor price line if and only if it is below the factor price line at the point where its tangent is parallel to it. Therefore, the two

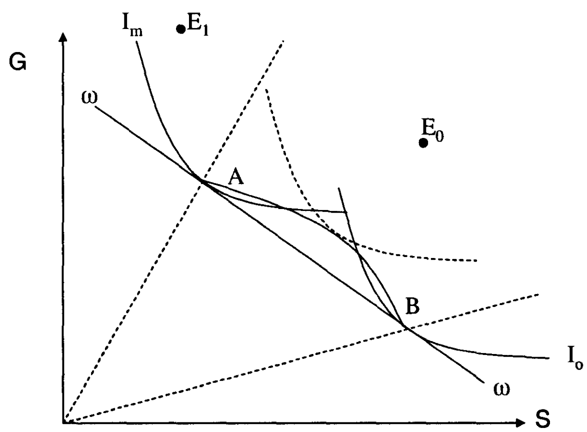

Figure 2-3: An Equilibrium Dual Economy

extreme technologies are used so long as the set of these points lies entirely above the factor price line, as in Figure 2.3. A sufficient but not necessary condition for characterising this set is that it be weakly concave. Economically, this requires that new technologies be "sufficiently" skill-biased.

Having established informally the key features of a dual equilibrium relating to factor endowments, factor prices and the characteristics of technology, I turn to the algebraic analysis.

2.2.2 Production

The economy consists of perfectly competitive firms, all of which have access to a well understood constant returns to scale production technology. The productivity of this technology is indexed by a^o. Production uses only specific skills and general skills. Consider the introduction of a continuum of new technologies, indexed by $a \in (a^o, a^m]$, where a^m is the most modern technology available, assumed to be the world technology frontier. The key assumption here is that firms face a menu of possible technological choices: they are not constrained to use either the frontier or the traditional technology. This assumption seems quite reasonable.

For example, it seems reasonable to suppose that technical progress continually expands the technological frontier, without eliminating the previous frontier technologies. This issue will be explored further in the dynamics.

Production by firm i takes the CES form:

$$\tilde{Y}_i = a_i[\lambda S_i^\rho + (1-\lambda)G_i^\rho]^{1/\rho},$$

where S_i and G_i are the effective inputs of specific and general skill labor demanded by the firm and a_i is the vintage of technology chosen by the firm. Note that $\rho \in (-\infty, 1)$ and that the elasticity of substitution is $\sigma = \frac{1}{1-\rho}$. For convenience, normalise output by a^o:

$$Y_i = \frac{\tilde{Y}_i}{a^o} = r_i\left[\lambda S_i^\rho + (1-\lambda)G_i^\rho\right]^{\frac{1}{\rho}}, \tag{2.1}$$

where $r_i \in [1, m]$ is equal to $\frac{a_i}{a^o}$, the ratio of output produced using the new technology to output produced using the old technology, for the same effective factor inputs to each.

Workers are endowed with either specific human capital or general human capital, or some combination of the two. To aid discussion, it is often convenient to assume that each worker is endowed with one unit of either specific or general human capital. Accordingly, the units of each type demanded by the firm can also be thought of as workers of each type. A worker endowed with one unit of general human capital can supply one efficiency unit to any vintage of technology. A worker endowed with one unit of specific human capital can supply one efficiency unit of specific capital to firms using a^o, but strictly less to firms using other vintages of technology. The number of efficiency units of specific human capital supplied by a specific-skill worker is described by the declining function $e(r_i)$, which by definition is equal to one when r_i is equal to one. This is the transferability function.

Denote by s_i the number of units of specific-skill capital that a firm hires, and g_i the number of general-skill units. Then the effective quantity of labor used by the firm is:

$$S_i = e(r_i)s_i \text{ where } e(1) = 1,\ e'(r_i) \leq 0\ \forall\ r_i \in [1, m] \tag{2.2}$$

$$G_i = g_i \tag{2.3}$$

The key assumption is that efficiency units of specific-skill labor supplied per worker decrease as the technology used becomes increasingly remote from the accustomed technology. Although it is not essential to the results that $e(r_i)$ be differentiable or continuous, this is assumed for now.

Since firms are perfectly competitive, they choose r_i, s_i and g_i to maximise

$$\pi = r_i[\lambda e(r_i)^\rho s_i^\rho + (1-\lambda) g_i^\rho]^{\frac{1}{\rho}} - (w_S s_i + w_G g_i), \tag{2.4}$$

where the price of output is normalised to one.

There are no externalities in this problem, and hence, although it is non-convex, the First Welfare Theorem applies. That is, the perfectly competitive equilibrium is Pareto optimal. The mathematical appendix shows that the competitive equilibrium allocation is the same as the social planner's allocation. This is true whether or not the economy lies within the cone of diversification defined by the available technologies, and hence whether or not the dual equilibrium is optimal.

2.2.3 Technology Choice

It is helpful to rewrite the production function, equation 2.1, as

$$Y_i = [\lambda \xi(r_i)^\rho s_i^\rho + (1-\lambda) r_i^\rho g_i^\rho]^{\frac{1}{\rho}}. \tag{2.5}$$

where $\xi(r_i) \equiv r_i e(r_i)$. This formulation highlights that, for given factor inputs, production is an increasing function of both r_i and $\xi(r_i)$. If $\xi'(r_i) > 0$, better technology automatically raises production, and all firms use the newest technology. Older technology is therefore only ever used under the following assumption.

Assumption A2: *Assume that* $\xi(1) = 1, \xi'(r) < 0, \xi''(r) > 0$.

Although the assumption that $\xi'(r) < 0$ is necessary in order to allow older technologies to be useful, the assumption of convexity is less obvious. So long as the efficiency of specific-skill workers is always strictly positive, some segment of the function will asymptote towards zero (or a positive constant) and hence be concave over this range. In fact, this is all that is required for a dual equilibrium.[9] Since it seems likely that an extra worker could always do something useful, this assumtion seems reasonable. In particular, if there are no problems in contracting, a worker can always pay for the cost of any training, so the remainder of their time must have weakly positive marginal product.

Associated with the CES production function is the unit cost function,

$$c_i(w_s, w_g, r_i) = \frac{1}{r_i}\left[\frac{w_s^\eta}{\lambda e(r_i)^\eta} + \frac{w_g^\eta}{(1-\lambda)}\right]^{\frac{1}{\eta}} \text{ where } \eta = \frac{\rho}{\rho - 1}. \tag{2.6}$$

This is a convenient way to consider the firm's optimal choice of technology, because it requires optimising only over r_i, rather than simultaneously over the factor input ratio. The following lemma and proposition establish the conditions under which a dual equilibrium exists.

Lemma 5 *If a dual equilibrium exists, the unit cost function must exhibit a turning point somewhere in the range $r \in [1, m]$.*

Proof. *The unit cost function is a continuous function in r whose domain is the compact subset $r \in [1, m]$. Therefore, by Weierstrass's Theorem, it must attain its global maximum and minimum on the domain. If a dual equilibrium exists, profits must be zero at $r = \{1, m\}$. Hence, by the mean value theorem, there must be a point $r^t \in (1, m)$ such that $c'(r^t) = 0$.* ■

Proposition 8 *If a dual equilibrium exists, Assumption A2 is a sufficient condition for it to be unique.*

Proof. *If the dual equilibrium exists, the turning point in the cost function must be a maximum. Differentiating the unit cost function, equation 2.6, shows that at a turning point,*

$$\frac{-1}{r^2}\left[\frac{w_s^\eta}{\lambda e(r)^\eta} + \frac{w_g^\eta}{(1-\lambda)}\right]^{\frac{1}{\eta}-1}\left\{\frac{w_s^\eta \xi'(r)}{\lambda e(r)^{\eta+1}} + \frac{w_g^\eta}{(1-\lambda)}\right\} = 0. \tag{2.7}$$

Equivalently, the turning point occurs when

$$\frac{\xi'(r)}{e(r)^{\eta+1}} = -\left(\frac{w_g}{w_s}\right)^\eta \left(\frac{\lambda}{1-\lambda}\right).$$

[9] For appropriate factor endowments, so long as there is some technology $\tilde{r}$ for which the assumption is true $\forall\ r \geq \tilde{r}$, a dual equilibrium exists, although the "traditional" technology in that case will be $\tilde{r}$. This property is discussed in more detail in the dynamics.

Evaluated at the turning point, the second derivative has the opposite sign to

$$\xi''(r)e(r)^{\eta+1} + \sigma e(r)^{\eta} e'(r)\,\xi'(r), \tag{2.8}$$

where $\sigma = -(\eta+1) > 0$. *But under assumption A2, this expression is always greater than zero. Therefore, any turning point must be a maximum, and must therefore also be unique.* ■

2.2.4 Static Equilibrium

Proposition 1 establishes that, if a turning point exists, it is unique and a maximum, and that therefore all firms will use either the traditional technology or the modern technology. However, the proposition does not establish the conditions under which such a turning point exists, since these depend on the equilibrium factor price ratio. To proceed, it is therefore most convenient to solve for a dual equilibrium on the assumption that one exists, and then solve for the conditions under which this is indeed the equilibrium.

Assuming that the dual equilibrium exists, what are its properties? The system contains six variables, the two wage rates and four employment outcomes, namely general and specific employment in the modern and traditional sectors. The traditional (old) sector is denoted with an o subscript, the modern sector with an m subscript. (With constant returns to scale, it is of course unnecessary to distinguish sectoral choices from firm choices). Efficiency of specific-skill labor in the modern sector, $e(m)$ is written e_m. Economy-wide supplies of specific and general skill labour are $\bar{s}$ and $\bar{g}$. The equilibrium conditions are:

$$\begin{aligned}
w_S &= \lambda s_o^{\rho-1}[\lambda s_o^{\rho} + (1-\lambda)g_o^{\rho}]^{\frac{1-\rho}{\rho}} && (2.9)\\
&= \lambda e_m^{\rho} s_m^{\rho-1} m[\lambda e_m^{\rho} s_m^{\rho} + (1-\lambda)g_m^{\rho}]^{\frac{1-\rho}{\rho}} && (2.10)\\
w_G &= (1-\lambda)g_o^{\rho-1}[\lambda s_o^{\rho} + (1-\lambda)g_o^{\rho}]^{\frac{1-\rho}{\rho}} && (2.11)\\
&= (1-\lambda)g_m^{\rho-1} m[\lambda e_m^{\rho} s_m^{\rho} + (1-\lambda)g_m^{\rho}]^{\frac{1-\rho}{\rho}} && (2.12)\\
s_o + s_m &= \bar{s} && (2.13)\\
g_o + g_m &= \bar{g} && (2.14)
\end{aligned}$$

Taking the ratio w_G/w_S with respect to both technologies and equating yields the equilibrium relationship between factor ratios in the two sectors. It is convenient to switch variables and work in terms of ϕ_i, the ratio of general-skill to specific-skill labour in each industry.

$$\left(\frac{\phi_m}{\phi_o}\right) = e_m^{\frac{\rho}{\rho-1}}. \tag{2.15}$$

It is clear from equation 2.15 that the modern sector is relatively general-skill intensive if and only if $e_m^{\frac{\rho}{\rho-1}} > 1$. Since $e_m < 1$, this can be true only if $\frac{\rho}{\rho-1} < 0$, that is $\rho > 0$ ($\sigma > 1$). Most existing estimates of labor substitution elasticities suggest that different types of labor have have a greater than unit substitution elasticity. For example, if general capital is associated with higher levels of education, the consensus estimates of the elasticity of substitution cluster around 1.5 or so.[10] Other types of human capital, such as different potential experience groups, appear to be still more easily substitutable.[11] Since the limited evidence on the topic accords with the intuition that holders of general skills should be those to gain from technical change, I make the following assumption.[12]

Assumption A3: *$\rho > 0$, or equivalently $\sigma > 1$: factors are gross substitutes.*

Given that $\rho > 0$, equation 2.15 shows that the frontier sector is more general-skill biased relative to the traditional sector when specific skills are relatively inefficient at the frontier. That is, a fall in e_m increases the factor bias of the frontier sector compared to the traditional sector. This expression can be used to gain quantitative estimates of the magnitude of differences in skill-ratios across sectors. If σ is about 1.5, approximately the rate of substitution estimated in various US studies between college graduates and high-school educated workers, then if $e_m = \frac{1}{2}$, the modern sector will be about 1.4 times as general-skill intensive as the traditional sector. If efficiency is much lower, say $e_m = \frac{1}{5}$, then the modern sector will be about 2.2 times as general-skill intensive as the traditional sector. Allowing for a larger elasticity of substitution produces larger effects; for $e_m = \frac{1}{2}$, the frontier sector is twice as general skill intensive as the traditional sector when $\sigma = 2$ and four times as intensive when $\sigma = 3$.

Equating specific and general skill wages across sectors yields the equilibrium factor ratio

[10] See Katz and Murphy (1992), Card and Lemieux (2001), and Johnson (1997).

[11] See Card and Lemieuz (2001) and Borjas (2003).

[12] The assumption is important for the interpretation of the results, but not for the results themselves. That is, a dual economy can certainly exist with a lower substitution elasticity, in which case the winners and losers from new technology are reversed relative to the analysis here.

in the traditional sector (for convenience, written to the power ρ):

$$\phi_o^\rho = \left(\frac{\lambda}{1-\lambda}\right)\left(\frac{1-(me_m)^{\frac{\rho}{1-\rho}}}{m^{\frac{\rho}{1-\rho}}-1}\right). \quad (2.16)$$

Equilibrium requires that $me_m < 1$, which is guaranteed by Assumption A2. Note that the equilibrium factor ratio is determined purely technologically: factor supplies play no role. Combining equations 2.15 and 2.16 to solve for ϕ_m shows that ϕ_m relies on exactly the same set of variables as ϕ_o, namely technological variables but not factor supplies.

Since relative factor supplies do not affect industry factor intensities, nor can they affect relative wages. The equilibrium wage ratio is a simple transformation of factor intensity in the traditional sector:

$$\frac{w_G}{w_S} = \left(\frac{1-\lambda}{\lambda}\right)\phi_o^{\rho-1} \quad (2.17)$$

The ratio of general to specific-skill wages is increasing in specific-skill intensity in the traditional sector. This is intuitive: relative wages of general skill workers (the "wage premium") are higher when they are relatively scarce in the traditional sector.

It remains to solve for the equilibrium allocation of workers between the two sectors. The market clearing relationships, equations 2.13 and 2.14, can be rearranged as follows:

$$\begin{aligned}(1-\delta)\phi_o + \delta\phi_m &= \bar{\phi} \text{ and} \\ (1-\gamma)\frac{1}{\phi_o} + \gamma\frac{1}{\phi_m} &= \frac{1}{\bar{\phi}},\end{aligned}$$

where δ is the fraction of specific skill workers employed in the modern sector, and γ is the corresponding variable for general skill workers.[13]

[13]These equations can be derived as follows. Dividing the g market clearing equation by $\bar{s}$ gives

$$\frac{\bar{g}}{\bar{s}} = \frac{g_m}{\bar{s}} + \frac{g_o}{\bar{s}}$$

Now just multiply the first term on the right by $\frac{s_m}{s_m}$, the second by $\frac{s_o}{s_o}$ and rearrange.

Using these as the market clearing variables,

$$\delta = \frac{\bar{\phi} - \phi_o}{\phi_o \left(e_m^{1-\sigma} - 1\right)}, \text{ and} \tag{2.18}$$

$$\gamma = \left(\frac{\bar{\phi} - \phi_o}{\bar{\phi}}\right)\left(\frac{e_m^{1-\sigma}}{e_m^{1-\sigma} - 1}\right) \tag{2.19}$$

Since factor intensity within each sector is unrelated to the economy-wide factor ratio, it is clear from equations 2.18 and 2.19 that both δ and γ are unambiguously increasing in $\bar{\phi}$. The more general-skill workers an economy possesses, the greater the fraction of both general and specific-skill workers allocated to the modern sector. While the economy maintains a dual structure, a greater fraction of general-skill workers than specific-skill workers will always work in the modern sector, since

$$\gamma = \delta \frac{\phi_o e_m^{1-\sigma}}{\bar{\phi}} = \delta \frac{\phi_m}{\bar{\phi}} > \delta.$$

In short, changes in factor supplies change the allocation of labor between the traditional and modern sectors, but do not affect relative factor intensities, just as the graphical analysis suggests. However, if factor supplies change to the point where the economy lies outside the cone of diversification, there are important consequences. An economy insufficiently endowed with general skill labor to lie within the cone of diversification uses only the traditional technology, and never adopts the new technology. Conversely, an economy abundantly endowed with general skill labor may also lie outside the cone of diversification, and close the traditional sector entirely.

Since δ and γ must both be between 0 and 1, the market clearing conditions formally provide the conditions under which the dual economy exists. The economy is backward, meaning no modern sector opens, if

$$\bar{\phi} \leq \phi_o = \left[\left(\frac{\lambda}{1-\lambda}\right)\left(\frac{1-(me_m)^{\sigma-1}}{m^{\sigma-1}-1}\right)\right]^{\frac{\sigma}{\sigma-1}}. \tag{2.20}$$

The economy is modern, meaning the traditional sector closes entirely, if

$$\begin{aligned} \bar{\phi} &\geq \phi_m = \phi_o e_m^{1-\sigma} \\ &= \left[\left(\frac{\lambda}{1-\lambda} \right) \left(\frac{1-(me_m)^{\sigma-1}}{m^{\sigma-1}-1} \right) \right]^{\frac{\sigma}{\sigma-1}} e_m^{1-\sigma}. \end{aligned} \tag{2.21}$$

Whether the economy lies within the cone of diversification is therefore determined by comparing the factor endowment of the economy with the required factor ratios of the traditional and modern sectors. Other things being equal, the cone of diversification is wider when e_m is lower, that is, there is a greater difference between the desired factor ratios in the two sectors.

2.2.5 Comparative Statics

How does the dual equilibrium change in response to parameter changes? Since the factor intensity of the traditional (and modern) sector is independent of factor supplies, it changes only with changes in technology. In particular, consider the effects of greater technological backwardness (a greater value of m), and of greater specific-skill efficiency (greater e_m). Since frontier efficiency depends on the distance to the frontier, $e_m = e(m)$, there is both a direct and an indirect effect of changes in m. In the comparative statics that follow, I consider the direct effect of m in isolation before considering the total effect of a change in m. In most cases the indirect effect offsets the direct effect, and without specifying a functional form for $e(m)$ it is not possible to determine which effect dominates.

Turning first to technological backwardness, take the partial derivative of equation 2.16 holding efficiency constant, in order to isolate the pure effect of being more relatively backward:

$$\frac{\partial \phi_o^\rho}{\partial m} = \left(\frac{\lambda}{1-\lambda} \right) \left(\frac{-(\sigma-1)\, m^{\sigma-2}}{(m^{\sigma-1}-1)^2} \right) \left[1 - e_m^{\sigma-1} \right] < 0.$$

Therefore, the general-skill intensity of the traditional sector declines in initial backwardness. If a relatively more advanced technology becomes available, or if a more backward country opens to world technology, the traditional sector becomes less general-skill intensive. Turning to changes in the efficiency of specific skills,

$$\frac{\partial \phi_o^\rho}{\partial e_m} = \left(\frac{\lambda}{1-\lambda} \right) \left(\frac{-(\sigma-1)\, m^{\sigma-1} e_m^{\sigma-2}}{m^{\sigma-1}-1} \right) < 0$$

When skills are less specific (e_m is relatively high), the general skill intensity of the traditional sector also falls. This result is more surprising: when specific-skill workers are more efficient at working in the modern sector, the traditional sector becomes less general-skill intensive, that is, employs relatively more specific-skill workers.

What about the total effect of a change in m? This is

$$\frac{d\phi_o^\rho}{dm} = \frac{-\lambda}{1-\lambda}\frac{(\sigma-1)\, m^{\sigma-2}}{\left(m^{\sigma-1}-1\right)^2}\left[e_m^{\sigma-2}\left(e_m + me'(m)\right)\left(m^{\sigma-1}-1\right) + 1 - \left(me_m\right)^{\sigma-1}\right]. \tag{2.22}$$

The direct effect dominates if equation 2.22 is less than zero. This is true if

$$e_m\left(\frac{e_m^{1-\sigma}-1}{m^{\sigma-1}-1}\right) + me'(m) > 0.$$

Intuitively, this condition says that the direct effect dominates so long as the indirect effect on efficiency, $me'(m) < 0$, is not too great. Recall that by assumption, $e_m + me'(m) < 0$. Therefore, the direct effect can dominate only if $\frac{e_m^{1-\sigma}-1}{m^{\sigma-1}-1} > 1$. However, assumption A2 also guarantees that this is true, since $(me_m)^{\sigma-1} < 1$. Therefore, no unambiguous prediction for the total effect of technological backwardness on factor intensity in the traditional sector is possible. For the other comparative statics, I therefore focus on the direct effect alone.

For the frontier sector, the corresponding partial derivatives are:

$$\begin{aligned}\frac{\partial\phi_m}{\partial m} &= \frac{\partial\phi_o}{\partial r_m}e_m^{1-\sigma} < 0\\ \frac{\partial\phi_m}{\partial e_m} &= \frac{\partial\phi_o}{\partial e_m}e_m^{1-\sigma} - (\sigma-1)e_m^{-\sigma}\phi_0 < 0\end{aligned}$$

The frontier sector is less general-skill intensive when relative backwardness is greater, and when specific-skills are more efficient at the frontier. Factor intensity in the frontier sector therefore behaves in the same way as factor intensity in the traditional sector, rather than, as perhaps may have been expected, offsetting it.

Since, holding efficiency constant, a more relatively backward economy (greater m) is less general-skill intensive in each sector, the general-skill intensive sector, namely the frontier, must be larger. Similarly, for a given level of relative backwardness, if specific-skill workers are more efficient at the frontier, the frontier sector is larger. These results are summarised below, with reference to the fraction of specific-skill workers (δ) and general-skill workers (γ) employed in the frontier sector.

$$\begin{aligned}
\frac{\partial \delta}{\partial m} &= \frac{-\bar{\phi}}{\phi_o^2}\frac{\partial \phi_o}{\partial m}\left(e_m^{1-\sigma}-1\right)^{-1} > 0 \\
\frac{\partial \delta}{\partial e_m} &= \frac{-1}{\phi_o\left(e_m^{1-\sigma}-1\right)}\left[\frac{\bar{\phi}}{\phi_o}\frac{\partial \phi_o}{\partial e} - \frac{e_m^{-\sigma}\left(\sigma-1\right)\left(\bar{\phi}-\phi_o\right)}{\left(e_m^{1-\sigma}-1\right)}\right] > 0 \\
\frac{\partial \gamma}{\partial r_m} &= \frac{-1}{\bar{\phi}}\left(\frac{e_m^{1-\sigma}}{e_m^{1-\sigma}-1}\right)\frac{\partial \phi_o}{\partial m} > 0 \\
\frac{\partial \gamma}{\partial e_m} &= \frac{-e_m^{-\sigma}}{\bar{\phi}\left(e_m^{1-\sigma}-1\right)}\left[e_m\frac{\partial \phi_o}{\partial e_m} - \frac{\left(\sigma-1\right)\left(\bar{\phi}-\phi_o\right)}{\left(e_m^{1-\sigma}-1\right)}\right] > 0
\end{aligned}$$

As with factor intensity, the offsetting effects of technological backwardness and specific-skill efficiency render impossible an unambiguous prediction for the total effect of technological backwardness. If, empirically, greater initial backwardness slows adoption of new technology, this suggests that the indirect effect through reduction of specific skills dominates. Leapfrogging-type hypotheses would be consistent with a smaller indirect effect.

Turning to relative wages, equation 2.17 establishes a simple relationship between the wage premium ($\frac{w_G}{w_S}$) and the general-skill intensity of the traditional sector. The wage premium is rising with technological backwardness, since

$$\frac{\partial(\frac{w_G}{w_S})}{\partial m} = \left(\frac{1-\lambda}{\lambda}\right)(\rho-1)\phi_o^{\rho-2}\frac{\partial \phi_o}{\partial m} > 0.$$

Similarly, with regard to the efficiency of specific skills,

$$\frac{\partial(\frac{w_G}{w_S})}{\partial e_m} = \left(\frac{1-\lambda}{\lambda}\right)(\rho-1)\phi_o^{\rho-2}\frac{\partial \phi_o}{\partial e_m} > 0.$$

The wage premium is therefore *increasing* in the efficiency of specific skills. This is surprising, since increases in the efficiency of specific-skill workers in the frontier sector (higher e_m) are effectively just specific-skill augmenting technical changes. In a standard setting with $\sigma > 1$, specific-skill biased technical change reduces the premium to general-skill labour. In this setting, by contrast, increases in the efficiency of specific-skill workers make the frontier sector relatively more productive, and hence expand the frontier sector. Since the frontier is relatively general-skill intensive, however, this requires reducing the general-skill intensity of both sectors. But this leaves specific-skill workers relatively worse off than before, as now in both sectors they have fewer complementary general-skill workers with whom to work.

While the comparative statics have focused on relative wages, in fact the wage of specific-skill workers falls in *absolute* terms. How can this be? As frontier efficiency increases, the general-skill intensity of both sectors falls. Workers in the traditional sector have not gained directly from increases in efficiency at the frontier; their technology is exactly as it was before. But now specific-skill workers in the traditional sector, using the same technology, have fewer complementary inputs, in the form of general-skill workers, with whom to work. Therefore, their marginal product, and thus also their wages, must have declined. Specific-skill workers are worse off in absolute terms than before the introduction of new technology throughout the entire range of e_m for which the dual economy persists.

Most models of biased technical change do not predict that wages of some workers actually fall. As noted in the introduction, those that do typically rely on settings deviating from perfect competition, or unusual production functions such as the O-ring production function. In contrast, the prediction is generated here using an entirely standard neo-classical setting and production function, and requires only that different types of workers are both required for production. The model is intuitive in that losers from technical progress are those who disproportionately continue for some time to use older technology. However, this is a more subtle prediction than simply that the losers will be those left behind using the old technology - the losers are those who have a comparative advantage in the use of old technology, whether or not they as individuals actually continue to use it. For example, the model predicts that the Lancashire handloom weavers lose wages during the Industrial Revolution. But more broadly, workers whose skills are close substitutes, such as many mill workers at the time, are also predicted to suffer wage losses relative to before the introduction of new technology.

2.3 Dynamics

The model discussed so far represents a one-time opening of an economy to a new technology. However, the broader questions are how countries modernize after they open, and what happens to wages over the course of modernization. I consider the dynamics first with a fixed frontier and then with a moving frontier. The first case is useful in fixing intuition and, it turns out, is a reasonable approximation to the case of a developing country opening to new technology.

The second case is particularly interesting in the context of developed countries, and allows comparison with some of the literature on GPTs. Moreover, it highlights the role of general human capital, not merely as a factor of production alongside specific skills, but in helping achieve a higher equilibrium level of efficiency when there is ongoing technical progress. This ties the idea of general human capital back to the Nelson-Phelps literature on human capital as aiding adaptability.[14]

2.3.1 A fixed frontier

Assume there is an overlapping generations economy, in which workers live for two periods. In the first, they either go to school to gain general skills or gain experience and specific skills on the job. Assume the choice between the two is exogenous, and that the marginal productivity of workers gaining specific skills on the job is zero in their first period of life. In the second period of their lives, everyone works and participates in training (costlessly) the members of the next generation.

The only choice made by any workers is made by specific skill workers, who choose in what sector to gain experience. Assume perfect foresight, so wages in the two sectors will be equal, exactly as in the static model. Since knowledge is cumulative and young workers receive training in their sector of choice, efficiency in the modern sector increases over time, towards its upper bound of unity. This learning-by-doing is assumed to be entirely external; alternatively, existing workers do not benefit from learning this period and do not take into account the utility

[14] Gellner highlights the importance of general human capital to modern, technologically dynamic societies in contrast to traditional societies: "A society ... based in the last analysis on economic growth ... is thereby committed to the need for innovation and hence to a changing occupational structure. From this it follows that certainly between generations ... men must be ready for reallocation to new tasks. Hence, in part, the importance of generic training (Gellner 1983), 32)."

of their descendants. Assume that knowledge is accumulated according to some function:

$$\begin{aligned} e_{m,t+1} &= \psi\left(\delta_t\left(\bar{\phi}_t, e_{m,t}, m\right), e_{m,t}\right), \qquad (2.23)\\ \text{where } \psi_\delta &> 0, \psi_{\delta\delta} < 0, \psi_e > 0, \psi_{ee} < 0 \\ &= f(e_{m,t}; \bar{\phi}_t, m), \qquad (2.24)\\ \text{where } f_e &> 0, f_{\bar{\phi}} > 0, f_m > 0. \end{aligned}$$

Next period's efficiency is an increasing function of this period's efficiency and of the fraction of specific-skill workers employed in the modern sector (since this increases the opportunities for learning). Although the returns to each of these are decreasing individually, this does not guarantee that the function $f(e_{m,t}; \phi_t)$ is concave in $e_{m,t}$. The learning function is guaranteed to be concave in e only if the cross partial, $\frac{\partial\psi}{\partial\delta\partial e}$, is negative. Otherwise, the learning function may be convex or S-shaped in e. Intuitively, the curve might be S-shaped if at low levels of $e_{m,t}$, learning is slow because relatively few workers are in the modern sector, and at high levels of $e_{m,t}$, learning is slow because of the decreasing rate of learning in $e_{m,t}$ despite having almost all labor allocated to the modern sector.

On the reasonable assumption that no learning takes place if the modern sector is not open ($\psi\left(0, e_{m,t}\right) = e_{m,t}$), it is clear that a factor endowment that lies below the cone of diversification now has serious implications for the long term growth of the economy. Any economy with a sufficient supply of general skills to open the modern sector at all eventually converges to complete modernization. By contrast, if the economy is insufficiently endowed with general skill labor for the modern sector to open, the economy stagnates, remaining an entirely traditional economy permanently, even with freely available modern technology.

Since, for a given level of efficiency, the rate of learning is determined by δ, the same comparative statics apply to the rate of learning as to δ. In particular, for a given level of initial efficiency, countries with a greater supply of general-skill labor modernize more quickly. This appears quite reasonable and is consistent with historical experience. For example, during the late 19th Century and early 20th Century, Germany and the US both experienced more rapid growth than Britain in the then modern sectors such as the chemical industry. Britain's relative failure to compete in these industries has been blamed in part on the failure of its education system to provide a sufficient supply of scientists and other skilled workers.[15]

Once the learning function is understood, the dynamic path of wages and the sectoral allocation of labor in the economy can easily be analyzed as a sequence of comparative statics. Considering first the input ratios, at the moment of introduction, the modern sector is much more general-skill intensive than either the economy as a whole or the traditional sector. As e_m increases, the general-skill intensity of both sectors falls, although the general-skill intensity of the modern sector falls more rapidly. Once efficiency rises to the point implicitly defined by equation 2.21, the desired skill intensity of the modern sector falls to the skill intensity of the economy as a whole: at this point, the traditional sector shuts down.

Turning to the wage premium, since opening reduces the general skill-intensity of the traditional sector, specific-skill wages fall on opening. The wage premium increases monotonically with e_m until the traditional sector shuts down. Only once the economy has fully modernized do further increases in e_m increase earnings of specific-skill workers. Wage inequality returns to its original level when specific-skill workers become fully efficient in the frontier sector. It is interesting to note in this context that the last of the handloom weavers (who perhaps can be taken as a proxy for the traditional sector as a whole) disappeared around the 1840s, around the same time the general level of wages started to increase more noticeably.[16]

Finally, throughout the period of adoption, the fraction of labor of both types working in the modern sector increases monotonically, although a greater fraction of general-skill workers always work in the modern sector, until the economy modernizes.

2.3.2 A moving frontier

The world technology frontier has so far remained fixed. Consider now exogenous technical progress at rate x, so that the frontier technology at time $t+n$ is described by

$$a_m(t+n) = a_m(t)\, x^n. \tag{2.25}$$

With a moving frontier, additional assumptions and notation are required concerning effi-

[15] See, for example, Broadberry and Ghosal (2002) on services and Chandler (1990) on manufacturing.
[16] See Blythell (1969) on the handloom weavers. Feinstein (1998) surveys the extensive literature on real wages and presents evidence that real wage growth only picked up from the 1850s or 1860s.

ciency across vintages. Efficiency in vintage i in period t is indexed as $e_{i,t}$. If a sector is the modern sector in period t, its efficiency in period t is denoted $e_{m,t}$. If it was the modern sector in period $t-1$, its efficiency in period t is denoted $e_{\bar{m},t}$. The transferability function generalizes to:

$$e_{m,t} = h\left(e_{\bar{m},t}, x\right) \tag{2.26}$$

$$\text{where } h_e > 0, h_x < 0, h_{ee} < 0, h_{xx} > 0, h_{ex} < 0$$

The transferability function stipulates that efficiency in the modern sector in period t is a function of two variables, current efficiency in the most advanced sector used last period, $e_{\bar{m},t}$, and the technological distance between that sector and the modern sector this period, x. The following assumption replaces Assumption A2:

Assumption A4: *Let* $\zeta\left(e_{\bar{m},t}, x\right) \equiv xe_{m,t} = xh\left(e_{\bar{m},t}, x\right)$. *Then* $\zeta\left(e_{\bar{m},t}, 1\right) = e_{\bar{m},t}, \zeta_x < 0, \zeta_{xx} < 0$.

As in the static model, this assumption guarantees that no intermediate vintages between last period's frontier and this period's frontier will be used in equilibrium. However, it does not guarantee that these two vintages will be used in equilibrium. This is because the older, fully-understood traditional technology is still available. From the unit cost function, equation 2.6, it can be shown that for any given technology, $\tilde{r}$, unit costs are equal to those in the modern and traditional sectors (dropping time subscripts) if

$$\tilde{r}e_{\tilde{r}}^{*} = \left[1 - \left(\frac{\tilde{r}^{\frac{\rho}{1-\rho}} - 1}{m^{\frac{\rho}{1-\rho}} - 1}\right)\left(1 - (me_m)^{\frac{\rho}{1-\rho}}\right)\right]^{\frac{1-\rho}{\rho}}. \tag{2.27}$$

If $e_{\tilde{r}} > e_{\tilde{r}}^{*}$, the traditional sector closes and the economy modernizes: it uses instead the current period's frontier technology and last period's frontier technology.

It follows that, with a moving frontier, an economy in which two sectors are open can be one of two types.

The first type is the developing country case. In this case, the two open sectors are the traditional sector and the current frontier sector, which is continually updated. The continual updating of the modern sector slows learning relative to the fixed frontier case. On the other hand, continual increases in technology favor expanding the modern sector.

If enough learning occurs, eventually the traditional sector will close. This occurs when efficiency in last period's frontier vintage is sufficiently great. In equation 2.27, this involves $\tilde{r} = \frac{m_t}{x}$. Rewriting the equation, the traditional sector is abandoned if

$$e^*_{\bar{m}} \geq \frac{m_t}{x}\left[1 - \left(\frac{m_t^{\frac{\rho}{1-\rho}} - x^{\frac{\rho}{1-\rho}}}{m_t^{\frac{\rho}{1-\rho}} - 1}\right)\left(\frac{1-(m_t e_{m,t})^{\frac{\rho}{1-\rho}}}{x^{\frac{\rho}{1-\rho}}}\right)\right].$$

It follows that positive (net) learning need not actually occur in this economy in order for it to eventually abandon the traditional sector, since continued technical advance reduces the critical value of $e^*_{\bar{m}}$ required in order to modernize.

In the modern economy, there are still two open sectors, but these are now the current frontier (denoted m) and last period's frontier (denoted o, since last period's frontier is the old sector this period). A balanced growth path is achieved when the loss of frontier efficiency caused by moving to the new frontier each period is exactly compensated for by learning over the period during which a particular technology is on the frontier. With constant relative efficiencies, the allocation of workers between the modern and the older technology is also constant. From equations 2.24 and 2.26, this can be seen as the value of e_m such that

$$\bar{e}_m = h\left(f\left(\bar{e}_m, \bar{\phi}, x\right), x\right).$$

A simple example of this fixed point is shown in Figure 2.4. The concave curve is the transferability function, mapping efficiency in this period's old sector to efficiency in this period's frontier sector. The other curve, is the learning function, which maps efficiency in the modern sector this period to efficiency in the old sector next period. In Figure 2.4, the learning curve is convex, and so the equilibrium is unique. As noted, however, the learning curve need not be strictly convex, and with a sufficiently S-shaped learning curve this economy can exhibit multiple equilibria for certain values of $\bar{\phi}$. Increases in the supply of general skills raise the learning function and can thereby eliminate the multiple equilibria.

The response of inequality to changes in parameters is quite complex. The ratio of factor

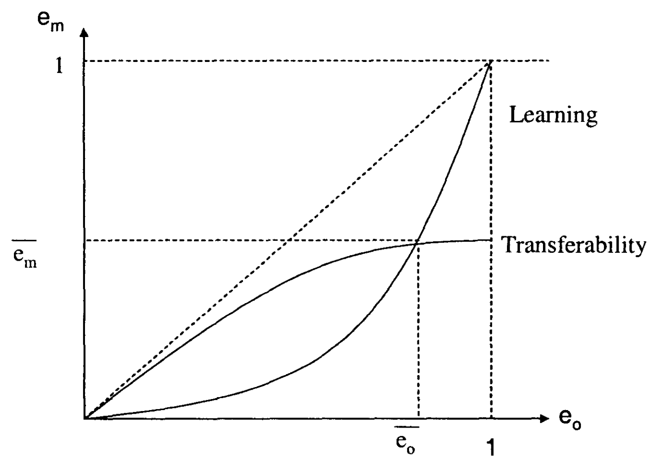

Figure 2-4: Balanced Growth with a Moving Frontier

intensities across sectors is now

$$\left(\frac{\phi_m}{\phi_o}\right) = \left(\frac{e_m}{e_o}\right)^{\frac{\rho}{\rho-1}} = \left(\frac{e_m}{e_o}\right)^{1-\sigma}. \tag{2.28}$$

This implies that skill intensity in the less-advanced sector is

$$\phi_o = e_o^{\frac{-\rho}{1-\rho}} \left[\left(\frac{\lambda}{1-\lambda}\right)\left(\frac{e_o^{\frac{\rho}{1-\rho}} - (xe_m)^{\frac{\rho}{1-\rho}}}{x^{\frac{\rho}{1-\rho}} - 1}\right)\right]^{\frac{1}{\rho}}. \tag{2.29}$$

Therefore, steady-state income inequality is

$$\frac{w_G}{w_S} = \left(\frac{1-\lambda}{\lambda}\right)^{\frac{1}{\rho}} \left[\frac{e_o^{\frac{\rho}{1-\rho}} - (xe_m)^{\frac{\rho}{1-\rho}}}{x^{\frac{\rho}{1-\rho}} - 1}\right]^{\frac{\rho-1}{\rho}} \tag{2.30}$$

$$= \left(\frac{1-\lambda}{\lambda}\right)^{\frac{1}{\rho}} \frac{1}{e_o} \left[\frac{1 - (x\frac{e_m}{e_o})^{\frac{\rho}{1-\rho}}}{x^{\frac{\rho}{1-\rho}} - 1}\right]^{\frac{\rho-1}{\rho}} \tag{2.31}$$

Equation 2.31 shows that inequality increases when e_m increases, and falls when e_o increases.

However, e_m and e_o may often change in the same direction. If e_o increases and $\frac{e_m}{e_o}$ falls, then inequality is unambiguously reduced, but if $\frac{e_m}{e_o}$ rises, there is no clear prediction relating to inequality.

Around the balanced growth path, it is possible to derive a rich set of comparative statics with respect to technical change. Inequality can respond to at least three types of changes in this economy. The first is an increase in skill transferability, such that the function $h(\cdot)$ is shifted up. A second is an increase in the rate of learning, such that the function $f(\cdot)$ is shifted out. Finally, there can be a change in the rate of technical progress, x.

Consider first an increased rate of transferability. This shifts up the transferability curve without affecting the learning curve. In the first generation after such a shock, only e_m increases, and hence inequality increases. In subsequent periods, however, e_o also rises, as the higher starting point feeds into higher efficiency in the less advanced sector. Whether the initial rise in inequality is entirely undone is not clear, since in equilibrium, e_o rises, but so does $\frac{e_m}{e_o}$, so long as the learning curve crosses the transferability curve from below, as shown in Figure 2.4. This is true of any stable equilibrium.[17]

By contrast, it is possible to sign the effect of an increase in learning. Shifting out the learning curve implies that equilibrium occurs at a point further along the concave transferability function. Hence, e_o increases and $\frac{e_m}{e_o}$ decreases, both of which reduce inequality. Adjustment towards the new steady-state is monotonic: inequality decreases gradually towards its new steady state.

The most obvious source of an increase in learning, for a given rate of technical advance, is an expansion in the supply of general-skill labour. This increases learning by increasing the size of the modern sector. Thus, the relative demand curve slopes down, effectively through a quasi-Rybczynski effect.

Finally, consider the effect of a change in the rate of technical progress. Holding relative efficiencies constant, an increase in the rate of technical progress increases the general-skill premium. However, both the learning curve and the transferability curve are drawn for given x, so this is not a satisfactory comparative static. Greater x shifts the transferability curve

[17]The initial shock to inequality when transferability increases is similar to the result in Aghion, Howitt and Violante (2002), although with fixed supplies they argue inequality increases permanently, because they have no general equilibrium effect on efficiency of older vintages.

down. Although this reduces both e_o and e_m, it also implies a fall in $\frac{e_m}{e_o}$, suggesting that the net effect is ambiguous. Similarly, for a given level of efficiency, more resources will now move to the modern sector. Hence, there will be greater learning for a given level of efficiency. It follows that the learning curve must shift outwards as a result of an increase in the rate of technical progress. This provides a further offsetting influence to the initial increase in inequality that occurs as a result of an increased rate of technical progress.

2.4 Extensions

2.4.1 Supply Responses

Endogenising the supply of general relative to specific skills adds little to the intuition of the model. The most important results, namely the dual economy and the changes in inequality, are not affected, since these are independent of factor supplies. The first order effect of an increase in e is to increase the premium paid to general skill workers. Therefore, if the supply of general skill workers is not completely elastic, endogenising supply merely hastens the rate at which the traditional economy converges to the modern economy. In the short run, this implies that the wage premium actually goes up faster, the greater the increase in general-skill workers. In other words, a time series regression of the wage premium on the relative supply of general skill workers would be perversely signed.

Another implication of the model is that the process of adopting a new technology creates large temporary quasi-rents to any specific skill workers who manage to acquire extra efficiency units of labour at the frontier. This suggests a possible explanation for the growth of technical education institutions in both developed and developing countries over recent decades, compared to gaining experience on the job as in a traditional apprenticeship. Suppose that instead of gaining experience on the job, it is possible to attend a technical college, which transfers a fixed amount of efficiency units to a specific-skill labourer. Then a specific skill worker gains more from going to technical college at a time of rapid technical change than when technologies are relatively well understood. This is consistent with the general move away from apprenticeship and other on-the-job training programs for specific-skill workers.

2.4.2 Experience Effects

The dynamic section assumed zero marginal product in the first period. Suppose instead that old workers gain some specific capital relative to young workers in the same sector, so that experienced workers supply $(1+\kappa)$ times as much effective labor as young workers. So long as young and old workers are perfect substitutes, however, the ratio of their wages must be fixed. Moreover, so long as κ is the same in each sector, the ratio of young to old earnings must be the same in both sectors. It follows that the lifetime path of earnings must be the same in both sectors. Hence, wages for specific-skill workers must be equalised across sectors in all periods. So a modification that allows young and old workers to be perfect substitutes would have no effect on the model. What if young and old workers are imperfect substitutes, as in Chari and Hopenhayn? For instance, suppose output is produced according to a nested CES function such as:

$$Y = a\left\{\lambda\left[\mu S_1^{\eta} + (1-\mu)\, S_2^{\eta}\right]^{\rho/\eta} + (1-\lambda)G^{\rho}\right\}^{1/\rho},$$

where S_1 is the effective supply of young specific-skill workers and S_2 is the effective supply of experienced specific-skill workers. So long as neither is more relatively substitutable for general skill workers than the other, as in the production function above, this would introduce some life-cycle dynamics along the lines of Chari and Hopenhayn without significantly affecting the relationships of interest here. In fact, this extension would strengthen the results, since the shortage of old workers in the frontier sector would further slow technology adoption. In short, generalising the assumption of zero marginal productivity in the first period would complicate the analysis without adding significantly to the insights of the model.[18]

2.4.3 More than one good, trade and empirical applications

The model can easily be extended to consider the effect of vintage-specific skills in a context where these skills are also industry-specific. Consider first the effect of technical change in only

[18] Moreover, the complementarity between experienced and inexperienced workers may not be very large. Estimates of the elasticity of substitution of workers across potential experience groups are around 5 to 10 (see Card and Lemieux (2001), and Borjas (2003)). Although this probably overstates the substitutability of workers with different actual experience within an industry, it is substantially different from Chari and Hopenhayn's central case elasticity of about 1.1 ($\eta = .1$). Chari and Hopenhayn's effects are smaller, the larger the elasticity of substitution.

one sector. This increases output in the sector, reducing the price of the good. This introduces an additional effect beyond that already modelled. The wage of specific-skill labour in this sector must fall, but general skill labour continues to be paid the same amount. Hence, the key prediction is that the wage premium in this sector increases. In general, with skills that are both vintage and sector specific, the model provides a framework for considering industry-specific skill differentials.

Price falls and/or output increases associated with new technology will be associated with two observable developments within the affected sector. First, as described above, intra-industry wage differentials rise. Second, the dispersion of skill-intensity across plants within the industry increases. In the most extreme case, a technological opening, plants will go from completely homogeneous to a split between modern, general-skill intensive plants and traditional, specific-skill intensive plants.

What if new technologies simultaneously become available in all industries? The basic prediction arising from a multi-good model is that the fastest-adopting industries will be those that have the greatest steady-state general skill bias. Therefore, the industries that start out as the most general-skill intensive will experience the greatest changes in intra-industry wage differentials and intra-industry plant dispersion early in the period of adoption, while the others follow.

Hence, if a developing country opens technologically, the most general-skill intensive goods should fall furthest in price/grow most in output. Obviously the price prediction is no different from Heckscher-Ohlin trade theory. But with vintage-sector-specific factors, no sector will close entirely, even if prices are set exogenously at the world level. If trade causes prices to fall most in relatively general-skill intensive sectors, specific-skill workers in that sector will be even more adversely affected. Unlike the simple Heckscher-Ohlin model, therefore, this model is consistent with increased inequality in poor countries that open their economies to trade (and other forms of technology transfer).

2.5 Conclusion

This paper has argued that a simple model of production in which general human capital and vintage-specific human capital are complements in production is useful in thinking about a number of issues in economic growth. The model builds on a couple of observations. First, different vintages of technology are often used at the same time in (roughly) the same place. Second, different workers are tied to particular tasks relating to a given level of technology to greater or lesser degrees, and these workers are not perfect substitutes. Although the ability to transfer skills across vintages is often related directly to ability, this paper has remained agnostic on this point, since endogenising supply makes little difference to the results. Thus, different degrees of transferability could relate to differences in innate ability, years of schooling, or even different types of schooling, such as apprenticeships and university degrees. Finally, new vintages of technology typically appear to benefit from learning-by-doing, and since workers differ in the degree to which their skills are vintage-specific, it would not be too surprising if learning-by-doing were also a biased process.

Developing these observations in the model leads to predictions consistent with some stylised facts of technical change. In the model, different vintages of technology coexist, and there is a dual economy if the economy starts out far enough behind the frontier. Technical change has strong effects on the distribution of wage income. In particular, the introduction of a new technology worsens the position of workers whose skills are relatively vintage-specific, both in relative terms, and, more notably, in absolute terms. This effect is not limited to the workers who continue to use the old technology, but is a general equilibrium effect common to workers with specific-skills in both the old and the new technology. Moreover, the negative effect on the wages of specific-skill workers actually increases as they become better acquainted with the new technology, and it is only after the old technology is abandoned that inequality starts to return to its initial level.

The dynamic version of the model extends these insights. Moreover, it illuminates the similarities between dual economies in developing countries and the ongoing process of technical development in developed countries. The eventual closure of the traditional sector suggests a systematic way to think of the difference between developing and developed economies. The developed country case of the model suggests a number of ways in which different types of technical change can affect inequality. For example, increases in the transferability of skills across vintages have an ambiguous effect in general equilibrium, contrary to what is often found in papers that adopt a partial equilibrium analysis.

The paper suggests that the Luddites can be right: new technology can hurt some workers. But it does not follow that policy should protect old industries. In fact, the reverse is true. Since learning is external, the economy modernises inefficiently slowly, and the best policy is to encourage the closure of old industries. Wages of lower skilled workers will not start to recover until these industries close. In this spirit, it is interesting that the general level of wages during the Industrial Revolution increased decisively only from around the 1840s, just as old technologies, such as handloom weaving, finally disappeared. In a more modern context, the model illuminates the decline of low-skilled wages in the US since the 1970s and the increase in inequality amongst poor countries opening their economies. It also, however, gives some grounds for hoping that these will be temporary phenomena, preceding, in both cases, sustained future wage increases.

2.6 Mathematical Appendix

2.6.1 Social Planner's Problem

In this appendix I sketch the proof that the resource allocation decision for the social planner is equivalent to the allocation of resources achieved in competitive equilibrium.

The social planner maximizes

$$\begin{aligned}\max_{f_s(r), f_g(r)} Y &= \int_r r[\lambda e(r)^\rho (f_s(r)\bar{s})^\rho + (1-\lambda)(f_g(r)\bar{g})^\rho]^{\frac{1}{\rho}} d(r) \\ \text{subject to } \int_r f_s(r)d(r) &= 1 \\ \text{and } \int_r f_g(r)d(r) &= 1\end{aligned}$$

where $f_s(r)$ is the density function of specialists in the economy using a given vintage of technology.

To see the equivalence of this problem to the perfectly competitive problem, note that the text has already shown the conditions under which it is true that, for *any* factor ratio, output is always maximised at either $r = 1$ or $r = m$. Since by definition the social planner wishes to maximize output, it follows that, if those conditions are satisfied, the social planner's allocation will necessarily take the form

$$f(s) = \begin{cases} c & \text{if} \quad r = 1 \\ 0 & \text{if} \quad 1 < r < m \\ 1-c & \text{if} \quad r = m \end{cases}$$

where c is a constant to be solved for, and $f_g(r)$ can be defined similarly.

It follows that the social planner's problem is a simple constrained maximisation problem as follows:

$$\begin{aligned}
\max_{s_0, s_w, g_0, g_w} Y &= [\lambda s_0^\rho + (1-\lambda) g_0^\rho]^{\frac{1}{\rho}} + m[\lambda e_w^\rho s_w^\rho + (1-\lambda) g_w^\rho]^{\frac{1}{\rho}} \\
\text{subject to } s_0 + s_w &= \bar{s} \\
\text{and } g_0 + g_w &= \bar{g}.
\end{aligned}$$

It is obvious that the first order conditions of this problem require that the marginal product of a particular type of worker must be equal in either sector. Since this is the same condition as in the perfectly competitive problem and there are no externalities, the social planner's allocation is identical to the perfectly competitive allocation.

Bibliography

[1] Acemoglu, Daron (1999). "Changes in Unemployment and Wage Inequality: An Alternative Theory and Some Evidence." *American Economic Review*, 89, 1259-78.

[2] Acemoglu, Daron and Robert Shimer (2000). "Wage and Technology Dispersion." *Review of Economic Studies*, 67, 585-607.

[3] Aghion, Philippe, Peter Howitt and Giovanni Violante (2002). "General Purpose Technology and Wage Inequality." *Journal of Economic Growth*, 7, 315-45.

[4] Banerjee, Abhijit and Andrew Newman (1998). "Information, the Dual Economy and Development." *Review of Economic Studies*, 65, 631-53.

[5] Bartel, Ann, Casey Ichniowski and Kathryn Shaw (2000). *New Technology, Human Resource Practices and Skill Requirements: Evidence from Plant Visits in Three Industries.* Mimeo.

[6] Basu, Susanto and David Weil (1998). "Appropriate Technology and Growth." *Quarterly Journal of Economics*, 113, 1025-54.

[7] Beaudry, Paul and Patrick Francois (2004). *Organizational Skills and Management in the Theory of Economic Development.* Mimeo.

[8] Beaudry, Paul and David Green (2002). "Changes in US Wages 1976-2000: Ongoing Skill Bias or Major Technological Change?" *NBER Working Paper* 8787.

[9] Blythell, Duncan (1969). *The Handloom Weavers: A Study in the English Cotton Industry During the Industrial Revolution.* Cambridge University Press.

[10] Borjas, George (2003). "The Labor Demand Curve is Downward Sloping: Reexamining the Impact of Immigration on the Labor Market." *Quarterly Journal of Economics*, 118, 1335-74.

[11] Broadberry, Stephen and Ghosal, Sayantan (2002). "From the Counting House to the Modern Office: Explaining Anglo-American Productivity Differences in Services, 1870-1990." *Journal of Economic History*, 62, 967-98.

[12] Card, David and Thomas Lemieux (2001). "Can Falling Supply Explain the Rising Return to College for Younger Men? A Cohort-Based Analysis." *Quarterly Journal of Economics*, 116, 705-46.

[13] Chandler, (1990). *Scale and Scope: The Dynamics of Industrial Capitalism.* Belknap Press.

[14] Chari, V.V. and Hugo Hopenhayn (1991). "Vintage Human Capital, Growth and the Diffusion of New Technology." *Journal of Political Economy*, 99, 1142-65.

[15] David, Paul (1975). *Technical Choice, Innovation and Economic Growth: Essays on American and British Experience in the Nineteenth Century.* Cambridge University Press.

[16] Feinstein, Charles (1998). "Pessimism Perpetuated: Real Wages and the Standard of Living in Britain during and after the Industrial Revolution." *Journal of Economic History*, 58, 625-658.

[17] Galor, Oded and Daniel Tsiddon (1997). "Technological Progress, Mobility and Economic Growth." *American Economic Review*, 87, 363-82.

[18] Gellner, Ernest (1983). *Nations and Nationalism.* Cornell University Press.

[19] Hall, Robert and Charles Jones (1999). "Why do some Countries Produce so much more Output per Worker than Others?" *Quarterly Journal of Economics*, 114, 83-116.

[20] Harris, John and Michael Todaro (1969). "Migration, Unemployment and Development: A Two-Sector Analysis." *American Economic Review*, 59, 126-42.

[21] Johnson, George (1997). "Changes in Earnings Inequality: The Role of Demand Shifts." *Journal of Economic Perspectives*, 11, 41-54.

[22] Katz, Lawrence and Kevin Murphy (1992). "Changes in Relative Wages, 1963-1987: Supply and Demand Factors." *Quarterly Journal of Economics*, 107, 35-78.

[23] Keller, Wolfgang (1996). "Absorptive Capacity: On the Creation and Acquisition of Technology in Development." *Journal of Development Economics*, 49, 199-227.

[24] Kremer, Michael (1993). "The O-Ring Theory of Economic Development." *Quarterly Journal of Economics*, 108, 551-75.

[25] Kremer, Michael (1997). "How Much Does Sorting Increase Inequality?" *Quarterly Journal of Economics*, 112, 115-39.

[26] Kremer, Michael and Eric Maskin (1995). *Globalization and Inequality.* Mimeo, Harvard.

[27] Kremer, Michael and James Thomson (1998). "Why Isn't Convergence Instantaneous? Young Workers, Old Workers, and Gradual Adjustment." *Journal of Economic Growth*, 3, 5-28.

[28] Krueger, Dirk and Krishna Kumar (2004). "Skill-Specific rather than General Education: A Reason for US-Europe Growth Differences?" *Journal of Economic Growth*, 9, 167-207.

[29] Kuznets, Simon (1955). "Economic Growth and Income Inequality." *American Economic Review*, 45, 1-28.

[30] Lewis, W. Arthur (1954). "Economic Development with Unlimited Supplies of Labor." *The Manchester School*, 22, 139-191.

[31] Lucas, Robert (1993). "Making a Miracle." *Econometrica*, 251-72.

[32] Parente, Stephen and Edward Prescott (1994). "Barriers to Technology Adoption and Development." *Journal of Political Economy*, 102, 298-321.

[33] Stokey, Nancy (1991). "Human Capital, Product Quality and Growth." *Quarterly Journal of Economics*, 106, 587-616.

[34] Zeira, Joseph (1998). "Workers, Machines and Economic Growth." *Quarterly Journal of Economics*, 113, 1091-117.

Chapter 3

When and Why did Technical Change Become Skill-Biased?

Summary 3 *This paper asks why, early in the modern era, technical change appeared to be primarily deskilling, while in the modern era it appears primarily skill-biased. Whereas previous explanations have focused on changes in technology, this paper suggests that changes in skills themselves may have been at least as important. If little technical change is expected, workers invest mainly in skills specific to a particular set of tasks. If they expect rapid technical change, they invest in skills that enable them to cope with an array of different tasks. The model also predicts that, with low rates of technical change, the modern sector is unskill-intensive, whereas the reverse is true when technical change is faster. However, the wage premium does not initially increase even after the modern sector has become skill-intensive, consistent with historical experience in the early part of the twentieth century.*

3.1 Introduction

Technical change in the modern world seems to be strongly skill-biased, in the sense that it tends to raise the wages of highly-skilled workers relative to those of less-skilled workers.[1] For

[1] Throughout this paper, I will use the term skill-biased as a convenient shorthand to describe the situation in which increases in the rate of technical change increase the skill premium. This is related to, but somewhat distinct from, the term's more usual meaning, which focuses on relative complementarity between skills and technical change in the production function.

example, the increased inequality experienced in the US over the past three decades has often been attributed to skill-biased technical change, particularly change associated with information technology and computers. Technical change has probably favored skilled workers more than unskilled workers for longer than the past thirty years: already in 1975, Tinbergen described the evolution of the wage structure as a race between technological development and educational advance. However, although these developments have by now inspired a large literature taking for granted the complementarity between technology and skill, the two have not always been partners. During the Industrial Revolution, for example, technical change primarily had the effect of reducing the value of skills. Even in the first decades of the twentieth century, perhaps the dominant forms of technical change were Fordism and Taylorism, approaches that explicitly emphasized reducing the demand for skilled labor.

This paper proposes a new theory to explain how and why technical change went from being primarily deskilling to primarily skill-biased. Instead of focusing on changes in the nature of technology, as in the previous literature, I argue that endogenous changes in the nature of skills themselves may have driven the changing effects of technology on the distribution of wages. The key idea is that technical change is a double-edged sword for workers who invest in education and skills: although it helps those with a capacity to learn new skills, it (relatively) harms those whose skills are rendered obsolete. The net effect on skilled workers as a group depends on the extent of their investment in these two different types of skills.

Intuitively, skills that can become obsolete are those that are specific to the performance of certain tasks such as operating a hand loom or a piece of machinery. The master pin-maker in *The Wealth of Nations* possessed these sorts of skills. Many other skills are less useful for the performance of any particular task, but are helpful in adapting to a range of situations. For example, the British empire was run largely by bureaucrats trained in classical Greek and Latin, rather than by engineers, lawyers or economists. Today, many college graduates gain a wage premium relative to their non-college educated peers despite not directly using the knowledge gained in their college major. In all of these cases, what is rewarded is not a special ability to perform any particular task but a general ability to perform a range of tasks or to adapt to new tasks as they arise.

If the acquisition of skills can take either of these two forms, then the effects of technical change on the wage premium will vary depending on the sorts of skills most skilled workers optimally choose to acquire. If the rate of technical progress is low, there is little chance that any particular set of skills will become obsolete, and hence it is optimal to specialize. If, on the other hand, there is rapid technical progress, and in particular rapid creative destruction, then any particular set of skills that a worker acquires is quite likely to be obsolete by the time they can be used, and accordingly workers will optimally choose instead to acquire more adaptable skills. If, moreover, highly skilled workers are required to create new technologies, it is possible to determine the rate of technical progress as a function of the productivity of the research sector relative to the productivity of acquiring specific skills for production. Because skilled wages fall with technical progress over some range, in a model with endogenous progress it is possible to generate multiple equilibria, in which the rate of progress depends on each agent's expectations of the skill investments made by other agents.

In this model, increases in the rate of creative destruction from a low base still leave most skilled workers preferring to acquire general skills, which are, however, more risky and hence less valuable than they were when there was a lower rate of creative destruction. Hence, it is mainly unskilled workers, with relatively little to lose, who work in the modern sector. However, once further increases in the rate of creative destruction entice skilled workers into acquiring general skills, the modern sector becomes skill-intensive. When the rate of technical progress is sufficiently great so that the modern sector absorbs all skilled workers, further increases in the rate of technical progress aid the relative position of skilled workers.

Most previous discussions of changes in the effect of technology on the demand for skilled labor have treated skills themselves as largely unchanged and focused on changes in technology itself. One view is simply that different versions of technology - different generations of General Purpose Technologies, perhaps - might exogenously differ in their skill requirements. Goldin and Katz (1998) make essentially this argument in describing the shift from an artisanal system, first to a factory system, and then to what they characterise as a more highly-skilled continuous process system. One possible driving force for such changes is the introduction of successive general purpose technologies, such as the electric dynamo, which Paul David and Gavin Wright have argued increased the demand for flexibility (David and Wright (2003)). A more subtle explanation is that the nature of technical change might be endogenous to other features of the economy. Acemoglu (2002) argues that, if technical change is "directed", then it makes sense to develop technologies that exploit primarily the relatively abundant factor.[2] Before the early twentieth century, this was unskilled labor; since then, it has been skilled labor (see also Acemoglu and Lester (2003)).

Although research on skill-bias and technology has, as noted, focused mainly on the effects of technology itself, there has also, of course, been research into changes in education and skills. Education has indeed changed dramatically over the past two centuries, in kind as well as in quantity. Ernest Gellner links the changing requirements of the education system explicitly to changes in the system of production. He notes that "The achievements of craft and art production in [mature agrarian] societies are extremely labour- and skill-intensive, and often reach levels of intricacy and perfection never remotely equalled by anything later attained by industrial societies... the major part of training in industrial society is *generic* training... (1983, pp26-7)." The history of education in America in the twentieth century is largely a history of two episodes of dramatic expansion of formal education in predominantly general rather than job-specific skills, these being the high school movement of the first quarter of the century and the expansion of tertiary education in the decades following the Second World War.[3]

In addition to the vast literature on technical progress and the return to skills, several other papers have considered technical progress and its relative effect on different types of skills. My previous paper (2005) shows how, in a more general model of production, workers with specific skills lose in absolute terms with the arrival of new technologies when they are to some extent complementary inputs to workers with more portable skills. However, the question of who optimally acquires which skills is not considered in that paper, and the rate of technical progress is treated as exogenous. Krueger and Kumar (2004) also consider a choice between what they label general and specific skills. In their model, however, there are no low-skilled agents, and nor is there any creative destruction of specific skills. The model is therefore not well suited to addressing changes in the wage premium that I consider here. Galor and Tsiddon (1997) and Galor and Moav (2000) are both concerned with how the return to ability changes with the rate of technical progress. In Galor and Tsiddon, specific capital is associated with

[2] This argument originates in the work of Habakkuk (1962) and the induced innovation literature of the 1960s.

[3] As Goldin (1998) has noted, the first of these was probably quantitatively substantially more important in lifting average educational attainment than the later expansion of college education.

the sector that employed a worker's parents, and technical progress tends to destroy some of that capital, as in the current model. However, technology is always skill-biased in the sense that high-ability workers always enjoy a comparative advantage in the modern sector.

Section Two of the paper presents the simplest possible model of the transition to skill-biased technical change, taking the rate of technical progress as exogenous. Section Three endogenises the rate of technical progress by incorporating the model of Section Two into an endogenous growth model similar to Caballero and Jaffe (1993) or Aghion and Howitt (1998). Section Four considers the implications of the model for the provision of education, and in particular the shift to formal pre-work, and often state-provided, education. Section Five concludes and discusses some possible directions for future research.

3.2 A Simple Model

3.2.1 The Environment

Consumption

Consider an OLG economy consisting of a fixed population of agents of unit mass, each of whom lives for two periods, consuming at the end of the second, and in which there is a continuum of intermediate goods, indexed on the unit interval. Utility is linear in the final good, which is produced (costlessly) as a CES aggregate of the intermediate goods,

$$C = \left(\int_{i=0}^{1} \tilde{c}_i^{\rho} di \right)^{\frac{1}{\rho}}.$$

Each good i is potentially available in a range of qualities indexed by k, all of which are perfect substitutes for each other, so that

$$\tilde{c}_i = \sum_k q_i(k) c_i(k).$$

where $q_i(k)$ is the quality of vintage k of good i and $c_i(k)$ is the quantity of it consumed. Note that, since different qualities of the same good are perfect substitutes, if the quality of one vintage is $\lambda > 1$ times that of an earlier vintage, the earlier vintage will be consumed only if its price is $1/\lambda$ the price of the newer vintage (or less).

Production

Perfectly competitive firms in industry i produce goods of quality $q_i(k) \leq q_i(k_i^{\max})$, where $q_i(k_i^{\max})$ is the quality of good i associated with the highest available vintage of good i. An innovation in industry i at time t allows production of good i at the frontier quality, $q_t^{\max}$, regardless of how many innovations had previously occured in the industry. The quality of the frontier increases exogenously at a constant proportional rate λ, so that

$$q_t^{\max} = \lambda q_{t-1}^{\max}.$$

Innovations are spread randomly across industries, and occur with probability z at the start of each period. Since in equilibrium only the best available quality of any good will be produced, I shall refer to any industry which in period t produces a good of quality $q_t^{\max}$ as a modern industry. Any industry in which in period t, $q_i(k_i^{\max}) < q_t^{\max}$ is called an old or a traditional industry. Since in equilibrium only the best available vintage of good i is produced, for notational simplicity I will generally denote that quality simply q_i.

Production of each good i requires labor and is carried out by perfectly competitive, constant returns to scale firms. One efficiency unit of any type of labor produces one unit of the intermediate good.

Labor Market

Of the unit mass of agents, $\bar{H}$ are high ability, while $\bar{L} = 1 - \bar{H}$ are low ability. In the first period of their life, workers costlessly acquire skills to be used in the second period of their life, at the end of which they consume their earnings.

As a result of their learning, each low ability worker supplies one efficiency unit of labor to any old industry (that is, $q_i < q^{\max}$) and $\delta < 1$ efficiency units of labor to a modern industry ($q_i = q^{\max}$). The loss of efficiency in the new industry reflects the lack of familarity of the low skill worker with the techniques used in the modern industry. Denote the wage of a low-skill worker working in sector i by $w^L(i)$.

Each high-skill worker can choose what sort of human capital to acquire. One choice is to specialize in producing good i using an existing technology. Clearly, in equilibrium this will be the best available technology, q_i. Working in this sector, each specialized worker provides $s > 1$ efficiency units of labor. If they work in any other industry or using any other technology, their effective labor supply is the same as that of an unskilled worker. Denote the wage of a skilled worker with specific skills by w_S^H.

The other choice is to acquire general training that is equally useful in working on any good or any technology, including modern technologies. A high-skill worker with general training provides g efficiency units of labor to any sector, where $1 < g < s$. Denote the wage of a general-skill worker w_G^H.

Since the labor market is competitive, each worker is paid the value of their production, so that a worker in industry i supplying e_i efficiency units of labor, which produces a good of price p_i, receives a wage

$$w_i(e_i) = p_i e_i. \tag{3.1}$$

3.2.2 Equilibrium

Goods Market

From the consumer's maximization problem, it is straightforward to show that demand for good i takes the form

$$c_i = p_i^{\frac{-1}{1-\rho}} q_i^{\frac{\rho}{1-\rho}} y$$

where y is the consumer's income. Since preferences are homothetic and there are constant returns to scale, aggregate demand must be

$$x_i = p_i^{\frac{-1}{1-\rho}} q_i^{\frac{\rho}{1-\rho}} E,$$

where E is aggregate expenditure. That is, consumption of each intermediate is declining in its price and increasing in its quality.

Labor Supply: Sectoral Choice

Since each worker (of a given type) can produce the same amount of any good using a technology that is at least one period old, all such goods will have the same price, which I normalize to unity. Similarly, all new goods will also have the same price as each other, p.

Consider now the choice of sector and technology for workers, taking prices as fixed. By equation 3.1, low skill workers receive a wage that is identically one by normalization in any old industry, and a wage $p\delta$ in any new industry. Therefore, low-skill wages are

$$w^L = \max\{p\delta, 1\}.$$

Accordingly, all low-skill workers work in the modern sector if $p > 1/\delta > 1$, while they are indifferent if $p = 1/\delta$. High skill workers whose technology has not become obsolete work in their own sector so long as

$$s > p\delta$$

while high-skill workers who have specialised in good i and whose technology has become obsolete work with the new technology if

$$p_i\left(q_i^{\max}\right)\delta > p_i\left(q_i\left(k_{i,t-1}\right)\right)s.$$

where $q_i\left(k_{t-1}\right)$ is the best quality of good i that was available in the previous period. If good i was of frontier quality both last period and this period, the greatest price the old vintage can command is p/λ. It follows that high-skill workers whose technology has become obsolete always work with the new technology if

$$s < \delta\lambda.$$

Finally, general skill workers work in the new sector so long as

$$pg > g.$$

The key point arising from these comparisons is that general skill workers are prepared to work in the new sector so long as goods in the new sector are at least as expensive as goods in the old sector, while low-skilled workers work in the new sector only if its goods command strictly higher prices than goods in the old sector.

It is now convenient to make the following two assumptions.

Assumption A5 $\lambda\delta > s$ This assumption ensures that even workers who have specialised in the most modern technology available in time $t-1$, but whose technology has been superseded, prefer to work as low-skill workers in the modern sector. Improvements in technology are sufficiently great so that any specialized high-skill worker whose technology has become obsolescent prefers to work as an unskilled worker in the new technology, rather than continue to use the old technology.

Assumption A6 $s/1 > g/\delta$ The advantage of specialized high-skilled workers over low-skilled workers in the old sector is greater than the advantage of generalist high-skilled workers over low-skilled workers working in the new sector.

Assumption A1 is important chiefly in simplifying the presentation of the results. Assumption A6 is more substantive, since without it high-skilled workers always enjoy a comparative advantage in the modern sector. One implication of the assumption will be seen to be that steady-state inequality is lower in states with high rates of technical progress than in those with low rates of technical progress. This seems to be a reasonable description of historical reality. Nevertheless, whether this assumption is true is ultimately an empirical question.

Labor Supply: Education Choice

Taking these decisions into consideration, it remains to solve for high-skilled workers' choice between general and specific education. Workers maximize their expected utility, which given the form of utility function is equivalent to maximizing their expected wage (workers are risk neutral with the CES utility function, since it is homogeneous of degree one). The expected wage of a specific-skill worker is

$$Ew_S^H = (1-z)\,s + zw^L.$$

A general skill worker works in whichever sector has the highest price, and their wage is

$$w_G^H = \max\{g, pg\}.$$

Although in steady state the rate of technical change will depend only on λ, for convenience I will refer to an increase in z interchangeably as either an increase in creative destruction or an increase in the rate of technical progress.

To solve for equilibrium, consider an economy in which there is no technical change. Then all goods will effectively be old goods, and obviously the return to acquiring specific skills is strictly greater than the return to acquiring general skills, since $s > g$. Since all workers are producing the same set of goods, the wage premium is just

$$\omega = \frac{Ew^H}{w^L} = s$$

Now consider some small rate of technical change, ε. The chance of a high-skilled worker being adversely affected by this small amount of creative destruction is very small, and hence high-skilled workers continue strictly to prefer to acquire specific skills. The new good can be produced by some low-skill worker, who can produce δ units of the good. It is clear that since any unskilled worker can produce the good equally well, all the unskilled workers must be indifferent between producing a new good and producing one of the old goods. Hence, in this case, which I shall refer to as case 1,

$$p_1 = \frac{1}{\delta}. \tag{3.2}$$

How does the wage premium change in response to this small deviation? Clearly, the wage of unskilled workers in terms of old goods is still unity. The expected wage of skilled workers in terms of old goods is, however, slightly less than s because of the creative destruction involved in technical change. Hence,

$$Ew^H = (1 - \varepsilon)\, s + \varepsilon 1 < s.$$

Note, however, that the expected wage is still strictly greater than the wage gained by a general-skill worker, so high-skill workers all acquire specific skills.

As z grows, one of two things may happen. First, all specific-skill workers may be drawn into the new sector, with the result that the relative price of new goods climbs above $1/\delta$. Second, at some point high-skill workers become indifferent between acquiring general and specific-skills, as they expect to lose more and more of their specific skills through obsolescence. This occurs at the value of z, z^*, such that

$$(1 - z^*)\, s + z^* w^L = pg.$$

Assuming that low-skill workers have not yet entirely specialised in the new sector, the critical value of z^* must be

$$z^* = \frac{\delta s - g}{\delta s - \delta}.$$

For values of $z > z^*$, high-skill workers will be indifferent between working in the new and old sector only if

$$p_2 = \frac{(1 - z)\, s + z}{g}. \tag{3.3}$$

In general, of course, equations 3.2 and 3.3 will not be simultaneously satisfied. It follows that either low-skilled or high-skilled workers can be indifferent between sectors, but not both. Since p_2 is declining in z but p_1 is not, it follows that for $z > z^*$, low-skill workers must strictly prefer to work in the old sector.

As z continues to increase above z^*, equilibrium requires that more and more high-skill workers work in the new sector. It may seem that for $z > z^*$, further increases in z should raise the wage premium, since the modern sector is the skill-intensive sector. This is not true, however, because skilled workers must still be indifferent between entering the modern sector and gaining specific skills in the traditional sector. Thus, for some range of rates of technical progress greater than z^*, relative prices are pinned down by equation 3.3 and thus are declining in z. Increases in z push more high-skilled workers into general education and the modern sector, but this reduces prices of new goods enough so that general skill workers are still left (relatively) worse off with faster technical progress. Hence, increases in the rate of technical progress continue to hurt the relative position of skilled workers even after they come to dominate production of new goods.

As z increases further, one of two things may happen. First, it may become so risky to acquire specific skills that all high-skill workers instead acquire general skills, even if they expect to work in the traditional sector. This transition occurs at z^{**} such that

$$\begin{aligned} (1 - z^{**})\, s + z^{**} &= g \\ z^{**} &= \frac{s-g}{s-1}. \end{aligned}$$

Increases in z beyond z^{**} obviously do not change the skill premium, since in this case it must be equal to g. Of course, in this case the price of new goods must also fall to unity, the same as the price of old goods.

Eventually, for some value of z, z^{***}, high-skill workers must specialise in working in the new sector. Only for $z > z^{***}$ does an increase in z increase the skill premium. This is because the effect on relative prices is now positive, as scarcity of skilled workers limits supply of new goods relative to old goods. However, note that the price of new goods can never rise above p_1, since that would cause low-skill workers to re-enter the market for modern goods. Hence, even very high rates of creative destruction, although they can increase the skill premium, cannot increase the skill premium beyond its initial level.

Therefore, the key result to emerge from the simple model is that the effect of changes in the rate of technical progress on the wage premium is U-shaped. For low rates of creative destruction, it is optimal for skilled workers to invest heavily in skills specific to a particular production process. For high rates of creative destruction, it is optimal for skilled workers to invest instead in skills that make them more adaptable. One implication of this simple model is that relative employment intensities in different sectors also depend on the rate of creative destruction. When creative destruction is low, the modern sector is unskill-intensive. When it is high, it is skill intensive. Interestingly, the turning point with respect to wages does not coincide with the turning point with respect to employment. There exist a range of values of creative destruction such that, although the modern sector is general-skill intensive, nevertheless further increases in the rate of creative destruction still reduce the relative wage of skilled workers. Increases in creative destruction increase the relative wage of skilled workers only after high-skilled workers have not merely joined the modern sector, but have actually specialised in it.

3.3 Endogenous Technical Progress

The simple model above suggested why the skill-bias of technological change might depend on the rate of creative destruction itself. However, it is not very satisfactory to treat the rate of creative destruction as entirely exogenous, particularly if costs of research depend partly on the cost of hiring high-skilled workers to perform the research. The following section therefore develops a more fully fledged model that endogenises the rate of creative destruction.

Consider the same productive structure as before, with two modifications. First, there is now also a research sector, which can produce new ideas. Research is assumed to be high skill intensive, and successful research is rewarded with a one-period patent. In particular, assume that each individual firm is small and that the probability of success is proportional to $\phi(H_R)$, where H_R is the mass of high-skilled workers in the research sector. $\phi(H_R)$ is a function that relates the individual firm's likelihood of discovering a patent to the overall amount of research being conducted in the economy, by H_R researchers. Assume that $\phi'(H_R) \leq 0$, which allows for social decreasing returns to scale in research, perhaps because of unmodelled heterogeneity between particular researchers.[4] The aggregate rate of creative destruction is therefore

$$z(H_R) = H_R \phi(H_R),$$

where $z' > 0$, and $z(0) = 0$, $\lim_{H_R \to \infty} z(H_R) = 1$.

Second, the rate of progress of the frontier is now determined by the exent of creative destruction, so that

$$q_{\max,t} = q_{\max,t-1} (1 - z)^{-\gamma}, \text{ where } \gamma > 1.$$

This specification is needed so that the rate of technical progress is determined by the rate of creative destruction. This specification is the discrete time analogue of that in continuous time considered by Caballero and Jaffe (1993) and Aghion and Howitt (1998). A similar argument to that found in Aghion and Howitt (1998) establishes that in the long run the cross sectoral

[4]This specification is similar to that in Acemoglu (1998) and for horizontal research in Howitt (1999).

distribution of industry productivity is determined by the cumulative distribution function

$$\Phi(a) = a^{\gamma} \tag{3.4}$$

where $a = \frac{q}{q_{\max}}$. Without the endogeneity of the rate of progress, the cdf converges to $a^{-\frac{\ln(1-z)}{\ln\lambda}}$. With a constant distribution of relative productivities, as occurs in either case in steady state, the growth rate is determined purely by the rate of growth of the frontier. Hence, unless the rate of the growth of the frontier responds to the rate of creative destruction, increases in the rate of creative destruction can only raise the level of income in the long-run, but not its growth rate. Importantly, with a steady-state distribution of productivities, it is straightforward to show that C must increase at the same rate as q, and hence demand for good i can be rewritten as

$$x_i(a) = p_i^{\frac{-1}{1-\rho}} a^{\frac{\rho}{1-\rho}} E \tag{3.5}$$

where E is aggregate expenditure, which is constant with respect to changes in quality in steady state. Demand for good i, x_i, of vintage $a = \frac{q_i}{q_{\max}}$ can therefore be expressed simply as a function of price and relative quality.

Monopolists maximize profits subject to the demand curve in equation 3.5. As usual, this leads to monopolists (who sell only new goods, x_N) setting $p_N = \frac{MC}{\rho}$, where MC is the marginal cost of production.[5] There are now four possible configurations of workers across industries, leaving aside high-skilled workers in the research sector. Each type of worker can specialize in one industry, for example all non-research high-skilled workers work as specialists in the traditional sector and all low-skilled workers work in the modern sector. Alternatively, one type of worker can specialize and the other can be indifferent between industries. As in the simple model, it should intuitively be clear that, in increasing order of creative destruction, first, high-skilled workers will specialize in the traditional sector and low-skilled workers will be indifferent between sectors (in equilibrium). Depending on parameters, there may exist a range of creative destruction rates over which low-skill workers all work in the modern sector.

[5] If the growth rate is sufficiently low, the size of each technical step may not be enough to allow all the new monopolists to price existing producers out of the market at this price, and hence some may adopt limit pricing. In what follows, I ignore this possibility, which in any event affects only the small number of sectors that have innovated both this period and in a recent period.

Past a critical value of z, low-skill workers switch and specialize in the traditional sectors, while high-skill workers may be indifferent between sectors. Finally, high-skill workers may specialize in the modern sector.

For simplicity, I assume that parameters are such that there are enough low-skill workers so that they never specialize in the modern sector. This assumption guarantees that the wage of low-skilled workers is equal to the price of old goods, and is therefore always equal to unity.

Equilibrium can therefore be described in terms of demand for, and supply of, skilled workers for the research sector. Before proceeding, it is worth noting the timing of events in the economy. In the first period of life, workers acquire skills. At the start of the second period, workers find employment in either a modern or traditional sector, or in the research sector. Research is then carried out, and its results are available immediately. Research workers, however, cannot work in production in the rest of the period. Production then occurs, and at the end of the period old agents consume. The convention that research occurs at the end of the period is convenient in that it enables equilibrium to be solved in a static setting, but is not otherwise important to the results.

3.3.1 Equilibrium

Demand for Research Labor

The demand for skilled labor in the research sector is pinned down by the zero profit condition,

$$\phi\left(H_R\right)\pi = w^H.$$

It is straightforward to show that *ex post* profits are

$$\pi = MCx_n\left(\frac{1-\rho}{\rho}\right) \tag{3.6}$$

where marginal cost of production is

$$MC = \min\left\{\frac{1}{\delta}, \frac{w^H}{g}\right\}. \tag{3.7}$$

Moreover, using equation 3.5, and setting $a = 1$, together with the fact that aggregate expenditure is equal to gross wages and that prices of old goods are equal to one, demand for skilled labor in the research sector can be solved for as

$$\phi(H_R)\, MC^{\frac{-\rho}{1-\rho}} (1-\rho)\, \rho^{\frac{\rho}{1-\rho}} \left(\bar{L} + w^H \bar{H}\right) = w^H$$

Therefore, the demand curve is

$$\phi(H_R) = \begin{cases} \frac{w^H}{\bar{L}+w^H\bar{H}} \frac{1}{(1-\rho)(\delta\rho)^{\frac{\rho}{1-\rho}}} & if \quad w^H \geq \frac{g}{\delta} \\ \frac{\left(w^H\right)^{\frac{1}{1-\rho}}}{\bar{L}+w^H\bar{H}} \frac{1}{(1-\rho)(g\rho)^{\frac{\rho}{1-\rho}}} & if \quad w^H < \frac{g}{\delta} \end{cases}$$

It is easy to verify that the demand curve is downward sloping so long as $\phi' < 0$. The standard scale effect is present, in that an increase in aggregate expenditure shifts the demand curve to the right. An increase in monopoly power increases the rate of technical progress, other things being equal.

Supply of Research Labor

Skilled workers will be prepared to work in the research sector so long as they earn their expected opportunity cost of working in some part of the production sector. This will be equal to the greatest of either their expected wage given that they have specialist skills, or their actual wage if they have general skills. The resulting supply curve will be very similar to that in the simple model in which the rate of technical progress was exogenous. As in that case, it is easiest to solve for the supply curve by considering progressively increasing rates of technical change.

At very low rates of technical change, demand for skilled labor in the research sector will be low, and demand for any type of labor in the modern production sector will also be low. Therefore, both high and low-skilled workers will work in the traditional sector, and the wages of each type will be determined by their expected wages in this sector. Hence, low skill workers receive a wage of unity and high-skilled workers receive a wage of s, although the expected wage of specialist high-skill workers is

$$w^H = (1-z)\, s + z. \tag{3.8}$$

For sufficiently low rates of technical progress, firms in the modern sector therefore hire low-skill workers and pay them their opportunity cost, unity. The supply of workers to the research sector is determined by equation 3.8.

As already highlighted in equation 3.7, for a sufficiently high rate of technical progress, it will be cheaper for the modern sector to employ generalist high-skilled workers than to employ low-skilled workers. This occurs at the value of z, z^*, such that

$$\begin{aligned} \frac{(1-z^*)\,s+z^*}{g} &= \frac{1}{\delta} \\ z^* &= \frac{\delta s-g}{\delta\,(s-1)}. \end{aligned}$$

However, the bulk of high-skill workers, who continue to work in the traditional sector, still prefer to acquire specific skills, since $g/\delta > g$. Hence, the supply condition for the research sector remains unchanged, despite the factor-intensity reversal in the modern sector.

The important feature of the supply curve up to the point $H_R\,(z^{**})$ is that it is downward sloping: an expansion of the research sector leaves skilled workers worse off. Intuitively, the reason for this is that skilled workers anticipate that there is an increasing chance that their skills will become obsolete next period. Therefore, they are prepared to work in the modern sector for a lower wage than at lower rates of technical progress.

The supply condition changes only once the rate of technical progress is sufficiently great that skilled workers working in the traditional sector prefer to acquire general skills than face the risk of obsolescence with specific skills. That is,

$$\begin{aligned} (1-z^{**})\,s+z^{**} &= g \\ z^{**} &= \frac{s-g}{s-1}. \end{aligned}$$

For rates of technical progress greater than z^{**}, the research sector therefore needs to pay skilled workers g. (Note that $z^{**} > z^*$ so long as $\delta < 1$.)

At sufficiently high rates of technical progress, the modern sector and research sector together demand all of the skilled labor available in the economy. However, expanding output in the modern sector would involve hiring low-skilled workers, driving the marginal cost of production up to $1/\delta$, and therefore, in a competitive labor market, high-skilled wages up to g/δ. Whether high-skilled wages go all the way to this level is determined by the zero-profit condition. That is, there is effectively a vertical segment of the supply curve at the critical value of technical progress. This is the value of H_R for which

$$\begin{aligned} z\frac{x_n}{g} + H_R &= \bar{H} \text{ or} \\ H_R\left[\phi\left(H_R\right)\left(\bar{L} + g\bar{H}\right) g^{\frac{\rho}{1-\rho}} \rho^{\frac{1}{1-\rho}} + 1\right] &= \bar{H}. \end{aligned}$$

Thus, to the left of the critical value of H_R, skilled labor is available to the research sector at a cost of g, while to the right it is available at a cost of g/δ. At H_R, the supply curve is vertical. Note that beyond H_R, the wage premium, g/δ, remains lower than the wage premium when there was no technical progress, by assumption A6, that $s > g/\delta$.

3.3.2 Discussion

One possible configuration of parameters is depicted in Figure 3.1. The supply curve is the light line, SS, while the demand curve is the heavy line, DD. There is a kink in the supply curve at H_R^*, the value of H_R associated with z^*, because marginal costs of production are constant up to this point (and reflect costs of unskilled labor) and falling after this point, reflecting the falling cost of skilled labor. Moreover, there is a discontinuity in demand at H_R^{***} because skilled workers absorb all productivity increases at this point, with the result that there is no change in the amount of labor devoted to research at this point.

The main comparative statics relate to the productivity of research and the endowment of skilled labor. Clearly, greater productivity of research or a greater endowment of skilled labor both shift the demand curve up, the latter purely through the standard scale effect, and therefore raise the rate of technical progress. Note that greater research productivity also shifts the supply curve to the left, since fewer research workers are needed in order to produce the same amount of technical change.

One consequence of the fact that over a certain range both the supply and demand curves slope down is that multiple equilibria arise quite naturally in this context, without any need to appeal to externalities. In Figure 3.2, a case is presented in which there are three equilibria.

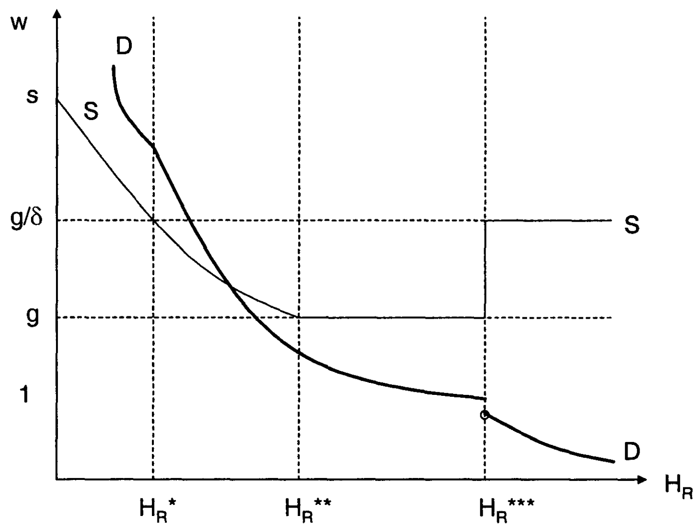

Figure 3-1: Equilibrium Rate of Technical Progress

The equilibrium at the intermediate rate of technical progress is of course unstable, while the other two equilibria are stable. In this setting, expectations of other agents' actions alone determine whether the economy experiences slow, deskilling technical change or intermediate, skill-neutral technical change.

3.4 Education Provision

One of the key insights of labor economics into the provision of education and training, dating from Becker (1964), is that employers will be prepared to pay for workers to acquire specific skills, but not for them to acquire general skills. Applying this insight to the current model suggests an explanation for the demise of apprenticeship-type training and the rise of pre-market training in the modern world. In particular, the timing of the high-school movement, at almost precisely the same time as it appears the modern sector first became relatively skill-intensive, is consistent with this model, as is the continued compression of the wage premium at that time.

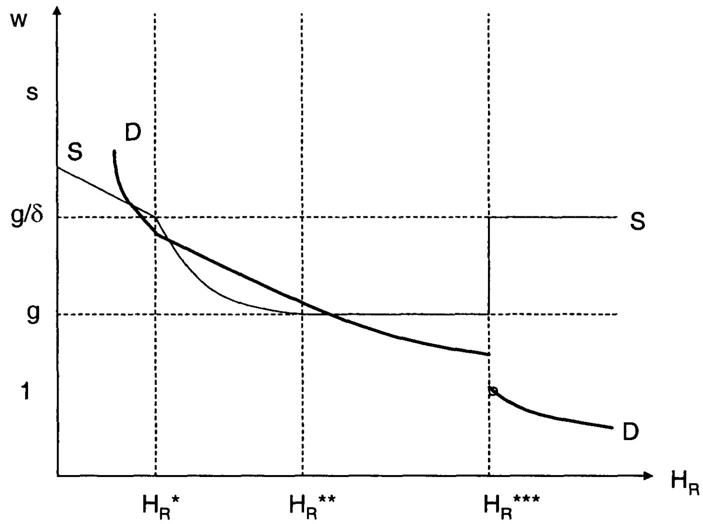

Figure 3-2: Multiple Equilibrium Rates of Technical Progress

The following discussion provides a sketch of the issues involved in this extension to the model.

To make the problem interesting, assume that there is no credit market, so workers cannot pay for the cost of any training they receive in the first period of their life. Moreover, assume that there are costs c_s and c_g of receiving training, that must be paid up front. Assume that contracts binding employers to pay a certain wage in the future, conditional on continuing in business, are binding, but that bonded labor is impossible.

In the product market, suppose that, although the patent expires after the first period, there is an ε entry cost into any particular industry. Then an equilibrium involves a monopolist pricing at the competitive price. However, the monopolist is now *ex post* a monopsony purchaser of the specific skills associated with their industry. Then the wage of each worker will be bid up to $s - c_s$ in the next period, conditional on the company remaining open.

Just as in the standard Becker model, however, firms will not be prepared to pay the cost of training general skill workers, since they can always bid them away from companies that have already trained them. The obvious solution to this problem is to impose a tax on production and subsidise the provision of general education. However, this only becomes necessary after high-skilled labor starts to work in the modern sector, that is, after the rate of technical progress has advanced beyond z^*.

3.5 Conclusion

This paper has used a simple model to argue that changes in skills may be just as important as changes in technology in explaining the effect of technology on the relative demand for skilled labor. In the simple model, there was a U-shaped relationship between the rate of creative destruction (and therefore also the rate of technical progress) and the return to skills. Low rates of creative destruction mainly had the effect of reducing the expected value of skills tied to specific tasks. High rates of creative destruction led skilled workers to invest instead in skills that would enable them to adapt easily to new technologies and circumstances.

This theory is consistent with the observed changes in the skill-bias of technical change over time. In historical perspective, the Industrial Revolution and even most of the Nineteenth Century involved surprisingly low rates of TFP growth, although even these rates were substantially higher than those that had gone before. Yet these changes were sufficient to cause large changes in the demand for labor and a general perception amongst observers as disparate as Ricardo and Marx that technology was deskilling. Certainly, workers in the modern sector, often women and children, tended to be less skilled than the experienced artisans they often replaced. Around the turn of the twentieth century, the rate of technical progress appears to have accelerated, and, as best we can judge, workers in the modern sector started to be better educated than the average of the workforce as a whole. Nevertheless, the early twentieth century was, if anything, a period when the wage premium continued to fall rather than rise. One reason that has often been cited for this is the contemporaneous dramatic expansion in high-school education. This paper suggests a new link between the accelerated rate of technical progress and the timing of the high-school movement, that is consistent with contemporaneous changes in both the skill-intensity of different industries and the wage premium. During the second half of the twentieth century, skilled employment continued to be concentrated in the modern sector, and TFP growth rates increased further still. This probably coincides with what appears in the model as specialization of skilled workers in general capital and the modern sector. Of course, the exact interpretation of increased inequality in recent decades depends on the interpretation of the evidence on technical change during this period.

There are a number of interesting directions for future research. Section Four of the paper sketched the first of these, which is to link changes in the rate of technical progress more explicitly with changes in the provision of education. In particular, applying the insights of Becker (1964), education can be expected to move from on-the-job training provided by employers to pre-market training, perhaps with an important role for the state.

Another interesting possibility is to consider the implications of trade in this model. In particular, unlike in Krueger and Kumar (2004), it is possible that differences in acquisition rates of general and specific skills could lead to differences in the level of output but not in the growth rate. Moreover, with technological spillovers, countries that are less technologically progressive might nevertheless achieve high *per capita* output by specialising in slightly older industries. This could potentially help explain the emphasis in some European countries such as Germany on specific skills, coupled with high rates of labor productivity, as against the emphasis on general skills in the more dynamic US labor market.

Finally, it would be interesting to reproduce the pattern of comparative advantage assumed here in Assumption A6 endogenously through complementarity in the production function. The most likely way in which this might be achieved is to create a dual index model of ability, whereby there is an idiosyncratic component in the choice of each worker as to whether to acquire general or specific skills. This would also yield additional insight into the pattern of relative wages within the class of skilled workers.

Bibliography

[1] Acemoglu, Daron (1998). "Why Do New Technologies Complement Skills? Directed Technical Change and Wage Inequality," *Quarterly Journal of Economics*, November, 113, 1055-1089.

[2] Acemoglu, Daron (2002). "Technical Change, Inequality and the Labor Market," *Journal of Economic Literature*, March , 40, 7-72

[3] Acemoglu, Daron and Ashley Lester (2003). "Technical Change and the Labor Market." Mimeo, MIT.

[4] Aghion, Philippe and Peter Howitt (1998). *Endogenous Growth Theory.* MIT Press.

[5] Becker, Gary (1964). *Human Capital.* University of Chicago Press.

[6] Caballero, Ricardo and Adam Jaffe (1993). "How High are the Giant's Shoulders: An Empirical Assessment of Knowledge Spillovers and Creative Destruction in a Model of Economic Growth." *NBER Macroeconomics Annual*, 15-74.

[7] David, Paul and Gavin Wright (2003). "General Purpose Technologies and Productivity Surges: Historical Reflections on the Future of the ICT Revolution." Mimeo, Stanford.

[8] Galor, Oded and Omer Moav (2000). "Ability-Biased Technological Transition, Wage Inequality and Economic Growth." *Quarterly Journal of Economics*, 115 (May), 469-497.

[9] Galor, Oded and Daniel Tsiddon (1997). "Technological Progress, Mobility and Economic Growth." *American Economic Review*, 87(3), 363-382.

[10] Gellner, Ernest (1983). *Nations and Nationalism.* Cornell University Press.

[11] Goldin, Claudia (1998). "America's Graduation from High School," *Journal of Economic History*, June, 58(2), 345-74.

[12] Goldin, Claudia and Lawrence Katz (1998). "The Origins of Technology-Skill Complementarity," *Quarterly Journal of Economics* 113 (June): 693-732.

[13] H.J. Habakkuk (1962). *American and British Technology in the Nineteenth Century.* Cambridge University Press.

[14] Howitt, Peter (1999). "Steady Endogenous Growth with Population and R&D Inputs Growing." *Journal of Political Economy*, 107(4), August, 715-730.

[15] Krueger, Dirk and Krishna Kumar (2004). "Skill-Specific rather than General Education: A Reason for US-Europe Growth Differences?" *Journal of Economic Growth*, 9, 167-207.

[16] Lester (2005) "Inequality and the Dual Economy: Technology Adoption with Specific and General Skills." Mimeo, MIT.

[17] Tinbergen, Jan (1975), Income Differences: Recent Research, Amsterdam, North Holland.